

City Research Online

City St George's, University of London

Citation: Arafat, Y. (2025). Application of computer vision and artificial intelligence techniques on histological and microscopical data. (Unpublished Doctoral thesis, City St George's, University of London)

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/35639/

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk



Application of computer vision and artificial intelligence techniques on histological and microscopical data

Youssef Arafat

First Supervisor: Dr Constantino Carlos Reyes-Aldasoro

Second Supervisor: Dr Robert Noble

This dissertation is submitted for the degree of
Doctor of Philosophy
at
City St George's, University of London

Department of Computer Science
School of Science and Technology

31st March, 2025

Acknowledgments

This work was conducted within the School of Science and Technology at City St George's, University of London.

Firstly, I would like to thank my first supervisor Dr Constantino Carlos Reyes-Aldasoro for his continuous support and guidance throughout my PhD. He has always gone above and beyond to assist me throughout my journey from the 1st year of my Biomedical Engineering degree at City till now. His expertise, and insightful feedback have been instrumental in shaping this work. Beyond his academic mentorship, I am truly grateful for his motivation and kindness, which have made this experience both enriching and inspiring. Thank you for your dedication and for being an exceptional inspirational mentor.

I would also like to thank my second supervisor Dr Robert Noble for his critical insights and guidance throughout the various stages of this work. I would also like to thank Dr Esther Castellano and Dr Cristina Cuesta-Apausa for providing the fluorescently labelled datasets that kick started my PhD journey and contributed to the formation of this thesis. I would also like to thank Dr Oscar Maiques for providing me with the collagen accumulation and Tissue Micro-Array datasets. His generosity in sharing knowledge, insights, and opportunities has been truly invaluable. I am especially grateful for his trust in bringing me into a groundbreaking project, which has greatly enriched my research and professional growth. I would like to thank Professor Victoria Sanz-Moreno on her insights and her trust for involving me in a groundbreaking project.

Throughout my PhD journey I have had the pleasure of having the support of many of my friends including; Riad, Daniel, Mariana, Iman, Parmis, Karina, Laura, Farouk, Ali, Tareq, Khalid, Abdelmoneom, Victor, Karim, Marwan, Mohamed, George, Raghda, Redjan, Amina, Alia, Hana, Faris, Dany, Mirela, Ananda, Kaleem, Halil, Jamal.

Finally, and most importantly I dedicate this PhD and all my accomplishments to my parents Reem and Amr and my brother Ahmed. To my parents, I thank you for your life lessons, your unwavering support throughout my whole life and your unconditional love that made this achievement possible. I am forever grateful for your guidance, inspiration, and endless motivation, which have shaped me into the person I am today. To my brother Ahmed, I thank you for your constant inspiration, support, guidance and encouragement.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text.

Youssef Arafat

Abstract

The morphological characteristics of the extracellular matrix play an important role in the progression of cancer and other health conditions. Yet, the extracellular matrix has been widely understudied in comparison to other factors such as vasculature, macrophages and soluble factors. This could be due to its complex three dimensional network that is composed of elements such as collagen, fibronectin and enzymes. Investigating these elements could yield valuable insights. This thesis investigates the nature of tissue and biological samples through the tracing of fibre-like structures of the extracellular matrix. To delineate, quantify and analyse fibres of the extracellular matrix, a novel tracing algorithm, Trace Ridges is presented.

Trace Ridges combines the methods of Watershed to detect ridges and Edge Detection to break any minor ridges that are running down from a main ridge towards the basins. The Watershed and Edge Detection methods are combined with morphological selection of length and brightness to favour brighter and longer ridges.

To compare Trace Ridges with state-of-the-art methods, an objective comparison is presented. Trace Ridges is compared to Edge Detection, CT Fire, Scale Space, Twombli, Graph Based and U-Net across 4 biological images of different backgrounds with three filtering options: no filtering, Gaussian filtering and DnCnn filtering. Accuracy is measured using distance error metrics that compare the fibre-like structures traced with the algorithms against a manually delineated ground truth. Trace Ridges outperforms all other algorithms by achieving the lowest total distance errors, and achieving the second lowest average distance error after Edge Detection, which has an affinity to over-segment and it is never far from the ground truth. In terms of speed, Trace Ridges was the second fastest algorithm and was four times faster than the next quickest method. Thus, Trace Ridges proved to be a fast, automatic and robust algorithm for tracing of fibre-like structures.

Trace Ridges was tested in several experimental settings. Firstly, it was used to analyse and quantify fibronectin fibres of mutant and wild-type populations. Morphological properties of the populations were extracted and compared, indicating that the number of fibres and average gap size were different between the populations. Secondly, Trace Ridges was used to investigate whether the accumulation of collagen leads to cytoskeletal remodelling and was reflected in the morphological properties. The samples compared were: (a) increasing concentrations of collagen, (b) control and protease inhibited collagen fibres, and (c) control and collagen fibres exhibiting Marangoni Flow. The results of the second experiment contributed to understanding the mechanism through which more aggressive tumours spread and metastasise. Thirdly, Trace Ridges was used to trace collagen fibres of normal and tumour breast tissues observed with Tissue Micro-Arrays. It was observed that the two populations express significant morphological differences.

List of Publications:

Journal Publications:

- Arafat Y, Cuesta-Apausa C, Castellano E, Reyes-Aldasoro C C. (2025) Fibre tracing in biomedical images: An objective comparison between seven algorithms. PLoS ONE 20(3): e0320006. https://doi.org/10.1371/journal.pone.0320006

- Maiques, O., Sallan, M.C., Laddach, R. et al. Matrix mechano-sensing at the invasive front induces a cytoskeletal and transcriptional memory supporting metastasis. Nat Commun 16, 1394 (2025). https://doi.org/10.1038/s41467-025-56299-7

- Cristina Cuesta, Marta Alcón-Pérez, Jie Zheng, Nicole Procel, Rosa Ramírez-Cota, Dirk Fennema Galparsoro, Alejandro Rosell, Diana Loa-Meson, Belén Martínez-Castedo, Youssef Arafat, Héctor Sanz-Fraile, Vinothini Rajeeve, Diego Alonso-López, Robert E. Hynds, Charles Swanton, Jordi Alcaraz, Pedro Cutillas, Constantino C. Reyes-Aldasoro, Haiyun Wang, Esther Castellano. RAS-PI3K Pathway in CAFs Shapes Physicochemical Properties of Tumor ECM to Impact Tumor Progression. bioRxiv 2025.01.20.633776; doi: https://doi.org/10.1101/2025.01.20.633776 (Submitted at Cancer Discovery).

Conference Publications:

- Arafat, Y., Reyes-Aldasoro, C.C. (2022). Computational Image Analysis Techniques, Programming Languages and Software Platforms Used in Cancer Research: A Scoping Review. In: Yang, G., Aviles-Rivero, A., Roberts, M., Schönlieb, CB. (eds) Medical Image Understanding and Analysis. MIUA 2022. Lecture Notes in Computer Science, vol 13413. Springer, Cham. https://doi.org/10.1007/978-3-031-12053-4_61

- Arafat, Y., Apausa, C.C., Castellano, E., Reyes-Aldasoro, C.C. (2023). Investigation of the structural characteristics of the extracellular matrix. 27th Conference on Medical Image Understanding and Analysis 2023

- Arafat,Y., Apausa, C.C., Castellano, E., Reyes-Aldasoro, C.C. (2024). Investigation of the organisation of the extracellular matrix using fibre tracing. In: A. Slingsby, D. Hunter. EG UK Computer Graphics & Visual Computing. CGVC 2024. The Eurographics Association. https://doi.org/10.2312/cgvc.20241238

Notations

Notation	Description
$a_i^{(\updownarrow-1)}$	Activation function
$\mathbf{q}$	Another neighbouring point
A_k	Area of current object
$\mathbf{a}$	Asymmetry parameter
b	Bias
J	Binary cross entropy
BW	Binary Image
$\mathbf{u}_i$	Boundary region of the neighbour of radius
$B(e, s_{\text{xbox}})$	Bounding box
δ_c	Cone Sector
ϕ	Cumulative distribution function
k	Current iteration of object/value
$\updownarrow$	Current layer
$\lambda_{\text{dirdecay}}$	Decay factor
$\oplus$	Dilation operator
D	Dilation
$\ominus$	Erosion operator
d	Distance
E	Erosion
EucD	Euclidean distance

Notation	Description
$\mathbb{L}$	Line shift
$\mathbf{Z}$	Linear combinations of linear features
w_L	Linear weight vector
ρ	Local distance parameter
LMP	Local Maxima Points
θ_p	Local orientation
M	Magnitude
$\mathbb{M}$	Marker
$\mathbf{m}$	Mask
M_G	Maximum Gabor image
μ	Mean
$\mathbf{v}$	Neighbouring point
$\boldsymbol{\delta}_n$	New direction
$\mathcal{N}$	Network
$NucP$	Nucleation Points
N	Number
N_c	Number of Columns
$\mathbf{K}$	Number of neighbours
N_r	Number of Rows
$\mathbb{I}'$	Output image
ϕ	Phase parameter
p	Pixel
x,y	Pixel Co-ordinates
$\hat{y}$	Predicted class
P	Probability
$\hat{L}_x$	Probability of class x
$\propto$	Proportional to

Notation	Description
r	Radius
$\mathbb{R}$	Real number
$\mathbf{R}$	Regions
r_b	Response of convolved line
Sat	Saturation
$\mathbb{S}$	Scale Space
i,j	Secondary pixel co-ordinates
g"	Second order Gaussian derivative
S	Set
S_p	Set of pixels
S_{ort}	Set of pixels that are perpendicular to the local orientation
S_{sym}	Set of symmetric pixels
skel	Skeletonisation
s_{xbox}	Size of box
ζ	Slack variable
ϵ	Small constant
SE	Structural element
T	Threshold
t	Time
Tree	Tree in Random Forest
L_x	True label
$\hat{\mathbf{d}}_{n+1}$	Updated direction vector
Val	Value
σ	Variance
$\mathbf{z}$	Vector of outputs
$\mathcal{V}$	Vertices
$\mathbf{v}_n^k$	Vertex identifiers
V	Voronoi
λ_n	Wavelength
w	Weight
∇_x	x-direction partial derivative
∇_y	y-direction partial derivative
$\mathbf{e}$	3D vector co-ordinates

Abbreviations

Abbreviations	Description
A	Accuracy
AFP	Alpha-Fetoprotein
BCB	Breast Cancer Biopsy
BRCA	Breast Cancer
CNN	Convolutional Neural Network
CRM	Confocal Reflectance Microscopy
CT	Computed Tomography
CTF	CT Fire
D	Dice Score
DIF	Distal Invasive Front
DME	Disease Mimicking ECM
DNA	Deoxyribonucleic Acid
DnCnn	Deconvolutional Convolutional Neural Network
ECM	Extracellular Matrix
ED	Edge Detection
ER	Estrogen Receptors
FC	Fluorescent Collagen
FF	Fluorescent Fibronnectin
FP	False Positive
FN	False Negative
GB	Graph Based
GT	Ground Truth
HPF	High Pass Filter
HSV	Hue, Saturation, Value
IF	Invasive Front
J	Jaccard Index

Abbreviations	Description
KNN	K Nearest Neighbour
LSTM	Long Short Term Memory
mask-RCNN	Region-Based Convolutional Neural Network
$MgCl_2$	Magnesium Chloride
MHSA	Multi Head Self Attention
MinAxMajAx	Aspect Ratio
MLP	Multi Layer Perceptron
MRI	Magnetic Resonance Imaging
mRNA	Messenger Ribonucleic Acid
PBS	Phosphate Buffered Solution
PI	Protease Inhibitors
PIF	Proximal Invasive Front
PR	Progesterone Receptors
PRISMA	Preffered Reporting Items for Systematic and Meta Analyses Reviews
PRISMA-ScR	PRISMA and Scoping Reviews
RBD	Ras Binding Domain
RBF	Radial Basis Function
ResNet	Residual Neural Network
RGB	Red, Green, Blue
ROI	Region Of Interest
SGDM	Stochastic Gradient Descent Momentum
SS	Scale Space
SVM	Support Vector Machines
TB	Tumour Body
TMA	Tissue Micro Array
TN	True Negative
TP	True Positive
TR	Trace Ridges
TW	Twombli
UMAP	Uniform Manifold Approximation and Projection
URL	Uniform Resource Locator
ViT	Vision Transformers
WT	Wild-Type

List of Figures

1.1 The structure of the thesis chapters. 6

2.1 Eukaryotic Animal Cell. 9
2.2 Hairy Cell Leukaemia stained with haematoxylin and eosin from a bone marrow biopsy. 9
2.3 Cell Membrane. 11
2.4 Cell Membrane Receptors. 11
2.5 Signal Transduction pathways. 12
2.6 Structure of the Extracellular Matrix. 15
2.7 How cancer arises 19
2.8 Cancer cell compared to normal cell. 27
2.9 PRISMA statement flowchart showing the details of the scoping review. Inclusion in the study was given when specific keywords were on PubMed for individual searches. Studies were not read individually. 30
2.10 Number of entries in PubMed of the data analysis techniques specified by the keywords. 33
2.11 Number of entries in PubMed of the specific keywords and the date of publication. 34
2.12 Relative number of entries in PubMed per keyword and date of publication. . 35
2.13 Number of entries in PubMed per specific keyword of all Machine Learning Techniques. 36
2.14 Number of entries in PubMed per year and each keyword of the Machine Learning Techniques. Most entries stayed the same but "Deep learning" has increased constantly. The vertical axis is logarithmic. 37
2.15 Number of entries in PubMed per specific keyword of the Deep Learning Techniques and architectures like AlexNet, VGG and U-Net. 37
2.16 Number of entries in PubMed for eight programming languages: Fiji, ImageJ, Matlab, R, Python, QuPath, ICY, CellProfiler and C. 38

3.1 Histopathology of well differentiated neuroendocrine tumour with salt and pepper chromatin. 41
3.2 Tissue Micro Array containing 120 cores of Normal Breast Tissue. 42
3.3 Limitations of the human eye. 44
3.4 Illustration of a simple microscope. 45

3.5 Compound Microscope 46
3.6 Köhler's illumination. 47
3.7 Illustration of Stoke's Shift. 49
3.8 Illustration of a Fluorescence microscope. 50
3.9 Illustration of FISH. 51
3.10 Datasets example: Four datasets are used throughout this thesis. 52
3.11 Dataset 1 example: Fluorescent microscopy image of a mutant (RBD) cell culture with 3 layers: DAPI, Phalloidin, Fibronectin. 54
3.12 Dataset 1: RBD images with DAPI, Phalloidin and Fibronectin. 55
3.13 Dataset 1: WT images with DAPI, Phalloidin and Fibronectin. 56
3.14 Dataset 2: Various images of extracellular matrix fibres from different microscopical methods. 57
3.15 Dataset 3: Increasing concentration of collagen, various images of the extracellular matrix collagen. 59
3.16 Dataset 3: Marangoni Flows, various images of the extracellular matrix collagen. 60
3.17 Dataset 3: Protease Inhibitors, various images of the extracellular matrix collagen. 62
3.18 Tissue Micro Array of tumour breast cancer tissue comprising of 120 cores. . 63
3.19 Normal Cores: Fibres and cells from cores of the original tissue micro arrays and the cores with only fibres. 65
3.20 Tumour Cores: Fibres and cells from cores of the original tissue micro arrays and the cores with only fibres. 66
3.21 Grayscale Image with size $N_r \times N_c$. 68
3.22 RGB Image with size $N_r \times N_c \times 3$ with three colour channels of red, green and, blue. 69
3.23 HSV Image with size $N_r \times N_c \times 3$ with three channels which are hue, saturation, value. 70
3.24 Image processing pipeline. 71
3.25 Illustration of Gaussian filtering on the basic fibre image. 73
3.26 Illustration of an averaging 3×3 kernel on a representative image. 74
3.27 Illustration of Mathematical Morphological operations and Otsu thresholding on a sample fibre image. 80
3.28 Illustration of labelling on the fibre image. 81
3.29 Illustration of Area opening on the fibre image. 82
3.30 Illustration of the Euclidean distance transform on the fibre image. 83
3.31 Illustration of the skeletonisation of the fibre-like structure. 84
3.32 Application of Voronoi partitioning and Delaunay Triangulation on segmented DAPI cells. 85
3.33 Illustration of segmentation of cells with CellPose on top row figure of (Fig. 3.19a). 87
3.34 Edge Detection pipeline follows: 90
3.35 Application of Edge Detection the fluorescently labelled collagen image (Fig. 3.14a). 90
3.36 Scale Space pipeline follows. 93

3.37 Application of Scale Space levels with increasing Gaussian filtering on (Fig. 3.14a). 94
3.38 Twombli pipeline follows. 99
3.39 Illustration of the processing of one image on (Fig. 3.14a) with Twombli. . . 100
3.40 CT Fire pipeline follows. 104
3.41 CT Fire algorithm applied on the FC image (Fig. 3.14a). 105
3.42 Graph Based pipeline follows. 108
3.43 Graph Based algorithm applied on the FC image (Fig. 3.14a). 109
3.44 Illustration of the U-Net architecture used to segment the fibres. 116
3.45 Illustration of the data augmentation methods used to train the U-Net. . . . 117
3.46 Illustration of the workflow of training and testing the U-Net models available on GitHub. 118
3.47 The VGG16 architecture. 121
3.48 Vision Transformer architecture. 123
3.49 Decision Trees representation. 125
3.50 Gini Impurity: Three tanks with stars of different colours to show the impurity levels. 126
3.51 Ground Truth generation 137
3.52 Ground Truth of dataset 2 (Fig. 3.14): Region of interest of original images and hand delineated tracing 138
3.53 Visualisation of three cases of two lines where the pixel-wise metrics of true positives, true negatives, false positives, false negatives are identical, but the distance between the lines is different. 141
3.54 Visualisation of the distance error measurement between the two lines of (Fig. 3.53c). 142
3.55 Illustration of the calculation of the error distance maps on the fluorescently labeled collagen image shown in (Fig. 3.14a). 143
3.56 Illustration of the Distance Metrics. 144

4.1 Processing of the DAPI channel to segment the cell nuclei 151
4.2 Processing of the DAPI channel to segment the cell nuclei continued 152
4.3 Pixel-Pixel Segmentation Comparison. 152
4.4 Phalloidin actin filaments island segmentation. 154
4.5 Phalloidin actin filaments island segmentation continued. 155
4.6 DAPI and Phalloidin combination. 156
4.7 DAPI and Phalloidin combination of all images. 157
4.8 Fibronectin Analysis. 158
4.9 Pixel-Pixel Segmentation of the Central Cells. The top five images are those of the RBD and bottom five images are those of the WT. 160
4.10 Boxplot illustration when comparing mean cell properties. 161
4.11 Boxplot illustration when comparing the standard deviation of the cell properties. 162
4.12 Phalloidin island morphological properties. 163
4.13 Number of Phalloidin Islands. 164

4.14 Boxplot illustration of the comparison between RBD and WT when comparing the Phalloidin islands metrics 164
4.15 DAPI and Phalloidin combination morphological properties. 165
4.16 Boxplot illustrations of the comparison between RBD and WT when observing the combined cell and Phalloidin islands metrics. 166
4.17 Fibronectin Edges Measurement & Major Axis Length. 167
4.18 Fibronectin Relative Gap Area Measurement. 167
4.19 Boxplot illustration of the comparison between RBD and WT when observing the differences in the fibronectin. 168
4.20 Illustration of the TMA cores preprocessing step to separate the cells from the fibres to be traced and segmented. 173

5.1 Illustration of the Trace Ridges methodology. 180
5.2 Edge Detection intensity profile and derivatives 181
5.3 Illustration of the Watershed and Thresholding methods. 182
5.4 Trace Ridges pipeline follows. 185
5.5 Trace Ridges algorithm steps to trace an image of fibre-like structures (Fig. 3.14a). Image used is the FC image from Dataset 2. 186
5.6 Trace Ridges algorithm steps to trace an image of fibre-like structures continued. 187
5.7 Images used in objective comparison, Dataset 2: Various images of extracellular matrix fibres from different microscopical methods. 189
5.8 The effect of applying a high pass filter to the DME, FC, FF images. 190
5.9 Comparison of tracing algorithms on FC image (full view). 198
5.10 Comparison of tracing algorithms on the FC image (region of interest). . . . 199
5.11 Comparison of tracing algorithms on FF image (full view). 201
5.12 Comparison of tracing algorithms on FF image (region of interest). 202
5.13 Comparison of tracing algorithms on Picrosirius red Breast Cancer Biopsy image (full view). 204
5.14 Comparison of tracing algorithms on Picrosirius red Breast Cancer Biopsy image (region of interest). 205
5.15 Comparison of tracing algorithms on DME image (full view). 207
5.16 Comparison of tracing algorithms on DME image (region of interest). 208
5.17 Implementation pipeline of Trace Ridges. 216
5.18 Processing the FF images through the Trace Ridges pipeline. 217
5.19 Mutant Fluorescent Fibronectin Fibres and Gaps detected by Trace Ridges. . 218
5.20 WT Fluorescent Fibronectin Fibres and Gaps detected by Trace Ridges. . . 219
5.21 Boxplots of RBD and WT morphological properties. 220
5.22 Scatter Plot of Gap size and the number of fibres. 221
5.23 Control samples of Collagen I before the application of Protease Inhibitors. . 223
5.24 Collagen I fibres after the addition of the Protease Inhibitors. 224
5.25 Boxplots of comparisons of control samples of Collagen and Collagen with the addition of Protease Inhibitors (PI). 225
5.26 Collagen I fibres at the concentration of 1.7 mg/ml. 227

5.27 Collagen I at the concentration of 6.6 mg/ml. 228
5.28 Boxplots of comparisons of the collagen morphology when the ECM concentration is increased from 1.7 mg/ml to 6.6 mg/ml. 229
5.29 Control samples of Collagen I before the application of $MgCl_2$. 231
5.30 Collagen I fibres after the addition of $MgCl_2$. 232
5.31 Boxplots of comparisons of Collagen fibres architecture and organisation with the addition of $MgCl_2$. 233
5.32 Examples of images used in the investigations of Tissue Micro-Arrays. . . . 235
5.33 Fibres and Gaps of the Normal TMA cores detected by Trace Ridges for Investigation 1. 237
5.34 Fibres and Gaps of the Tumour TMA cores detected by Trace Ridges for Investigation 1. 238
5.35 Boxplots of morphological metrics that showed statistical significance when comparing the Normal to Tumour. 239
5.36 Cells of the Normal TMA cores detected by CellPose. 242
5.37 Boxplots of cell nuclei metrics when comparing Normal and Tumour cell nuclei.243
5.38 UMAP representation of fibres and cells of TMA cores. 245
5.39 Confusion Matrices of the classification tasks on dataset 4a. 249
5.40 Confusion Matrices of the classification tasks on dataset 4a continued. 250
5.41 Confusion Matrices of the classification tasks on dataset 4b. 253
5.42 Confusion Matrices of the classification tasks on dataset 4b continued. . . . 254
5.43 Boxplots of statistically different morphological metrics when comparing normal and tumour populations observed with ER and CERB2. 256
5.44 Boxplots of statistically different morphological metrics when comparing Normal and tumour populations observed with PR. 256
5.45 Boxplots of statistically different morphological metrics when comparing Normal and tumour populations observed with lymph node involvement. 257

B.0.1 Boxplots of morphological metrics that showed statistical significance when comparing the Normal to Tumour. 311
B.0.2 Boxplots of morphological metrics that showed statistical significance when comparing the Normal to Tumour. 312

C.0.1 Boxplots of increasing cell nuclei metrics when comparing Normal and Tumour cell nuclei. 315

List of Tables

2.1 Cancer types and their descriptions. 21
2.2 Specific keywords used in the search strategy to explore Image Processing and Computer Vision, the Machine Learning and Deep Learning Techniques. . . 31

3.1 Outline of datasets. 53
3.2 U-Net Model Training 115

4.1 DAPI segmentation measurements of Accuracy, Jaccard Index and Dice Score. 159

5.1 Quantitative Comparison by total error of distance in pixels. 192
5.2 Quantitative Comparison by average error of distance in pixels. 193
5.3 Quantitative Comparison by maximum error of distance in pixels. 195
5.4 Quantitative Comparison by time taken in seconds. 196
5.5 Quantitative Comparison by total error of distance in pixels with the addition of high pass filtering. 210
5.6 Quantitative Comparison by average error of distance in pixels with the addition of high pass filtering. 211
5.7 Quantitative Comparison by maximum error of distance in pixels with the addition of high pass filtering. 212
5.8 Quantitative Comparison by time taken in seconds with the addition of high pass filtering. 213
5.9 Number of feature sets and images used for training and testing for the classification task of investigation 1. 246
5.10 Classification metrics of machine learning and deep learning algorithm on images of investigation 1. 247
5.11 Number of feature sets and images used for training and testing for the classification task of investigation 2. 251
5.12 Classification metrics of machine learning and deep learning algorithm on images of investigation 2. 252

A.1 Morphological analysis on the five sets of mutant RBD and WT images. . . . 310

Contents

1 Introduction **1**
- 1.1 Overview: . . . 1
- 1.2 Motivation and Relevance: . . . 3
- 1.3 Problem Statement: . . . 3
- 1.4 Aims & Objectives . . . 4
- 1.5 Contributions: . . . 5
- 1.6 Thesis Structure: . . . 5

2 Biological Background: **8**
- 2.1 Overview: . . . 8
- 2.2 Cell and Cell Structure: . . . 8
  - 2.2.1 Cell Morphology: . . . 12
  - 2.2.2 Cytoskeleton Image Analysis: . . . 13
- 2.3 Extracellular Matrix: . . . 14
  - 2.3.1 Fibre Morphology: . . . 16
  - 2.3.2 Image Analysis of the ECM: . . . 16
- 2.4 Cancer: . . . 16
  - 2.4.1 How Cancer Arises: . . . 18
  - 2.4.2 Cancer from a Molecular Perspective: . . . 18
  - 2.4.3 Risk Factors: . . . 20
  - 2.4.4 Breast Cancer: . . . 21
  - 2.4.5 Cancer Diagnosis: . . . 23
  - 2.4.6 Cancer Treatment: . . . 24
  - 2.4.7 Image Analysis in Cancer: . . . 26
- 2.5 Scoping Review: . . . 28
- 2.6 Summary: . . . 39

3 Methods & Materials **40**
- 3.1 Overview: . . . 40
- 3.2 Histopathology: . . . 40
  - 3.2.1 Tissue Microarray (TMA): . . . 41
  - 3.2.2 Staining Protocols: . . . 42
  - 3.2.3 Microscopy: . . . 43

3.3 Datasets: . . . 52
3.3.1 Dataset 1: . . . 52
3.3.2 Dataset 2: . . . 54
3.3.3 Dataset 3: . . . 57
3.3.4 Dataset 4: . . . 63
3.4 What is an Image: . . . 67
3.4.1 Grayscale Images: . . . 67
3.4.2 Colour Images: . . . 67
3.4.3 HSV Images: . . . 68
3.4.4 Binary Images: . . . 70
3.5 Image Processing Techniques: . . . 71
3.5.1 Filtering: . . . 71
3.5.2 Segmentation: . . . 76
3.5.3 Tracing Algorithms: . . . 87
3.6 Classification Approaches: . . . 119
3.6.1 VGG-16: . . . 119
3.6.2 Tiny Vision Transformer (ViT): . . . 121
3.6.3 Decision Trees: . . . 123
3.6.4 Naive Bayes: . . . 126
3.6.5 Random Forest: . . . 128
3.6.6 Support Vector Machines: . . . 129
3.6.7 Logistic Regression: . . . 131
3.6.8 K Nearest Neighbour: . . . 133
3.6.9 Multi-Layer Perceptron: . . . 134
3.7 Visualisation Technique: . . . 135
3.7.1 UMAP: . . . 135
3.8 Metrics: . . . 136
3.8.1 Cell Morphological Metrics: . . . 136
3.8.2 Ground Truth: . . . 137
3.8.3 Pixel-wise Metrics: . . . 138
3.8.4 Fibre Metrics: . . . 139
3.8.5 Object-wise Metrics: . . . 146
3.9 Summary: . . . 147

4 Exploration of Data: 149
4.1 Overview: . . . 149
4.2 DAPI Stained Images: . . . 149
4.2.1 Segmentation: . . . 149
4.2.2 Ground Truth Segmentation Comparison: . . . 150
4.3 Phalloidin Stained Images: . . . 153
4.4 DAPI + Phalloidin Combinatory Properties: . . . 156
4.5 Fibronectin Images: . . . 158
4.6 Segmentation Performance Results & Morphological Metrics Comparisons: . 159

4.6.1 DAPI Images: . . . 159
4.6.2 Phalloidin Images: . . . 163
4.6.3 DAPI & Phalloidin Combination: . . . 164
4.6.4 Fibronectin Images: . . . 166
4.7 Image Preprocessing Pipelines: . . . 170
4.7.1 No Preprocessing: . . . 170
4.7.2 Fluorescent Fibronectin Images: . . . 170
4.7.3 Increasing Collagen, Protease Inhibited, Marangoni Flow images: . . . 170
4.7.4 TMA Cores: . . . 170
4.8 Summary: . . . 174

5 Results: 175
5.1 Overview: . . . 175
5.2 Trace Ridges: . . . 175
5.2.1 Algorithm Steps: . . . 175
5.2.2 Watershed Transform: . . . 176
5.2.3 Edge Detection: . . . 177
5.2.4 Watershed Combined with Edge Detection: . . . 178
5.2.5 Tracing clean-up: . . . 179
5.2.6 Missed ridges . . . 183
5.2.7 Ridges Combination: . . . 184
5.2.8 Trace Ridges Implementation: . . . 184
5.3 Tracing Algorithms Comparison: . . . 188
5.3.1 Introduction: . . . 188
5.3.2 Quantitative comparison: . . . 191
5.3.3 Visual Assessment: . . . 196
5.3.4 High Pass Filter investigation: . . . 209
5.3.5 Comparison Summary: . . . 213
5.4 Modified Cell Line Investigation: . . . 215
5.5 Matrix Mechano-sensing: . . . 221
5.5.1 Protease Inhibitors: . . . 222
5.5.2 Increasing Collagen Concentration: . . . 226
5.5.3 Marangoni Flows: . . . 230
5.5.4 Conclusion: . . . 234
5.6 Evaluation of Morphological Properties and Machine Learning Approaches for Tissue Micro-array Image Classification: . . . 235
5.6.1 Investigation 1 Metrics: . . . 235
5.6.2 UMAP Representation: . . . 245
5.6.3 Investigation 1 classification: . . . 245
5.6.4 Investigation 2 classification: . . . 251
5.6.5 Investigation 1 Clinical Information: . . . 255
5.6.6 Conclusions: . . . 257
5.7 Summary: . . . 259

6 Conclusion **261**
6.1 Summary 261
6.2 Major Contributions: 262
6.3 Major Conclusions: 263
6.4 Discussion & Reflections: 264
6.5 Limitations: 267
6.6 Future Work: 268

Appendices **309**

A RBD and WT Morphological Metrics: **310**

B Significant fibre morphological features: **311**

C Significant cell nuclei morphological features: **314**

D Previous publications: **316**

Chapter 1

Introduction

1.1 Overview:

Image analysis is the process of extracting useful information from digital images. Commonly, the images are observed and investigated by humans with their vision and brain power. This has been the case in many of the important discoveries from micro-organisms and cell structures seen by Hooke and Leeuwenhoek [1] to the findings of the retinal arrangements and neural networks by Santiago Ramon y Cajal [2]. With advancements in the technological fields, images are now recorded digitally, and it has now become extremely easy to acquire more sophisticated, high quality large datasets for processing. These images have become difficult for just one person to process in a reasonable time. To address the limitation of a single observer analysing an image, a distributed approach where multiple individuals perform part of the analysis has been called "crowdsourcing" or "citizen science" has been used with various levels of success [3]. However, the problems of inter- and intra-observer variability [4] remain and can even be amplified as in the cases of citizen science, the observers are not always experts in the area. Furthermore, some images may contain characteristics and features that are far too subtle to be detected and recognised by the human eyes as well as our eyes may deceive us. Therefore, the application of image analysis and image processing techniques may help disclose hidden features that would have remained undiscovered. Furthermore, an algorithm will provide consistent reproducible results, even for very large amount of data. With the current computing power of advanced hardware such as CPUs, GPUs and RAM, multiple images can be processed simultaneously and efficiently for example by using parallel processing with GPUs, therefore the analysis of images can be done faster and with reproducible results than if a human expert were to process a set of images.

Cancer is a disease where cells in the body grow abnormally and uncontrollably and could spread to other parts of the body. In a normal healthy body, cells live through a cycle known as cell cycle. Cells grow and multiply to help the body with its processes and eventually cells that are old or damaged die. Along the journey, the process can become altered as the damaged cells grow and multiply. These clusters of damaged cells can form

tumours and tissue lumps, these could be cancerous these cells are also known as malignant cells [5, 6]. Cells from tumours can then spread to nearby tissues and invade as well as travel to further parts of the body and form new tumours, in a process known as metastasis. Cancerous tumours are known as malignant tumours while non-cancerous tumours are known as benign tumours and these tumours do not spread and invade other tissue of the body [5, 7].

The way cancer tumours can thrive in the body is through a series of molecular, biochemical, and cellular capabilities commonly known as traits or hallmarks of cancer [5, 8, 9, 10]. Cancer cells are able to divide by producing their own growth signals [11]. They are also able to ignore signals that usually tell the cells to stop growing and developing by evading the growth suppressors [12]. Cancer cells are able to evade programmed cell death, which consequently allows them to survive even when damaged [13]. Cancer cells are able to divide indefinitely by stabilising telomeres [14]. They can instruct the blood vessels to supply them with blood which provides them with the necessary oxygen and nutrients to thrive [15]. Cancer cells are able to spread from their original location and into other organs [16]. They can deceive and hide from the immune system; they deceive it by making the immune system protect it instead of attacking it [17]. Cancerous cells can produce energy which is vastly different to the energy produced by normal cells therefore they can grow faster [18].

Cancer is a genetic disease that could also be caused due to environmental factors. Cancer can be the result of damage to the DNA caused by environmental factors such as smoking tobacco [19] as well as ultraviolet rays from the sun [20]. Cancer can also be caused by inheritance from parents and errors in the cell division process [7]. Estimates indicate that approximately one in five men or women develop cancer in a lifetime, while around one in nine men and one in 12 women die from it [21].

The morphological characteristics of cells and their surrounding have been observed and analysed with interest in Cancer [22, 23, 24] and other conditions [25]. The extracellular matrix (ECM) which is an interstitial area surrounding cells has shown profound influence in the development and progression of diseases [26] and in particular cancer [27, 28, 29, 30]. The ECM is a driving force and a crucial factor for the essential structural supports of the tissue [31], where each organ in the human body contains a different ECM structure that provides essential purposeful support. The ECM is produced by fibroblasts, and they contribute to the dynamic remodeling of the ECM by producing both structural ECM components and proteinases like MMPs which are a network of proteins that plays an important role in processes like tissue repair and development [32]. These enzymes allow the ECM to continuously reshape itself, which is crucial. The ECM has been understudied [33] in comparison to other components of the micro-environment like macrophages, vasculature, or soluble factors. This is due to its intricate 3D network [34, 35] which compromises of many elements such as enzymes, fibronectin, glycoproteins, and collagen.

Whilst analysing the cell nuclei and ECM structures visually can reveal important features as commonly done by pathologists, it is time consuming and often difficult to compare two populations. Structural characteristics for both the cell nuclei and ECM such as length, shape, distribution, and orientation can reveal hidden valuable information on the cellular

level environment when the body is subjected to a disease or a treatment [36, 37]. This promotes further understanding of the effects of the disease and its progression or its treatment effects and can assist in tasks such as early detection and diagnosis. Many manual, semi-automatic and automatic algorithms and methods have been introduced to segment cells [38, 39, 40] and delineate and trace elongated structures of the ECM [41, 42]. Quantification of the cell nuclei and ECM fibres and their characteristics is essential and for this objective a robust cell detection and fibre tracing algorithm is necessary. In addition to this, mathematical and statistical processes must be implemented to compare the extracted features.

The work in this thesis focuses on exploring established methods of segmentation and tracing of medical and biomedical images to extract meaningful findings from structures of the cell nuclei and the extracellular matrix fibres. In addition to investigating established methods, work presented in this thesis aims to find areas of improvement as well as develop a state-of-the-art algorithm that outperforms established methods. In this thesis a statistical comparison of populations of data and their extracted features is conducted for example. samples of cells or extracellular matrix fibres before and after treatment as well as cells and fibres in normal and tumour samples. Thus, extracting useful information that would assist in the diagnosis of patients and shed light on the key differences between the populations.

1.2 Motivation and Relevance:

Cancer is a disease that affects many people [43] and has been widely studied and is continued to be researched, therefore it is crucial to understand the causes of cancer. The extracellular matrix is understudied [33] yet it plays a crucial role in the progression of disease and in specific cancer. Morphological features have proven to be a source of hidden information of biological structures. Understanding the morphological properties of the cell and its components and the extracellular matrix may reveal hidden information that would have otherwise remained hidden to the human eye. Morphological properties provide a clear picture on the progression of disease and treatment effects.

1.3 Problem Statement:

Many algorithms exist that can be used to trace extracellular matrix fibres. Some of these methods are more general and can be used for other tasks such as Edge Detection [44] and Scale Space [45, 46] while other methods are specific for fibre tracing such as Twombli [47], CT Fire [48] and Graph Based [49, 50]. These algorithms tend to exhibit a high computational time to process just one image, whilst the main objective is to process many images at the same time and faster. Furthermore, all these algorithms require parameter hyper tuning, while it is our aim to create an algorithm that requires no parameter input. The algorithm must be able to support the extraction of morphological features which is essential to create a profile of characteristics for the different populations of conditions. The

problem to be solved in this thesis is the slow computational speed and high distance errors of tracing algorithms.

1.4 Aims & Objectives

The primary aim of this thesis is to develop an algorithm to analyse and compare biological datasets on the basis of their morphological characteristics. In the case of this thesis, the focus was the fibre-like structures of the extracellular matrix of cancer-related datasets. The developed algorithm would segment fibre-like structures of fibronectin and collagen and produce reproducible results, where if it was performed by a human would not be guaranteed. The algorithm should also outperform state-of-the-art methods in terms of accuracy and speed. The algorithm will be used on cancer-related datasets, where fibres are segmented and morphological features are extracted to compare the populations, this is to further our understanding on the behaviour of cancer, a disease that affects millions of individuals yearly on the extracellular matrix. Work conducted in this thesis aims to provide a state-of-the-art method that provides a clearer picture on cancer in the hope to provide better treatment plans for patients.

To achieve this aim, key objectives must be met:

1. Explore image processing techniques and how commonly used they are in cancer research through conducting a scoping review that identifies the total observations of techniques in literature as well as current and past trends.

2. Determine which biological component (cell nuclei, actin filaments and ECM fibronectin fibres) from a set of cell culture images has the highest deterministic factors to differentiate between populations of the mutated Ras Binding Domain (RBD) and the Wild Type (WT) populations through experimenting with traditional image processing techniques to segment and trace and extract morphological features for comparison.

3. Develop a tracing algorithm that delineates fibre-like structures that outperforms current state-of-the-art methods.

4. Compare the performance of the established tracing algorithm against the performance of state-of-the-art methods to determine if the algorithm improves on current performances.

5. Implement the novel algorithm on cancer-related populations of the extracellular matrix fibres to extract meaningful and distinguishable morphological characteristics.

The objectives aim to answer questions about the role of the extracellular matrix in the progression of cancer and the effects of treatments.

1.5 Contributions:

The original contributions of this thesis are:

1. For the objectives above, a comprehensive scoping review of image analysis techniques in cancer research and the most common computing languages and tools used.

2. Objective comparison of tracing algorithms across three filtering options on images of different backgrounds.

3. Development of algorithm that traces fibrial structures and in particular fibres of the extracellular matrix published in PLOS One.

4. The identification and extraction of features from data populations, analysing tumour behaviour on the extracellular matrix, its modifications, and the effects of treatment.

1.6 Thesis Structure:

Chapter 1: The **Introduction** provides an overview of the topic of research and its motivation and defines the problem. The aims and objectives are outlined and a statement on the contributions of this thesis is stated. Finally, the structure of the thesis is outlined seen in (Fig. 1.1).

Chapter 2: Biological Background provides an overview of key mechanisms of cells and the extracellular matrix which the work of this thesis focuses on. Additionally, the role of image analysis in the context of cells and the extracellular matrix is discussed. Cancer, a major topic in this thesis is described and in particular breast cancer and its influence on the extracellular matrix, image analysis and its uses in cancer research is also described. Finally, a scoping review is presented that investigates image analysis techniques in cancer research and the computational languages used.

Chapter 3: Methods & Materials, In this chapter, the acquisition of images and the different techniques of microscopy observed in this thesis are presented. Datasets used throughout the thesis are presented. Dataset 1 contains a set of 10 fluorescent images, 5 for wild type and 5 for mutant cells. Each image contains three channels, one observing cell nuclei, one observing actin filaments and a channel observing fibronectin fibres. Dataset 2 contains four fibrial images of different origins and acquisitions that were used to compare the performance of tracing algorithms. Dataset 3 contains three smaller datasets. Firstly, images of standard bovine collagen I at a concentration of 1.7 mg/ml and standard bovine collagen I at a concentration of 6.6 mg/ml. The second sub dataset contains control images of collagen I and collagen I after the addition of $MgCl_2$ to observe Marangoni flows. The final sub dataset contains control images of collagen I and images of the addition of protease inhibitors. The final dataset contains two sub datasets compromised of tissue micro array cores. The datasets contain images of two populations which are normal breast tissue and tumour breast tissue. The fist sub dataset observes only the collagen fibres of those TMA

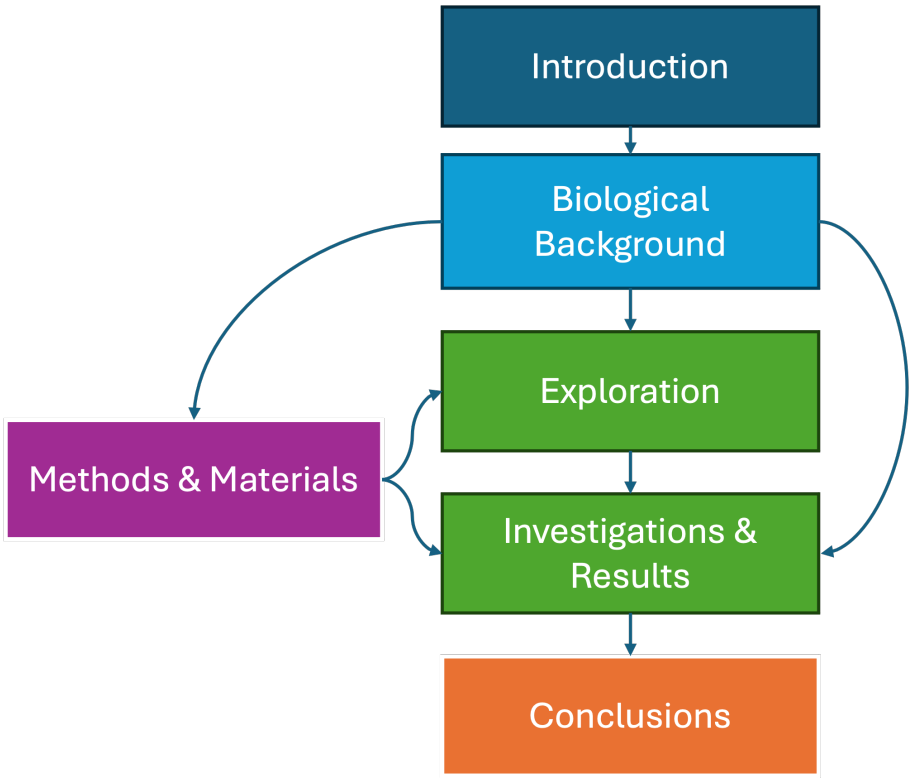


Figure 1.1: The structure of the thesis chapters. The thesis starts with chapter 1 the introduction, after which chapter 2 the background. The background feeds into chapter 3 the materials and chapter 4 the methods. Chapter 3 and chapter 4 provide the building blocks of chapter 5 the exploration and chapter 6 the investigations and the results. Finally, chapter 7 outlines the conclusions and future work.

cores whilst the second sub dataset observes both cell nuclei and collagen. Furthermore, an image is defined, and traditional image analysis techniques are explored. This is followed by cell segmentation techniques. Modern tracing algorithms are outlined, and the novel algorithm *Trace Ridges* is introduced. All the image processing techniques are explained and explored with a sample image. Classification techniques are introduced. Furthermore, distance metrics that measure distances between a trace and a ground truth are presented as well as classification metrics.

Chapter 4: Exploration of Data, in this chapter, images of the first dataset are explored and analysed with traditional image processing techniques to decide the direction of the thesis. It was chosen that the extracellular matrix is the preferred biological component to conduct further research on considering its impact in the development and progression of cancer as well as the lack of research in the field. The pre-processing of the tissue micro array cores and other pre-processing pipelines that utilise traditional image processing techniques.

Chapter 5: Results, this chapter contains the results of four experiments conducted for the purpose of this research but firstly, it introduces *Trace Ridges* our novel tracing approach.

The first experiment is the objective tracing algorithms comparison which is conducted to compare the novel algorithm to state-of-the-art techniques. The second experiment focuses on investigating the difference in morphological features between the fibronectin fibres of wild type cells and modified cell line cells. The third experiment investigates matrix mechanosensing, if the accumulation and abundance of collagen I promotes cytoskeletal remodelling by increasing collagen concentration, adding protease inhibitors, and observing the collagen fibres under Marangoni flows. The fourth experiment focuses on investigating the differences between normal and tumour breast tissue, first just the fibres then fibres and cell nuclei. Additionally, the features by the novel algorithm are trained and tested with classification methods including deep learning methods. Finally, patient clinical features are analysed.

Chapter 6: Conclusions, this chapter contains a thesis summary and a section outlining the major contributions of the thesis. This chapter also contains a section that outlines the major points learnt or major conclusions from this thesis. Moreover, there is a section on the discussion and reflections of the thesis, what went right and what went wrong is outlined as well as a discussion on the impact of this work. Furthermore, a section on the limitations faced throughout the work carried out is outlined. Finally, a section on future work and improvements that could have been done is presented.

Chapter 2

Biological Background:

2.1 Overview:

This chapter introduces key biological concepts and processes related to this research. The cell and its processes are outlined and an introduction of the microenvironment that surrounds these cells, the extracellular matrix (ECM) is presented. The ECM plays an important role in the development of this thesis. This chapter also introduces cancer, its diagnosis methods, and the role of image analysis in Cancer. Lastly, a scoping review is conducted to provide insight on the most popular image processing techniques in cancer research as well as the most used computational languages. This scoping review was previously published in MIUA 2022.

2.2 Cell and Cell Structure:

Cells are the smallest units that make up tissues of the body and are the basic component of all living organisms. All human cells can only be looked at through a microscope. The cell is made up of three important parts. The cell membrane, the cell cytoplasm, and the cell nucleus [51]. A cell and its features can be seen in (Fig. 2.1).

The cell nucleus [53] as illustrated in (Fig. 2.2) oversees and co-ordinates cellular activities (e.g. growth and metabolism) and gene expression [54]. The cell nucleus is separated from the rest of the cell by a double layer known as nuclear membrane. Nucleoli are the small bodies that are inside of the nucleus, which are responsible for ribosomes biogenesis. Components inside the nucleus are immersed in a gel like solution, the nucleoplasm [53]. As the nucleus contains an organism's genetic material, which controls gene expression [55], this decides the sequence of amino acids of proteins which are important for day-to-day functionality. Information in the Deoxyribonucleic acid (DNA) which is the genetic material is transcribed or copied, into messenger ribonucleic acid (mRNA) [56] molecules each of which encode the information for one protein. The mRNA molecules move through the nuclear envelope into the cytoplasm where they are translated.

The cell cytoplasm is the fluid that forms the rest of the cell. The cell cytoplasm acts

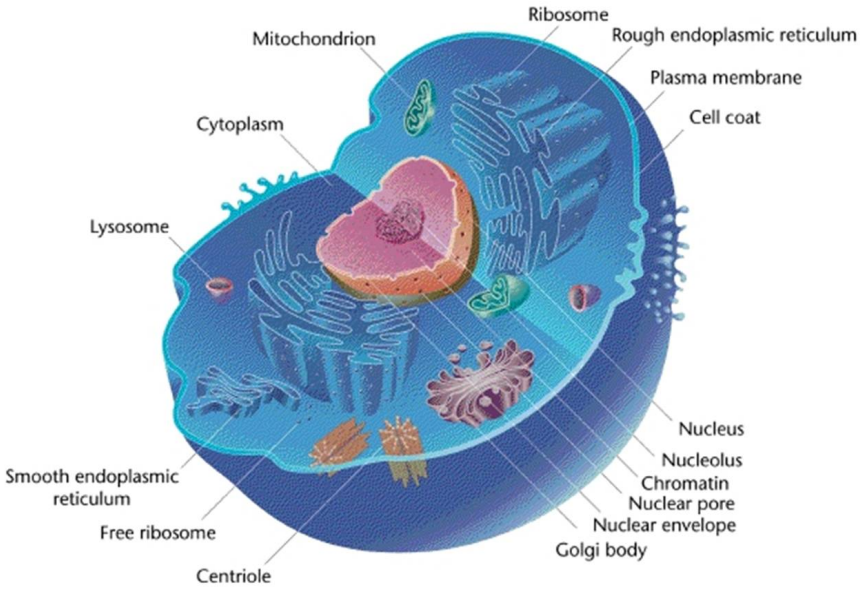


Figure 2.1: Eukaryotic Animal Cell. This cell contains the basic components that make up a cell. It is made up of the nucleus, the nucleolus, ribosomes, mitochondria, and the cell membrane [52]. This image is available through Wikimedica Commons.

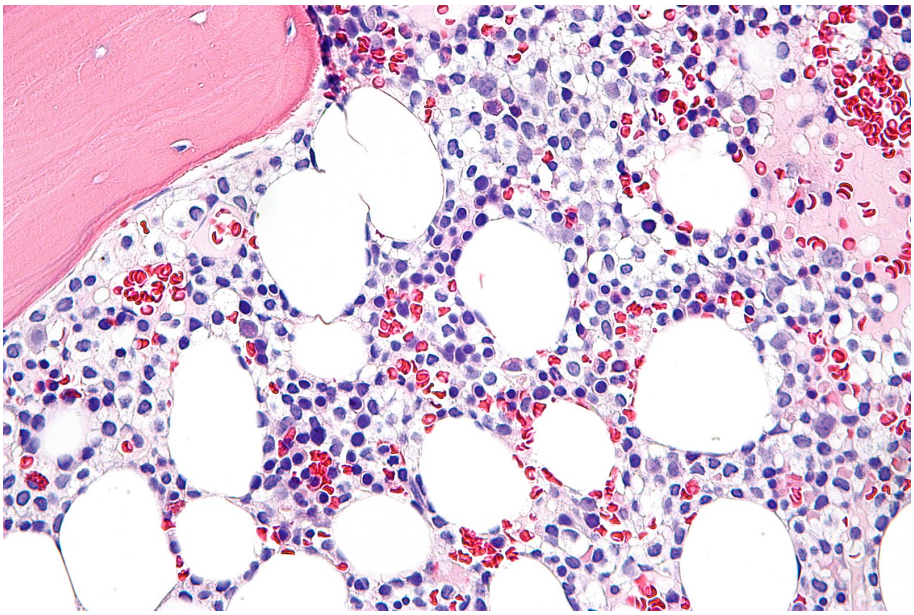


Figure 2.2: Hairy Cell Leukaemia stained with haematoxylin and eosin from a bone marrow biopsy. The cell nucleus appears in the colour purple. Cytoplasm is white and cell membrane is the thin lining encompassing the cytoplasm and the nucleus [57]. This image is available on Wikimedia Commons.

as the transfer medium for chemical reactions between the nucleus and what is outside of the nucleus. The cytoplasm plays a particularly key role in cell growth, expansion, and replication. The cell cytoplasm as illustrated in (Fig. 2.2), contains many important components known as organelles. These organelles are the endoplasmic reticulum, Golgi apparatus, mitochondria, ribosomes and the lysomes [58].

The cytoplasm also contains the cytoskeleton which is a network of protein fibres that provide structural support, shape, movement, and intracellular transport. The cytoskeleton is made up of three components the actin filaments, the intermediate filaments, and microtubules [59].

The most common cytoskeletal protein for most cells is actin [53, 60], which when polymerised makes up thin, flexible fibres known as actin filaments. Within the cell, actin filaments are bundled up to form three dimensional networks. The bundling of the actin filaments and their interactions with other cell structures such as the plasma membrane is controlled by different actin-binding proteins. Actin filaments are usually found in abundance below the cell membrane [61]. The filaments form a network that facilitates mechanical support, decides the cell shape, and facilitates the movement of the cell surface, this allows the cells to migrate and divide.

The cell membrane is the structure that encompasses the cell cytoplasm and the cell nucleus. The cell membrane acts as the transfer medium between the extracellular material outside of the cell and intracellular material inside of the cell. The cell membrane is formed of a phospholipid bilayer. The proteins on the cell membrane act as receptor sites, providing pathways for material to enter and exit the cell. They also act as carrier molecules and provide structural support [62]. An example of a cell membrane can be seen in (Fig. 2.3).

Cell membrane receptors [64] are proteins on the surface of the membrane which receive signals. Usually, a protein ligand binds to the cell receptor and a cellular reaction is ensued, for example changing cell morphology and many more other reactions. The ligands that bind to the cell membrane receptors are incapable of passing through by themselves. The interactions between the ligands and the cell membrane receptors result in signals being sent to the nucleus that alter cell functioning. The cell receptors on the cell membrane are specific to that type of cell therefore they do not allow any ligand to interact with the cell. The ligand's shape must perfectly fit onto the receptor's shape. The cell receptor is made up of three components. The first component is the intracellular component in charge of cascading a second messenger signal. The second component is the hydrophobic region which lines up the cell membrane. The third component is the extracellular ligand binding component. There are three types of cell membrane receptors, and these are GPCRs, Ion channel receptors and enzyme linked receptors [65]. An example of cell membrane receptors can be seen in (Fig. 2.4).

Cell signalling is the process through which the cell can receive and provide signals between the outside of the cell and inside of the cell [67]. Cell signalling can occur at any distance. Signal transduction, which is the cell signalling pathway, starts when a signal is transformed into a chemical cellular reaction, this can activate an ion channel, or it can also start a second messenger system that can move the signal through the cell. The second

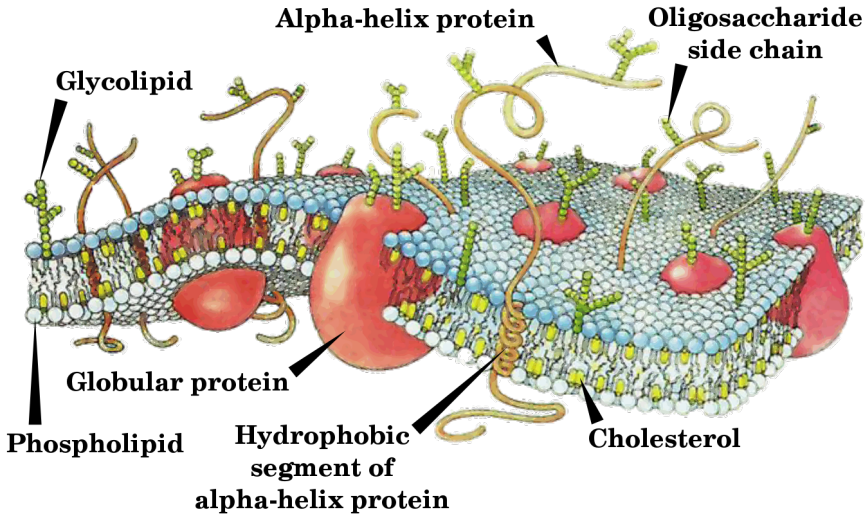


Figure 2.3: Cell Membrane. An illustration of the phospholipid bilayer of the cell membrane which facilitates the movement of molecules inside and outside of the cell with the help of the globular proteins external to the membrane [63] Image obtained from Wikimedia Commons.

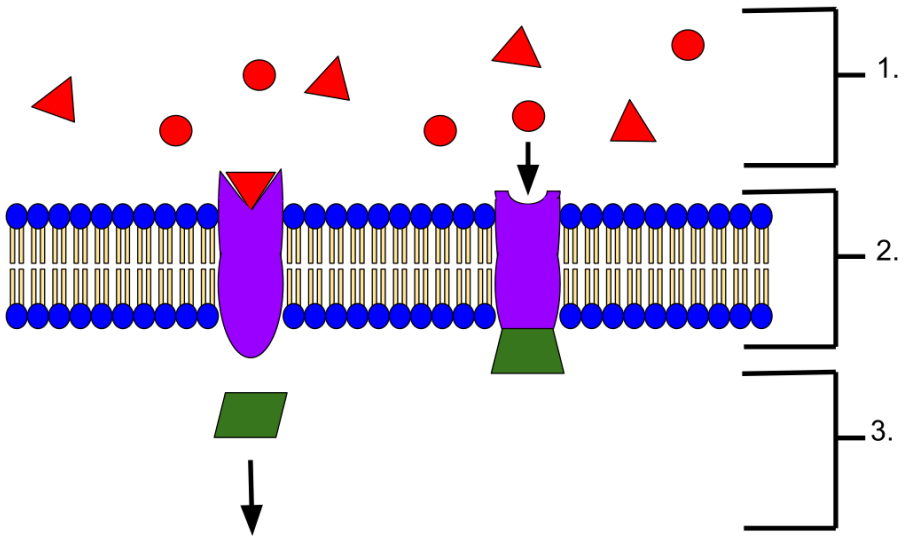


Figure 2.4: Cell Membrane Receptors, the enzyme that is attached to the receptors advances through multiple processes to be converted into a ligand. Step 1 illustrates the enzymes of different shapes; step 2 illustrates the phospholipid bilayer and the lock-like receptors that only accept an enzyme of a particular structure (key). Step 3 illustrates the conversion of the enzyme into a ligand shown in green [66]. Image obtained through Wikimedia Commons

messenger system can make the signal bigger and have a larger effect which can result in the initiation of many other receptors and reinforces the signal from the initial messenger. The initiation of the proteins is done by the removal or the addition of phosphate or other ions and molecular groups and these can act as messengers. There are five types of cells signalling and these are autocrine, juxtacrine, intracrine, pracrine and endocrine [68, 69]. An example of cell signalling pathways can be seen in (Fig. 2.5).

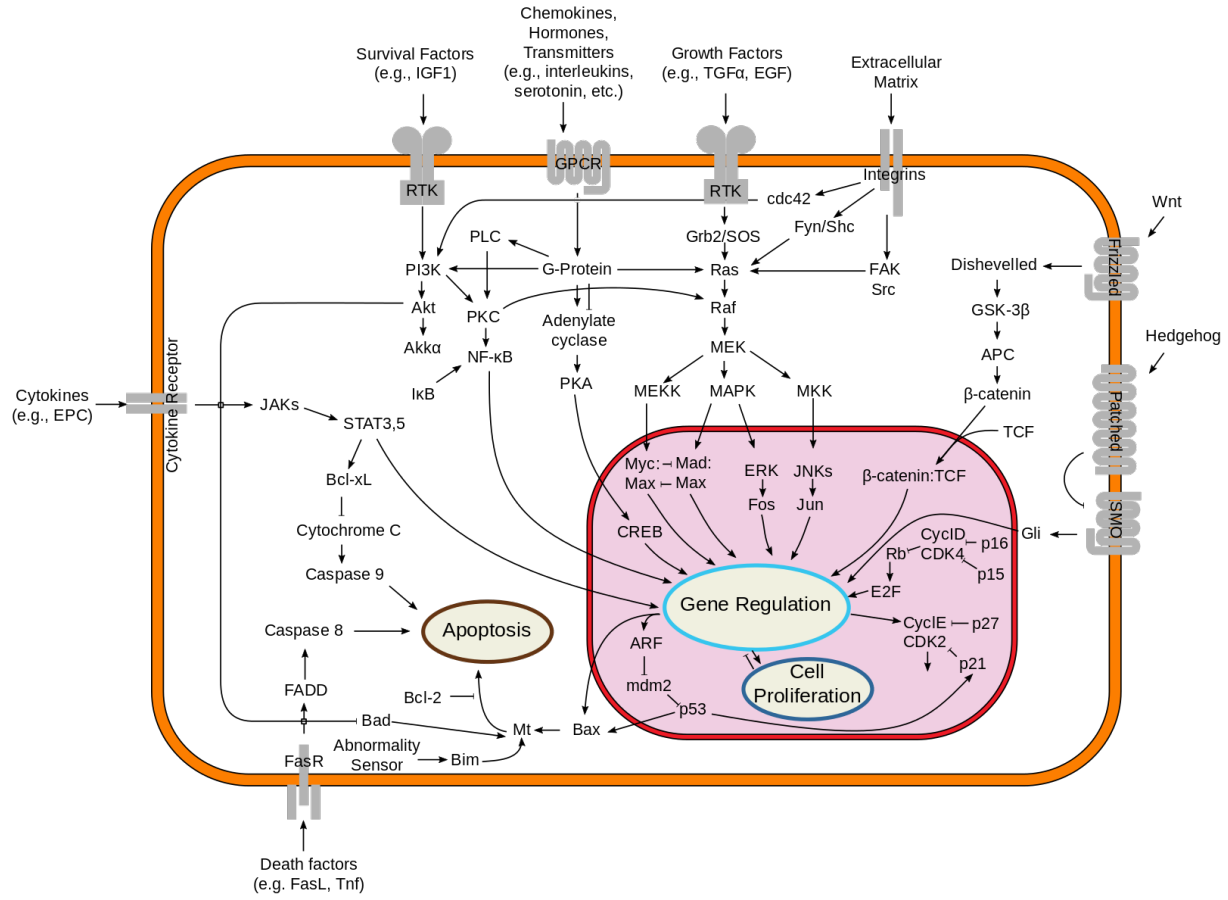


Figure 2.5: Signal Transduction pathways. This diagram illustrates the different enzymes that enter the cell through the cell membrane and the process of their conversion into different ligands. As well as the form at which they reach the cell nucleus [70]. This image is available through Wikimedia Commons.

2.2.1 Cell Morphology:

Cell morphology [71] plays a significant role in the analysis and in the detection of cancer as morphological changes of the nucleus have been detected in most cancers. Cells can regulate their own nuclear and cellular structures. Cell morphology refers to the size, shape, structure, and form of a cell. There is a plethora of information that can be extracted from

cell morphology such as circularity, solidity, intensity, orientation, area, and perimeter. It can also reveal information such as cell function and signal responses. These measurements are critical as they explain the behaviour of a cell and can be used to distinguish diverse types of cells from each other such as the mutated RBD and the normal wild-type cells. They can also be used to distinguish damaged cells. In the case of the RBD, the normal cell has been altered therefore with cell morphology profiling, it is easy to observe how the cells shape has changed and how the new cells behave in comparison to the WT cells. The morphological properties of cells have been used to describe conditions and effects of treatments previously [72, 73, 74], they have also been used compare difference between two populations of cells in [75].

Cell Image Analysis:

Traditionally, cells are observed by a human with a microscope, who conducts the analysis. As previously mentioned, this has its shortcomings, particularly "reproducibility". Two different observers may observe the same cell but come back with different conclusions, as our eyes may deceive us. With the advancements in technologies many algorithms now exist that segment, trace, track cellular objects and conduct morphological analysis that would produce reproducible findings [71]. Image analysis plays a great role in the analysis of cells and its structures.

Morphological cell analysis has been used in tasks such as automatic segmentation and analysis of histological tumour sections [76], [77], detecting the boundary of cervical cell nuclei [78], other biomedical cells in [72], [74], understanding drug influences on cells [79], [80], [81] as well as identifying the morphology of cells at different stages of the cell cycle [82]

2.2.2 Cytoskeleton Image Analysis:

Image processing has been used on the cytoskeleton in many cases in order to further our understanding of the actin filaments and microtubules under certain conditions. Authors in [83] validated a linear feature detection algorithm in order to quantify changes in the actin cytoskeleton organisation from confocal images. The algorithm was able to distinguish bundled actin filaments from diffuse staining. Authors in [84] utilised an image processing technique to segment and analyse actin stress fibres in 2D fluorescence images and they extracted their length, width, and orientation under drug treatment. Using a napari plug-in [85] authors were able to reveal how microtubules remodel in T-cell immunological synapses, where they discovered an unexpected pattern that is critical for vesicle transport during immune responses.

2.3 Extracellular Matrix:

The extracellular matrix (ECM) [31, 86, 34, 87] is a non-cellular component present in all tissues and organs. It provides support to the cellular components by acting as a scaffold. It also initiates important biochemical and biomechanical processes required for tissue morphogenesis, differentiation, and homeostasis [88]. The basic building blocks of the ECM are water, proteins, and polysaccharides, but each tissue has an ECM with a distinct composition and topology that is created during tissue development through the biochemical and biophysical interactions of the epithelial, fibroblast and other cellular components. Cell adhesion to the ECM is activated by the ECM receptors such as integrins. As the cells attach to the ECM, the ECM promotes cell migration. The ECM is constantly being remodelled, and its molecular components are liable to be modified. Every organ and tissue's physical and biochemical properties such as elasticity and tensile strength are controlled by the ECM [89]. The ECM is also in charge of morphological organisation by binding with growth factors and its interactions with cell surface receptors that initiate signal transduction and promote gene transcription.

Fibroblasts are cells that are found in connective tissue that produce and maintain ECM and create signalling cues through biophysical and biochemical cues [90].

Fibroblasts express certain genes encoding ECM proteins such as collagen, elastin, fibronectin [31]. mRNA is then transcribed and translated into polypeptides, after which it goes through critical post-translational modifications [91] such as hydroxylation, glycosylation, disulfide bond formation and crosslinking. Procollagen and proelastin precursors self-assemble within the cell or in the pericellular region [92]. Proteoglycans grow by enzymatic addition of GAG chains attached to core proteins [93]. These precursors are placed into vesicles which are then secreted through exocytosis into the extracellular environment. Often, fibronectin and laminin are polymerised at the cell surface through integrins [94]. Procollagen and proelastin are processed enzymatically to remove pro-sequences, after which they self-assemble into the mature collagen fibrils and elastin fibres. Fibronectin assembles into fibrils through integrin facilitated tension, which reveals the binding sites and forms the extracellular network [94]. The ECM matures into an organised, three-dimensional scaffold due to the interactions amongst the several proteins [95].

The main fibrous ECM proteins are collagen, elastin, fibronectin, and laminin. The interstitial space between the ECM proteins is filled with a hydrated gel that is the macromolecule proteoglycans.

Collagen [96] is the most common fibrous protein of the ECM and is the main structural element of the ECM. It provides tensile strength, promotes cell adhesion and migration as well as tissue development. Most of the collagen is created and secreted by fibroblasts in the stroma or nearby tissue. With the application of tension on the matrix, the fibroblasts can organise collagen fibrils and influence their alignment. Collagen interacts with elastin to provide recoil to tissues.

Fibronectin [97] is another major ECM protein that is responsible for the organisation of the interstitial ECM and helps in mediating cell attachment and function. Fibronectin plays a role in cell migration during development and has been associated with tumour metastasis

and other conditions.

The extracellular matrix plays a vital role in disease progression and treatment effects [98]. Cancer can be explained as the loss of tissue organisation and the change in behaviour of cellular components due to mutation. Stromal Tumours are stiffer than their counterparts in healthy tissue. The stiffness of the stroma is attributed to the remodelling of fibroblasts. Changes in the ECM and the structure of its proteins have been studied to determine its effects [28]. The datasets used in this thesis are primarily images of fibres of collagen and fibronectin of the extracellular matrix. (Fig. 2.6) is an illustration of the Extracellular matrix that surrounds cells.

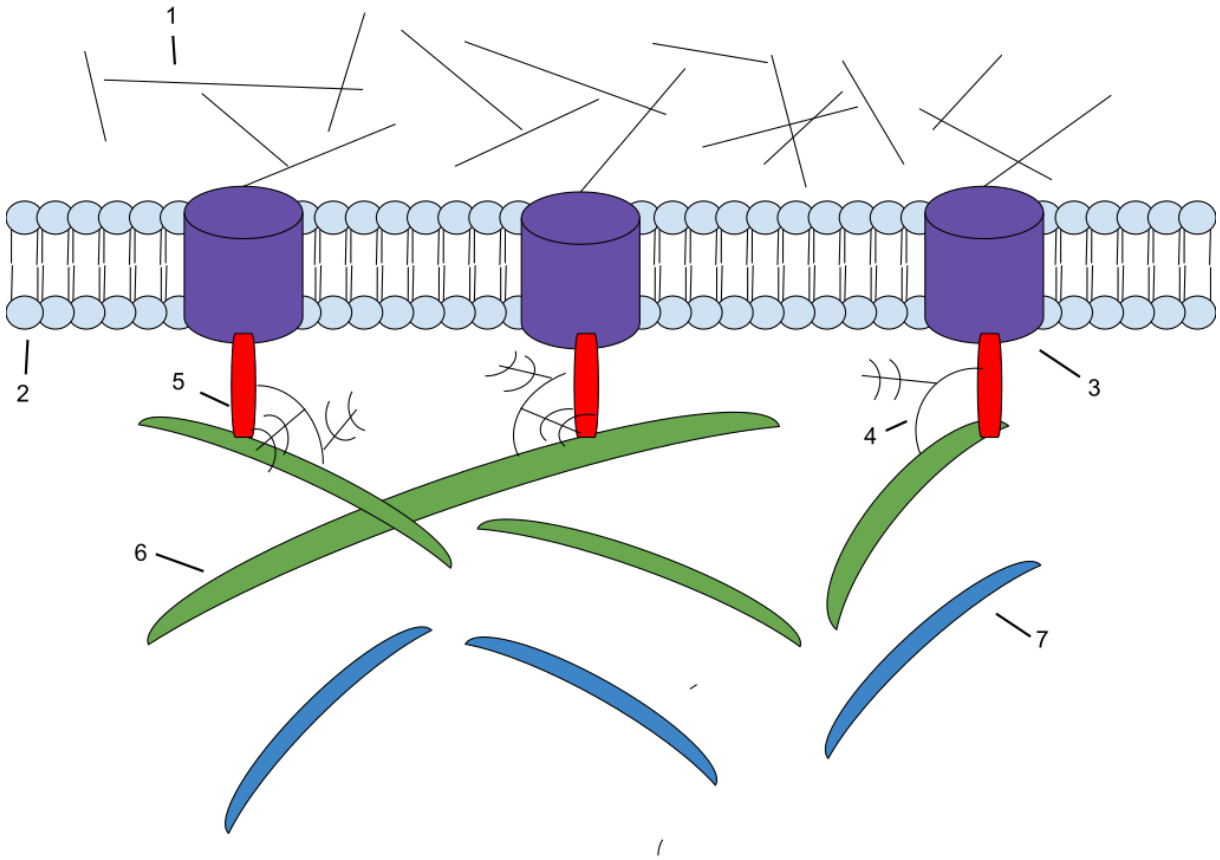


Figure 2.6: Structure of the Extracellular Matrix that surrounds cells. 1. Microfilaments, 2. Phospholipid Bilayer of a cell. 3. Integrin. 4.Proteoglycan. 5. Fibronectin connected to the integrin. 6. Collagen. 7. Elastin. [99] This image is available through Wikimedia Commons..

2.3.1 Fibre Morphology:

Like cell morphology, fibre morphology refers to the characteristics of fibres such as length, width, orientation, and alignment with other fibres. Fibre morphology contributes to the understanding of many biological processes relating to the extracellular matrix. Many studies have focused on changes of the extracellular matrix fibres such as collagen and fibronectin to explain the effects of treatments or the progression of illnesses [100, 101]. Chronic or severe tissue injuries can lead to excessive ECM production and deposition without balancing the ECM with degradation this can result in fibrosis. This healing process that deviates from the standard way can be very severe could lead to organ failure, such as cirrhosis in the liver [32]. Fibrosis also increases the risk of cancer for example fibrosis could increase the mammographic density which would indicate the amount of collagen in the breast and would reflect an increased risk of breast cancer [102]. Ageing comes with changes in the ECM composition. A decrease in collagen and elastin may influence the body over time. Skin dehydration, thinning of joint cartilages and reduction in the elasticity of tendons, ligaments and skin which can become stiffer [103]. ECM can be improved through physical activity or medical procedures. Exercise can influence the shape of bones and size. Exercise can also improve functions of tendons and ligaments which in turn improves the ECM [104]. The ECM composition has also seen changes in chronic lung diseases [105], the ratio of components becomes unusual and there is an increase ECM deposition during fibrosis but a degradation with emphysema. There is also evidence of unique fragments released from the ECM that shows active disease processes which reflect on cell responses to the ECM. The ECM of idiopathic pulmonary fibrosis exhibits stiffness which regulates cell adhesion, migration, and proliferation.

2.3.2 Image Analysis of the ECM:

Image analysis has been used to quantify the abundance and organisation of ECM proteins such as collagen, fibronectin and laminin as seen in [106], where image analysis was used to quantify collagen alignment in breast cancer. Image processing also enables the analysis of fibre alignment, density and network topology as seen in [107], where image analysis was applied to study ECM architecture in desmoplasia. This is also important in cancer biology as changes in collagen fibres are linked to tumour invasiveness. Time-lapse imaging allows the tracking of changes of the ECM during wound healing, fibrosis, and tissue development, this was investigated in [108], where the authors demonstrated collagen separation in fibrotic tissues through advanced imaging. Recent studies employ machine learning to classify ECM patterns, predict disease, and model cell-ECM interactions as seen in [109], where authors used machine learning techniques to analyse ECM features from histological images.

2.4 Cancer:

Cancer [5, 6] is a disease characterised by the abnormal and uncontrolled growth of cells, which can spread to other parts of the body. In a healthy body, cells follow a regulated cycle

of growth, division, and programmed death, that ensures proper functioning. However, when this process is disrupted, damaged cells continue to grow and multiply instead of dying. These clusters of abnormal cells can form lumps or tumours, some of which can be cancerous. Cancerous cells can invade surrounding tissue and spread to distant parts of the body through a process called metastasis. Tumours that travel and invade other tissues are classified as malignant, while non-cancerous tumours, are known as benign tumours and do not spread further from their original location [5, 7].

The survival and progression of cancer cells is dictated by a set of molecular, biochemical, and cellular characteristics known as the hallmarks of cancer [5, 8, 9, 10]. Cancer cells achieve uncontrolled growth by generating their own growth signals [11]. Normal tissues systematically control the production of growth promoting signals that instruct the progression of the cell cycle, this ensures homeostasis of cell number which preserves normal tissue architecture and function.

Cancer cells can evade regulatory mechanisms that normally subdue cell division [12]. Pathways such as $TGF-\beta$ are known to be antiproliferative and appear to evade cancer cells [110]. Many late-stage tumours deviate the $TGF-\beta$ signalling from suppressing cell proliferation and alternatively is found to be activating a cellular program.

Cancer cells also evade programmed cell death, allowing them to stay alive despite cellular damage [13]. Cancer cells can evade programmed death as a response to many physiological stresses encountered by cancer cells during tumour progression or it could also be attributed to anticancer treatments. Among the key stresses are the disruptions of signalling balance caused by excessive oncogene activity, another stress is DNA damage caused by uncontrolled cell proliferation.

Additionally, cancer cells maintain indefinite division by stabilising their telomeres [14]. This is different to the behaviour of normal cells, which are able to pass only through a certain number of cell growth and division processes.

Cancer cells promote their survival by forming new blood vessels and instructing the blood vessels to supply them with oxygen and nutrients [15]. Like normal cells tumours depend on nutrients and oxygen for survival as well as to remove metabolic waste and carbon dioxide. These requirements are met through tumour blood vessels formed by angiogenesis. In embryonic development, the vascular system is established through two mechanisms. Firstly, new endothelial cells are formed and organised into blood vessels. The second mechanism is angiogenesis, where new blood vessels are formed from existing one. When this network is formed, normal vasculature becomes mostly inactive as it is generally activated in physiological processes such as wound healing. However, when a tumour is developing the "angiogenic switch remains active which in turn the blood vessels keep forming and supporting the growth of tumours [111].

Through metastasis, they spread from their original site to distant organs [16].

Moreover, cancer cells can elude immune system attacks by convincing the body to protect them rather than kill them [17]. Expressions of genes in charge of cell to cell and cell to ECM adhesion molecules are altered in highly aggressive carcinomas. Adhesion molecules that are linked to cell migration that happen during inflammation are upregulated.

Unlike normal cells, cancerous cells generate energy by modifying the metabolic processes, which allows them to grow faster [18]. Normally cells process glucose by glycolysis and eventually into carbon dioxide. In the presence of oxygen, cancer cells are able to reprogram the glucose metabolism [112].

Two traits exist that are supportive of enabling cancers. Firstly, genome instability and mutation. Tumour progression can be attributed to the succession of clonal expansions which are each triggered by enabling mutant genotype [113]. Secondly tumour promoting inflammation, cancer cells target inflammation to support its growth and survival [114].

2.4.1 How Cancer Arises:

Tumour development starts when some cells that are within a normal population experiences a genetic mutation that increases its inclination to proliferate when it normally would not [115]. The mutated cell and its descendants continue to appear normal, but in fact they are producing too much this is known as hyperplasia. After a number of years, one in a million of these cells experiences another mutation that further depreciates cell growth control. Besides the excessive proliferation, the offspring of this cell may appear abnormal in shape and orientation, this is known as dysplasia. After some time, another mutation further alters the cell behaviour. The affected cells become more abnormal in growth and appearance. If this clump of cells, the tumour has not broken any tissue boundaries, it is referred to as in situ cancer. This tumour may not spread to other areas indefinitely; although some of the cells may experience additional mutations. If the genetic alterations promote the tumour to invading other tissues as well as deposit cells into the blood or lymph, the tumour mass is now considered to have become malignant. These cells have a higher chance on establishing new tumours (metastases) throughout the body, which may become deadly by disturbing the vital organs [115].

2.4.2 Cancer from a Molecular Perspective:

Genetic alterations that lead to oncogene inception [8], a mutated gene that can lead to cancer and other genetic disorders can include chromosomal translocation such as gene Bcr and oncogene Abl in chronic blood cancer, deletion with Erb-B gene in breast cancer, amplification with N-myc in neuroblastoma and insertion activation with C-myc in acute blood cancer. Chronic blood cancers may arise in the elderly due to a trade-off of the genetic material between chromosomes 9 and 22, which can lead to the production of the biomarker ph1, this biomarker is found in 95% of the patients which helps with accurate diagnosis. The fusion of the Bcr gene with the Abl oncogene leads to the formation of a new gene combination that produces a protein with kinase activity [116, 117].

Mutations in the p53 gene results in the production of an abnormal protein that significantly disrupts molecular processes normally regulated by p53 [8]. These disruptions contribute to the development of cancer cells, which is why p53 is closely linked to cancer progression. Research shows that p53 abnormalities are present in approximately 60% of cancer cases. Under normal circumstances, p53 is vital in cell division, programmed

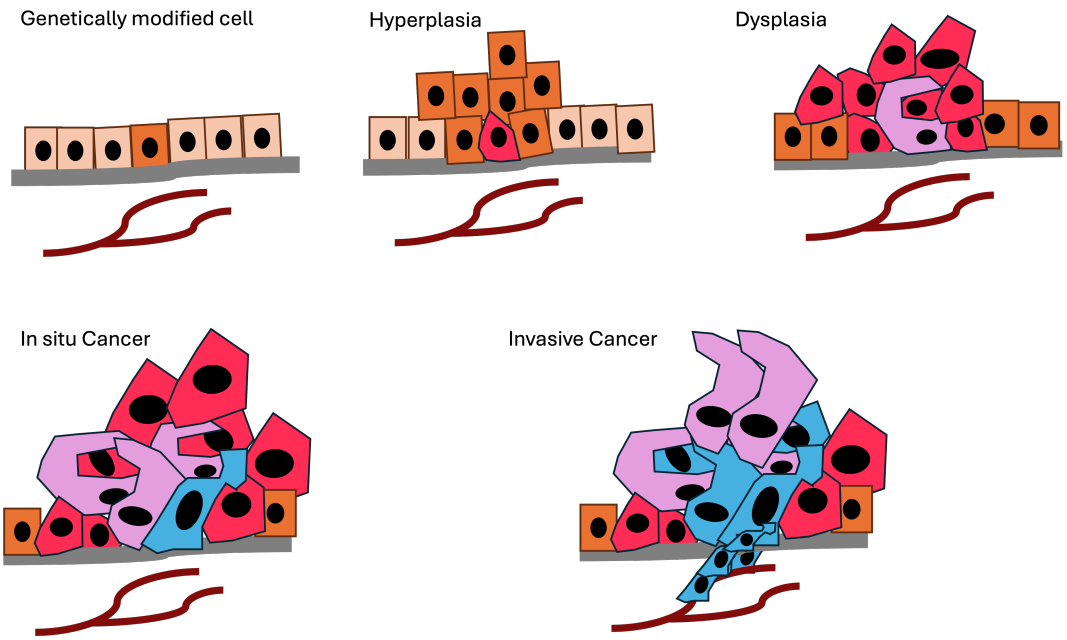


Figure 2.7: How cancer arises from a cell becoming altered (orange) within a normal population (beige) to hyperplasia and then the cell experiences another mutation (pink) which causes further turbulence in cell growth control. Followed by dysplasia where the offspring of the cells appears different in shape and orientation. After some time, the cell experiences another mutation (purple). As the tumour has still not broken any boundaries of the tissue it is commonly known in situ cancer, but some cells may experience further mutations (blue). These mutations might result in the tumour to break the tissue barrier, and some cells might enter the bloodstream or the lymphs, they are considered malignant, and they are likely to metastasise.

cell death (apoptosis), cellular ageing (senescence), angiogenesis, differentiation, and DNA metabolism. Moreover, most of the genetic alterations related to the p53 gene happen in the DNA-binding site, impairing its ability to control genes involved in DNA replication. Additionally, p53 interacts with CDK1-P2 and CDC2 to hold and maintain cancer cells in the G1 and G2 phases of the cell cycle [118]. Interestingly, p53 can act both as a tumour suppressor and, under certain conditions, as a promoter of cancer cell development. Following DNA damage caused by other genes, the p53 protein attaches to DNA and activates the WAF1 gene [119]. This activation promotes the interaction between p53 and CDK2, which prevents the function of p21 in advancing to the next stage of the cell cycle. The tumour suppressive functions of p53 operate through three main mechanisms, activation of DNA repair proteins, induction of apoptosis, and halting the cell cycle at the G1/S phase [120].

In general, hypomethylation, which is the reduction in the number of methyl groups attached to the DNA mainly occurs in repetitive DNA sequences, leading to increased genomic instability, including gene deletions, translocations, and chromosomal instability [121]. A well-known example is the hypomethylation of L1 elements from the LINE family which

has been seen in a range of cancer including breast, lung, and bladder cancer [122]. Additionally, hypomethylation at specific gene promoters can result in the abnormal activation of oncogenes. For instance, this has been observed with MASPIN, a tumour suppressor gene, in breast and prostate cancers, as well as with S100P in pancreatic cancer, SNCG in breast and ovarian cancers, and MAGE and DPP6 in melanoma [123].

In contrast, hypermethylation typically occurs in specific CpG-rich regions, particularly within gene promoters, and often this results in the silencing of gene expression. This type of transcriptional repression commonly affects genes responsible for DNA repair (such as HMLH1, WRN and BRCA1), vitamin response (CRBP1, RARB2), cell cycle regulation (P16INK4b, P16INK4a) and apoptosis (TMS1, DAPK1, and WIF-1), all of which play vital roles in cancer prevention [124]. As a result, hypermethylated promoters have gained attention as potential biomarkers for cancer diagnosis and prognosis, with research primarily focusing on promoter CpG islands. Moreover, abnormal hypermethylation within CpG regions is a frequent event in cancer occurring in approximately 45-65% of cases [125, 126].

Disruptions in DNA methylation patterns often stem from the dysregulation of DNA methyltransferases (DNMTs). Elevated expression levels of DNMT1 and DNMT3b have been reported in many tumours. Additionally, miRNAs particularly the MIR-29 family, play a regulatory role by directly suppressing DNMT3a and DNMT3b expression and indirectly downregulating DNMT1 [127].

2.4.3 Risk Factors:

Smoking is the leading preventable cause of cancer which is associated with lung, bladder, head and neck, oesophageal, pancreatic, liver, kidney, stomach, cervical, and colorectal cancers [128]. Cigarette smoke contains a number of carcinogens including nitrosamines which can lead to oxidative stress and inflammation [128]. High body mass index is linked to a higher risk of multiple cancers such as colon, postmenopausal breast, ovarian and others. Low activity lifestyle may raise the risk of breast, colon, endometrial, and ovarian cancers [129]. Alcohol is categorised as a group 1 carcinogen and may be linked to cancers of the mouth, liver, colon, breast and other [130]. Poor dietary habits such as highly processed food, red meat, salt, refined carbs, low in fruits and vegetables are linked to increased risks of stomach cancers [131]. Infections such as H.pylori and certain parasites contribute to 15-20% of global cancer cases [132]. Environmental pollutants such as arsenic, dioxins, and polycylic aromatic hydrocarbons can also increase cancer risk [133]. Inherited mutations such as BRCA1 and BRCA2 elevate the individual risk particularly for breast and ovarian cancers [134].

There are several types of cancers based on the type of cells, including Carcinomas, the most common type of cancer. They are from the epithelial cells. There are different types of epithelial cells and therefore diverse types of Carcinomas. Types of cancer are outlined in Table. 2.1.

Cancer Types	Description
Basal Cell Carcinoma	Basal cell carcinoma usually starts and is seen on the outer layer of the skin.
Squamous Cell Carcinoma	Squamous cell carcinoma is prevalent in the lining of organs such as the kidneys, stomach, intestines.
Adenocarcinoma	Adenocarcinoma are formed from the cells that produce fluids or mucus. Cancers such as colon, breast and prostate are all adenocarcinoma.
Transitional Cell Carcinoma	The final type of cancers that are from epithelial cells is the Transitional cell carcinoma, they are usually associated with cancers of the ureters and bladders.
Sarcoma	Another type of cellular cancer is Sarcomas, these are usually formed in the soft tissues such as muscle and fat as well as bones.
Leukemia	Leukaemia are cancers that are formed in the bone marrow therefore they are cancers of the blood, where there is a large number of white blood cells.
Melanoma	Melanomas are prevalent with skin cancer and intraocular melanoma.

Table 2.1: Cancer types and their descriptions. Basal Cell Carcinoma [7, 135]. Squamous Cell Carcinoma [7, 136]. Adenocarcinoma [7, 137]. Transitional Cell Carcinoma [7, 138]. Sarcoma [7, 139]. Leukaemia [7, 140]. Melanoma [7, 141].

2.4.4 Breast Cancer:

The primary image data used in this thesis consists of breast cancer biopsy samples. Breast cancer [142, 143] is caused by mutated cells of the breast that multiply and produce tumours and spread to the other regions of the breast. Breast Cancer typically affects women over the age of 50 but can also affect younger women. Men may also develop breast cancer [144]. There are three main types of invasive Breast Cancer [145]: Invasive ductal carcinoma, Lobular breast cancer, and inflammatory breast cancer. Invasive ductal carcinoma starts in the milk ducts and progresses to nearby breast tissue. Lobular breast cancer begins in the milk producing glands (lobules) and spreads to nearby tissue. Inflammatory breast cancer is an aggressive type of cancer that occurs when the cancer cells infiltrate the skin and lymph vessels of the breast. Other non-common breast cancer types include Triple-negative breast cancer, Inflammatory breast cancer and Paget's disease of the breast. Non-invasive breast cancers include Ductal carcinoma in situ, which starts in the milk ducts but does not spread beyond.

Risk factors [146] may increase the chances of developing breast cancer, these are Age (above 55), Sex (Female), Family history (Close relatives who have developed breast cancer), Genetics, Smoking, Obesity, Radiation exposure (prior radiation therapy) and, Hormone replacement therapy.

Breast Cancer from a molecular perspective:

Breast cancer occurs due to a series of complex genetic and epigenetic events that lead to a malignant transformation of a normal epithelial cell [147]. Genetic mutations where certain nucleotide base pairs of a gene are altered or lost are the most common causes of these events. Epigenetic changes cause alterations in the transcription of a gene without changing its normal base-pair sequence. An example that is directly related to breast cancer initiation is hypermethylation of the promoter region of the gene [147].

For breast cancer to initiate and progress it must be able to invade, recruit a vascular supply, and proliferate [148]. For these phenotypes to happen, activation of oncogenes and deactivation of tumour suppressor genes must occur which inhibit many of these malignant traits and functions to maintain the cells in most cases. Tumour suppressor genes play an essential role in maintaining cell cycle control and genomic integrity. Some of these genes ensure proper checkpoints during the cell cycle, preventing the transmission of genetic errors to daughter cells during mitosis [149]. Others, are responsible for detecting, repairing, or managing DNA damage. When damage is beyond repair, certain tumour suppressor genes can trigger apoptosis, a programmed cell death mechanism to eliminate the compromised cell [150].

Mutations in tumour suppressor genes are frequently observed in breast cancer and are thought to be early and pivotal steps in the transformation from normal to malignant cells. Notable tumour suppressor genes that are directly linked to breast cancer include p53, BRCA1, BRCA2, PTEN and possibly ATM [151].

These mutations can be either be inherited (germline) or acquired over a person's lifetime (somatic). Every gene has two copies (alleles), one from each parent [152]. In cases of germline mutations linked to breast cancer, individuals inherit one mutated allele and one normal allele. Often, the normal copy produced enough protein to maintain normal cell function. However, a second mutation known as "second hit" in the remaining normal allele can lead to the loss of gene function, significantly increasing the risk of cancer development. Women who inherit a faulty allele are more vulnerable to acquiring this second hit.

Somatic mutations, on the other hand, arise in individual cells during a person's life and are passed on only to the descendants of those cells. It is now understood that most breast cancers result from somatic rather than inherited mutations [153].

Another major genetic mechanism contributing to breast cancer is the activation of oncogenes [147]. Oncogenes are often mutated or otherwise altered versions of proto-oncogenes, which normally regulate healthy cell growth. Once activated, oncogenes promote cancerous behaviour such as unregulated proliferation [53]. Unlike tumour suppressor genes, oncogene mutations result in a gain of function and typically they only require only one allele to influence the phenotype, these mutations are always somatic.

Oncogenes can produce proteins that promote tumour growth by enhancing cell invasion, accelerating cell cycle progression, or stimulating the development of new blood vessels to support the tumour. Several proto-oncogenes have been linked to breast cancer development, including HER2/neu, EGR, Ras, Myc, and $\beta - catenin$ [147].

In recent years, significant progress has been made in identifying inherited genetic conditions that could increase the risk of breast cancer. Initial discoveries came from studies of families where nearly half of the women developed breast cancer, a pattern consistent with autosomal-dominant inheritance [147]. These investigations revealed that inherited mutations in tumour suppressor genes such as BRCA1, BRCA2, p53, and PTEN are strongly associated with breast cancer [151]. Individuals who inherit these mutations may face a high likelihood of developing the disease.

Breast Cancer and the ECM:

There is a great connection between the extracellular matrix fibres such as collagen and its architecture and breast cancer. In tumour progression, the ECM tends to become stiffer and denser as there is an increase collagen and fibronectin deposition by the cancer associated fibroblasts. The stiffened ECM fibres promote mechanotransduction signalling, which encourages tumour invasion and reduced drug accessibility [29]. Furthermore, the dynamic crosstalk between tumour cells and the stromal cells such as immune cells remodels the extracellular matrix fibres, this establishes a supportive function that enhances cancer cell survival, metastasis, and resistance to therapy [154]. Changes to the ECM allow epithelial to mesenchymal transition, this promotes cancer cells to invade and migrate. In vitro studies with decellularised ECM show its significant role in allowing epithelial to mesenchymal transition through changes in E-cadherin, vimentin and N-cadherin expression [155]. Certain ECM pattern such as aligned collagen fibres, have a direct connection with poor prognosis and higher risks of metastasis, recurrence, and therapy resistance [156]. Treatment plans are now focused on targeting ECM remodelling enzymes such as lysyl oxidase as well as mechanical stiffness, and ECM signalling interaction to improve treatment response and impede metastasis [157]. This is important as researchers can look at biophysical cues such as stiffness which activates pathways in tumour cells supporting growth, invasion, and survival. It is also important to understand the ECM heterogenity which is specific to different types of breast cancer subtypes such as luminal and triple negative, this can have an effect on the type of therapy the patient should take [158]. Understanding how the ECM remodels is of great importance as it is both a driver and a marker of tumour aggressiveness, which can be a potential diagnostic and targeting indicator.

2.4.5 Cancer Diagnosis:

Cancers are diagnosed with multiple tests from non-invasive examinations to invasive examinations [159]. Lab tests include blood tests, Complete blood count mainly to screen for leukaemia and lymphoma [160]. Urinalysis is used to diagnose bladder cancer [161], and tumour markers such as Alpha-fetoprotein (AFP) is used to analyse blood and diagnose liver

cancer, ovarian cancer, and germ cell tumours [162]. Another way to diagnose cancers is through diagnostic imaging, transmission imaging such as X-ray is used to detect cancers in bones, lungs, stomach, kidneys, and breasts [163]. Computed Tomography (CT) scan is used to diagnose cancers of the lung, colorectal, pancreatic, kidney, bladder, ovarian, stomach and a number of liver cancers [164]. Mammograms (4 breast X-rays) are used to detect and diagnose breast cancer [165]. Reflection imaging such as Ultrasound may be used to diagnose cancers found in the liver, pancreas, gallbladder, and prostate [166]. Emission imaging such as Magnetic Resonance Imaging (MRI) are used to detect cancers of soft tissue for example cancers of the brain, spinal cord, pelvic organs including cancers of the prostate, bladder, uterus, and ovaries [167]. Some endoscopic examinations such as Cystoscopy for bladder cancer and Colonoscopy for colorectal cancer are used as well [168, 169]. Finally, tumour biopsies are used to diagnose cancers. Tumour biopsy is a procedure where a tissue sample or a cell sample is removed from the body to be examined under the microscope. Usually, the sample of tissue or cell is stained using substances such as Haematoxylin and Eosin. The staining process enhances contrast of specific components, so it makes it easier for the pathologists and the clinicians to determine whether the tumour is malignant or benign. Common biopsies include but are not limited to Bone Marrow biopsy, Fine needle aspiration biopsy which includes breast tissue biopsies and Skin biopsy [170, 171, 172].

2.4.6 Cancer Treatment:

For many years, cancer treatment options were limited to surgery, radiation therapy and chemotherapy, either used alone or in combination [173]. However, recent advancements have significantly expanded the understanding of the biological pathways involved in cancer progression and how they can be effectively targeted. As a result, combination therapies, using multiple targeted agents or a combination of targeted therapies with traditional chemotherapeutic drugs like taxanes and platinum compounds have shown a larger effect [174]. Moreover, new treatment strategies, including novel drugs, biological agents, and immune-based therapies, are being introduced. Although these therapies have yet to fully achieve the desired impact on reducing mortality and significantly extending survival in patients with metastatic cancer, they represent a progress in cancer treatment.

The most commonly recommended traditional cancer treatment is surgical removal of tumours; this is often followed by radiation therapy using X-rays and/or chemotherapy [175]. Among these options, surgery is typically most effective when cancer is detected at an early stage. However, radiation therapy can harm not only the cancer cells but also the surrounding healthy tissues and organs. While chemotherapy has contributed towards a lower cancer-related illness and death rates, it still affects healthy cells, particularly those that divide and grow rapidly [176]. A significant challenge with chemotherapy is drug resistance, where cancer cells that were initially responsive to treatment become resistant over time. This resistance is largely due to decreased drug absorption and increased expulsion of the drug from cancer cells [177]. Other drawbacks of traditional chemotherapy include difficulty in determining optimal dosage, lack of target specificity, rapid breakdown of drugs in the body, and severe side effects.

One of the major challenges in cancer treatment is drug resistance and the limitations of drug delivery systems, which significantly hinder effective cures and symptom management. Despite these obstacles, numerous approved drugs and treatment strategies are now available. The effectiveness of traditional cancer therapies is often compromised by the unique structure and abnormal blood vessel architecture within tumour tissues [178].

Stem Cell therapy is a type of advanced therapy. Stem cells are undifferentiated cells found in the bone marrow. They have a unique ability to develop into various types of cells in the body. Using stem cells as a therapeutic approach for cancer is considered a promising and potentially effective and safe option. However, stem cell therapies are still largely in the experimental phase and are being tested out in clinical trials, particularly for their potential to regenerate damaged tissues. Mesenchymal stem cells, which can be derived from bone marrow, fat tissue, and connective tissue, are currently being tested in clinical trials for this purpose [179].

Targeted cancer therapies often refer to drugs or substances designed to specifically interfere with molecules involved in cancer growth and progression. These therapies are often referred to as "molecularly targeted drugs". Their primary mechanism of action involves blocking signals that promote tumour growth and spread [180].

The development and progression of tumours is influenced by the tumour microenvironment, which consists of various cell types, including endothelial cells, smooth muscle cells, fibroblasts, and cancer stem cells. These cells interact with cancer cells through complex signalling pathways, supporting cancer cell survival and proliferation. As a result, targeting the tumour microenvironment has become a major area of research to improve cancer treatment outcomes [181].

Conventional chemotherapy struggles to selectively target cancer cells due to their similarities with normal cells. Targeted therapies help overcome this by using cellular mechanisms such as inducing cell cycle, triggering apoptosis, preventing cell proliferation, and disrupting cancer cell metabolism [182]. Additionally, modifying or targeting the tissue microenvironment for drug delivery offers another promising approach to treating cancer [183].

Unlike traditional chemotherapy, targeted therapies aim to selectively attack cancer cells, causing minimal damage to healthy cells [184]. These treatments have significantly improved survival rates in certain cancers. For instance, adding erlotinib to standard chemotherapy increased the survival rate for advanced pancreatic cancer from 17% to 24%. Similarly, imatinib has shown remarkable effectiveness in treating chronic myeloid leukaemia, while rituximab, sunitinib and trastuzumab have transformed treatment approaches for renal cell carcinoma and breast cancer [185].

Target therapies can be classified based on how they work or their specific target. Some act as enzyme inhibitors, blocking enzymes that signal cancer cells to grow. By cutting off these signals, cancer growth and spread can be suppressed. Other targeted therapies, known as apoptosis inducing drugs, act on parts of the cell that regulate survival, promoting the death of cancer cells. For example, inhibitors of serine/threonine kinase or protein kinase B (PKB/Akt), which support cell survival, are currently being tested in preclinical trials [186].

Additionally, some targeted therapies inhibit the formation of new blood vessels (angio-

genesis), depriving tumours of their blood supply and limiting their growth [187]. This can be achieved by blocking factors such as vascular endothelial growth factor or its receptors. Studies have shown that combining bevacizumab with 5-fluororacil based chemotherapy can extend survival in patients with advanced colorectal cancer by several months [188].

Gene therapy involves introducing a healthy version of a faulty gene into a person's genome to treat a specific disease. The first successful application of this approach was in 1990 when a retrorival vector was used to deliver the adenosine deaminase gene into T cells of patients with severe combined immunodeficiency. Several gene therapy strategies are being explored for cancer, including introducing genes that promote cell death (pro-apoptotic genes), enhance sensitivity to chemotherapy, restore normal tumour suppressor gene function, stimulate anti-tumour immune responses, and silence oncogenes that drive cancer development [189].

Natural antioxidants like vitamins, polyphenols, and plant-based bioactive compounds are commonly used as preventive and therapeutic agents due to their anti-inflammatory and antioxidant effects, which help protect the body from harmful molecules. Their anti-proliferative and pro-apoptotic properties have also led to their inclusion in cancer treatment research. Various natural antioxidants, including vitamins, alkaloids, flavonoids, carotenoids, curcumin, berberine, quercetin, and others, have been extensively studied through in vitro and in vivo experiments [190].

2.4.7 Image Analysis in Cancer:

Image analysis has been widely used and is pivotal in the progress of our understanding of cancer biology, and in particular the interactions between cancer cells and the extracellular matrix. With the quantification of the ECM structure and cell behaviour in tumour environments, image-based methods have revealed how ECM organisation, stiffness, and composition can influence actors such as tumour progression, metastasis, and therapeutic resistance. Collagen alignment and tumour invasiveness was investigated in [191], where SHG imaging and vector analysis were used to quantify collagen orientation, revealing the invasive paths along aligned fibres. Aligned collagen fibres at the tumour-stroma boundary facilitate cancer cell invasion in particular TACS-3 which is the radially aligned collagen of the Tumour-Associated Collagen Signatures (TACS) is indicative of poor prognosis in breast cancer. Image analysis has facilitated the quantification of how ECM stiffness and density regulate cell signalling pathways which influence cell migration and proliferation. This is illustrated in [192], where it was found that denser ECM, quantified through 3D confocal and SHG microscopy, enhances integrin signalling and induces an invasive phenotype. In another example, authors used microscopy, combined with image segmentation and tracking, which allowed the observation of real-time ECM remodelling by cancer-associated fibroblasts (CAFs) and its influence on cancer cell migration [193].

Similarly in terms of cells, cancer cells often exhibit characteristics that are different to those of a normal cell (Fig. 2.8) such as increased nuclear size [194], irregular shape and contours, increased nucleus-to-cytoplasm ratio and abnormal mitotic figures and increased proliferation [9]. That is when a technique such as nuclear segmentation can identify and

isolate individual nuclei in microscopical images. Once the cells are segmented, morphological features such as area, perimeter, eccentricity and textual features such as entropy and contrast and spatial features such as the clustering of cells can be extracted. These features can be used to train machine learning classifiers that can differentiate between malignant and benign cells based on their measurements. This was presented in [195], where deep learning was applied on segmented nuclei to classify tumour and normal regions in colon cancer.

Nuclear segmentation allows the extraction of key morphological and textural features that are hallmarks of malignant transformation. Through quantifying these features across thousands of cells, image analysis pipelines are able to identify abnormal patterns that distinguish cancerous from benign cells with high sensitivity and reproducibility.

As seen through the examples mentioned that high quality imaging and robust algorithms that produce reproducible results are imperative to the advancement of our knowledge in cancer and other conditions. The algorithms and languages used in image analysis of cancer were investigated in a scoping review (Section 2.5).

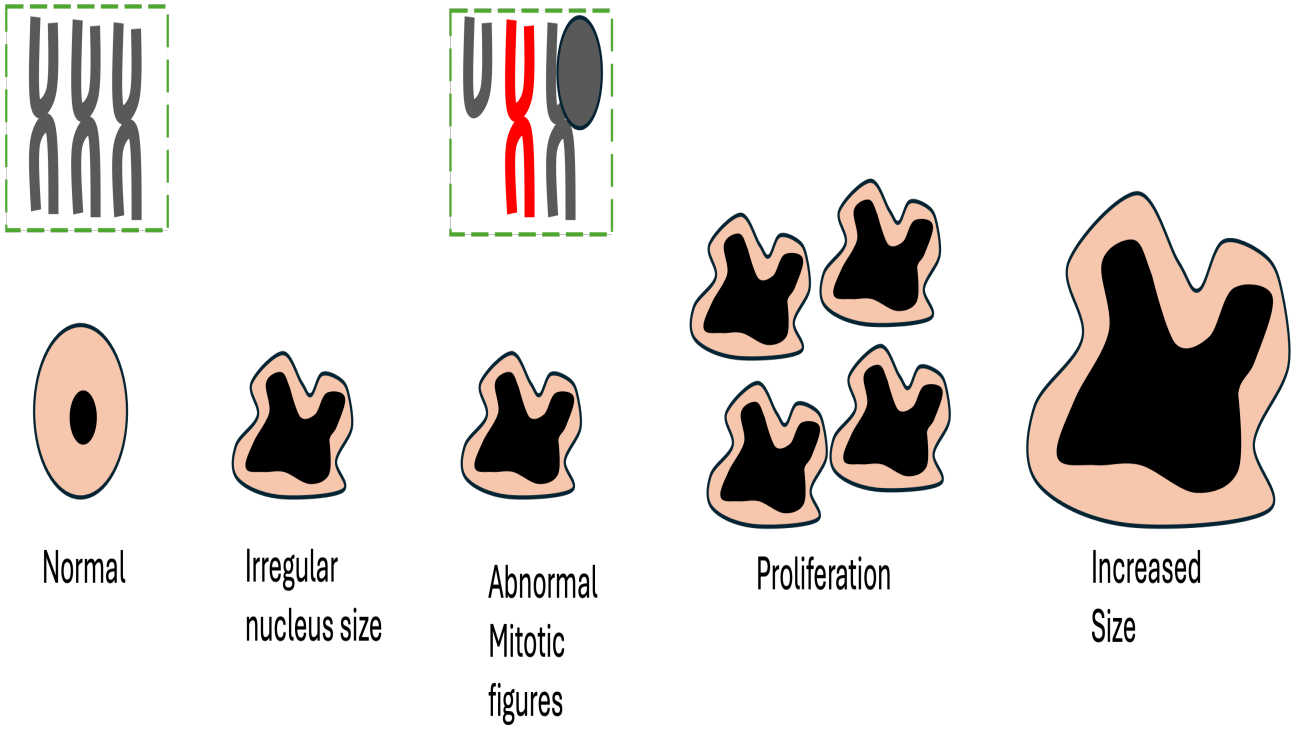


Figure 2.8: Cancer cell compared to normal cell. Cancer cells often exhibit characteristics that make them different than normal cell such as irregular nucleus size and nucleus to cytoplasm ratio, abnormal mitotic figure, proliferation, and increased size.

2.5 Scoping Review:

Excerpts of this section have been published as a conference article in MIUA 2022, Cambridge, UK [196].

As this thesis focuses on image analysis of cancer-related datasets, a scoping review was conducted to asses the popular methods used as well as the computational languages.

A wide range of image analysis techniques are available to analyse cancer imaging data. Therefore, a review was conducted to identify popular methods and examine trends in image processing techniques.

Cancer topics have flooded medical research literature. At the time of writing of this section (March 2022), a total of 4,574,067 entries out of 33,844,076 entries on the search engine of the United States National Library of Medicine MEDLINE had cancer as a keyword, it does not mean that it is the main topic, the main topic can be surgery or otherwise. This accounts for 13% of entries in PubMed. In the 1950s cancer-related entries accounted for 6% of all yearly entries, to exceed 16% seventy years afterwards [197]. With the rise of technological advancements, many contributions to medicine and cancer have been made. Computational data analysis techniques have been vital and grew in popularity as new imaging technologies produced increasingly new data every year. Diagnostic images acquired with Computed Tomography [198], Magnetic Resonance Imaging [199], Ultrasound [200], and pathology staining [201, 202] are constantly being produced and analysed, usually by experts that have undergone vigorous training throughout many years to produce an accurate diagnosis. Computational techniques have also been of high importance in pre-operative planning [203], decision making [204] and survival predictability [205].

Recently, computational techniques that are encompassed under the umbrella term of Artificial Intelligence have advanced rapidly. Historically, a wide range of techniques from the fields of image processing and computer vision have been widely used to analyse data from many Cancer datasets. Approaches such as Fourier analysis [206], mathematical morphology [207], and fractals [208] among others, have been successfully applied in the analysis of cancer treatment and drug development [209, 210, 211]. This raises the question: are artificial intelligence techniques now replacing these earlier image processing and computer vision methods in cancer image analysis?

This section presents a review of the data analysis techniques applied to cancer research. The aim of this review is to map the existing literature and evaluate the volume of work, based on the number of entries in PubMed, rather than to compare outcomes or conduct a detailed systematic review and synthesis of evidence [212, 213]. Therefore, a scoping review, rather than a systematic review, was conducted. Data mining techniques were used to extract entries of the various methods applied in Cancer image analysis. A series of keywords were combined to select different techniques to track their yearly entries in PubMed.

The scoping review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [214, 215] along with the PRISMA extension for Scoping Reviews (PRISMA-ScR) [212, 213]. The process involved had the following stages: (1) defining the research questions, (2) identifying the data sources and selecting keyword combinations to retrieve relevant entries, (3) mine the number of entries published each year, (4) visualise the

trends, (5) interpret the results. Unlike traditional reviews, this review used an automated process for screening and eligibility. Instead of manually excluding entries, a series of specific keywords were sequentially added to a basic search, with inclusion only based on the presence of those keywords in the query as described in (Fig. 2.9).

The following questions were addressed: *Which traditional and deep learning computational techniques have been most used in the analysis of cancer-related image analysis? Which tools have been the most frequently utilised to implement these computational techniques?*

The review only considered the PubMed database which has been recognised as "the most widely used database of biomedical literature" [216]. The first keyword used in the search was *Cancer*, this yielded over 4 million results. Adding *Pathology* with a logical AND narrowed the results to 1.6 million entries. To specifically target image processing/analysis, the terms *images AND imaging* were included, reducing the number of entries to 277,302. Out of these entries, 254,658 entries were published between 1990 and 2022. This formed the basic dataset, to which additional specific keywords were added to further analyse the trends in cancer image analysis.

Three groups of specific keywords were employed to explore the trends: traditional Image Processing and Computer Vision, Machine Learning, and Deep Learning techniques. The specific keywords for each group are listed in Table 2.2.

The first group of keywords was used to observe the historical and current trends in traditional image processing / computer vision techniques.

Some keywords are one worded and specific therefore only required a singular term to be added, for example *Fractal*. Other keywords require a combination of 2 or three words to be specific, for example texture, this can be interpreted as the texture of the image sample not texture analysis the technique. Thus, *texture analysis* was used.

Some ambiguity remains possible; for example, the keyword texture can be associated with Fourier transforms, and entries could include works by an author named Fourier, even if Fourier analysis was not applied. However the query results were manually reviewed, and the keywords were adjusted to minimise such errors. The second and third keyword groups focused on Machine Learning and Deep Learning techniques, which have seen significant growth in recent years.

The number of article entries was observed for each year for the analysis period. For the yearly analysis, in addition to examining the absolute number of entries, the relative number of entries was also considered. There are two methods that can be used to calculate relative values. First, as a percentage of the total number of entries of all keywords in a given year. The total number of keywords is summed, and the proportion of entries for each keyword is calculated by dividing its count by the total number of entries across all keywords. Second, as a percentage of the overall number of entries for the year, regardless of specific keywords, meaning entries that do not include any of the keywords are also factored into the total.

A final set of queries was conducted to explore the programming languages, software, and environments used for image analysis in cancer and pathological datasets. The most used options within both "point and click" and computational environments, as defined in [217] were considered. These included general programming languages *Matlab*, *R*, *C*, and *Python*,

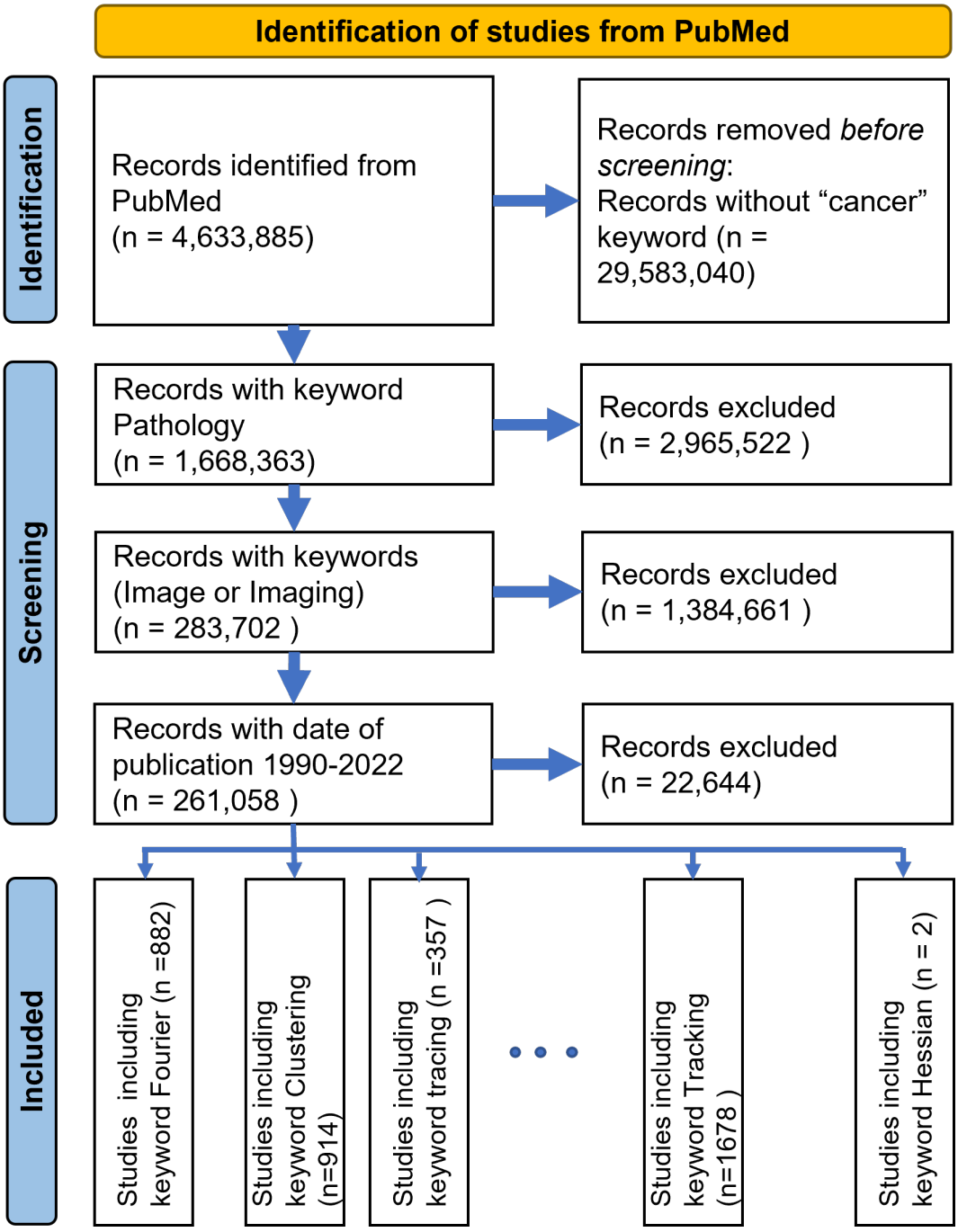


Figure 2.9: PRISMA statement flowchart showing the details of the scoping review. Inclusion in the study was given when specific keywords were on PubMed for individual searches. Studies were not read individually.

as well as specialised software such as *QuPath* [218], *CellProfiler* [219], *Fiji* [220], *ImageJ* [221] and *ICY* [222].

Image Processing / Computer Vision	**Machine Learning**	**Deep Learning**
Texture analysis	Classification	Deep belief
Fourier	Regression	Recurrent neural networks
Geometric	Dimensionality reduction	Auto-Encoders
Tracing	Ensemble learning	Multilayer perceptron
Linear discriminant analysis	Reinforcement learning	Generative adversarial network
Thresholding	Supervised learning	Convolutional neural network
Feature extraction	Bayesian	U-Net
Tracking	Decision tree	Transformer
Clustering	Linear classifier	Fully convolutional
Scale space	Unsupervised learning	ResNet (Residual neural network)
Hessian	Artificial neural networks	VGG
Self-organizing	Hierarchical clustering	Mask-RCNN (Region-Based
Region growing	Cluster analysis	Convolutional Neural Networks)
Mutual information	Anomaly detection	LSTM (Long short-term memory)
Wavelet	Semi-supervised learning	GoogleNet
Multiresolution	Deep learning	AlexNet
Principal component analysis	Support Vector Machines	Inception-v3
Filtering	Naive Bayes	DenseNet
Active contour	Nearest Neighbor	Inception-ResNet-v2.
Fractal	Discriminant Analysis	
Linear regression	K-Means	
Ensemble	Hidden Markov Model	
Transfer learning	Feature Selection	
Convolutional neural	Feature Engineering	
Machine learning	Random Forest.	
Deep learning		

Table 2.2: Specific keywords used in the search strategy to explore Image Processing and Computer Vision, the Machine Learning and Deep Learning Techniques.

A number of these names, such as CellProfiler and QuPath can be used directly as keywords, as they specifically refer to the tools being analysed and do not introduce any ambiguity. However, it was observed that in some instances, searching by keyword yielded fewer results than the number of citations to a specific publication on the software. For example, a PubMed search for QuPath (*https://pubmed.ncbi.nlm.nih.gov/?term=QuPath*) returned 86 entries, whereas citations to the original paper describing QuPath (*https://pubmed.ncbi.nlm.nih.gov /?linkname=pubmed_pubmed_citedin&from_uid=29203879*) resulted in 780 entries. In contrast, for tools like Matlab or R, searches using their names as keywords returned results,

but there was no single original paper or reference that could be used for comparison.

To address this, two distinct search strategies were used. Entries for *Matlab, R, C, and Python* were obtained directly through searches using specialised keywords in the URL. Meanwhile, entries for *Fiji, ImageJ, ICY, QuPath and CellProfiler* were gathered from citation data. Although these approaches are not entirely equivalent, they offer a broad overview of the tools utilised in research published in PubMed. In both cases, yearly entry counts were available through the field *yearCounts*, as outlined in the mining strategy section below.

To determine the specific keywords for programming languages, several factors were considered. *Matlab* could refer to both the programming environment and a rural region in Bangladesh, so the query was structured as *Matlab NOT ((rural) AND (bangladesh))* to exclude unrelated entries. Mining entries for the languages R and C, was more challenging, as single letter searches are ineffective in PubMed. To address this, the query for *R* was expanded to *(("R project") OR ("R package") OR ("Rstudio") OR ("R/Shiny))* and *("C programming") OR ("C language") OR ("C package")*. For Python, it was necessary to filter out results related to python snakes, leading to the query *(Python) NOT (snake) NOT (python regius)*.

Since *Fiji* was not mined using keywords, doing so would have introduced significant ambiguity. It would have required filtering out entries related to the country, authors affiliated with universities in Fiji, or cities. A strategy that would help mitigate these issues would be: *(Fiji) NOT (suva) NOT (pacific) NOT (Samoa) NOT (Palau) NOT (Nausori)*.

To investigate the key trends of the analysis techniques, the PubMed database was queried using a combination of keywords, following the methodology outlined in [197, 223].

The PubMed mining process was conducted by constructing a Uniform Resource Locator (URL) using a combination of keywords to query the database. The URL started with the PubMed address (i.e. *https://www.ncbi.nlm.nih.gov/pubmed/*), followed by the search term, *?term=*, and then the keywords were concatenated. Since URLs cannot contain certain characters like spaces or quotes, these were converted to their ASCII equivalent (space = %20, quotes = %22). The initial keywords used were *cancer* and *pathology*. After which to further restrict the entries to those related with images, the terms *((image) OR (imaging))* were added. The search was further refined by limiting it by date, initially from 1990 to 2022 and later from 2010 to 2022 to focus on more recent entries. These keywords combined with logical AND operators formed the **basic URL** (e.g. *https://pubmed.ncbi.nlm.nih.gov/?term=(pathology)%20AND%20(cancer) %20AND%20 ((image)+OR+(imaging))%20AND%20(2010:2022[dp])*). Specific keywords for example. Fourier were concatenated to the URL to explore the research questions.

Yearly entries were extracted from the query results for all keywords, as PubMed provides a field named *timelineData* which contains pairs of values [year, entries] for all years that contain entries. For instance, a user can search a term like *Cancer* in PubMed, view the page source (*view-source:https://pubmed.ncbi.nlm.nih.gov/?term=cancer*), and search for the string *yearCounts* to locate the relevant data.

All the code used for the scoping review was developed as Matlab® (The Mathworks™, Natick, USA) functions and Python Jupyter Notebooks.

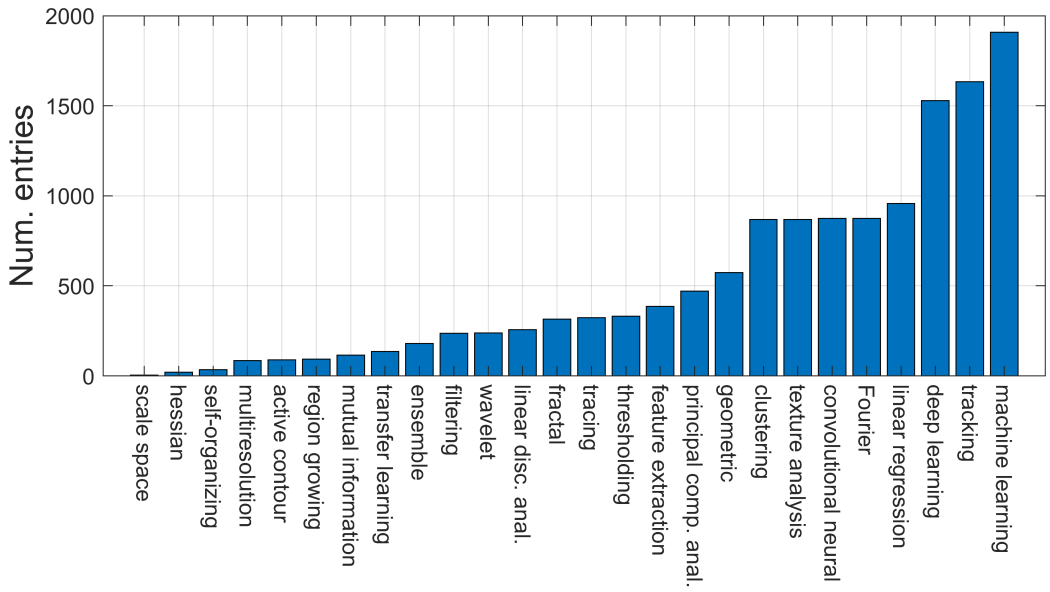


Figure 2.10: Number of entries in PubMed of the data analysis techniques specified by the keywords. The queries were created with the combination of the following terms (*keyword*) AND (pathology) AND (cancer) AND ((image) OR (imaging)) AND (1990:2022[dp]). This last term is of the date of publication.

From 1990 to 2022, the analysis techniques with the highest number of entries in PubMed were associated with the keywords *machine learning, tracking*, and *deep learning*, each exceeding 1,500 entries. These were followed by five keywords, *linear regression, convolutional neural, Fourier, texture analysis* and *clustering* which had less than 1,000 entries. In contrast the keywords that returned the fewest entries included *mutual information, region growing, multi resolution, self-organizing, hessian* and *scale space* (Fig. 2.10).

The year-on-year analysis of these keywords revealed some notable trends and painted a different picture (Fig. 2.11). It is expected that the number of entries per keyword would sequentially increase as the number of entries in PubMed have risen annually, from 410,894 entries in 1990 to 1,763,742 entries in 2021 (*https://pubmed.ncbi.nlm.nih.gov/?term=(1990[dp])*, *https://pubmed.ncbi.nlm.nih.gov/?term=(2021[dp])*), certain techniques have grown at a faster rate. Notably, *machine learning, deep learning and convolutional neural* have seen a rapid increase in their number of entries recently, unlike *tracking, linear regression*, and *Fourier*. Although *Clustering, texture analysis* and *feature extraction* have also shown growth, but it is far less pronounced and difficult to interpret in this graph. This highlights the importance of viewing trends in relative terms. (Fig. 2.12) shows the ratio of entries for each keyword divided by the total number of entries for all keywords in each year, illustrating the relative contribution of each keyword. This makes it clear that certain techniques, such as *Fourier* and *linear regression* have seen a relative decline. In the early 1990s, these two techniques accounted for nearly 30% of entries. *Tracking* gained most of its entries in the 2000s but has

since experienced a sharp decline.

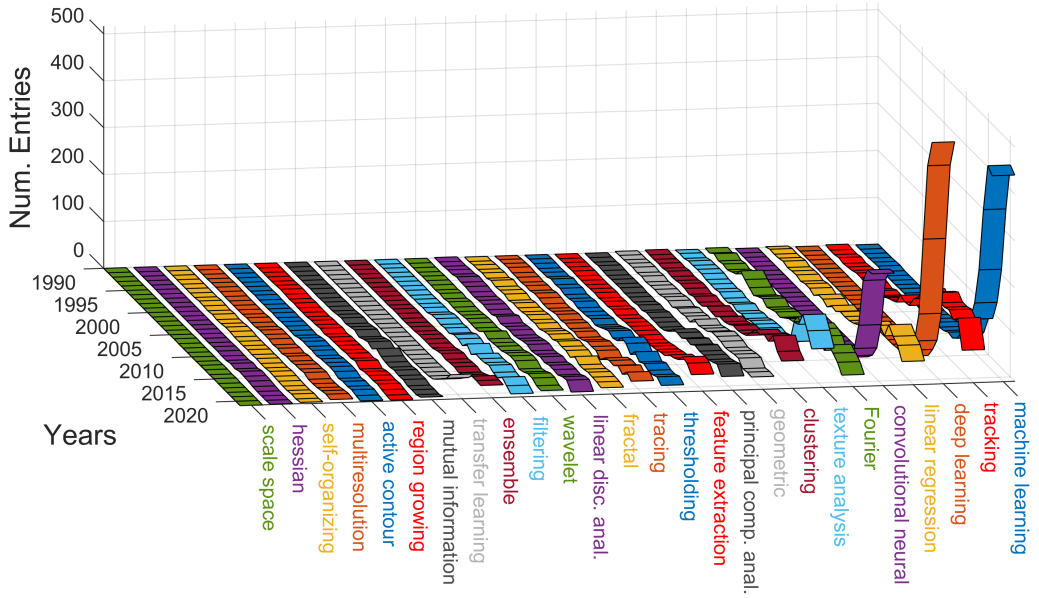


Figure 2.11: Number of entries in PubMed of the specific keywords and the date of publication. Colours have been assigned to the ribbons and the names to assist in the visual discrimination of the computational techniques.

The number of entries for machine learning related keywords is shown in (Fig. 2.13). The vertical scale is logarithmic, as the number of entries for *Regression* and *Classification* is significantly higher than for the other keywords. This is expected as, as these terms encompass broader concepts, such as classification with random forests. Nonetheless, these values provide useful context for comparing the remaining keywords. *Deep learning* has entries one order of magnitude higher than the next keywords; *Feature Selection, Random Forest, Discriminant Analysis* and *Cluster Analysis*. On the lower end, the keywords with the fewest entries were *Feature Engineering, Reinforcement learning, Anomaly detection* and *Hidden Markov Model*. It is important to emphasise that the queries only captured what is indexed in PubMed. If a technique, like feature engineering was mentioned in a paper's methods section but not listed in the PubMed entry fields, that entry would not be counted towards the results of that query. An example of the data extracted through the queries, such as the title, abstract and, MESH terms, can be seen in this URL: *https://pubmed.ncbi.nlm.nih.gov/31783023/?format=pubmed* which corresponds to the work by Lee *et al* [224].

When the trends are analysed on a yearly basis, the trends reveal some interesting patterns (Fig. 2.14). Most of the keywords show stable numbers over time, with one notable exception, *deep learning*. Since 2010, entries for deep learning have surged significantly and are now approaching the numbers for regression and classification. It is likely that in the coming year, deep learning will surpass the other keywords in terms of entries.

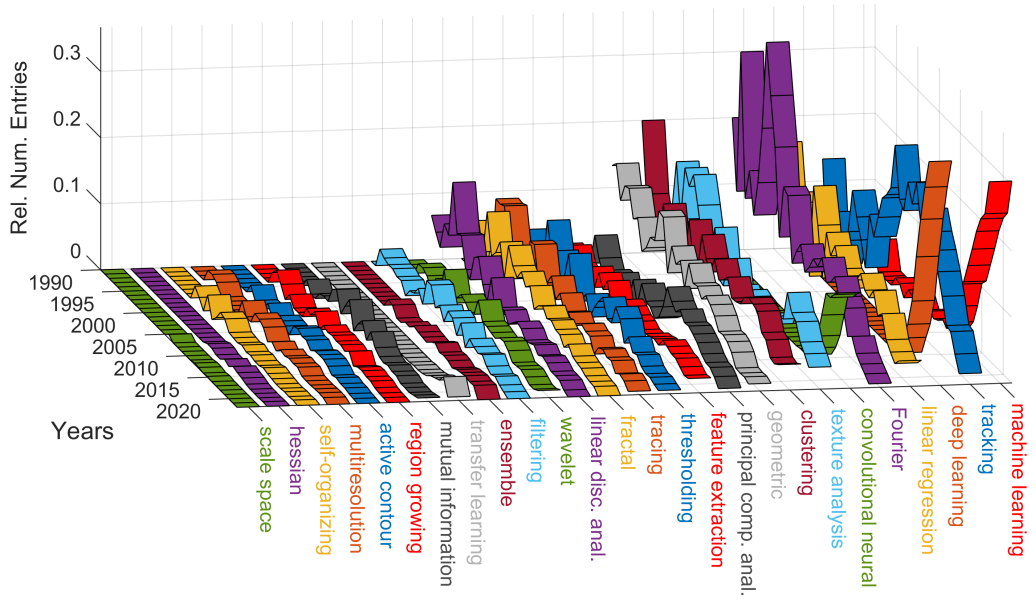


Figure 2.12: Relative number of entries in PubMed per keyword and date of publication. The number of each entry was divided by the sum of all the entries of the keywords of a given year and illustrates the relative decrease of certain techniques. Colours have been assigned to the ribbons and the names to assist in the visual discrimination of the computational technique.

The results for specific deep learning keywords are displayed in (Fig. 2.15), where the most frequently mentioned term is *convolutional neural network*, followed by *U-Net* and *ResNet*. Convolutional Neural Networks (CNN's) have nearly ten times more entries than the other methods. The keywords with the fewest entries are *Auto-Encoders, Transformers* and *Mask-RCNN*. It is important to note that some of these terms are new, so they have not yet accumulated many entries in PubMed.

The entries related to software and programming languages is presented in (Fig. 2.16) It is displayed in two formats: the total number of entries (left) and the yearly count (right), with logarithmic scales on the vertical axes in both cases. *Fiji* and *ImageJ* recorded the highest number of entries this was followed by *Matlab*, *R*, and *Python*. One order of magnitude lower were *QuPath*, *ICY*, and *CellProfiler*, which is expected since the former are general-purpose platforms and languages applicable to various tasks, whereas the latter are more specialised for specific applications in Quantitative Pathology, Bioimage informatics, and cell image analysis. As previously noted, the query searches were conducted using different approaches and did not include the keywords "cancer", "pathology", or "image analysis", therefore they cover all entries indexed in PubMed.

The lowest number of entries was attributed to *C*. The relatively low number could be due to *C* not being commonly used by researchers whose work is indexed in PubMed. Additionally, projects developed in *C* may lead to larger software tools that are recognised

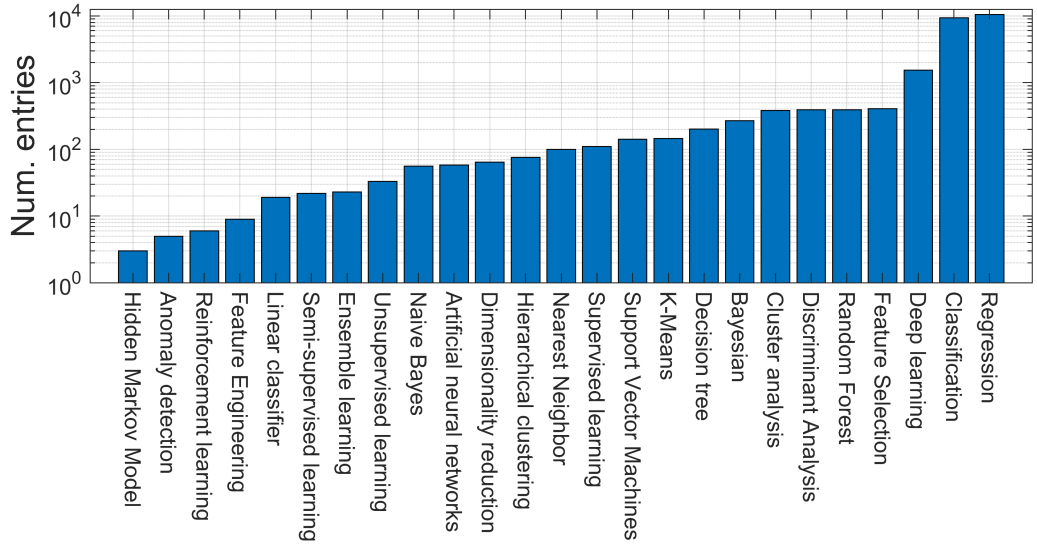


Figure 2.13: Number of entries in PubMed per specific keyword of all Machine Learning Techniques. The terms "Regression" and "Classification" include other keywords (e.g., Classification with Random Forest) and therefore are more frequent. It is illustrated that the term "Deep learning" is one order of magnitude higher than all others. The vertical axis is logarithmic.

under different names. For example, *ICY* is java-based software platform.

The yearly entries reveal that Fiji has surpassed ImageJ in recent years, while a similar pattern is evident with R and Python overtaking Matlab. Notably, QuPath has experienced the most rapid growth, outpacing all other software and programming languages. If this trend persists, it is likely that QuPath will soon reach the same number of entries as the programming languages and may eventually approach the levels of ImageJ and Fiji.

This scoping review offers a broad overview of the computational techniques used to analyse images in Cancer research. Machine learning and Deep learning methods have become dominant in recent studies indexed in PubMed. Although other techniques are still present in the database of PubMed, their relative proportions are declining compared to the growing use of Machine and Deep learning approaches.

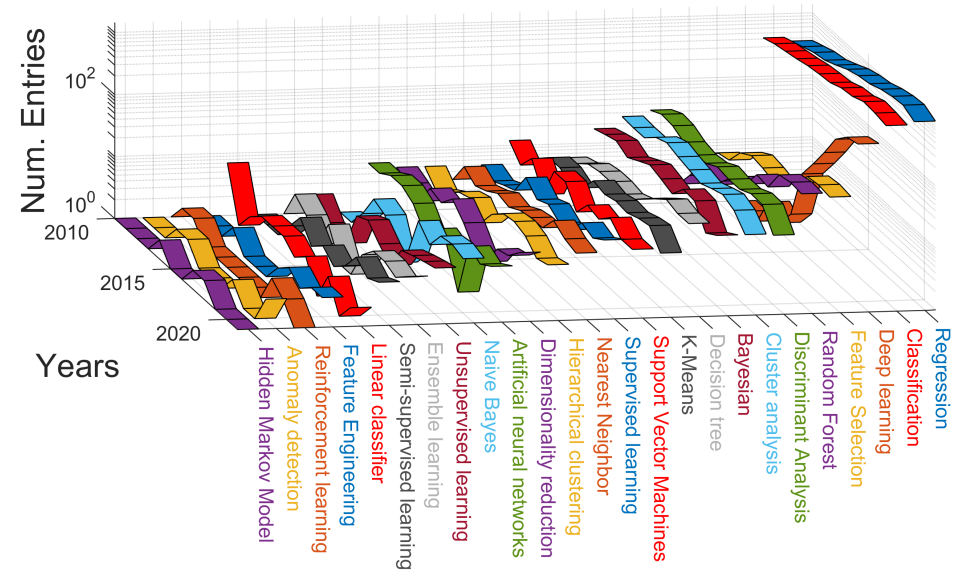


Figure 2.14: Number of entries in PubMed per year and each keyword of the Machine Learning Techniques. Most entries stayed the same but "Deep learning" has increased constantly. The vertical axis is logarithmic.

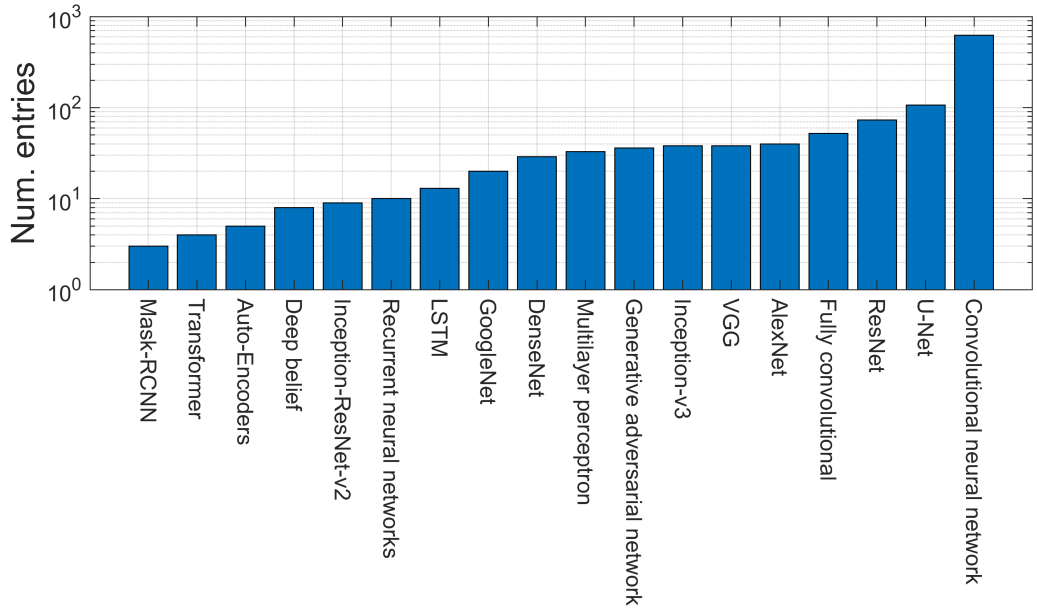


Figure 2.15: Number of entries in PubMed per specific keyword of the Deep Learning Techniques and architectures like AlexNet, VGG and U-Net. The vertical axis is logarithmic.

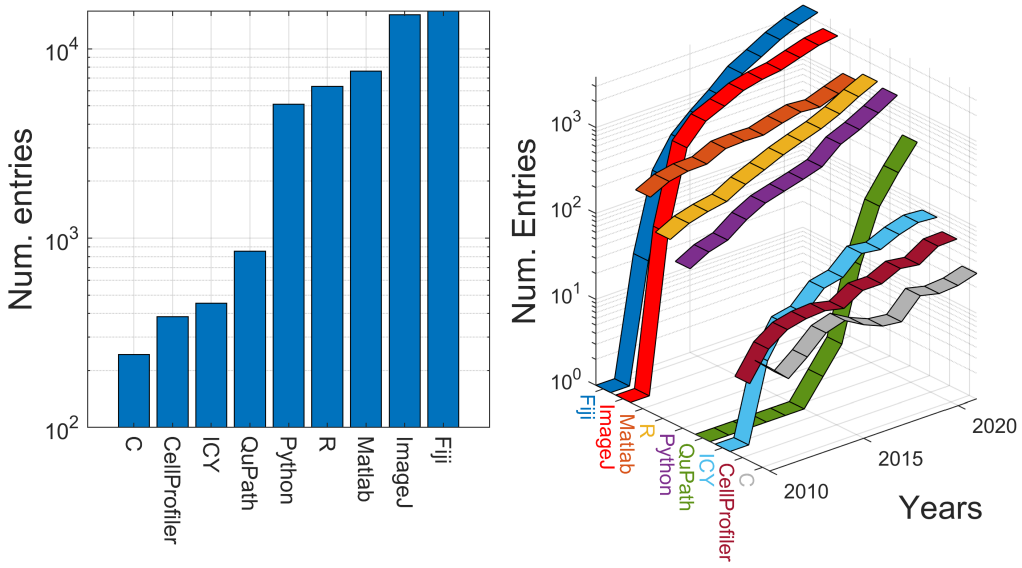


Figure 2.16: Number of entries in PubMed for eight programming languages: Fiji, ImageJ, Matlab, R, Python, QuPath, ICY, CellProfiler and C. On the left are the entries for the keywords for the programming languages and all the previous keywords *(pathology) AND (cancer) AND ((image) OR (imaging)) AND (2010:2022[dp])* and on the right just for the programming languages across the years. The vertical axis is logarithmic in both graphs.

2.6 Summary:

In this chapter the definition of cells, cellular processes and the extracellular matrix are outlined and their role in the progression of diseases and in specific cancer. This leads to a definition of cancer and the diagnosis of it as well as the role of image analysis in cancer. The scoping review explores the image processing techniques used to investigate cancer images. It is found that at the time of writing, cancer research articles make up 13% of PubMed entries. With the advancements of technologies, it is important to understand the trends of which computational techniques have been used in cancer image analysis as it contributes to operative planning, decision making and survival predictability. Recently with the rise of computational powers, techniques of machine learning and deep learning have been widely used, this raises the question if artificial intelligence techniques are replacing earlier image processing and computer vision methods in cancer image analysis. Within years of analysis between 1990 and 2022, the highest number of entries in PubMed were related to machine learning, tracking and deep learning with each method exceeding 1500 entries. The keywords that provided the least number of entries are mutual information, region growing, multi resolution, self-organising, hessian and scale space. The year-on-year analysis illustrated a different picture. Certain techniques have exponentially grown at a faster rate notably machine learning, deep learning, and convolutional neural networks, while other techniques such as tracking, linear regression and Fourier have seen a decline over the years. In terms of machine learning, deep learning has the highest number of entries with an order of magnitude higher than the next keywords of feature selection and Random Forest. On the lower end, keywords that had the fewest entries were Feature Engineering, Reinforcement Learning, Anomaly Detection and Hidden Markov Model. When reviewing the year-on-year trends, all keywords showed stable entries over time except for deep learning which has significantly surged in numbers over the years. Deep learning keywords entries were analysed. Convolutional neural networks exhibited the highest number of entries followed by U-Net and ResNet, CNNs have almost ten times more entries than the next keyword. The keywords with the fewest entries were Auto-Encoders, Transformers and Mask-RCNN. Fiji, ImageJ observed the highest number of entries followed by Matlab, R and Python. The lowest entries were observed by C, CellProfiler and ICY. Looking at the year-on-year change the use of QuPath is exponentially growing.

Chapter 3

Methods & Materials

3.1 Overview:

In this chapter different methodologies and techniques used throughout this thesis are explored. Firstly, acquisition methods are outlined which leads to the concept of staining biopsies and their image acquisition through microscopy. This is followed by the four datasets that will be used in the rest of the thesis. An image is defined, and traditional image processing techniques are outlined and investigated. The definition of segmentation is presented and Tracing algorithms that are used to delineate and detect fibrial like structures which are used to trace fibres of the extracellular matrix are investigated. Classification algorithms are outlined followed by UMAP [225]. The definition of ground truth and its uses are described. Finally, the metrics used to measure performance for the various investigations in this thesis are outlined. Additionally, the chapter explores cell morphological metrics, pixel-wise metrics, fibre metrics and finally object-wise metrics.

3.2 Histopathology:

Histopathology [226] is the study of the indicators and signs of a disease existing in a sample of tissue or cell using microscopy. Staining is used to show different cellular components, but before staining could occur the sample of tissue must first be prepared for staining. Firstly, the tissue goes through fixation, a process through which chemicals are used to maintain the structure of the tissue. By fixating the sample also becomes hardened which helps in sectioning later. Neutral Buffered Formalin is common for this task. After fixation, the tissue sample goes through dehydration. The sample is treated with ethanol to remove water. Once the tissue is completely dehydrated, xylene is used to remove the alcohol [227]. Once dehydration is completed the tissue is embedded. Embedding is the process in which the tissue sample is submerged into a paraffin wax or a plastic resin. If the goal is to immunostain the tissue sample must be deparaffinised or antigen retrieval must be introduced to ensure precise antibody binding. Sectioning follows embedding, where the tissue sample is cut into sections to be placed under the microscope for examination. Antigen retrieval involves

heating the tissue sample in a buffer and revealing the target proteins [226]. Staining is done after the antigen retrieval. Staining aims to highlight different features of the tissue also intensifying the contrast. Many staining protocols exist but perhaps the most common staining process is using Haematoxylin and Eosin [228], cell nuclei appear as purple/blue, Cytoplasm and connective tissue appear as clear/pink [229, 170]. An example of a stained sample is seen in (Fig. 3.1). Many of the images used throughout this thesis are stained.

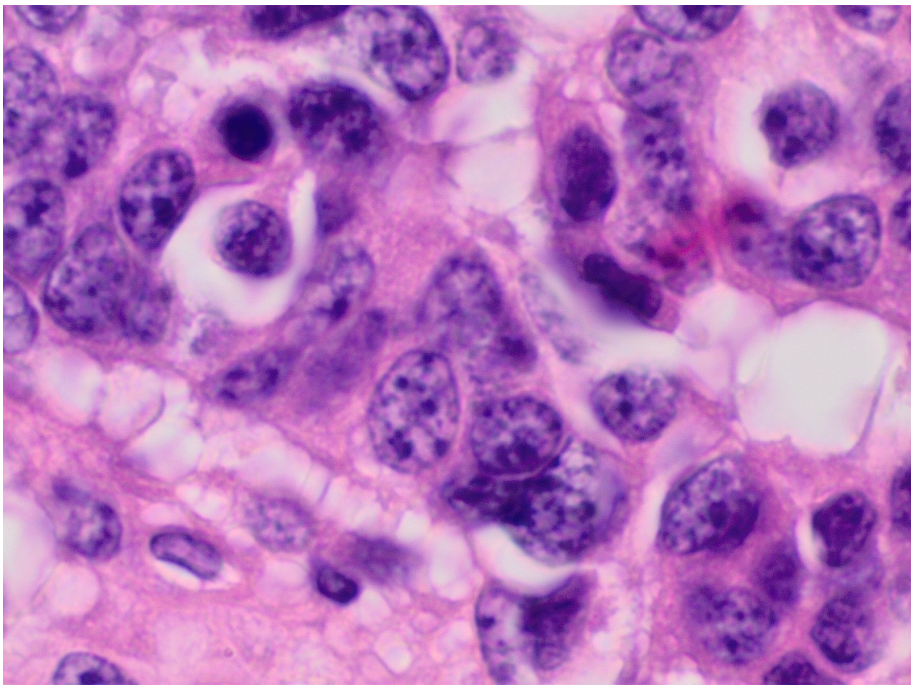


Figure 3.1: Histopathology of well differentiated neuroendocrine tumour with salt and pepper chromatin. The nucleus exhibits a deeper purple colour which is different from all the different components of the cell [230]. Image is available through Wikimedia Commons.

Immunohistochemistry [231] is a technique that uses antibodies which bind to various substances known as antigens on the cells. The binding is not exclusive to antibodies but also to a large number of agents like virus or bacteria. Multiple chemicals can be added to the antibodies, directly or through a second antibody which provide colour [232].

3.2.1 Tissue Microarray (TMA):

Tissue microarrays [233] were developed as a technique to evaluate many samples of tissue in a single slice, thus instead of using a whole section of tissue, only small portions of it are used but in turn a large number of samples can be included in a single block. Tissue microarrays use many tissues in a single paraffin slide using an exact size and shape to prepare the recipient block. This has become an essential tool in the evaluation and research of cancer and other diseases [234, 235]. This technology allows researchers to evaluate diseases at

an early stage. Above 100 tissue samples are placed in a single block. Different sections can be cut from the array block, allowing hundreds of tissue cores to be placed on a single viewing slide. Many staining procedures can be used with this technique. TMAs ensure standardisation as all the tissue samples are treated and pre-processed in the same manner before their evaluation. An example of a TMA of Normal Breast Tissue can be seen in (Fig. 3.2). This sheet includes 120 cores, and it has a size of $35200 \times 48000 \times 3$ pixels. When zoomed in, each core can be observed as shown in (Fig. 3.20). One of the main datasets used in this thesis is a Tissue Micro Array dataset.

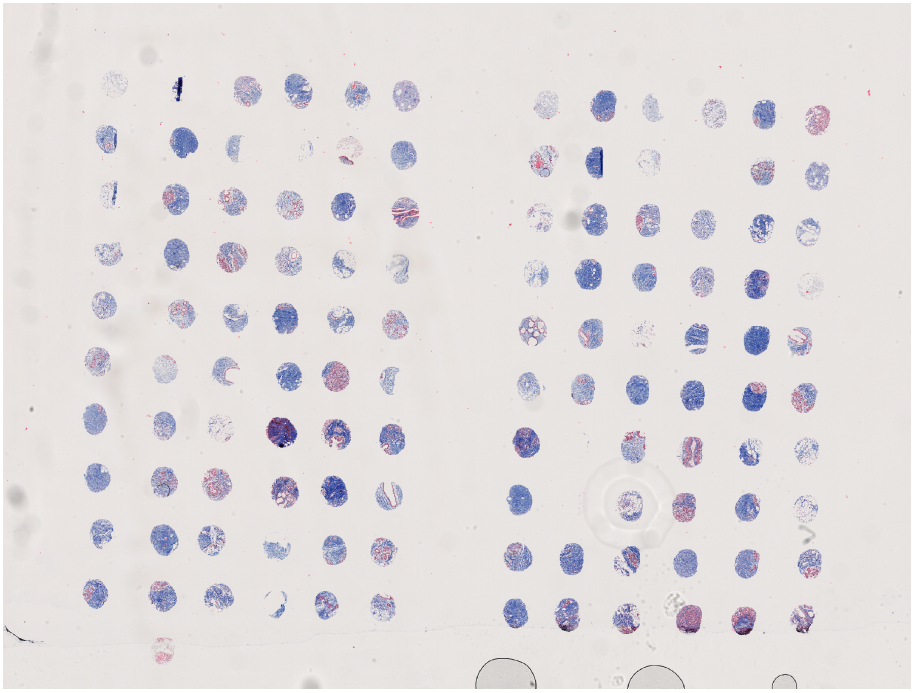


Figure 3.2: Tissue Micro Array containing 120 cores of Normal Breast Tissue. When zoomed in each core reveals details of the structure of the extracellular matrix where collagen is dyed in blue and cell nuclei is in red. Image courtesy of Dr Oscar Maiques, Barts QMUL.

3.2.2 Staining Protocols:

Staining protocols are the techniques of staining biological samples to enhance contrast and visualisation under a microscope; the following protocols have been used on datasets used throughout this thesis.

Masson's Trichrome:

Masson's Trichrome [236] is a histological staining protocol that particularly dyes collagen fibres. It uses three stains which are Weigert's Haematoxylin, Biebrich scarlet-acid fuchsin solutions, and Aniline blue. The stain is used for collagen fibre detection particularly in areas

such as the breast, skin, heart, and muscles [237, 238]. Weigert's Haematoxylin is used to stain the nuclei. Beibrich scarlet-acid fuchsin solution can stain all the acidic tissues such as the cytoplasm, and muscle. Aniline blue stains the collagen. The collagen fibres appear blue; nuclei appear black with a red background and the cytoplasm appears red/pink. The benefit of using Masson's Trichrome is its ability to differentiate the collagen fibres by producing clear images of the collagen compared to other staining agents. An example of a tissue slide observed with Masson's Trichrome staining can be viewed in (Fig. 3.19).

Picrosirius red:

Picrosirius red [239] is a histological staining agent that is used to visualise collagen. It can be used to differentiate between type I and type III collagen fibres. Collagen and fibrial structures appear in red. Nuclei appear black but could also appear brown or grey. An example of collagen stained with Picrosirius red can be found in (Fig. 3.14c).

Fluorochromes:

Fluorochromes are molecules used to label cells, tissues, and proteins to be examined by a fluorescence microscope. Fluorochromes work by absorbing light of a particular wavelength (excitation wavelength) and re-emit that energy at another wavelength (emission wavelength) depending on the contrast [240].

An example of a cell under fluorescence microscopy can be clearly seen in (Fig. 3.11). This is an example of a RBD cell culture from dataset 1, where the nuclei are seen in blue as they were dyed with DAPI, the actin filaments can be seen in red as it was dyed with Phalloidin, the fibronectin fibres can be seen in grey.

DAPI also known as 4',6-diamidino-2-phenylindole is a blue fluorescent DNA stain, due to it having a high affinity for DNA it is used for counting cells and in nuclear segmentation [241].

Phalloidin is a fluorescent dye if it is conjugated, that has a high affinity for the F-actin, which are linear polymers which form the microfilaments in the cytoskeleton as well as thin filaments which are a component of the contractile components of muscle and non-muscle cells. The Phalloidin stained actin filaments in the work of this thesis appear in red [242].

3.2.3 Microscopy:

Microscopy is an optical technique that is used to visualise structures that cannot be seen with the naked eye which are the objects that are not within the resolution range of the normal eye [243].

The human eye will only respond to colour and brightness of the electromagnetic radiation. The wavelength response is limited to light in the visible region of 380 and 780 nm of the electromagnetic spectrum [244]. The nearest an object can be brought to an eye and maintain focus is known as Least Distance of Distinct Vision [245], this is commonly taken to be 250 mm. The closer an object is brought to the eye, the larger it appears. The size of the

object remains unchanged, the image is spread over a larger area of the retina. The human can see objects that are approximately 0.2 mm in size and larger, therefore microscopes are used to see objects smaller (Fig. 3.3). The limit of the light microscope is three orders of magnitude smaller around 200 nm [246].

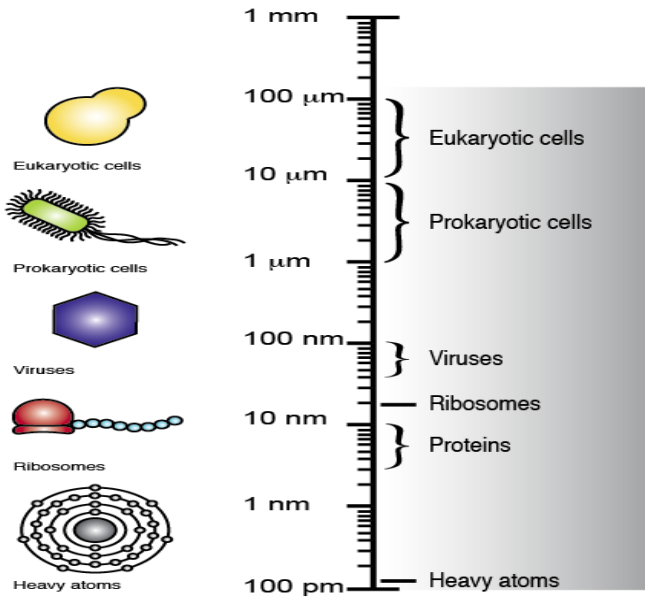


Figure 3.3: Limitations of the human eye, which can only see objects larger than 0.2 mm in size. Eukaryotic cells exist between 100 μm and 10 μm, Prokaryotic cells between 10 μm and 1 μm, viruses are below 100 nm, ribosomes just larger than 10 nm and proteins below that. While heavy atoms are just larger than 100 pm. Image is available through Wikimedia Commons [247].

There are several types of microscopies which could be classified into light microscopy, electron microscopy, and scanning probe microscopy. The images provided and used throughout the thesis have been observed with light microscopy techniques.

A light microscope is any type of microscopy that utilises photons to illuminate and visualise a specimen [248]. Light microscopy includes white light with all the wavelengths, as well as light from specific wavelengths as used in fluorescent microscopy. Many types of light microscopy exist those include perhaps the most popular is Brightfield microscopy, as well as fluorescent microscopy and confocal reflectance microscopy.

A simple microscope (Fig. 3.4) uses a single convex lens which acts as a magnifying glass to magnify objects up to approximately 10x [249]. Its key features include using a single lens for magnification as well as its portability and low cost. Its limitations include low resolution, and it is unsuitable for observing any sub-cellular structures [250].

More commonly compound microscopes (Fig. 3.5) are used which use multiple lenses with a light source to observe and visualise specimens with a much higher magnification [250].

Usually, a sample is illuminated by a light source followed by a condenser lens, where

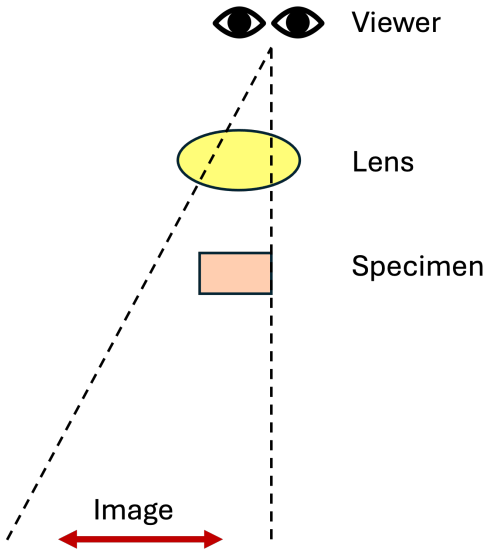


Figure 3.4: Illustration of a simple microscope, which uses a single convex lens to act as a magnifying glass.

small features in the sample diffract light. The light is bent outwards by diffraction into a "fan". The objective, which is a component made up of lenses captures the image detail carried by the diffracted light, which forms the primary image. The primary image is then magnified by the eyepiece.

The major components of a light microscopy include the light source which provides illumination on the sample to observe the specimen, depending on the type it can be in the form of a light emitting diode, halogen lamp or a mirror that is directing external light. Some microscopes may allow certain adjustments to the light intensity for optimal contrast. A condenser lens focuses and concentrates light onto the specimen, and it is located below the stage. The condenser lens is often used with an iris diaphragm to control the diameter of the light beam, this is to enhance the image clarity as well as the contrast [251]. The stage is a flat platform where the specimen slide is to be placed. It contains clip or mechanics to contain and accurately move the slide. The objective lens is the main magnifying lens placed closest to the specimen. It is available in multiple magnifications 4x for scanning, 10x low power, 40x high power and 100x oil immersion [252]. High quality objective lenses often correct for optical aberration for example achromatic, plan, or apochromatic lenses. The eyepiece also known as the ocular lens is a secondary lens through which the observer views the image, it is commonly in 10x magnification. The total magnification is the objective lens multiplied by the eyepiece lens [253]. The coarse and fine focus knobs components adjust distance between the objective lens and the specimen. Coarse focus is generally used for large and quick adjustments with low power objectives. Fine focus is used for precise, small adjustments used with higher magnifications. The body tube connects the eyepiece to the objective lenses and maintains the correct alignment of optical components. The turret is the

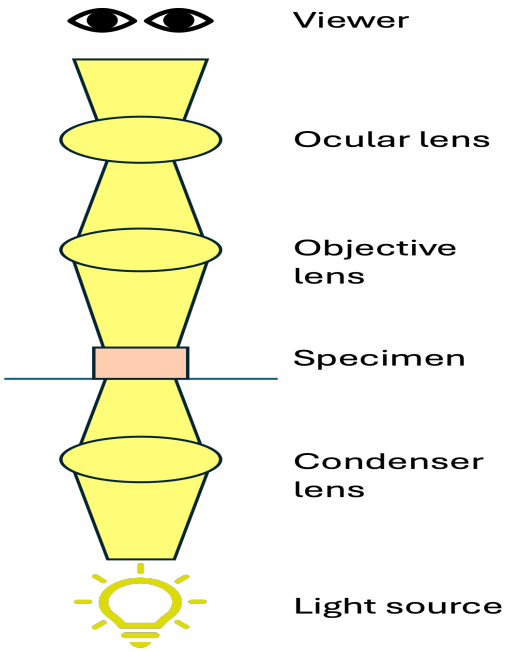


Figure 3.5: Compound Microscope. A sample is illuminated by a light source which then passes through a condenser lens, which focuses the light into a cone or "fan" which lights the sample evenly. The light then passes through the objective lens where the image is both inverted and magnified, this forms the primary, real image inside of the microscope. The ocular lens acts as a simple magnifying glass which magnifies the image produced by the objective lens. The final image seen by the eye is a virtual, magnified, and inverted image of the sample.

rotating mount that hold many objective lenses, it also allows for quick interchange between the different magnifications. Finally, the arm and based provide the structural support for a microscope.

One of the most fundamental optical techniques used in light microscopy is Köhler illumination [254], it is a method that provides uniform, high illumination on the field of view. It utilises both contrast and resolution by ensuring that the sample is evenly lit, thus eliminating artifacts as well as uneven brightness or glare. The light source is focused on the plane of the condenser diaphragm which controls the angle of illumination and thus directly affecting the contrast and depth of field [255]. The field diaphragm which controls the area of illumination reducing stray light is focused at the plane of the specimen. The end result is that the illumination is bright, even as well as free of artifacts. This technique is essential as it enhances contrast and resolution, as well as allows reproducibility and standardisation of microscopic condition. Köhlers illumination is essential for advanced microscopical techniques such as fluorescence microscopy. Köhlers illumination can be seen in (Fig. 3.6), where the top part represents the illumination beam path which shows how the light travels from its source through the microscope to illuminate the specimen sample. The light source is

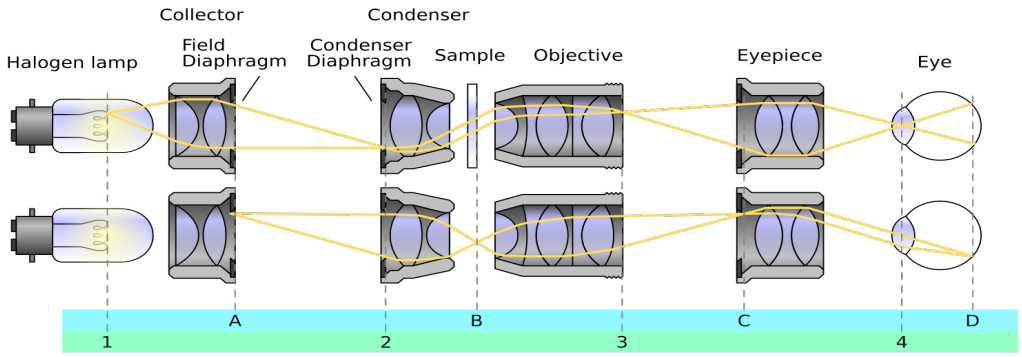


Figure 3.6: Köhler's illumination. The top part represents the illumination beam path where the light travels from its source through the microscope to illuminate the sample. The light is defocused at the specimen plane to result in uniform illumination, the conjugate planes involved are indicated with the light green bar. The bottom part represents the light path after passing through the specimen to form the image, the conjugate planes involved are indicated in the blue bar. This image is copyright free and was made available through Wikimedia Commons [256].

deliberately defocused at the specimen plane to result in uniform, even illumination. The conjugate planes involved in this part are the light source, field diaphragm, condenser aperture and the specimen plan as indicated with the light green bar. The bottom part represents the light path after passing through the specimen to form the image. The path is in charge of producing sharp, detailed images at the camera or eyepiece. The conjugate planes include the specimen, the objective's intermediate image plane as well as the eyepiece and camera sensor plane, as indicated in the blue bar.

Microscopes must provide resolving power, which is the ability to carry information about fine detail of the specimen to the image [257]. They must also provide contrast which is the differences in the image between features and their surroundings. Most biological samples are thin and translucent, as well as difficult to stain if they are living. Therefore, techniques such as Brightfield and Fluorescence are used to increase contrast. Finally, microscopes must also provide magnification to make the images large enough for the eye to view the detail of the sample. The following microscopical techniques have been used to visualise and capture images used throughout this thesis.

Brightfield:

Brightfield microscopy is a simple type of light microscopy, where illumination light from a halogen or LED lamp is transmitted through the tissue sample [258]. Contrast is created when dense areas in the sample absorb the light and light is attenuated, which is collected by an image sensor. The Tissue micro arrays were observed with a Brightfield microscope seen in (Fig. 3.19). In Brightfield microscopy the specimen appears dark on a white background while on the contrary in Darkfield microscopy, the specimen appears light on a dark background

[258]. In Brightfield the sample is illuminated from below while in Darkfield the sample is illuminated using a special condenser at an angle. A condenser is a lens system or one lens that is located above the light source but below the sample. Their purpose is to gather light from the light source and confine it through a cone to shine on the specimen [259].

The advantages of Brightfield microscopy [254, 251] is in its simplicity and accessibility as it is very easy to use and is widely available, it also requires minimal training in comparison to the more advanced techniques. In addition to this, it is relatively cheaper in comparison. Brightfield microscopy works well with stained biological specimens such as H&E stained tissue samples, providing a better view of the structural details [260]. Furthermore, this technique allows live viewing of samples and could be beneficial for quick assessments. For some specimens, the amount of preparation is minimal with this technique they can also be fitted with cameras to capture images of the specimen for analysis. The disadvantages of Brightfield microscopy [254, 251] are its limited contrast for unstained samples, poor visualisation of transparent and colourless specimens that have not been stained. It also has a low optical resolution compared to more advanced techniques, where it is limited by the diffraction of light and therefore it cannot visualise some subcellular structures such as organelles or macromolecular complexes. Brightfield microscopy only provides flat, two-dimensional images therefore, it can lack depth of information. Finally, thick or incorrectly prepared samples can often lead to poor image clarity and focus issues.

Fluorescence Microscopy:

Fluorescence microscopy deploys the use of fluorochromes to enhance contrast [261]. Fluorescent microscopes use high intensity light sources such as laser, mercury, or xenon lamps, filtered to achieve the required wavelength. As mentioned earlier fluorochromes absorb photons of a specific wavelength and later emit photons of another wavelength, as energy is lost, the emitted photons have experienced a decrease in energy compared to the absorbed photon. This difference is also known as Stokes shift (Fig. 3.7).

The aim of fluorescence microscope is to separate the emitted dim light from the bright excitation light. Separation of the light is done with the use of optical filters.

A dichroic mirror is used to pass the excitation light towards the sample, while allowing the emitted light to pass through to the detector. This process ensures clear imaging by blocking the excess excitation light. The emitted photons are detected by highly sensitive photomultiplier tubes and more commonly cameras, which convert the light signal into digital data. This allows the microscope to build the image of what is being viewed under the microscope. The confocal microscope uses a pinhole which blocks out of focus light. Here the pinhole is in front of the detector. This process enhances the image contrast and lowers the background noise. An example of a fluorescence microscope is illustrated in (Fig. 3.8).

The advantages of using fluorescence confocal microscopy are that the use of fluorochromes can target specific structures inside complex samples. Detecting small structures is also enabled through the emitted photons. The disadvantage of fluorescence microscope is the phenomenon known as photobleaching, where fluorochromes are degraded over time when exposed to the light for a long time, resulting in signal loss. Samples being observed can

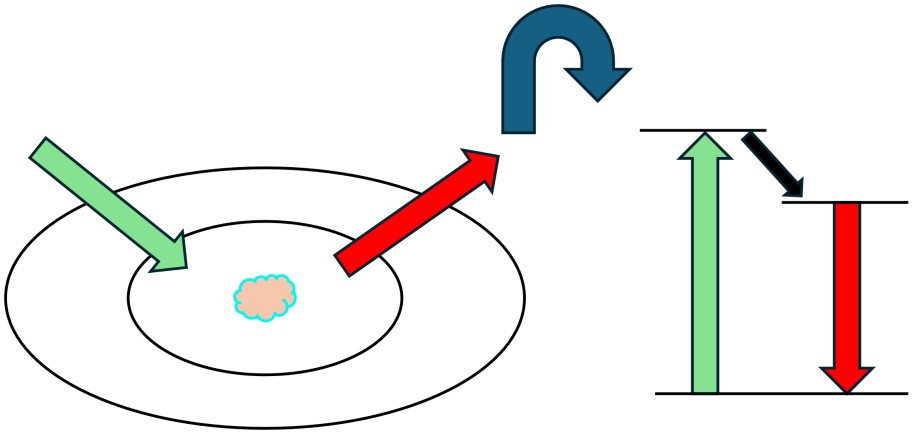


Figure 3.7: Illustration of Stoke's Shift. Photons of high energy are absorbed by the sample and photon of a lower energy are emitted. The difference in the wavelength is known as Stoke's Shift (difference between red and green lines).

also be damaged due to prolonged exposure of light [262].

Some of the applications of fluorescence microscopy include visualising cellular structures such as the nucleus and mitochondria [264], protein localisation [265], tracking dynamic processes in live cells [266] and diagnosing diseases through fluorescent markers [267]. As well Fluorescence in situ hybridisation (FISH), a technique commonly used to locate and detect specific DNA sequences on chromosomes [268], an example of FISH can be seen in (Fig. 3.9)

Confocal Reflectance Microscopy:

Confocal Reflectance Microscopy (CRM) uses a diode laser as a source of light [270]. The light travels through a beam splitter and a focus lens where it illuminates the tissue being studied. Light from the focal point is reflected through the lens and focuses through a pinhole which forms the image. As it is confocal the pinhole only allows light from the focal point to pass through as explained earlier. CRM depends on the back scattering off light from the varying contrast cellular structures of tissue. Reflectance happens at the edges of two structures with different refractive indices.

The advantage of CRM is that it can control the depth of field and removes unwanted background information by focusing the focal plane away from it. An example of collagen observed with confocal reflectance microscopy can be seen in (Fig. 3.15) [271].

Some of the applications of confocal reflectance microscopy is in dermatology where non-invasive imaging is used on skin lesions [272]. Confocal reflectance microscopy is also used in cancer diagnosis with early detection of oral cancers [273] as well ophthalmology for corneal imaging [274].

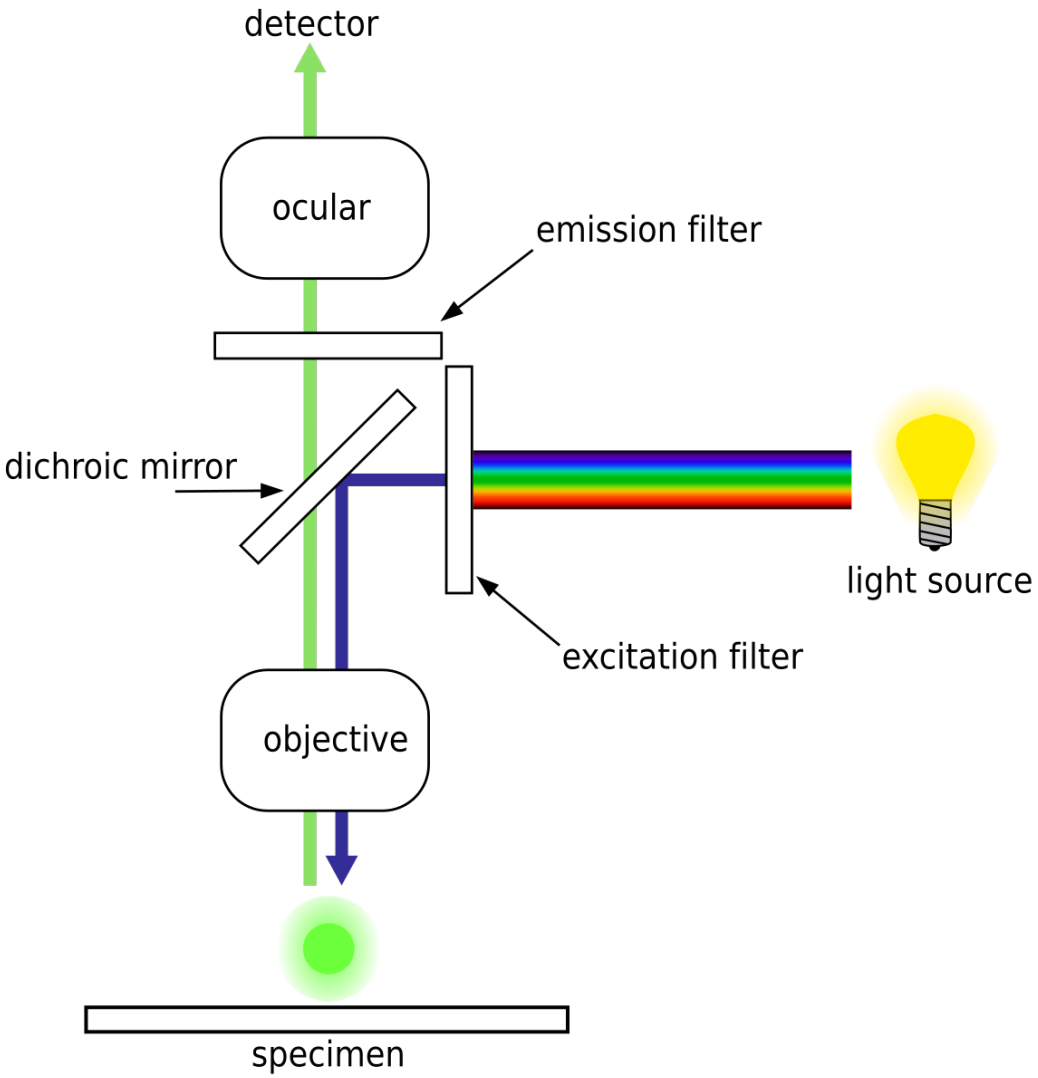


Figure 3.8: Illustration of a Fluorescence microscope. Light source emits photons which passes through the excitation filter to limit the light to a specific wavelength, the photons are then passed to the specimen through the objective lens via the dichroic mirror. Photons of a different wavelength are emitted from the specimen which pass through the emission filter towards the ocular lens and finally to the detector where the light signals are converted into digital signals [263].

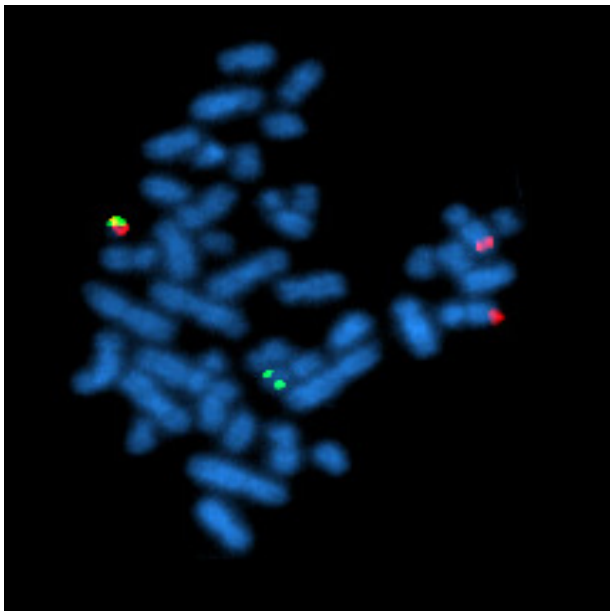
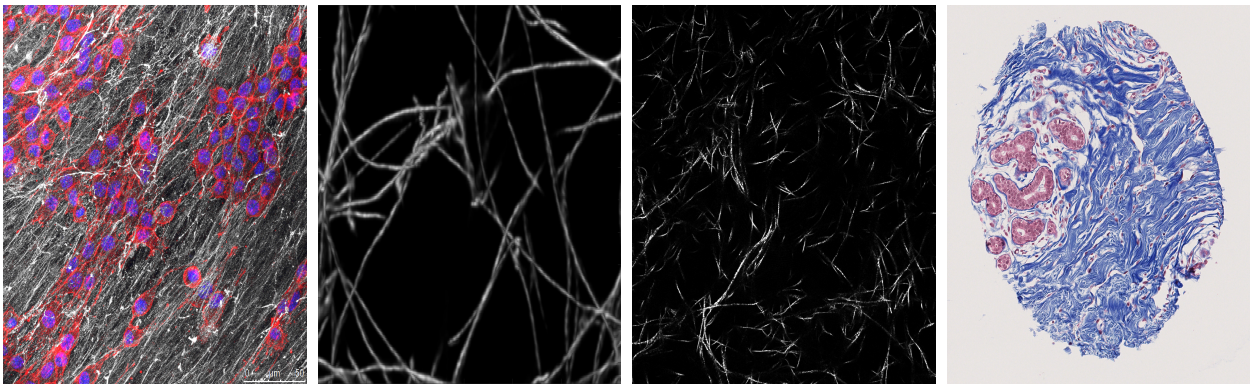


Figure 3.9: Illustration of FISH. The rearrangement of bcr and abl genes in the chromosomes (blue) of a cell. The red green colour in the top left of the image indicated the presence of the bcr-abl 1 fusion gene [269].

3.3 Datasets:

In this section, the four datasets used throughout this thesis will be described (Fig. 3.10). The datasets are also outlined in Table. 3.1, The first dataset provided in the thesis set the tone of the work conducted in the thesis. Initially, they were experimented with traditional image processing techniques to segment the objects, extract morphological features and compare populations. Once it was decided that the work of the thesis will focus on the ECM fibres, the fibronectin fibres images were re-used with the developed algorithm. The images in dataset 2 were used to compare the developed algorithm 'Trace Ridges' with other state-of-the-art fibre tracing methods to determine which algorithm performs best. The third dataset was used to investigate whether the accumulation of collagen fibres results in cytoskeletal remodelling. Fibres from those images were extracted using Trace Ridges. The fourth dataset was used to investigate whether there are differences between Normal and Tumour cores using Trace Ridges and CellPose. Classifiers were also trained on those images.



(a) (b) (c) (d)

Figure 3.10: Datasets example: Four datasets are used throughout this thesis. (a) Dataset 1: Fluorescent three channel image that is broken down into the DAPI stained cell nuclei, Phalloidin stained actin filaments and extracellular matrix Fibronectin channels. (b) Dataset 2: this dataset is made up of four images of fibrial-like structures from different backgrounds. (c) Dataset 3: this dataset is made up 3 sets of images. All the images are of the extracellular matrix collagen under different conditions. Dataset 4: this dataset is made up Tissue Micro-Arrays of breast tissue.

3.3.1 Dataset 1:

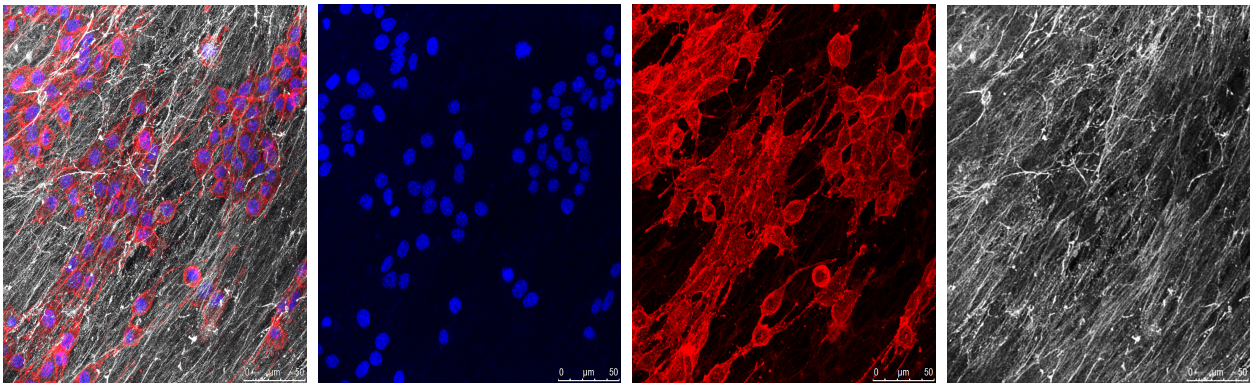
The dataset contained a set of ten images with three different channels (Fig. 3.11a), the DAPI channel (Fig. 3.11b), the Phalloidin channel (Fig. 3.11c) and the Fibronectin channel (Fig. 3.11d). Each channel contained five images for cells altered with a Receptor Binding Domain (RBD) and five for the Wild Type (WT) cells. The cells from the RBD domain have

Datasets	Type	Populations	Number of images
1	Cells - DAPI Channel	Ras Binding Domain (mutated RBD)	5
		Wild Type (WT)	5
	Actin Filaments - Phalloidin Channel	Ras Binding Domain (mutated RBD)	5
		Wild Type (WT)	5
	Fibronecin Fibres	Ras Binding Domain (mutated RBD)	5
		Wild Type (WT)	5
2	fluorescently Labelled Bovine Collagen I (FC)	\	1
	fluorescently Labelled Fibronectin (FF)	\	1
	Breast Cancer Biopsy (BCB)	\	1
	Disease Mimicking ECM (DME)	\	1
3	Concentration	1.7 mg/ml	198
		6.6 mg/ml	198
	Marangoni Flows	Control	134
		Marangoni Flow induced	150
	Protease Inhibitors	Control	156
		Addition of protease inhibitors	180
4	Tissue Micro Arrays (TMAs)	Normal	120
		Tumour	120

Table 3.1: Outline of datasets. Dataset 1 contained Fluorescent three channel image. Dataset 2 is made up of four images of fibrial-like structures from different backgrounds. Dataset 3 is made up 3 sets of images. All the images are of the extracellular matrix collagen under different conditions. Dataset 4 is made up Tissue Micro-Arrays of breast tissue.

been altered to fight against cancer. They were altered on the cell membrane on how their cell receptors work and interact with cancerous cells. They are known as mutated cells. They mutant images were from an LKR13 cell line. A cell line is a cell culture that was grown from one cell only. LKR13 cell line is extracted from a K-ras driven mouse lung tumour [275]. Wild types are the usual or normal configuration of a cell or protein as found in nature. Therefore, it is not altered in any way. Images are courtesy of Dr Esther Castellano, Universidad de Salamanca.

As it is shown in (Fig. 3.12a) for the RBD and Fig. 3.13a) for the WT, the DAPI stain has highlighted the nuclei in a bright blue fluorescent. These images were used as the primary input for the DAPI analysis and cell segmentation. The Phalloidin has highlighted the actin filaments in red as seen in (Fig. 3.12b) for the RBD and Fig. 3.13b) for the WT, these images act as the input for the Phalloidin channel actin filaments segmentation and analysis. The last layer or the third channel for analysis is the Fibronectin. Fibronectin is an extracellular matrix component. It is also a medium of communication between the intra and extracellular environments and it does this through binding integrin receptors to the cell surface [276]. The preparation of the fibronectin was briefly outlined in [277]. The Fibronectin fibres exhibit a hair-like structure, and these fibres have gaps between them (Fig. 3.12c) for the RBD and Fig. 3.13c) for the WT. These images act as the input for the Fibronectin analysis.



(a) (b) (c) (d)

Figure 3.11: Dataset 1 example: Fluorescent microscopy image of a mutant (RBD) cell culture with 3 layers: DAPI, Phalloidin, Fibronectin. (a) RBD sample where three colours can be seen for the three layers. (b) The nuclei can be seen in blue as it was dyed with DAPI. (c) The actin filaments in red as it reacted to the Phalloidin dye. (d) The extracellular matrix fibronectin fibres can be seen in grey. Images courtesy of Dr Esther Castellano, and Dr Cristina Cuesta-Apausa, Universidad de Salamanca.

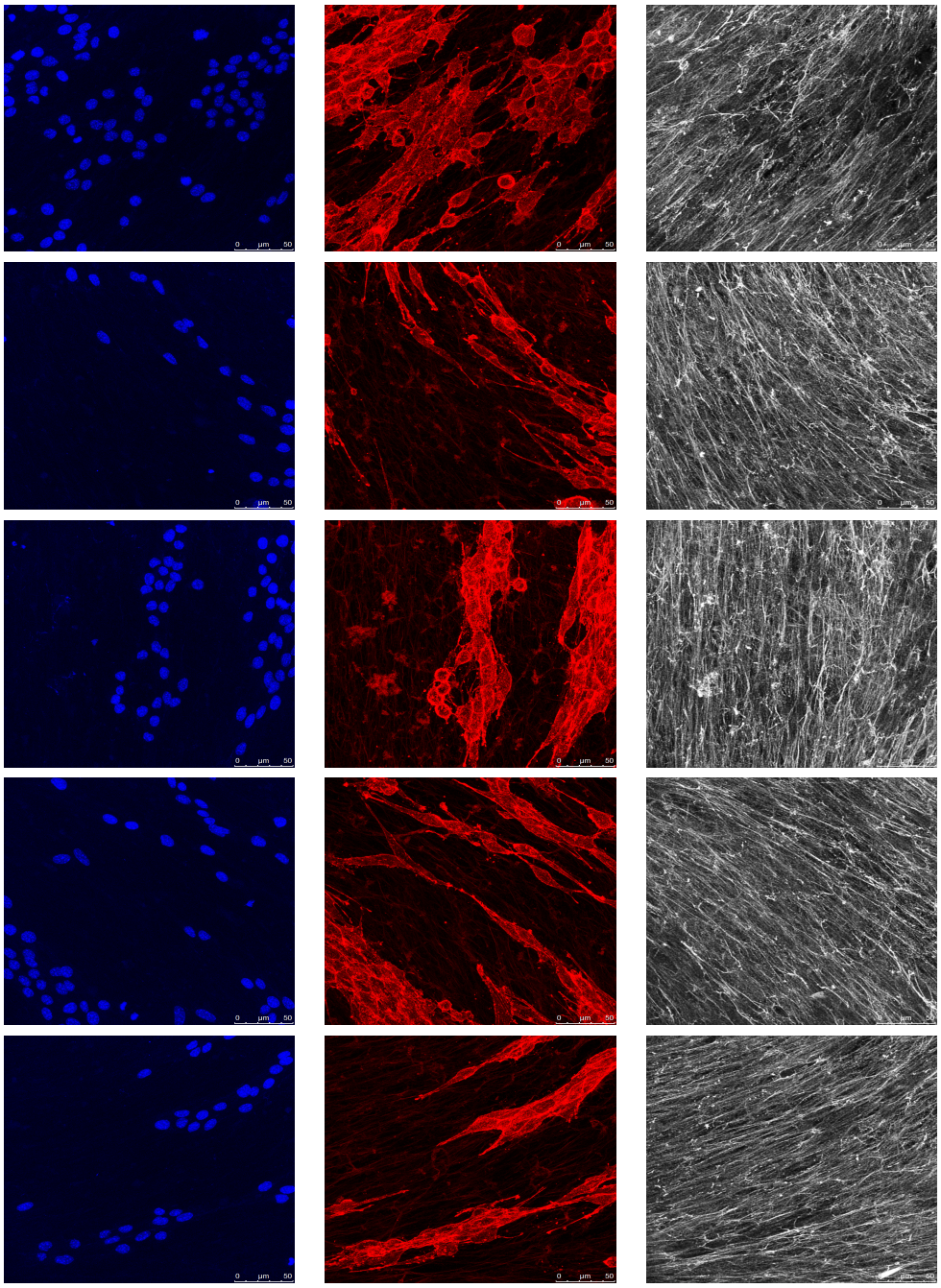
3.3.2 Dataset 2:

Four biological datasets of fibre-like structures from different origins were used to test and compare the differences in performance of the tracing algorithms (Fig. 3.14). The images have completely different visual properties with some images being noisy while others are clean, and their fibres are easier to distinguish by eye.

- Fluorescently labeled bovine collagen I (FC) (Fig. 3.14a) type with TAMRA [278, 48, 279]. The dye was left for 12-24 hours. The collagen TAMRA solution was dialysed again with a soloution of 0.2% acetic acetic acid in deionised water, with pH adjusted to 4.0. A leica SP5 resonant confocal microscope was used. The excitation wavelength was 543nm and the emitted photon were captured between 550 and 650nm. Image is available online: (https://github.com/LindaNiu/Curvelets/tree/master/FIRE/example/images).

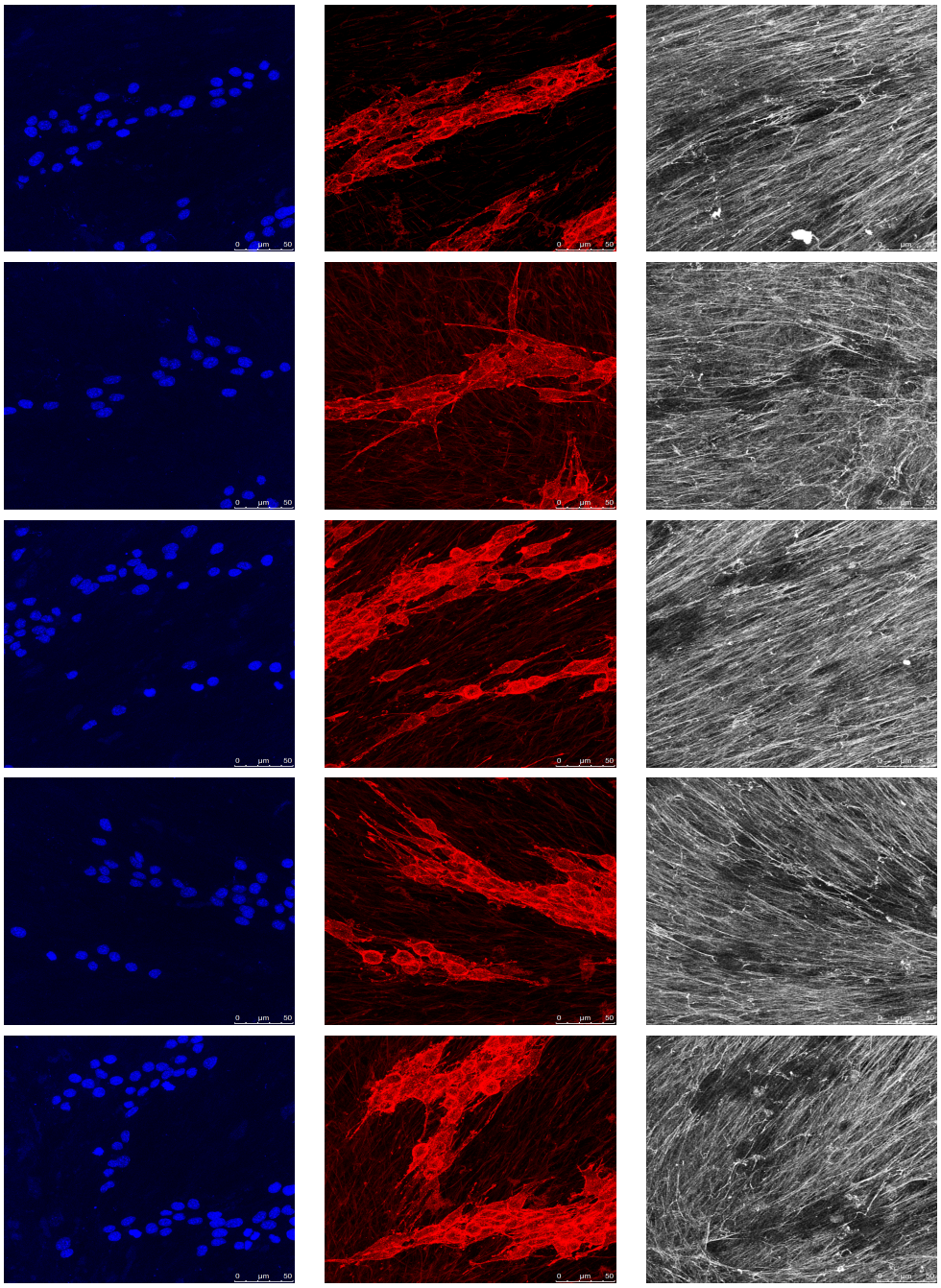
- Fluorescently labelled fibronectin images (FF) (Fig. 3.14b). Fibronectin fibroblasts were cultivated on a gelatine coated surface bonded with glutaraldehyde, used to create the extracellular matrix. Fluorescent images were captured to observe the fibronectin expression. Fibres appear bright and have high intensity, background is darker with lower intensity. Image is available at *github.com/youssefarafat/Trace_Ridges/blob/main/Fluorescent_Fibronectin*.

- Breast Cancer Biopsy (BCB) images of collagen (Fig. 3.14c) were published with the Twombli algorithm [47]. The samples were stained with Picrosirius red. The slides were deparaffinised and hydrated prior to the application of Picrosirius red mixture for 60



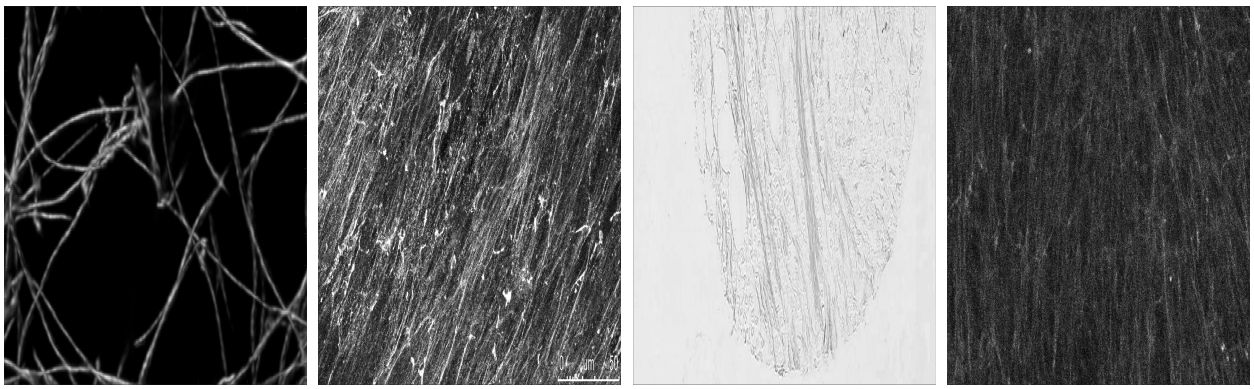
(a) (b) (c)

Figure 3.12: Dataset 1: RBD images with DAPI, Phalloidin and Fibronectin. (a) Images of the DAPI channel highlighting the cell nuclei in blue. (b) Images of the phalloidin channel, highlighting the actin filaments in red. (c) Images of the Fibronectin channel, which show fibres of the extracellular matrix fibronectin. The extracellular matrix is complex, making it difficult to analyse by eye.



(a) (b) (c)

Figure 3.13: Dataset 1: WT images with DAPI, Phalloidin and Fibronectin. (a) Images of the DAPI channel highlighting the cell nuclei in blue. (b) Images of the phalloidin channel, highlighting the actin filaments in red. (c) Images of the Fibronectin channel, which show fibres of the extracellular matrix fibronectin. The extracellular matrix is complex, making it difficult to analyse by eye.



(a) (b) (c) (d)

Figure 3.14: Dataset 2: Various images of extracellular matrix fibres from different microscopical methods. These images are used to compare the tracing algorithms (a) Fluorescently labeled Collagen (FC). (b) Fluorescent Fibronectin (FF), image courtesy of Dr Esther Castellano, Universidad de Salamanca. (c) Breast Cancer Biopsy slide (BCB). (d) Disease Mimicking Extra Cellular Matrix (DME). Images (a), (c) and (d) are open source.

minutes, after which the samples were rinsed twice in acetic acid followed by alcohol dehydration and mounting. Finally, slides are scanned using a Zeiss Axio Scan.Z1 at 10x magnification. This dataset was chosen as the images are stained with Picrosirius red, including variety to the previous monotone images. Fibres have a filamentous structure, which makes it unique. Image is available online: (https://github.com/wershofe/TWOMBLI/blob/master/TWOMBLI_v1/tutorial/ OriginalFiles/Eligible/2L_4D.png).

- Disease mimicking ECM (DME) (Fig. 3.14d) images of fibroblasts were published with the Graph based algorithm [50]. Normal resting fibroblasts were treated with TGF-B1 cytokine to activate the fibroblast in the tumour micro-environment. The Fibronectin rich matrice was created by exposing recombinant cFn to the Fibronectin null mouse embryo fibroblasts. The cultures finally are decellularised after 7 days. The resultant matrices are viewed with a confocal microscope after they are fluorescently stained, at a 10x/0.45 magnification. Images from this data set are noisy. Brightness variation is low. Image is available online: (https://github.com/aigrapa/ECM-fibergraph/blob/main/sample_img/A%2B_N.tif).

3.3.3 Dataset 3:

This dataset is comprised of three categories of images, Concentration Collagen, Marangoni Flow and Protease Inhibitors. Images courtesy of Dr Oscar Maiques, Barts QMUL.

Concentration Collagen:

Images of increased collagen concentration were used throughout this thesis. Images were captured and observed with a confocal reflectance microscope (Zeiss LSM). Single confocal z-slices were captured with the properties (1024 x 1024 pixels, 8-bit; 0.7x zoom) and taken at the mid-plane of the cells.

Standard collagen of fibrillar bovine dermal atelo Type-I Collagen solution (10mg/ml) was prepared at 1.7 and 6.6 mg/ml. Changing and preparing the concentrations of collagen was done by using Phosphate Buffered Saline (PBS), water and the pH of the mixture was adjusted using a manufacturer's neutralising solution.

There were 198 images of collagen cultures with a matrix concentration of 1.7 mg/ml and 198 images of collagen cultures with a matrix concentration of 6.6 mg/ml. Images were of size (1024×1024). Sample of those images can be seen in (Fig. 3.15).

Marangoni Flows:

Marangoni Flow can be explained as the instance where mass transfer is driven by surface tension gradient initiated by the evaporation variation across a droplet's surface. These differences can be due to temperature or concentration gradients. Cell-laden hydrogels can be created using evaporating droplets of collagen.

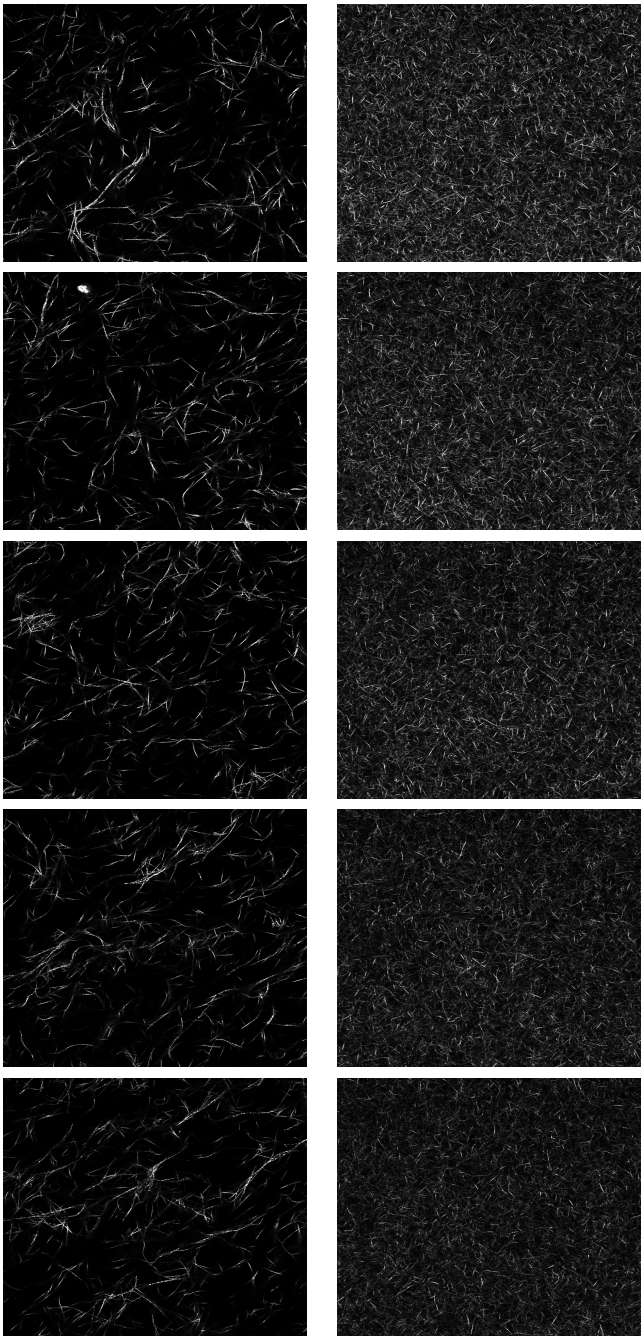
Preparing the Marangoni Flow cultures involves drop casting neutralised solutions of collagen onto ultraviolet/ozone-treated glass bottom culture dishes. The relative humidity in the culture dish is controlled using a saturated solution of $MgCl_2$, which is placed inside the perimeter of the dish before drop casting the collagen. The culture dish is then placed in an incubator to start the self-assembly of collagen. In the incubation period, collagen self assembles as water evaporates from the droplet. The Marangoni Flow procedure increases the alignment of fibres and can be used as a control when observing comparisons of increasing collagen concentrations.

Collagen observed with Marangoni Flows was observed with confocal reflectance microscopy. Protocol to prepare the Marangoni flow cultures followed the protocol described in Nerger et al [280]. T24 or T6 well plates were prepared by making a circle with hydrophobic pen in the centre of the well, to facilitate some space to add the saturated $MgCl_2$ solution around the well.

The Marangoni flows dataset had 150 images of control and 134 images of the collagen fibres after the addition of $MgCl_2$. Images were of size (1024×1024). A sample of collagen images before the addition of $MgCl_2$ (control) and after (Marangoni) can be seen in (Fig. 3.16).

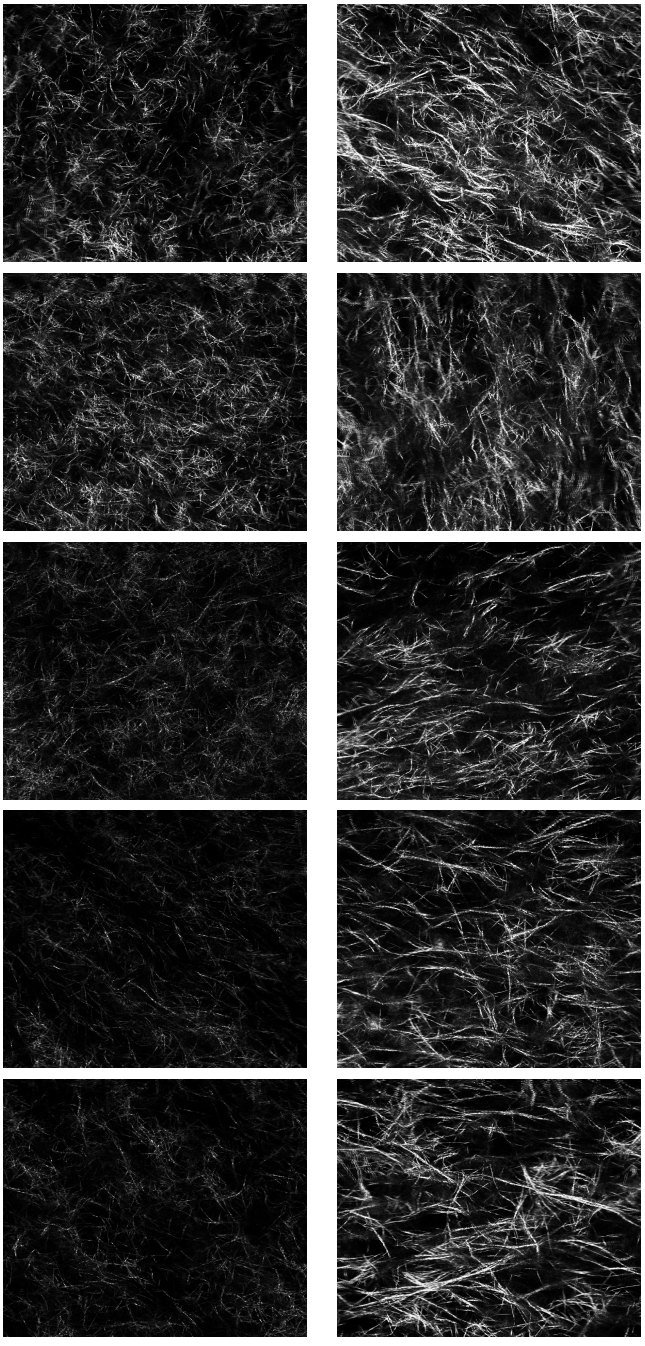
Protease Inhibitors:

Protease Inhibitors [281] are molecules which suppress the function of proteases. Proteases are enzymes that aid in the breakdown of proteins. Tumour masses of non-malignant cells exist close to neoplastic cells (abnormal cells). Non-malignant cells have an active contribution



(a) (b)

Figure 3.15: Dataset 3: Increasing concentration of collagen, various images of the extracellular matrix collagen. (a) Collagen fibres images at an ECM concentration of 1.7 mg/ml. (b) Collagen fibres images at an ECM concentration of 6.6 mg/ml.



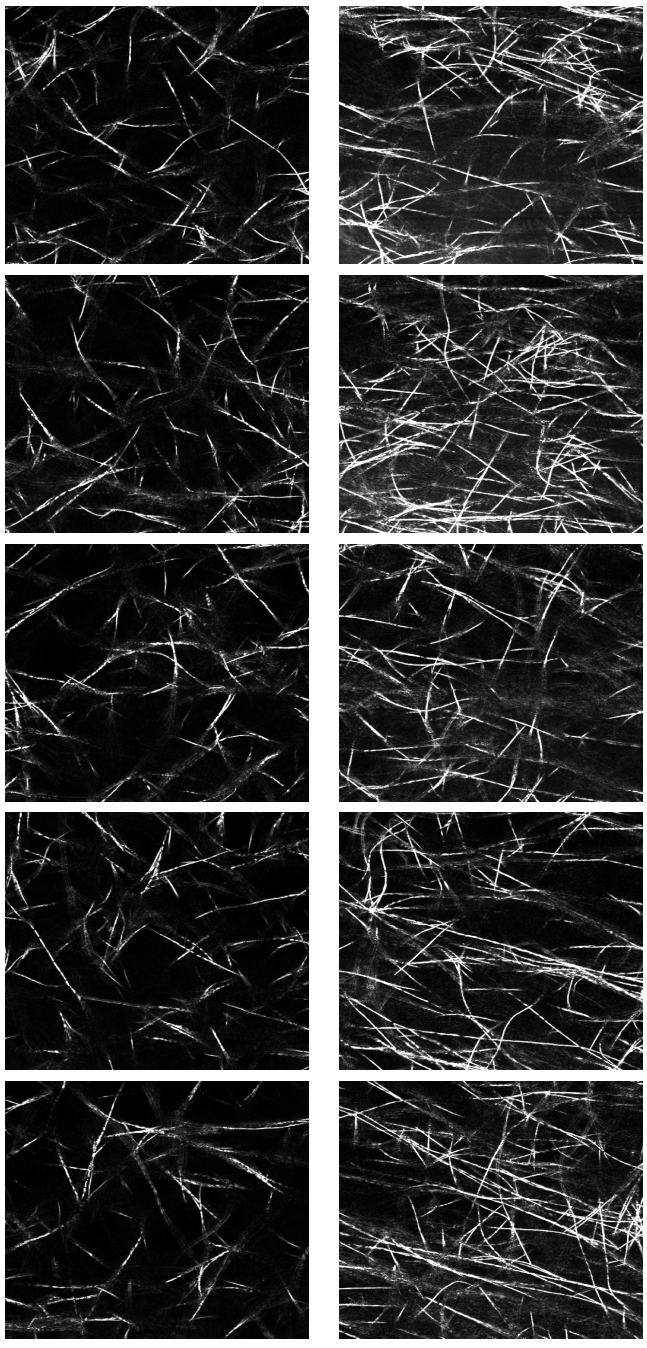
(a) (b)

Figure 3.16: Dataset 3: Marangoni Flows, various images of the extracellular matrix collagen. (a) Images belonging to control, which are the collagen fibres before the addition of MgCl2. (b) Images belonging to the Marangoni Flows, which are the same collagen fibres of the control but after the addition of MgCl2.

to tumour progression. Fibroblasts within the tumour stroma constitute a microenvironment that can alter the behaviour of malignant cells. Proteins of the extracellular matrix form a non-cellular region on the tumour micro-environment that is constantly remodeled and modified by proteases secreted by those neoplastic cells. Morphological and biochemical interactions of the extracellular matrix alter and these alterations influence important cell behaviours such as proliferation and survival. For collagen, the presence of protease inhibitors enhances the extracellular fibril deposition. This causes the fibres to experience increased alignment as it prevents the degradation of the fibres. This can be used as a control when comparing populations of increasing collagen concentrations.

Collagen observed with protease inhibitors was observed with confocal reflectance microscopy.

The dataset contained 156 control images of size (512×512), and 180 Protease inhibited images. The control samples belong to observations of collagen I before the addition of protease inhibitors. A sample of the control and the collagen after the addition of the protease inhibitors can be seen in (Fig. 3.17).



(a) (b)

Figure 3.17: Dataset 3: Protease Inhibitors, various images of the extracellular matrix collagen. (a) Images belonging to control, which are the collagen fibres before the addition of the protease inhibitors. (b) Images belonging the Protease inhibitors, which are the same collagen fibres of the control but after the addition of the protease inhibitors.

3.3.4 Dataset 4:

This dataset contained the TMA cores of Normal and Tumour Breast tissue. Images courtesy of Dr Oscar Maiques, Barts QMUL. Breast cancer (BRCA) samples of 30 patients were obtained from collaborators. A tissue arrayer device was used to build the tissue microarrays (TMA). Cylinders of 0.4 mm in size were used from distinctive regions for Tumour Body (TB) and Invasive Front (IF) which have Proximal Invasive Front (PIF) and Distal Invasive Front (DIF).

For the Masson Trichrome staining of the samples and they were rehydrated and stained with Weigert's iron haematoxylin solution for 10 minutes, then the samples were stained in Biebrich scarlet-acid fuchsin solution for 10-15 minutes. After which it was differentiated in phosphomolybdic-phosphotungstic acid for 10-15 minutes. The samples were then stained with an aniline blue solution for 5-10 minutes, after which the sample was rinsed with distilled water for 2-5 minutes.

The TMA blocks were then sectioned to the size of 3 μm thick and dried for 1 hour at 65° before they are deparaffinised and rehydrated. The TMA cores were viewed and captured with a Brightfield Microscope. This resulted in 120 Normal cores and 120 Tumour Cancer cores. The normal cores can be seen in (Fig. 3.2) and tumour cores can be seen in (Fig. 3.18). Samples of the normal cores can be viewed in (Fig. 3.20a) and tumour in (Fig. 3.20b).

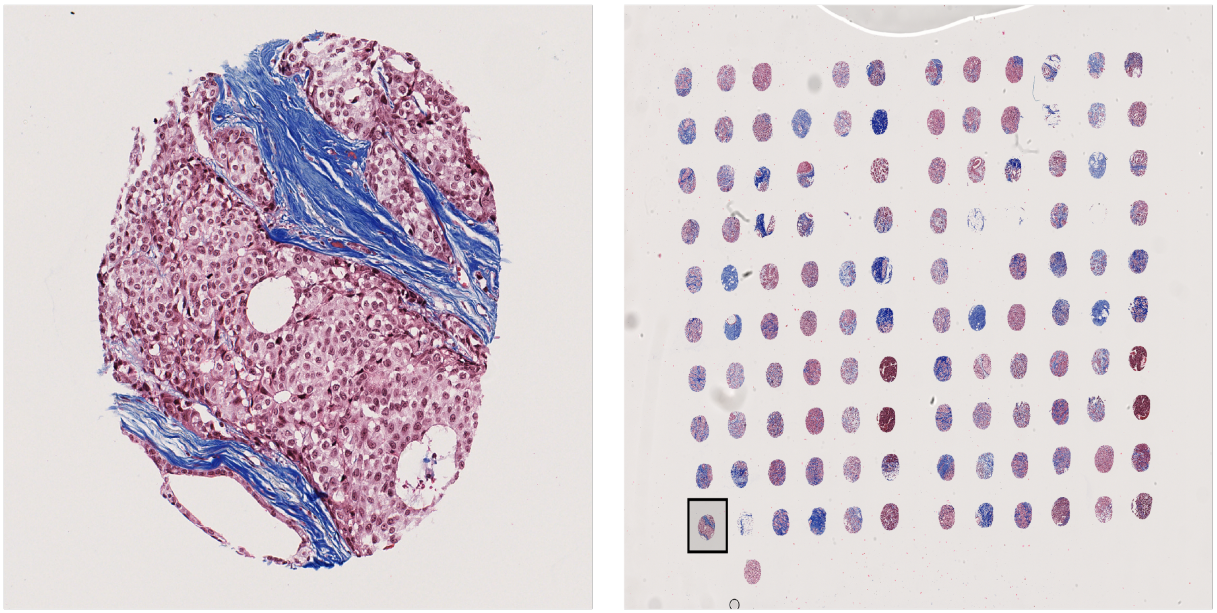
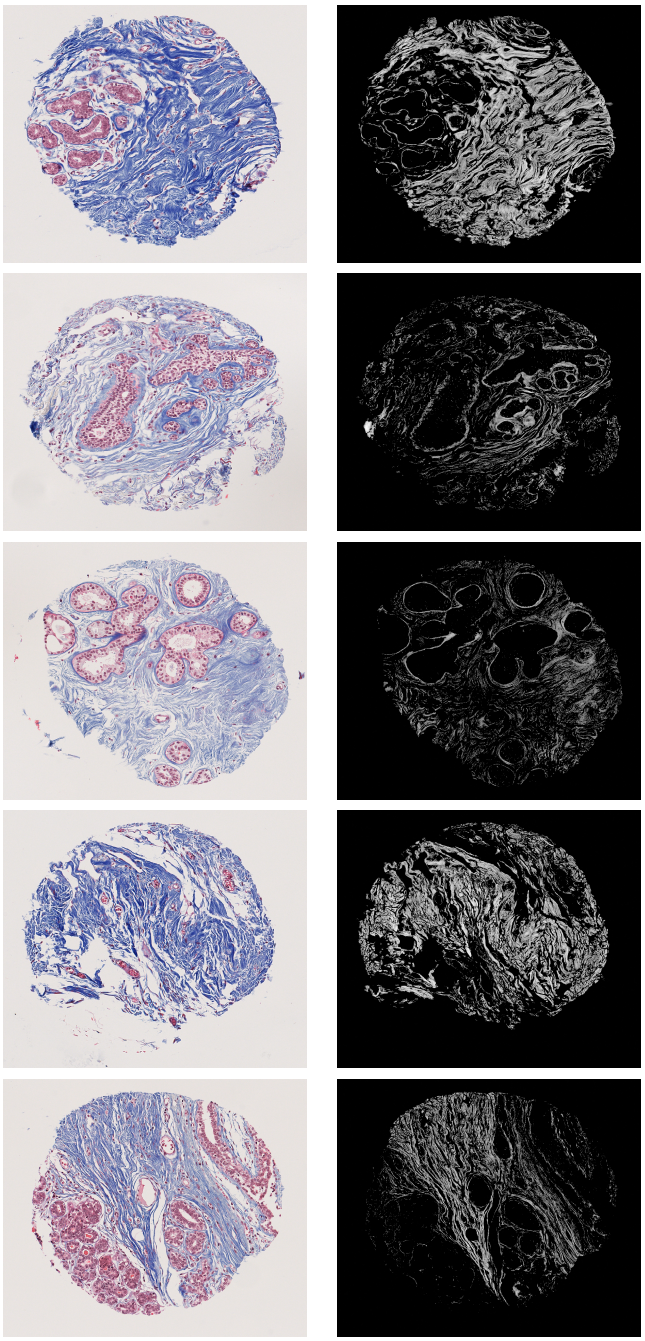


Figure 3.18: Tissue Micro Array of tumour breast cancer tissue comprising of 120 cores. When zoomed in the cores reveal details of the extracellular matrix structure. Collagen fibres are dyed in blue while cell nuclei are dyed in red.

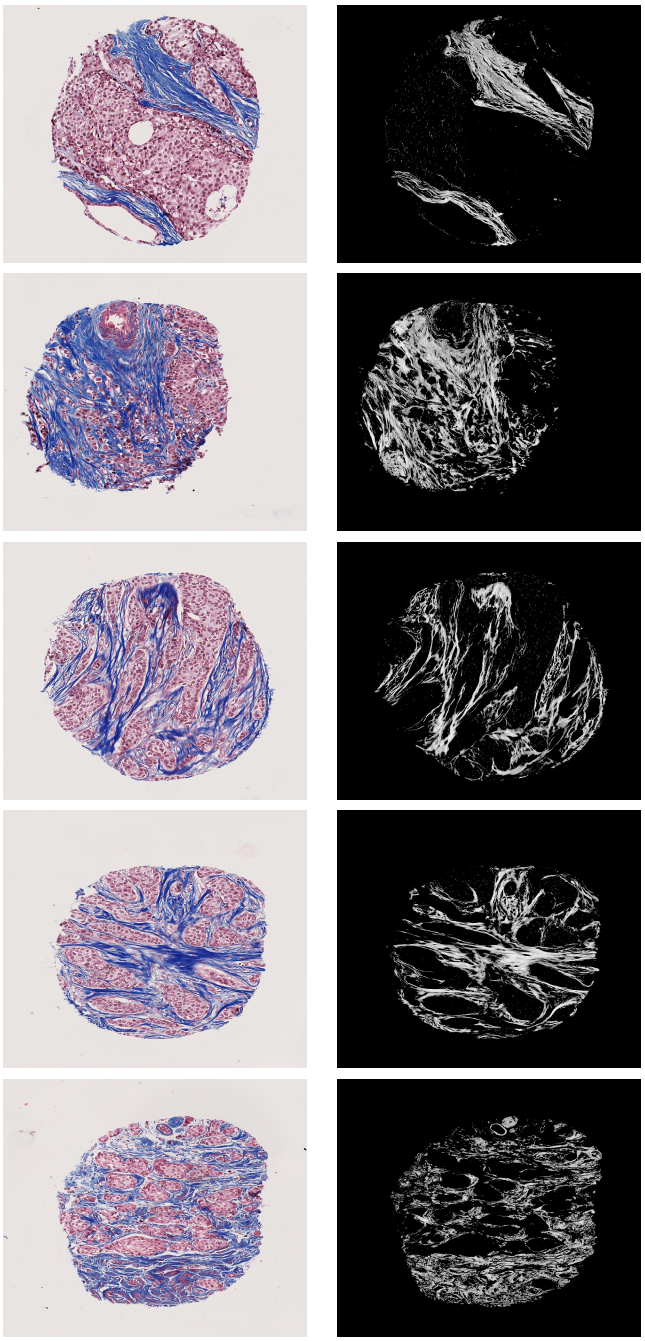
Throughout this thesis, the TMA cores will be used for two investigations. One investigation will focus on the whole cores and will include both fibres and cell nuclei. The other investigation will focus only on the collagen fibres. The normal tissue micro array cores can

be seen in (Fig. 3.19a), the processed fibres of those cores can be seen in (Fig. 3.19b). The tumour tissue micro array cores can be seen in (Fig. 3.20a), the processed fibres of those cores can be seen in (Fig. 3.20b).



(a) (b)

Figure 3.19: Normal Cores: Fibres and cells from cores of the original tissue micro arrays and the cores with only fibres. (a) Cores representing the original samples. (b) Cores processed by keeping only the fibres.



(a) (b)

Figure 3.20: Tumour Cores: Fibres and cells from cores of the original tissue micro arrays and the cores with only fibres. (a) Cores representing the original samples. (b) Cores processed by keeping only the fibres.

3.4 What is an Image:

A digital image $\mathbb{I}$ is a digital representation of visual information. They are typically captured by a camera; images used throughout this thesis have been captured with a microscope's camera. Digital images consist of a grid (matrix) of pixels p, which have coordinates of their positions in the image based on the spatial arrangement of rows, columns, slices and timepoints. Each pixel contains information about its light intensity I. Images are often represented as a two-dimensional array for grayscale images and three-dimensional array for colour images commonly red, green, blue (RGB) and hue, saturation, value (HSV) images. Digital images are created when an array of light sensitive arrays captures the light entering them and the number of pixels in an image is dependent upon the resolution of the sensor array. Digital images representation allows a structured way to process the images mathematically in computer vision and image processing. Images can be represented as a function of intensities (Eq. 3.1).

$$\mathbb{I} = I(x, y) \tag{3.1}$$

$x = 1, 2, ...N_r$ where N_r is the number of rows in an image. $y = 1, 2, ..., N_c$ where N_c is the number of columns. The intensities of an image can be quantised into a finite set $I = 0, 1, 2, \ldots, I_{\max}$ when 8 bits are used to encode the intensities, these are within the range of $I = 0, 1, 2, ..., 255$. The digital co-ordinates of an image (x, y) can represent the image as a matrix $\mathbb{I}$ of size $N_r \times N_c$. Pixels can be represented as (Eq. 3.2).

$$p = \mathbb{I}(x, y) \tag{3.2}$$

3.4.1 Grayscale Images:

Grayscale images are a type of digital images that represent visual information in various shades of gray and they range between black and white. Each pixel p in the grayscale image $\mathbb{I}$ has information about its intensity value. Grayscale images have one channel which means that each pixel has one intensity value. Thus, grayscale images can be described as a matrix $\mathbb{I}(x, y)$ of size $N_r \times N_c$. The pixel intensity can be represented as $I(x, y) \in \{0, 1, 2, \ldots, 255\}$ for an 8-bit image but greater than 8-bit is also common. Images used in this thesis were mainly 8-bit images. Pixels with intensity value 0 represent black while 255 represent white, values in between are shades of gray. An example of a grayscale image and its intensities is illustrated in (Fig. 3.21). For most acquisition systems and processing pipelines grayscale is the default.

3.4.2 Colour Images:

Colour is coded in digital images as a combination of three channels, most commonly red, green, and blue (RGB) [282]. Each pixel in an image will contain intensity I values in each channel. Specific levels of intensities of red, green, and blue in a pixel determine its resulting colour. An RGB image is three-dimensional and can be represented as (Eq. 3.3).

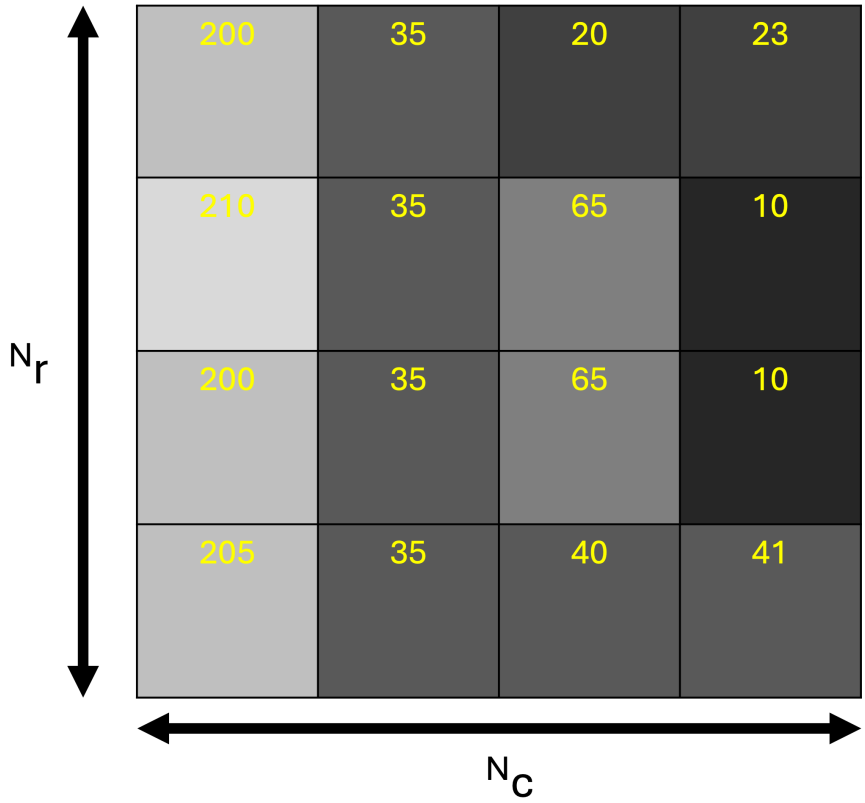


Figure 3.21: Grayscale Image with size $N_r \times N_c$. Intensities range between 0 and 255 where 0 is black and 255 is white, values in between are shades of grey.

$$\mathbb{I}(x, y, c) \tag{3.3}$$

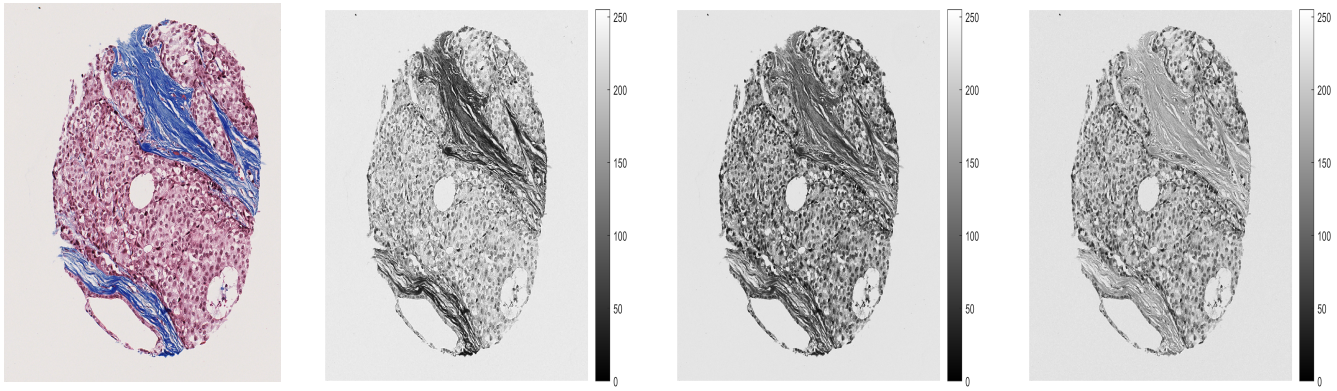
of size $N_r \times N_c \times 3$. The colour of a pixel can be represented as outlined in (Eq. 3.4).

$$p(x, y) = (I_r(x, y), I_g(x, y), I_b(x, y) \tag{3.4}$$

Where I_r, I_g, I_b are intensity values of their respective channels. When using 8 bits to code the intensities, it is between the ranges of $I(x, y) \in \{0, 1, 2, \ldots, 255\}$. An example of an RGB image is illustrated in (Fig. 3.22).

3.4.3 HSV Images:

HSV Images [283] are a type of digital images that represent the visual information in three channels (hue, saturation, value). This is an intuitive model that is more aligned with how humans perceive and interpret colours. It can separate the colour information such as hue and saturation from the brightness (value), which allows easy manipulation in image



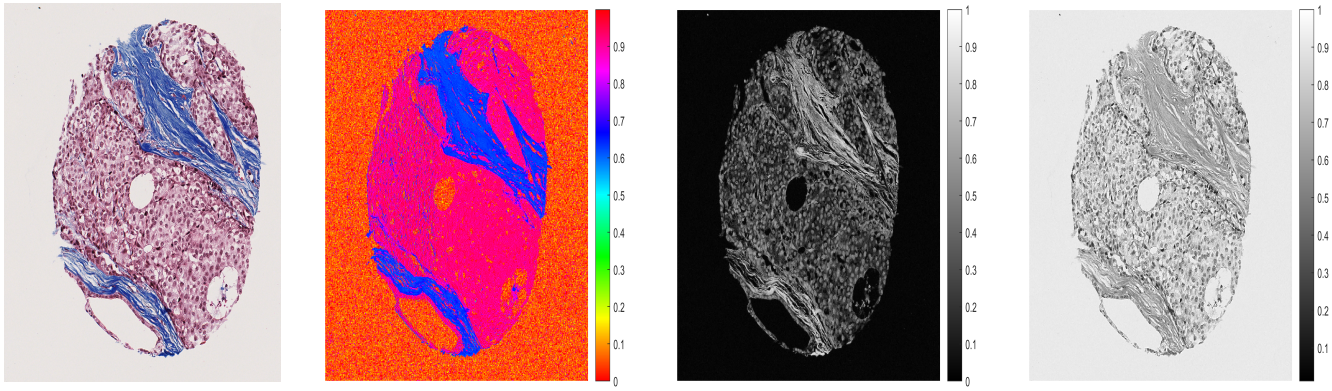
(a) (b) (c) (d)

Figure 3.22: RGB Image with size $N_r \times N_c \times 3$ illustrated in (a) with three colour channels of red (b), green (c) and, blue (d). Values of pixels reflect the intensity of the respective colour at that pixel location.

processing tasks. Hue is the colour type, and it is expressed as an angle in degrees on the colour wheel from 0 to 360. Saturation is the intensity of the colour it ranges from 0 to 1 or 0 to 100 depending on the representation. Value is the brightness of the colour ranging from 0 to 100. An HSV image can be represented as a matrix $\mathbb{I}(x, y, c)$ just as (Eq. 3.4), of size $N_r \times N_c \times 3$. The colour of a pixel can be represented as (Eq. 3.5).

$$p(x, y) = (I_H(x, y), I_S(x, y), I_V(x, y)) \tag{3.5}$$

Where I_H, I_S, I_V are intensity values of their respective channels. Intensity values can be normalised as such that intensity is between the ranges of $I_H(x, y) \in \{0, 1, 2, \ldots, 360\}$, $I_S(x, y) \in \{0, 1, 2, \ldots, 255\}$ and, $I_V(x, y) \in \{0, 1, 2, \ldots, 255\}$. An example of a HSV image is illustrated in (Fig. 3.23).



(a) (b) (c) (d)

Figure 3.23: HSV Image with size $N_r \times N_c \times 3$ illustrated in (a) with three channels which are hue (b), saturation (c), value (d). Hue is the colour type; saturation is the intensity of the colour and value is the brightness of the colour.

3.4.4 Binary Images:

Images often have objects and surrounding empty areas. The pixels of importance are those of the objects and can be thought of as the foreground while the empty areas contain pixels that are of less importance and are considered background. Pixels in the image are classified as either positive or foreground "1" or negative or background "0". The pixels of interest can be chosen manually by assigning a pixel to one of the classes or it can be done by setting a threshold T which sets an intensity value I below which is the background, and above which is the foreground. Binary Images can be represented as a two-dimensional matrix as in (Eq. 3.6), the intensity becomes either 0 or 1.

$$BW(x, y) \in \{0, 1\} \tag{3.6}$$

3.5 Image Processing Techniques:

Image processing in the simplest terms involves applying a set of techniques and manipulating an image to highlight and extract key features [284]. Image processing can be used to detect and distinguish objects in an image, create an enhanced cleaner image, and recognise patterns in an image. Many techniques have been introduced over the years that help with identifying, segmenting, and detecting objects in an image. A typical image processing pipeline consists of multiple steps including filtering the image to enhance features or eliminate noise, followed by segmentation of the objects in an image and finally, extracting morphological features of extracted properties (Fig. 3.24).

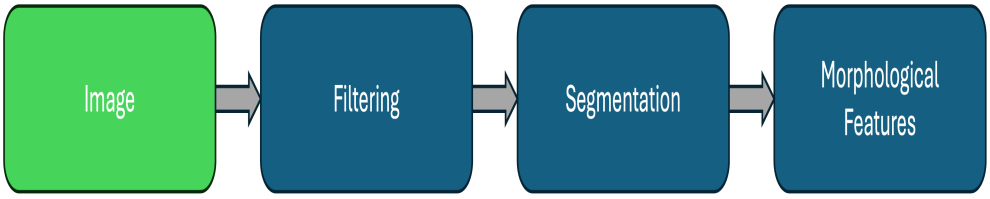


Figure 3.24: Image processing pipeline. The workflow consists of filtering an image to enhance its features or remove noise, followed by segmentation of the objects in an image and finally, extracting morphological features of segmented objects.

3.5.1 Filtering:

This section provides an outline on filtering; an image processing technique used throughout this thesis. Filtering and the use of filters have been commonly used to modify an image into a new image where the filter has modified the data. Filters can be represented in the form of kernels which are matrices with size N. Kernels can be in different shapes dependent on their use. Filters can enhance relevant features of an image such as edges of objects or suppress unwanted characteristics, for example noise. Commonly filters [285] have been either spatial or frequency [286] domain but with the advancement of artificial intelligence convolutional neural networks can be trained to remove noise from images and highlight target features. Filters have been used in many imaging domains such as facial images [287] or MRI denoising [288].

Spatial Filters:

There are many types of filters. The most common type is spatial filters, which multiply the value of pixels of a region with a kernel $\mathbb{K}$ of size N, then the kernel moves across the image where the convolution process is repeated until all the pixels have been multiplied, and a new filtered image is produced. A kernel representing the filter is used to convolve the image. A kernel is a matrix that applies the transformations to the images. Kernels often are squares $N \times N$ and odd-sized to ensure a centre pixel in the kernel exists for convolutional

operations, as this allows the kernel to be centred over a pixel in an image when applying the filter. $N = 2a + 1$ where $a \in \mathbb{N}$. Kernels are often centred around (0,0).

An example of this is the Gaussian filter [289] which is used to smooth an image and reduce or eliminate noise. A Gaussian filter has a bell shape. The values follow a discrete 2D Gaussian of the size of the kernel. The filter exhibits zero mean (0) and variance (σ), as well as the size of the filter (N). The width of the kernel is determined by the variance width. The Gaussian filter assigns weight of the kernels based on a Gaussian distribution, giving more importance to the pixels closer to the centre.

A Gaussian filter function can be represented as (Eq. 3.7).

$$G(i, j) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{i^2 + j^2}{2\sigma^2}\right) \tag{3.7}$$

where (i, j) are the pixel co-ordinates relative to the centre of the kernel. The convolution of the Gaussian filter with an image can be represented as such (Eq. 3.8).

$$\mathbb{I}'(x, y) = \sum_{i=-a}^{a} \sum_{j=-a}^{a} \mathbb{I}(x + i, y + j) \cdot G(i, j) \tag{3.8}$$

Here $\mathbb{I}'(x, y)$ is the output pixel value after applying the Gaussian filter. $\mathbb{I}(x + i, y + j)$ is the pixel value from the input image at the shifted position. $G(i, j)$ is the value of the Gaussian kernel at the position (i, j).

For a 3x3 Gaussian Filter, the co-ordinates range from -1 to 1 in both the x and y directions, with the centre of the kernel being at (0,0) as such

$$\begin{bmatrix} (-1,-1) & (0,-1) & (1,-1) \\ (-1,0) & (0,0) & (1,0) \\ (-1,1) & (0,1) & (1,1) \end{bmatrix}$$

With a σ value of 1, the Gaussian filter function can be simplified to (Eq. 3.9).

$$G(i, j) = \frac{1}{2\pi} \exp\left(-\frac{i^2 + j^2}{2\sigma^2}\right) \tag{3.9}$$

By substituting the co-ordinate values of the Gaussian kernel into the Gaussian function. The non-normalised Gaussian kernel is the following

$$\mathbb{K}_G = \begin{bmatrix} 0.0585 & 0.0965 & 0.0585 \\ 0.0965 & 0.1592 & 0.0965 \\ 0.0585 & 0.0965 & 0.0585 \end{bmatrix}$$

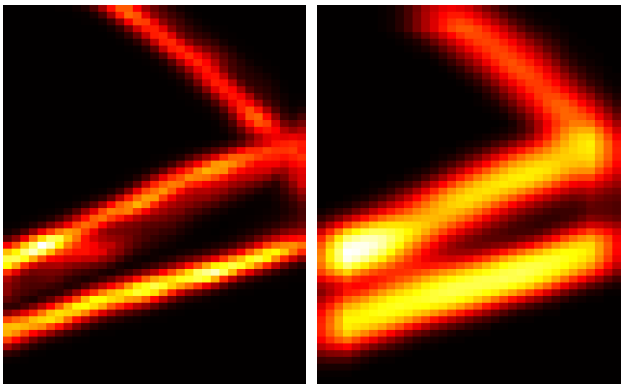
In many implementations for simplicity, the matrix can be approximated as such

$$\mathbb{K}_G = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

To ensure that the Gaussian kernel does not change the overall brightness of the image. The kernel is normalised by dividing each value by the sum of all elements in the matrix. In this case the sum is 16. The normalised kernel takes shape as such

$$\mathbb{K}_G \begin{bmatrix} 0.0625 & 0.125 & 0.0625 \\ 0.125 & 0.25 & 0.125 \\ 0.0625 & 0.125 & 0.0625 \end{bmatrix}$$

Gaussian filtering can be seen applied to the fibre image in (Fig. 3.25). For illustration purposes a Gaussian filter of $\mathbb{K} = 7$ and $\sigma = 20$ was used.



(a) (b)

Figure 3.25: Illustration of Gaussian filtering on the basic fibre image. (a) Original fibres image. (b) Gaussian filtering application. It can be seen that the structures become more smooth.

Another type of smoothing filters is the averaging filter seen in (Fig. 3.26) which is a simple filter. Depending on the filter size N. The kernel moves across the image averaging N pixels where the pixels are multiplied by the values of the kernel. Those values are summed and averaged. A new pixel output is created where each output value is the average of its neighbouring pixels. This helps in reducing noise and fluctuations in data. An example of an averaging filter is the disk filter. The disk filter has a circular shape, where it only covers a set of pixels within a specific radius around the target pixel. The size of the kernel can be represented as $2r + 1$ where r is the radius. As explained earlier the filter convolves around the image. For each convolutional operation it adds the intensities of these pixels inside the disk and divides by the number of pixels in that area.

Gaussian filtering is what is known as a low pass filter where it removes high frequency features and retains only low frequency features. High pass filtering on the other hand removed low frequency features which are the smooth variations and retains the high frequency features where there are sharp transitions. A high pass filter aims to sharpen an image and its features. High pass filters sharpen details and textures by disregarding smooth gradients in return high pass filters may at times increase noise visible in the image. Typically, a high pass filter can be applied in the frequency domain using fast Fourier transform or it can be

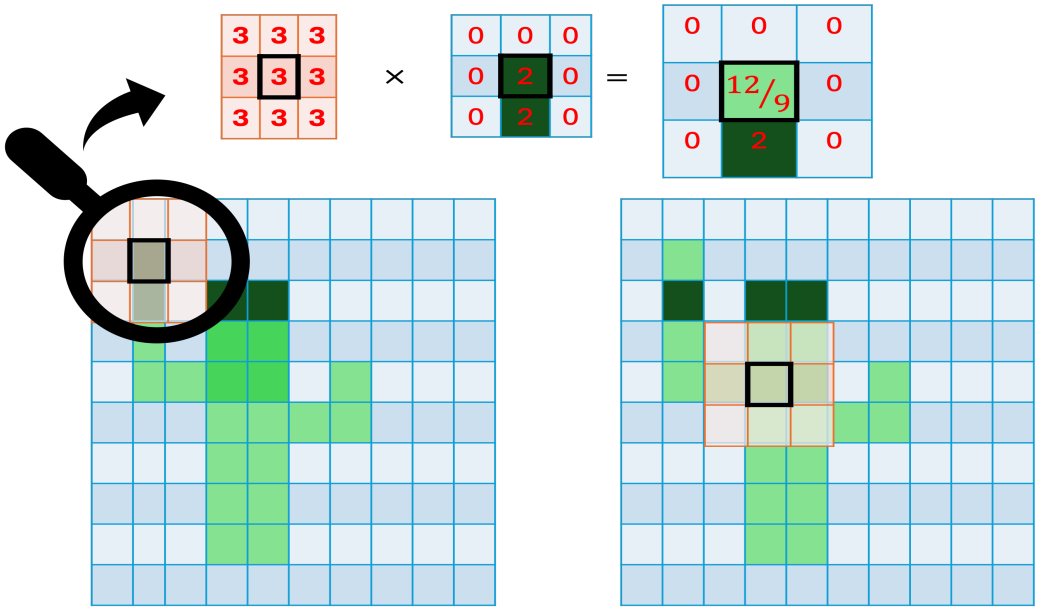


Figure 3.26: Illustration of an averaging 3×3 kernel on a representative image. Value of the original middle pixel of the image is averaged by the values of the operation of the kernel on the neighbouring pixels.

applied using convolutions with a kernel in the spatial domain. For this thesis, the high pass filter applied uses kernel convolutions in the spatial domain.

The high pass filter used in this thesis is of size 3×3. The shape of the kernel is

$$\mathbb{K}_{HP} \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

the central pixel has a weight of 4 therefore it amplifies the current pixel value, while the four neighbouring pixels exhibit a weight value of -1, therefore it darkens the regions around the central pixel. If the neighbouring pixels have similar intensity values to the current central pixel their subtraction would reduce their current intensities which would result in lower values and therefore dark areas. If the neighbouring pixels have a vastly different intensity than the current central pixel for example if it is at an edge, the difference would remain strong which would result in high values meaning the presence of bright edges. The sum of the kernel values is 0, which ensures that low frequencies are removed while edges which are regions of high frequencies are retained.

Fourier Filters:

Frequency or Fourier filters convert an image into the frequency domain using the Fourier Transform or other methods like the Wavelet or Cosine transform [290]. After converting the image into the frequency domain, frequency filters are applied, followed by an inverse

transform to convert the processed image back to the spatial domain. The shape of the filter function depends on its intended purpose. For instance, filters can attenuate high frequencies to smooth the image and reduce noise, or they can suppress low frequencies to emphasise edges.

DnCnn:

DnCnn [291] is a deep convolutional neural network designed to denoise images. It uses residual learning and batch normalisation to efficiently remove noise from input images. Residual learning allows the model to learn the difference between the input and the output by understanding the small changes required to convert the input to the desired output. This allows the network to understand what needs to be changed rather than starting from scratch. DnCnn is particularly effective in managing blind Gaussian denoising, even when the noise level is unknown. The DnCnn model used throughout the work of this thesis was trained using a publicly available dataset containing images of fibre-like structures. The dataset images were divided into two parts. One for training and the other for testing. Images were converted to 2-D grayscale when necessary to match the model's input format. Key training parameters included a learning rate of 0.001 and a batch size of 32.

3.5.2 Segmentation:

Segmentation can be placed into two categories, and these are semantic segmentation [292] and instance segmentation [293]. Semantic segmentation refers to assigning each pixel to a particular class, either a foreground pixel or a background pixel [294]. Instance segmentation aims to identify different instances of the same class by classifying every single pixel separately [295].

Traditional Segmentation:

Traditional segmentation techniques can be edge based which include canny, gradient, and Laplacian. Edge-based segmentation extracts edges of the objects in the image and reduces their size. Edges are identified based on contrast, colour, saturation, and texture. Therefore, borders of objects can be easily extracted. Thresholding based segmentation includes mean shift and Otsu's. Thresholding is the most basic form of segmentation; it is based on pixel intensity; pixels will fall into one of two categories based on the selected thresholding intensity. The pixels will either fall into the foreground or background. Region based segmentation includes region growing and region split and merge methods, this technique works by grouping regions of an image together based on similarities of characteristics. The algorithm can find those regions based on seed points, and it can shrink or grow the regions. Cluster based segmentation methods include K-means and fuzzy K-means, this is a form of unsupervised segmentation. It works by isolating structures and clustering them based on similar characteristics of the pixels. Watershed based segmentation includes watershed and marker-controlled watershed methods, this technique uses pixel intensity to determine whether a pixel is elevated or not. Then ridges and lines form between the regions to split them based on whether they are elevated or not [296].

Non-traditional Segmentation:

Non-traditional segmentation techniques include deep learning, such as U-Net. For instance segmentation in deep learning, Deep watershed is used for contour-aware approaches and R-CNN, centreMask are used for object-detection-based approach. Other techniques include the use of Generative Adversarial Networks also known as GAN. They all have their advantages and disadvantages. Deep learning techniques such as CNN, use multiple layers in its architectures such as convolutional, pooling, fully connected layers. The convolutional layer is used for feature detection. The max pooling layer is used for feature selection. The outputs of the convolution and max pooling layers is fed into the fully connected layers which is used for the classification [297, 295].

Otsu Thresholding:

In Otsu threshold there is a single intensity threshold where each pixel in the image is either in the foreground class or the background class. Otsu threshold aims to minimise the spread of summing of the foreground and background pixels [298]. The Threshold (T) is the value

that separates the grayscale image into two classes. Where Class 1: Pixels (p) with intensity values from 0 to T, whereas Class 2 is: Pixels (p) with intensity values from T + 1 to 255 in an 8-bit image. There are two ways to find the threshold, maximising between class variance $\sigma_B^2(T)$ and minimising within class variance. Between class variance aims to maximise the variance between the two classes that the threshold creates. This can be expressed as in (Eq. 3.10).

$$\sigma_B^2(T) = w_1(T) \cdot w_2(T) \cdot [\mu_1(T) - \mu_2(T)]^2 \tag{3.10}$$

Here $w_1(T)$ is the weight or probability of class 1 where pixels value range between 0 and T, similarly for $w_2(T)$, where pixel values range between T+1 and 255. $\mu_1(T)$ is the mean intensity of class 1 and similarly $\mu_2(T)$ is the mean intensity of class 2. The weights $w_1(T)$ and $w_2(T)$ can be represented as in (Eq. 3.11).

$$w_1(T) = \sum_{I=0}^{T} P(I) \quad \text{and} \quad w_2(T) = \sum_{I=T+1}^{255} P(I) \tag{3.11}$$

Where $P(I)$ is the probability of intensity I. This is represented in (Eq. 3.12).

$$P(I) = \frac{N(I)}{N} \tag{3.12}$$

Here $N(I)$ is the number of pixels in the image with the intensity I and N is the total number of pixels in an image across all intensities (Eq. 3.13).

$$N = \sum_{I=0}^{255} N(I) \tag{3.13}$$

The mean intensities of the two classes can be thus represented as (Eq. 3.14).

$$\mu_1(T) = \frac{\sum_{I=0}^{T} I \cdot P(I)}{w_1(T)} \quad \text{and} \quad \mu_2(T) = \frac{\sum_{I=T+1}^{255} I \cdot P(I)}{w_2(T)} \tag{3.14}$$

Maximising the between class variance is equivalent to minimising the within class variance where the total variance is the between class variance and the within class variance without considering a threshold (Eq. 3.15).

$$\sigma_T^2 = \sum_{I=0}^{255} (I - \mu_T)^2 \cdot P(I) \tag{3.15}$$

The application of Otsu threshold can be seen on a simple fibre image in (Fig. 3.27b), the threshold resulted in the pixels of the image to belong to either the foreground or the background.

Mathematical Morphology:

Mathematical Morphology is used to manipulate binary images. It is the theory for the analysis of spatial structures that are mainly in digital images [299] based on set theory, integral geometry, and lattice algebra. Mathematical morphology is commonly used in image processing and computer vision tasks [300]. It works on binary images, so pixel values must be converted to 0's and 1's based on whether they belong to the foreground or the background. Commonly a structuring element (SE) is used which is a kernel that moves over the image and interacts with the pixels of an image. Structural elements may have different shapes such as square, disk, rectangle, diamond and cross. SE's may also have different sizes that determine the effect on the image. The basic mathematical morphology operations are erosion, dilation, closing, opening.

Erosion:

In Erosion, objects are reduced in an image by excluding the boundary pixels. An image can be represented as a function of intensities $I(x, y)$ where (x, y) are coordinates. SE is the structural element which can be thought of as a kernel K that is utilised for morphological operations. This kernel defines the neighbourhood shape that is used during the dilation and erosion processes. $\ominus$ is the erosion operation and erosion could be represented simply as (Eq. 3.16) and pixel wise as (Eq. 3.17).

$$\text{Erosion}(I, SE) = (I \ominus SE) \tag{3.16}$$

$$E(x, y) = \min\{I(x + x_i, y + y_i) \mid (x_i, y_i) \in SE\} \tag{3.17}$$

In (Eq. 3.17), $E(x, y)$ is the eroded image value at the coordinates(x,y). When erosion is applied the minimum value of the kernel operation at the coordinates of the image (x, y) and its neighbours based on the kernel size is chosen and becomes the new value of the central pixel of the kernel. Erosion can be seen applied on a sample fibre image in (Fig. 3.27c), where the fibres lost pixels at their boundaries.

Dilation:

In Dilation, objects are expanded in an image by adding pixels to the boundary. $\oplus$ is the dilation operation and dilation could be represented simply as (Eq. 3.18) and pixel wise as (Eq. 3.19).

$$\text{Dilation}(I, SE) = (I \oplus SE) \tag{3.18}$$

$$D(x, y) = \max\{I(x + x_i, y + y_i) \mid (x_i, y_i) \in SE\} \tag{3.19}$$

In (Eq. 3.19), $D(x, y)$ is the dilated image at coordinates (x, y). $I(x + x_i, y + y_i)$ refers to the pixel values in the image $\mathbb{I}$ at the positions offset by the coordinates (x_i, y_i) of the

structuring element SE. Dilation can be observed on a sample fibre image in (Fig. 3.27d), where they have gained pixels on their boundaries.

Closing:

Closing is when the objects go through dilation than erosion. The dilation is used so that the image can be filled within and then the erosion makes sure that only the edges become smaller in size ensuring that no objects merge with each other and become one object, erosion also ensures that there are no gaps that could have risen from the dilation [301]. Closing can be represented as seen in (Eq. 3.20).

$$\text{closing}(I, SE) = (I \oplus SE) \ominus SE \tag{3.20}$$

The morphological operation "closing" can thus be represented as (Eq. 3.21) which expands the foreground pixels followed by (Eq. 3.22) which removes some pixels of the foreground.

$$D(x, y) = (I \oplus SE) \tag{3.21}$$

$$E(x, y) = (D \ominus SE) \tag{3.22}$$

A structuring element with the shape of a disk and a value of three was used to show the effect of closing on a simple fibre image (Fig. 3.27e), where the fibres joined from the dilation then thinned at the boundaries from erosion.

Opening:

Opening is erosion followed by dilation. The purpose of opening is to clear away any impurities or small objects and the dilation allows the smoothing of edges. Opening can be represented as (Eq. 3.23)

$$\text{opening}(I, SE) = (I \ominus SE) \oplus SE \tag{3.23}$$

The morphological operation "opening" can thus be represented as (Eq. 3.24) which erodes the foreground pixels followed by (Eq. 3.25) which expands some pixels of the foreground. .

$$E(x, y) = (I \ominus SE) \tag{3.24}$$

$$D(x, y) = (E \oplus SE) \tag{3.25}$$

Opening applied on a sample fibre image can be seen in (Fig. 3.27f), where the fibres first underwent erosion hence the breaking of the larger fibre after which they were dilated at the boundaries.

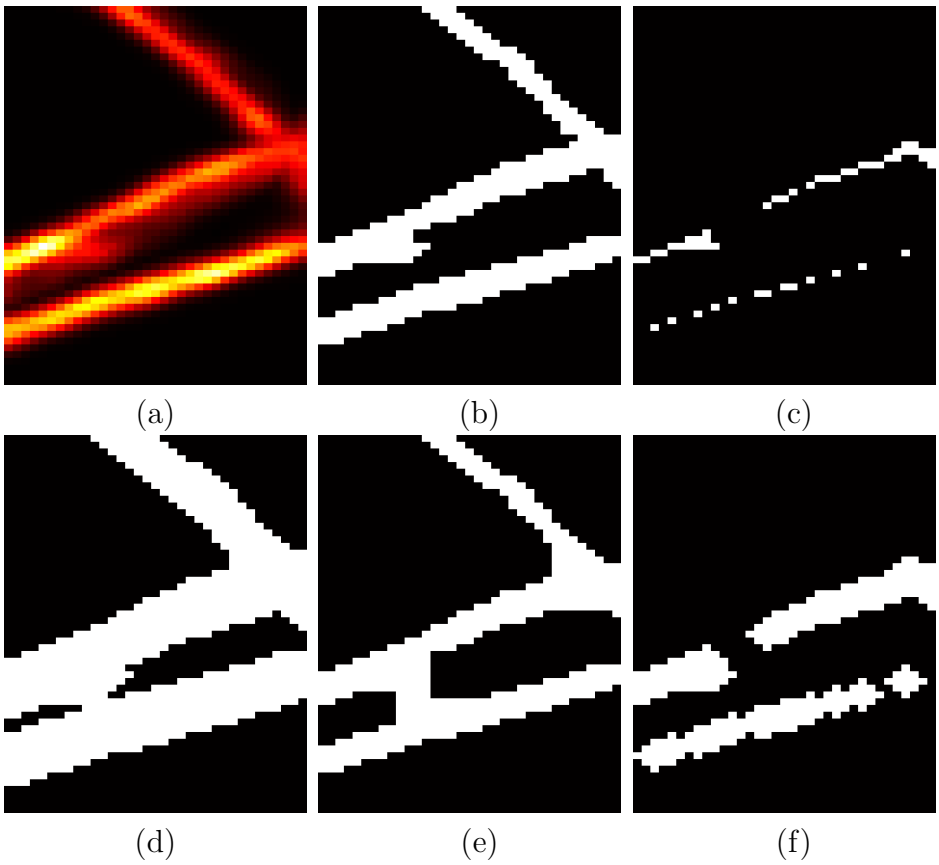


Figure 3.27: Illustration of Mathematical Morphological operations and Otsu thresholding on a sample fibre image. (a) Sample fibre image. (b) Otsu threshold applied on sample fibre image. Fibre structure appears in white and belongs to the foreground, the rest is in the background. (c) Application of erosion on thresholded fibre image. Fibres appear to have lost pixels on their boundaries. (d) Application of dilation on the thresholded fibre image. Fibres appear to have gained pixels on their boundaries. (e) Application of closing on thresholded fibre image. Fibres have dilated hence the two fibres joining and then they were thinned after erosion. (f) Application of opening on thresholded fibre image. The fibres appear to have been thinned by erosion which is observed by the larger fibre breaking into two after which the fibres were smoothed with the addition of pixels on their boundaries from dilation.

Labelling:

The way in which objects in images are labelled is when there a minimum of 8 pixels connected to each other in the foreground, the algorithm considers it as one object therefore it gets labelled. Labels are connected when their edges or corners touch. Two different pixels are the same object if they are both connected on the horizontal, vertical, or diagonal directions [302]. Assume (x, y) are co-ordinates in a binary image BW, for 8 connectivity the pixels observed are illustrated in (Eq. 3.26).

$$(x-1,y-1),(x-1,y),(x-1,y+1),(x,y-1),(x+1,y-1),(x+1,y),(x+1,y+1) \quad (3.26)$$

The labelled image can be initialised as such $L(x,y)$ as the same size if BW is initialised at zero (Eq. 3.27). If $BW(x,y) = 1$ which is the foreground as seen in (Eq. 3.28), the 8 connected neighbours are checked for connectivity, if none of the neighbouring pixels are assigned a label the labelled image will update as such (Eq. 3.29).

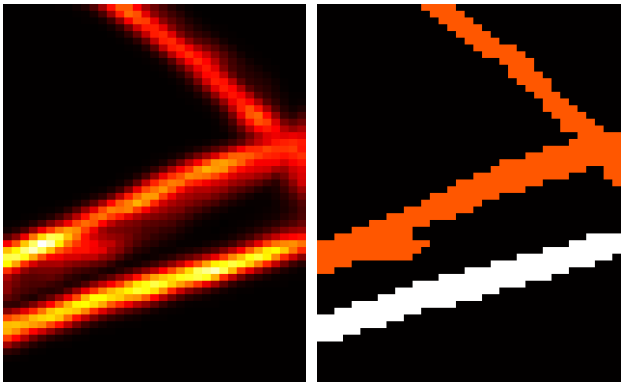
$$L(x,y) = 0 \quad \forall\,(x,y) \tag{3.27}$$

$$BW(x,y) = 1 \tag{3.28}$$

$$L(x,y) = k \tag{3.29}$$

If the connected neighbour already has a label l it can be assigned l in $L(x,y)$. If the connected neighbours have different labels they can merge into one label. New labels are assigned $(k+1)$ where k is the current label of the labeled image. In the background $L(x,y) = 0$ at (x,y).

Labelling can be seen in (Fig. 3.28) on the fibre image. Pixels that were connected to each other formed an object that was in a colour different from the other object and fibre in the image.



(a) (b)

Figure 3.28: Illustration of labelling on the fibre image. (a) Original fibres image (b) labelling of fibres, where each connected object is labelled.

Area Opening:

Area opening is the removal of objects or labeled objects with a particular number of pixels [303]. A labelled image can be defined as stated in (Eq. 3.30). Each unique label is k and the area can be found as such (Eq. 3.31). The condition "1" is to check the foreground pixels

where $(x, y) = 1$. A Threshold T in pixels is set and a new labeled image can be defined as (Eq. 3.32).

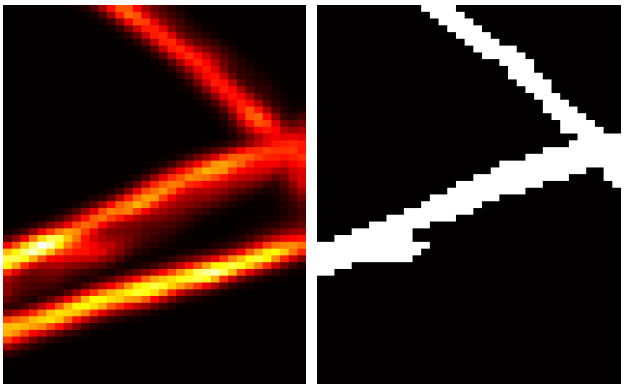
$$L(x, y) = k \tag{3.30}$$

$$A_k = \sum_{(x,y)\,:\,L(x,y)=k} 1 \tag{3.31}$$

$$\begin{cases} L(x, y) & \text{if } A_{L(x,y)} \geq T, \\ 0 & \text{if } A_{L(x,y)} < T. \end{cases} \tag{3.32}$$

Only objects that have more pixels than or equal to the threshold T are kept, the other objects are discarded and are converted into background. $A_{L(x,y)}$ is the area of the object to which the pixels (x, y) belong to.

This can be seen in (Fig. 3.29), where area opening of the fibre image removed a fibre that was below the threshold set.



(a) (b)

Figure 3.29: Illustration of Area opening on the fibre image. (a) Original fibres image (b) Area opening on the fibres, where one of the fibres below the threshold was removed.

Distance Transform:

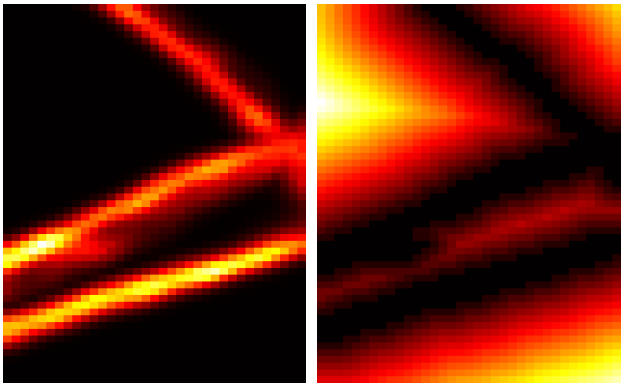
The Euclidean distance transform [304] works by assigning a number which is equivalent of the distance measurement of each pixel to the closest non-zero pixel.

Given a binary image, $BW(x, y) = 1$ represents foreground pixels whereas $BW(x, y) = 0$ represents background pixels. The Euclidean Distance Transform can be represented as $EucD(x, y)$ and it can be defined as (Eq. 3.33).

$$EucD(x, y) = \min_{(x_f,y_f):\, BW(x_f,y_f)=1} \sqrt{(x - x_f)^2 + (y - y_f)^2} \tag{3.33}$$

Here (x, y) are the pixel co-ordinates of the binary image BW. (x_f, y_f) are the co-ordinates of a foreground pixel in BW where $BW(x_f, y_f) = 1$. The Euclidean Distance Transform $EucD(x, y)$ assigns to each pixel (x,y) the smallest Euclidean Distance from the pixel to any of the foreground pixels (x_f, y_f). If a pixel is a foreground pixel the Euclidean distance transform assigns it a value of zero.

The distance transform algorithm was also applied to the fibre image in (Fig. 3.30), pixels further away from the fibres are yellow while pixels close to the fibres are red. The fibres themselves appear in black.



(a) (b)

Figure 3.30: Illustration of the Euclidean distance transform on the fibre image. (a) Original fibres image (b) Area opening on the fibres, where one of the fibres below the threshold was removed.

Skeleton map:

The skeleton map is useful as it shows the points where the objects in an image are joined together [305, 306], so that those areas could be targeted, and the objects of an image could be isolated from each other. The skeleton map is produced by reducing the image into 1-pixel wide curved lines that also does not change the structure of the objects, the centre line is extracted while the Euler number and topology stay the same. More generally, skeleton maps allow for the representation of objects as graphs or networks, highlighting branches, endpoints and the connections that exist between them.

Skeletonisation can be explained as a medial axis transform (Eq. 3.34).

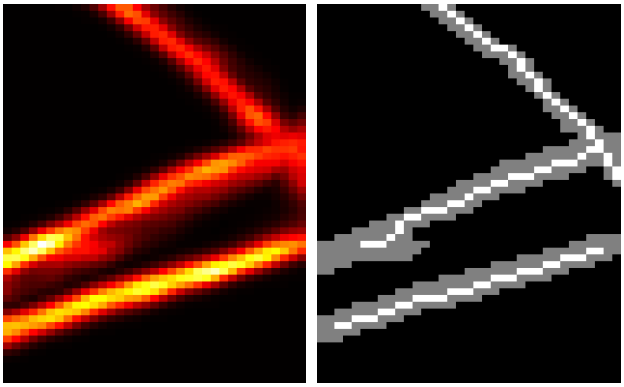
$$\text{skel}(BW) = \{(x, y) \mid EucD(x, y) = \max(EucD(x_f, y_f)) \text{ for neighbours } (x_f, y_f)\} \quad (3.34)$$

This suggests that the pixel coordinates (x, y) are set only if the condition (Eq. 3.35) is satisfied.

$$EucD(x, y) = \max\left(EucD(x_f, y_f)\right) \text{ for neighbours } (x_f, y_f) \quad (3.35)$$

(x, y) belongs to the skeleton if the $EucD(x, y)$ is equal to the maximum distance $\max(EucD(x_f, y_f))$ among the pixel neighbours (x_f, y_f). The pixel (x, y) exhibits the highest distance to the nearest background pixel among its neighbours, which means it is in the centre or the peak of the distance map. This indicates it is part of the medial axis of the shape. The distance map computed here gives values to pixels based on their distance to the nearest background pixel.

This can be seen in (Fig. 3.31), where the skeleton map algorithm was applied to the binary fibre image, the skeletal structure of the fibres was extracted.



(a) (b)

Figure 3.31: Illustration of the skeletonisation of the fibre-like structure. (a) Original fibres image (b) Skeletonisation of the binarised fibres. Fibre structure was reduced to one pixel wide of the medial axis.

Voronoi Diagram and Delaunay Triangulation:

Voronoi works by partitioning the objects or points in an image as suggested in [307]. The partitions are based on how close the originating point in the partition is to the points outside the partition, the more crowded the area in term of cell nuclei the smaller the partition of the originating point. The Delaunay triangulation is the same as the nerve of the points in the Voronoi diagram [308] and as used in [309]. With the condition of that none of the triangles vertices should be inside the circumcircle of any of the other triangles in the image as in (Fig. 3.32). For a set of N points (Eq. 3.36) in a Euclidean space, the Voronoi cell $V(M_i)$ for a location in the image M_i can be defined as (Eq. 3.37).

$$S = M_1, M_2, ..., M_N \tag{3.36}$$

$$V(M_i) = \{(x, y) \in \mathbb{R}^2 \mid d((x, y), M_i) < EucD((x, y), M_j) \, \forall j \neq i \tag{3.37}$$

(x, y) must be points on a two-dimensional Euclidean space to the point M_i. $EucD((x, y), M_i)$ represents the distance between (x, y) and point M_i. The distance function can be computed

as $EucD((x,y),M_i) = \sqrt{(x-x_i)^2+(y-y_i)^2}$, (x_i, y_i) are the co-ordinates of point M_i. The condition that must be satisfied is (Eq. 3.38).

$$EucD((x,y),M_i) < EucD((x,y),M_j) \tag{3.38}$$

This condition ensures that the point (x,y) is closer to M_i than any other point M_j.

Voronoi partitioning and Delaunay Triangulation were applied to the segmented cells from the DAPI images (Fig. 3.32b). The partitions of the Voronoi diagram are computed based on the distance of each nuclei to its neighbours (Fig. 3.32d). The lengths of the sides of the triangles correspond to the distance between the detected objects (Fig. 3.32e). The Voronoi diagram and Delaunay Triangulation has been used in this thesis to visualise the distribution of cell nuclei.

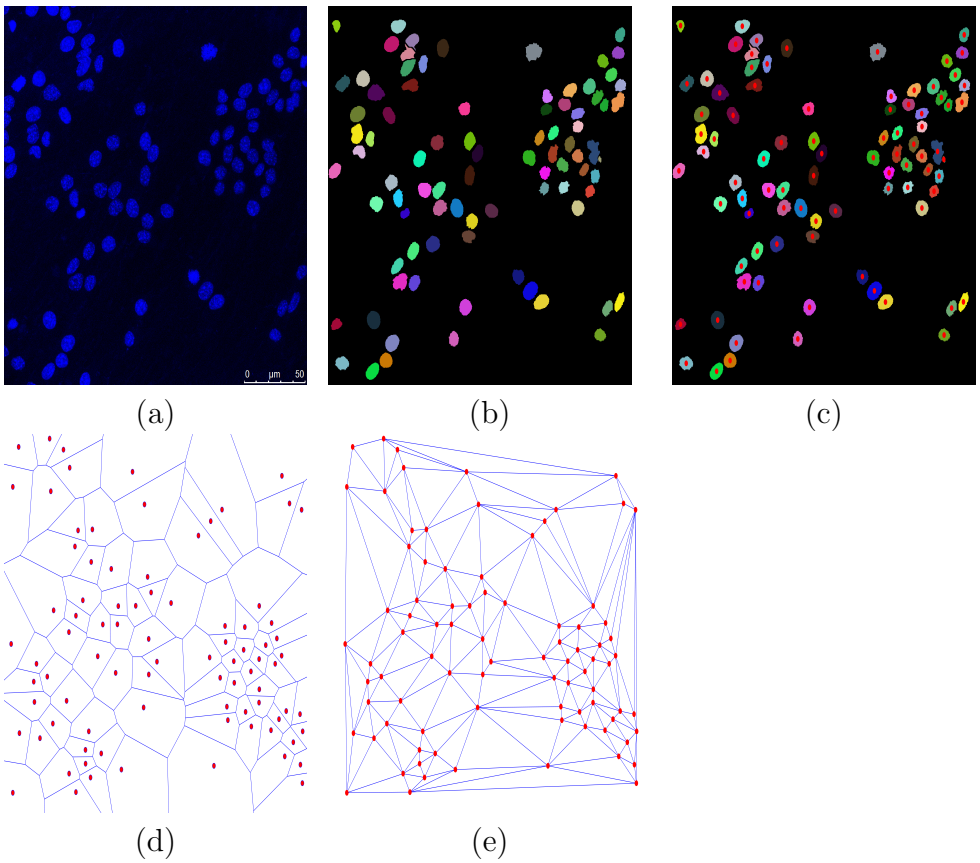


Figure 3.32: Application of Voronoi partitioning and Delaunay Triangulation on segmented DAPI cells. (a) Cell nuclei stained with DAPI. (b) Segmented labelled cell nuclei. Each nuclei is assigned a random colour. (c) Centroids positions of the labelled cells appear in red. (d) Voronoi partitioning of the cell nuclei, the centroid of each nuclei appears in red. (e) The Delaunay triangulation of the cell nuclei. The centroid positions of the nuclei appear in red. It can be seen that the size of the triangles relates to how the originating point of the objects relate to each other.

Cell segmentation:

Cell segmentation refers to the process by which an image of cells is split into segments, or the cells are identified based on their morphology [310, 311]. With cell segmentation it became easier to detect and predict cancerous cells. With the development of microscopy, high quality images are easily generated in substantial numbers and with the advancements in staining, biopsy samples are cellularly stained to be analysed. Cell segmentation was seen in [312].

CellPose:

CellPose [38] is a cell deep learning based segmentation method, that is capable of segmenting cells of several types of images. CellPose has been trained on a dataset containing 70,000 segmented objects. The CellPose method requires no training or parameter adjustments such as weights and can be used as is but fine-tuning of the parameters may also be used. Its architecture is remarkably like that of a U-Net, where it contains a downsampling path, a bridge and an upsampling path with skip connections, connecting the same sized decoder and encoder layers. The training of CellPose included altering the manually labelled objects into vector flow representations, that are predicted by the neural network. A simulated diffusion process is used that starts from the centre of the mask, this extracts the spatial gradients which point to the centre of the cell. The gradients are then normalised. During testing, the predicted flow fields are used to build a system that has fixed points, where its basins of attraction are those of the predicted masks, subsequently all the pixels that converge to the same fixed point are given the same mask label where they are labelled as the same object. The cells from the Normal Breast tissue core at the top of (Fig. 3.20a) were segmented with CellPose and can be seen laid over the original image core (Fig. 3.33a) and the labelled cells can be seen in (Fig. 3.33b). CellPose has been used on one of the datasets of this thesis to segment cell nuclei.

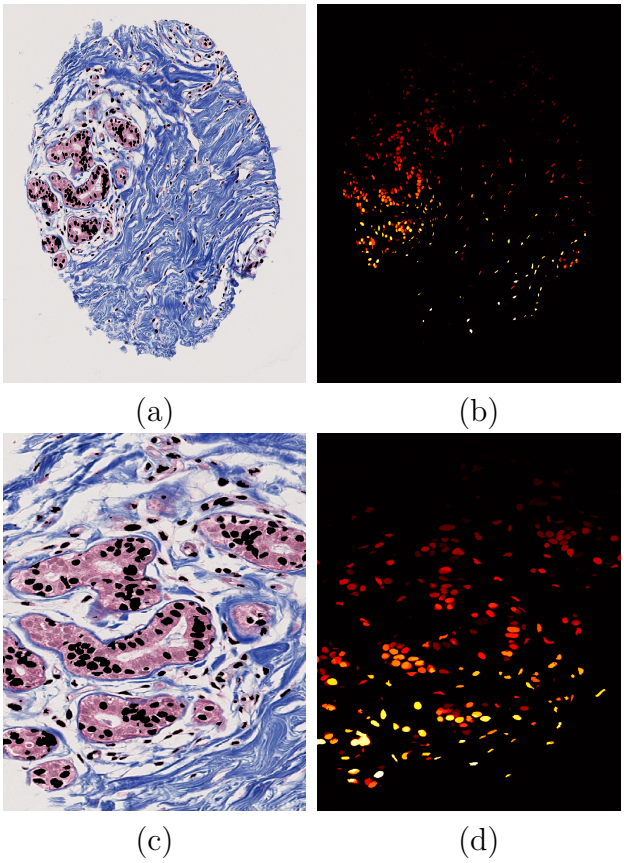


Figure 3.33: Illustration of segmentation of cells with CellPose on top row figure of (Fig. 3.19a). (a) Segmented cells have been removed and appear in black over the original core image. (b) Isolated labelled segmented cells. (c)-(d) Zoomed in ROI.

3.5.3 Tracing Algorithms:

Tracing algorithms are algorithms that can trace and detect thin fibrial like structures. A robust tracing algorithm is important for the identification of singular fibre-like objects that can reveal hidden characteristics.

Various manual, semi-automatic methods have been developed to delineate and trace elongated structures. Some of these approaches rely on more general algorithms that can be applied in tasks other than tracing, such as skeletonisation [313, 314, 315], watershed transform [316, 317], medial axis transform [318, 319, 320], Scale Space [45, 46], and Edge Detection [44]. However, many of these methods have been designed for specific applications, which could imply that these methods have been optimised to perform well on datasets with particular characteristics for example neuron tracing [321, 322, 323], retinal images [324, 325, 326], angiography [327, 328, 329], intravital vasculature seen with confocal, fluorescence or light microscopy [330, 331, 332], and extracellular matrix [333, 47]. Additionally, deep learning architectures like U-Net [334] have been extensively used in image analysis, delivering excellent results without the need for hand-crafted features or algorithms. Instead,

these models learn from labelled data pairs, allowing them to be trained for specific tasks, which for the purpose of this thesis is delineating fibre-like structures.

Edge Detection

The Canny Edge Detection [44] identifies object edges in an image by detecting changes in pixel intensities. Firstly, it applies Gaussian filtering to minimise noise as mentioned earlier (Eq. 3.7), $G(i,j) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{i^2+j^2}{2\sigma^2}\right)$ and the smoothed image is the product of the Gaussian filter on the image as such (Eq. 3.39).

$$\mathbb{I}'(x,y) = \sum_{i=-a}^{a} \sum_{j=-a}^{a} \mathbb{I}(x+i, y+j) \cdot G(i,j) \tag{3.39}$$

After which it calculates the intensity gradient of the smoothed image by applying a Sobel kernel in both the horizontal and vertical directions to get the first derivatives in both directions. The measurement of intensity changes in the x and y direction can be thought of in partial derivatives as such (Eq. 3.40) and (Eq. 3.41).

$$\nabla_x \mathbb{I}(x,y) \approx \frac{\partial \mathbb{I}(x,y)}{\partial x} \tag{3.40}$$

$$\nabla_y \mathbb{I}(x,y) \approx \frac{\partial \mathbb{I}(x,y)}{\partial y} \tag{3.41}$$

This could be done instantly if the image is treated as a continuous function, but digital images are discrete, so they consist of pixels rather than a continuous function. An approximation of these derivatives could be derived by convoluting the image with discrete convolutional kernel such as the Sobel. Sobel kernels have different magnitudes for the x direction and the y direction. The Sobel kernel that detects the horizontal changes $\mathbb{K}_x$ has the following magnitudes

$$\mathbb{K}_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

and the kernel that detects the vertical changes $\mathbb{K}_y$ has the following magnitude

$$\mathbb{K}_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

The convolution of the smooth image and the Sobel kernel in the x direction can be written as (Eq. 3.42) and for the y direction (Eq. 3.43).

$$\nabla_x(x,y) \approx \sum_{i=-a}^{a} \sum_{j=-a}^{a} \mathbb{K}_x(i,j) \cdot \mathbb{I}(x+i, y+j) \tag{3.42}$$

$$\nabla_y(x,y) \approx \sum_{i=-a}^{a} \sum_{j=-a}^{a} \mathbb{K}_y(i,j) \cdot \mathbb{I}(x+i, y+j) \tag{3.43}$$

These convolutions give approximations of the partial derivatives ∇_x, ∇_y. Here $\mathbb{I}(x+i, y+j)$ represent the intensity of the neighbouring pixels around (x,y) and $\mathbb{K}_x(i,j)$ and $\mathbb{K}_y(i,j)$ are the values of the Sobel kernel at the position (i,j). Sharp changes in intensity where there are high derivative values indicating the presence of edges. The magnitude of the gradient which is the strength of the change at each pixel can be computed as such (Eq. 3.44) and the gradient direction as such (Eq. 3.45).

$$M(x,y) = \sqrt{\nabla_x^2 + \nabla_y^2} \tag{3.44}$$

$$\theta(x,y) = \tan^{-1}\left(\frac{\nabla_y}{\nabla_x}\right) \tag{3.45}$$

The magnitude reveals which edges are strong and which edges are weak. The direction indicates the orientation of the edge at that pixel measured in degrees or radians. Non-maximum suppression is then applied to thin the edges by retaining only the local maxima along the gradient direction. This is done by checking the magnitude of the current pixel (x,y) in the gradient direction for example if $\theta(x,y)$ is $0°$, $M(x,y)$ is compared to $M(x-1,y)$ and $M(x+1,y)$, if $M(x,y)$ does not exhibit a higher value than the pixels in its direction, it is set to 0. Thresholds are then applied to assign the edges to three classes (strong edges, weak edges, and non-edges). A high T_h and low T_l threshold are set. Strong edges can be represented as (Eq. 3.46), weak edges as (Eq. 3.47), and non-edges as such (Eq. 3.48).

$$M(x,y) \geq T_h \tag{3.46}$$

$$T_l \leq M(x,y) < T_h \tag{3.47}$$

$$M(x,y) < T_l \tag{3.48}$$

Hysteresis is then applied to determine which edges to keep or discard, if a weak edge is connected to a strong edge pixel, it is preserved as part of an edge, else the edge is discarded. The steps of Edge Detection are simplified in (Fig. 3.34). Edge Detection detects two edges for each fibre of the FC image of dataset 2 (Fig. 3.35a), this is seen in (Fig. 3.35b).

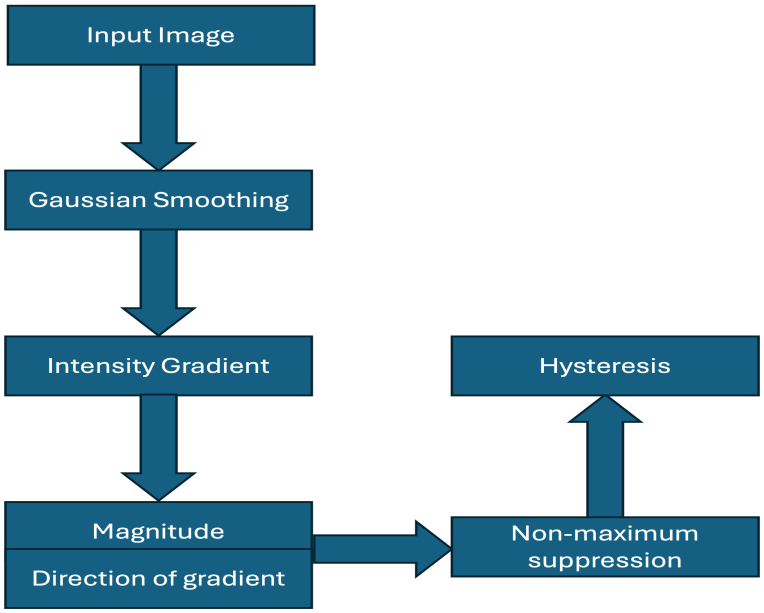
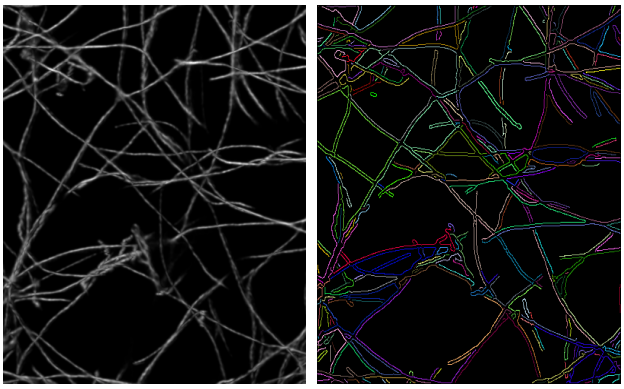


Figure 3.34: Edge Detection pipeline follows: Gaussian filtering is applied to the image after which the intensity gradient is extracted as well as the magnitude and the direction of the gradient. Non-maximum suppression is applied followed by hysteresis.



(a) (b)

Figure 3.35: Application of Edge Detection the fluorescently labelled collagen image. (a) FC image. (b) Edge Detection's application on the FC image, each fibre resulted in two edges detected for it. Colours are assigned randomly to distinguish between the edges.

Watershed Transform:

The Watershed Transform [316, 317] is a powerful image processing technique that views an image as a topographic surface where high intensity pixels (ridges) are regions of high elevations like peaks. Low intensity pixels represent low elevations such as basins. The name Watershed is derived from the concept of geographical watersheds, where water moves down from the top of the mountain ridges towards different basins. Watershed aims to find the

boundaries between the regions in an image. Watershed views the regions in the image as catchment basins, areas where water would accumulate if it were poured from the top of the image.

Firstly, the input image $\mathbb{I}(x, y)$ is converted to a gradient magnitude image seen in (Eq. 3.44) but for simplicity it is repeated here in (Eq. 3.49) .

$$M(x, y) = \sqrt{\nabla_x^2(x, y) + \nabla_y^2(x, y)} \tag{3.49}$$

∇_x and ∇_y are the partial derivatives of the image with respect to x and y. As mentioned earlier the partial derivatives highlight regions of intensity changes which correspond to the ridges or edges.

A set of markers is then initialised where each marker corresponds to the different edges or regions in the image. This can be done by using the inverse of the Euclidean distance transform to highlight regions where objects exist, it could also be found using thresholding. Markers can be set as such (Eq. 3.50).

$$\mathbb{M}(x, y) = \mathbf{R} \tag{3.50}$$

where pixel (x_R, y_R) belongs to the region $\mathbf{R}$.

The image is then "flooded", imagine pouring water into the low intensity regions the basins from the regions highlighted by the markers. As the water rises inside the marked regions, the regions begin to fill up. Once water from different regions or markers meet which are the basins, a boundary is placed to separate the regions. This process is continued until all regions have been flooded with water and the boundaries between the regions have been identified. The watershed lines or ridges represent the boundaries between regions. One of the limitations of the Watershed Transform is its affinity to over-segment if an image is noisy.

Scale Space:

Scale Space [45, 46] blurs an image across levels of scales to represent its characteristics. Crucial features emerge at lower scales, while finer details are visible at first scales and are progressively smoothed out. Scale Space progressively smooths an image through Gaussian blurring using an increased standard deviation or scale σ. As mentioned before a Gaussian kernel can be represented as such (Eq. 3.7) $G(i, j, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{i^2+j^2}{2\sigma}\right)$ and given an input image $\mathbb{I}(x, y)$. The Scale Space output at position (x, y) and scale σ can be represented as a convolution of the original image with the Gaussian kernel at various stages as such (Eq. 3.51).

$$\mathbb{S}(x, y; \sigma) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} I(x + i, y + j)G(i, j; \sigma) \tag{3.51}$$

The shape of the Gaussian kernel as mentioned earlier is determined by the variance σ value. A small σ preserves the finer details in the image while a larger σ results in the smoothing of finer structures and highlighting larger structures. The semicolon (;) suggests that the convolution is only performed on variables (x, y). A Scale Space of an image is constructed when the image is progressively blurred with increasing values of σ and thus images are created. This is represented as such (Eq. 3.52).

$$\{\mathbb{S}(x, y, \sigma_1), \mathbb{S}(x, y, \sigma_2), \ldots, \mathbb{S}(x, y, \sigma_n)\} \quad \text{where } \sigma_1 < \sigma_2 < \ldots < \sigma_n \tag{3.52}$$

To detect the edges, a gradient magnitude of the image at different scales can be computed, as mentioned earlier gradient changes correspond to regions where there is a change of intensities in an image which are an indication of an existence of an edge. Like Edge detection the partial derivatives or the change in intensity is found as such (Eq. 3.53) and the magnitude is found similarly as such (Eq. 3.54).

$$\nabla_x(x, y, \sigma) = \frac{\partial \mathbb{S}(x, y, \sigma)}{\partial x}, \quad \nabla_y(x, y, \sigma) = \frac{\partial \mathbb{S}(x, y, \sigma)}{\partial y} \tag{3.53}$$

$$M(x, y, \sigma) = \nabla_x^2(x, y, \sigma) + \nabla_y^2(x, y, \sigma) \tag{3.54}$$

Steps of Scale Space are shown in (Fig. 3.36). Scales of 10 levels of Gaussian filtering with increasing kernel sizes from 1 till 10 and the trace of Scale Space at each level is shown in (Fig. 3.37). This method was implemented as seen in [332].

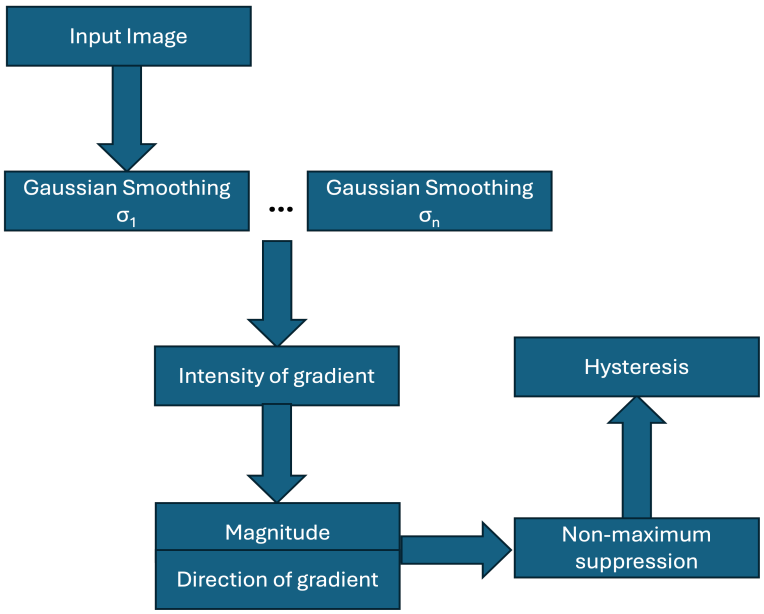
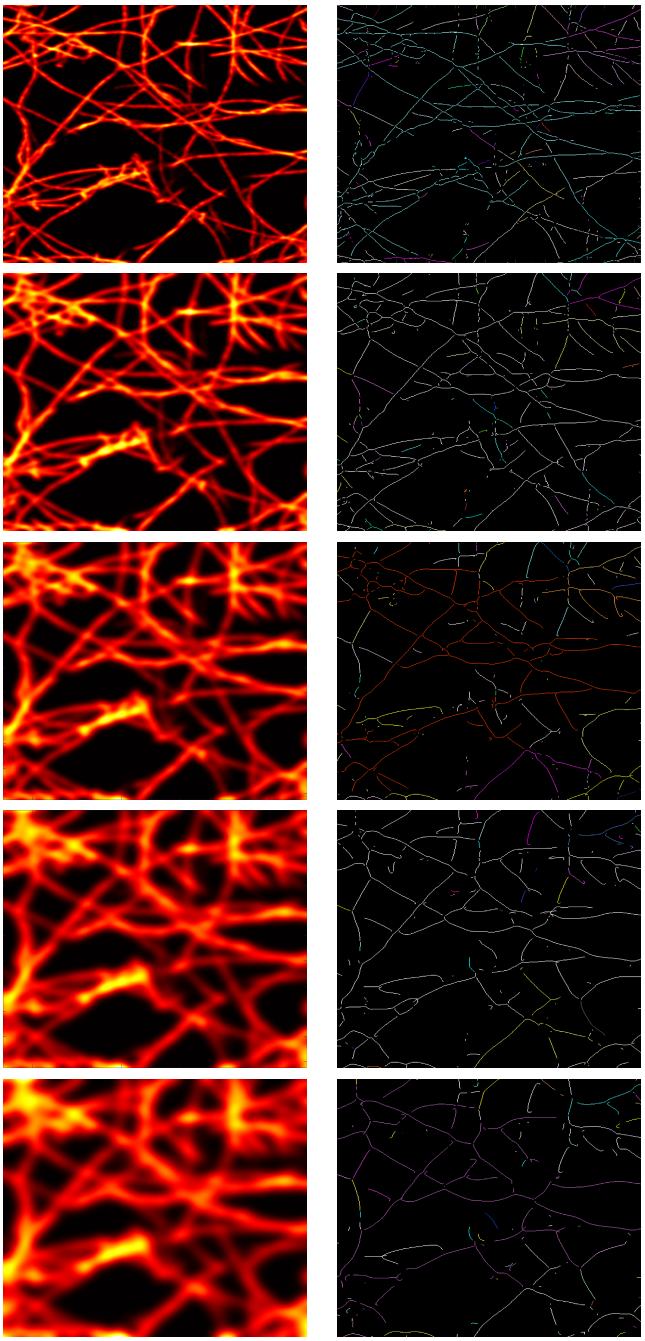


Figure 3.36: Scale Space pipeline follows: Gaussian filtering of multiple sizes is applied to the image to reveal details at multiple scales. The intensity magnitude and direction of the gradient is extracted. After which non-maximum suppression is applied followed by hysteresis.



(a) (b)

Figure 3.37: Application of Scale Space levels with increasing Gaussian filtering on (Fig. 3.14a). (a) Gaussian filtering with increasing kernel sizes of 2, 4, 6, 8 and 10. (b) Scale Space tracing across Gaussian filtering of multiple levels.

Twombli:

Twombli [47] is a FIJI macro tool for quantifying fibre-like structures. Twombli uses existing techniques, including Ridge Detection [335] and AnaMorf [336]. Ridge Detection assesses first and second directional derivatives using Gaussian kernels to detect line points and extract line width with sub-pixel accuracy. The symmetrical line profile is presented as (Eq. 3.55).

$$f_b(i) = \begin{cases} I, & \text{if } |i| \leq wid \\ 0, & \text{otherwise} \end{cases} \tag{3.55}$$

Here wid is half of the line width, and I is the intensity of the line. The function f_x is a model of the shape of the line in one-dimensional space using a bar-shaped line profile. This model assumes a line has the same intensity or contrast on both sides, but that is rarely true in real images. Therefore, an asymmetrical line profile can be defined as (Eq. 3.56) where in (Eq. 3.57), $\mathbf{a}$ is the asymmetry parameter.

$$f_{\mathbf{a}}(i) = \begin{cases} 1, & \text{if } -wid \leq i < 0 \\ \mathbf{a}, & \text{if } 0 \leq i \leq wid \\ 0, & \text{otherwise} \end{cases} \tag{3.56}$$

$$\mathbf{a} \in [0, 1] \tag{3.57}$$

If $\mathbf{a} = 1$, the line is symmetrical which means both sides of the line have the same contrast, while $\mathbf{a} < 1$, the line is asymmetrical which means one side of the line is darker than the other. In asymmetrical lines, one side might have a larger gradient compared to the other, which as mentioned earlier is a transition in the intensity of the pixels of the line. This asymmetry affects the position of the detected line as the side with the weaker gradient will pull the detected centre towards it. After which to improve the robustness of the image Gaussian kernels are used as before to smooth the image which helps detect the changes of a line which are the edges. The Gaussian function is defined as (Eq. 3.7), $G(i, j, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{i^2+j^2}{2\sigma}\right)$ but for simplicity here and for the 1D line model (Eq. 3.58).

$$G(i, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{iwid^2}{2\sigma}\right) \tag{3.58}$$

Therefore, the Gaussian first and second derivatives can be (Eq. 3.59) and (Eq. 3.60).

$$g'_\sigma(i) = -\frac{i}{\sigma^2} G_\sigma(i) \tag{3.59}$$

$$g''_\sigma(i) = \left(\frac{i^2 - \sigma^2}{\sigma^4}\right) G_\sigma(i) \tag{3.60}$$

These derivatives help locate the edges where the intensity changes, which indicate the presence of a line.

The response of the convolved line is as such (Eq. 3.61).

$$r_b(i, \sigma, wid, I) = I\left(\phi_\sigma(i + wid) - \phi_\sigma(i - wid)\right) \tag{3.61}$$

$\phi_\sigma(i)$ is the cumulative distribution function of the Gaussian function $g_\sigma(i)$ which helps to smooth out transitions as it captures the cumulative effect of the Gaussian blur up to point i. For curvilinear structure detection, $\phi_\sigma(i)$ assists in describing how a line profile changes when it is smoothed by the Gaussian function. The difference $(\phi_\sigma(i + wid) - \phi_\sigma(i - wid)$ represents how the profile changes from one side of the line to the other. This gives insight into the positioning and the intensity of the line. $i + wid$ is a point on the right edge of the line, while $i - wid$ is a point on the left edge of the line. If there is a significant difference between the two points $(\phi_\sigma(i + wid) - \phi_\sigma(i - wid)$, it indicates that a line is present at that location. If the difference is small, it suggests that the area around point x is more uniform and there is no profound change in intensity.

When the Gaussian function is integrated from $-\infty$ to i as seen in (Eq. 3.62), the cumulative Gaussian function is extracted. This is an important concept in image processing for smoothing operations as it helps in computing the cumulative intensity response when an image is smoothed with a Gaussian filter. It assists in detecting intensity transitions in an image.

$$\phi_\sigma(i) = \int_{-\infty}^{i} G_\sigma(t)\, dt \tag{3.62}$$

The first and second order derivatives can be represented as such (Eq. 3.63) and (Eq. 3.64).

$$r'_b(i, \sigma, wid, I) = I\left(G_\sigma(i + wid) - G_\sigma(i - wid)\right) \tag{3.63}$$

$$r''_b(i, \sigma, wid, I) = I\left(g'_\sigma(i + wid) - g'_\sigma(i - wid)\right) \tag{3.64}$$

The first derivative response measures the response rate of change at point x. This indicates the existence of the line's centre. If $r'_b = 0$ is an indication of the centre of line. If r'_b changes from positive to negative this suggests a local maximum which is an indication of the centre of a bright line while if r'_b changes from negative to positive suggesting a local minimum which could be the centre of a dark line.

The second derivative response determines the strength and the sharpness of the detected line and the saliency.

As it uses Scale Space the variance σ of the Gaussian kernels must be determined and here it is based on the line width where (Eq. 3.65).

$$\sigma \geq \frac{wid}{\sqrt{3}} \tag{3.65}$$

Handling the asymmetrical line profile involves finding the response of the asymmetrical lines as we have assumed that lines in images have equal contrast on both sides when in it is not true in most cases.

The response of the asymmetrical bar shaped lines is represented as seen in (Eq. 3.66).

$$r_{\mathbf{a}}(i, \sigma, wid, \mathbf{a}) = \phi_\sigma(i + wid) + (\mathbf{a} - 1)\phi_\sigma(i - wid) \tag{3.66}$$

$\phi_\sigma(i)$ represents the cumulative Gaussian integral, which outlines how the intensity is distributed after smoothing. The term $\phi_\sigma(i + wid)$ represents the response from one side of the line while $(\mathbf{a} - 1)\phi_\sigma(i - wid)$ adjusts the response based on the contrast asymmetry $\mathbf{a}$. This means that if $\mathbf{a} = 1$ the equation will just simplify to the symmetrical case. In the case that $\mathbf{a} < 1$, one side of the line will have less intensity, which will shift the detected position.

The first derivative response can then be computed as such (Eq. 3.67).

$$r'_{\mathbf{a}}(i, \sigma, wid, \mathbf{a}) = G_\sigma(i + wid) + (\mathbf{a} - 1)G_\sigma(i - wid) \tag{3.67}$$

The first derivative response detects where the intensity changes the most which can help identifying the edges. $\mathbf{a} - 1$, the asymmetry factor can shift the edge response. If $\mathbf{a} < 1$, one edge is weaker which would make the detected centre shift towards the stronger side.

The second derivative of the response can be computed as such (Eq. 3.68).

$$r''_{\mathbf{a}}(i, \sigma, wid, \mathbf{a}) = g'_\sigma(i + wid) + (\mathbf{a} - 1)g'_\sigma(i - wid) \tag{3.68}$$

The second derivative finds the exact centre of the line.

When lines have a different contrast on each side hence the asymmetry, the detection could be biased which means that the detected centre of the line is shifted. The shift of the detected centre from the actual line centre can be represented as (Eq. 3.69). In a symmetrical case, the point where the function changes sign is at the true centre. While in an asymmetrical case, that point is shifted which indicates that the detected centre is biased. The line shift equation shows how the estimated position is shifted due to this asymmetry.

$$\mathbb{L} = \frac{\sigma^2}{2wid} \ln(1 - \mathbf{a}) \tag{3.69}$$

Bias is reduced by inverting the Scale Space behaviour of the asymmetric line model. Bias inversion is the process of correcting this systematic error by adjusting the estimated line position back to where it should be. Ridge detection uses the estimated values of a, σ and w, to determine how the detected line position is shifted back by the value of l to correct for the bias, this ensures that the final detected position of the line aligns more closely with the true centre of the line.

For two-dimensional lines, they are detected by analysing the derivatives in different directions. The Hessian matrix is used, which involves second order partial derivatives as such (Eq. 3.70).

$$H(x, y) = \begin{bmatrix} r_{\mathbf{xx}} & r_{\mathbf{xy}} \\ r_{\mathbf{xy}} & r_{\mathbf{yy}} \end{bmatrix} \tag{3.70}$$

$r_{\mathbf{xx}}$, $r_{\mathbf{xy}}$, $r_{\mathbf{xy}}$ and $r_{\mathbf{yy}}$ are the second derivatives of the image in the x and y directions. The Eigenvalue of this matrix indicates the characteristics of the curvature which helps in determining the direction of the line and its intensity changes.

Once the individual lines are detected, they are linked together to form continuous lines, this is done by following the direction of the line and connecting nearby points that share similar orientations. The choice of neighbour to add is based on which neighbour would minimise the difference in subpixel distance and orientation between the two points.

Points of lines or ridges are chosen based on whether they are above or below a threshold. An upper threshold T_u is based on the highest grayscale value b_u of the line which estimates the upper threshold as such (Eq. 3.71).

$$T_u = \left\lfloor 0.17 \cdot \frac{2 \cdot b_{\mathrm{u}} \cdot \frac{wid}{2}}{\sqrt{2\pi}\sigma^3} e^{-\frac{\left(\frac{wid}{2}\right)^2}{2\sigma^2}} \right\rfloor \tag{3.71}$$

Line points that have a response as large as this threshold are accepted, while the lower threshold is set as such (Eq. 3.72).

$$T_l = \left\lfloor 0.17 \cdot \frac{2 \cdot b_{\mathrm{l}} \cdot \frac{wid}{2}}{\sqrt{2\pi}\sigma^3} e^{-\frac{\left(\frac{wid}{2}\right)^2}{2\sigma^2}} \right\rfloor \tag{3.72}$$

Here T_l is the low threshold, and b_l is the lowest grayscale value of the line. AnaMorf quantifies the structures identified. AnaMorf is a post processing step that ensures the detected structures are actual fibres and not skeleton artefacts. It returns the average of properties of filamentous structures in an image. It requires a minimum branch length in microns to recognise the branch as a true branch and not the result of skeletonisation. In addition to this a maximum circularity must be stated to act as a morphological constraint and threshold to distinguish between background and foreground. Objects with a circularity value greater than this are ignored. Minimum area in microns is another threshold parameter where only objects of specific areas are included. Touching objects may also be separated by Watershed segmentation.

Twombli also incorporates OrientationJ to obtain the global alignment metric, determining the dominant direction of the fibrial structure. By analysing changes in intensity signal as a function of nearby pixel distances, Twombli constructs segmentation masks using fibre edges, along with a skeletonised version with branch-points and endpoints highlighted.

Twombli's full steps have been simplified and illustrated in (Fig. 3.38). In the implementation of Twombli (Fig. 3.39), the minimum fibre length and minimum branch length must be entered by the user. In all cases, a minimum fibre (Fig. 3.39b) and branch length (Fig. 3.39c) of 10 pixels is chosen. The final Twombli trace can be seen in (Fig. 3.39d).

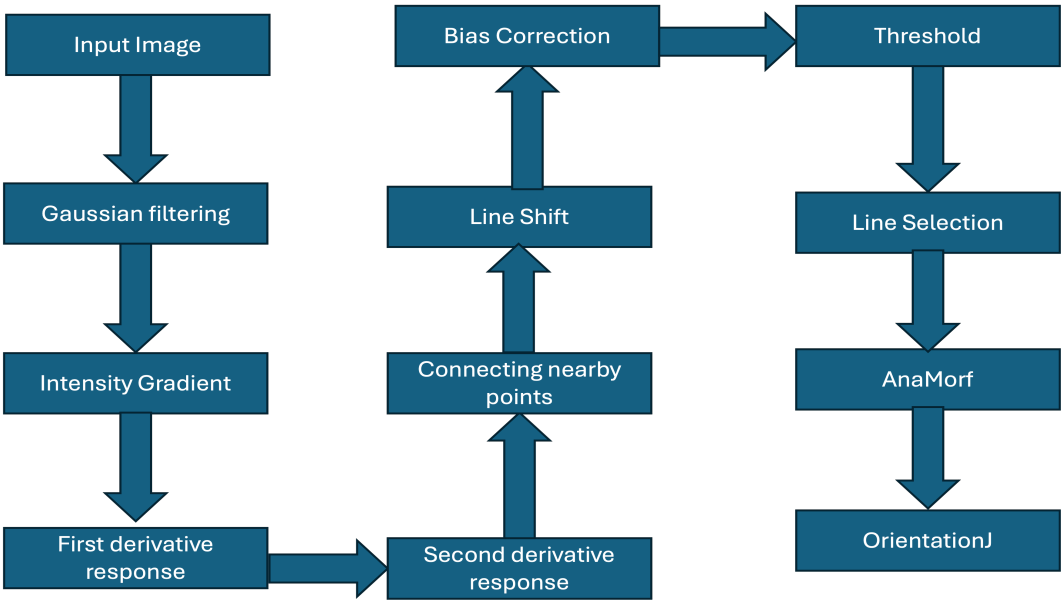


Figure 3.38: Twombli pipeline follows: Input image is Gaussian filtered, where the first and second derivative gradient intensity responses are extracted. Neighbouring points are connected. Fibre lines are shifted, and the bias is corrected. Thresholds are applied and lines are selected. AnaMorf is applied to extract morphological features and OrientationJ for orientation features.

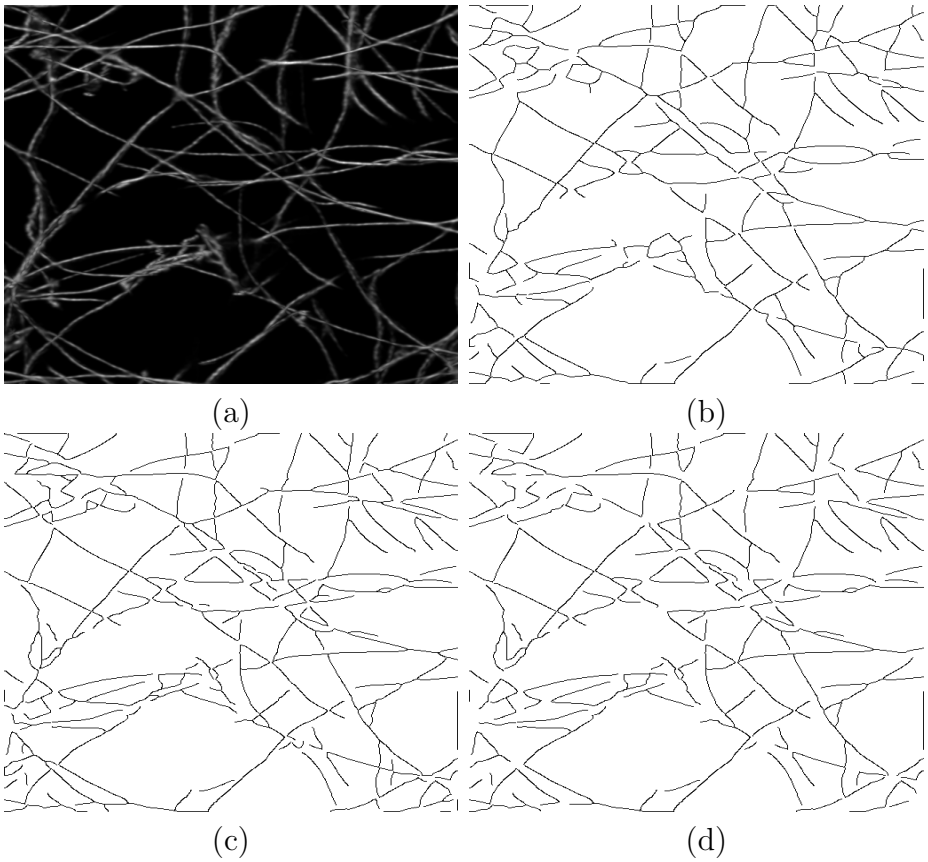


Figure 3.39: Illustration of the processing of one image on (Fig. 3.14a) with Twombli. Twombli traces require the user to enter a minimum fibre and branch length to reach the final tracing. (a) SHG fibre image. (b) Minimum fibre length of 10 pixels. (c) Minimum branch length of 10 pixels. (d) Final tracing of Twombli.

CT Fire:

CT Fire [48, 279] was developed to extract and quantify collagen fibres from CRM and Second Harmonic Generation images. It utilises a curvelet transform (CT) [337] de-noising filter. The curvelet filter represents the image as stacked elements following ridge lines and wavelets in the orthogonal direction, with a frequency wrapping based de-noising technique. Curvelets commonly follow the rule in (Eq. 3.73). The denoising technique here implements a frequency-wrapping version of fast discrete CT and reconstructs the images using the highest % of the curvelet coefficients from scales 4, 5 and 6 out of 7 total scales. The curvelet coefficient was optimised to give the best results.

The fibre extraction algorithm [278, 48, 279] processes the binary image, with foreground pixels representing fibres. It computes a distance transform to determine the proximity of the foreground pixels to the nearest background pixel, identifies maximal ridges of the smoothed image, and generates nucleation points. After which fibres extend from each nucleation point based on trajectory, followed by removal of short branches. The algorithm includes many parameters which must be set by the user.

$$\text{width} = \text{length}^2 \tag{3.73}$$

The Fire algorithm is developed to extract the network architecture of three-dimensional biopolymer networks from confocal microscopy images. It improves upon previous algorithms and is designed to reconstruct both the geometry and mechanical properties of the biopolymer networks.

Firstly, a network of fibres is defined as (Eq. 3.74).

$$[(\mathcal{N} = \mathcal{V}, \mathcal{X}, \mathcal{F}]) \tag{3.74}$$

which consists of a list of n vertices $\mathcal{V} = 1, 2, ..., n$ with co-ordinates $\mathcal{X} = \mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n$ and n fibres $\mathcal{F} = \mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n$. Each vertex $\mathcal{V}_n$ has a 3D vector $\mathbf{e}_n = (x_n, y_n, z_n)$. Each fibre $\mathbf{f}^n$ is a vector of vertex identifiers $\mathbf{v}_n^k$. $\mathbf{f}^n = (\mathbf{v}_1^k, \mathbf{v}_2^k, ..., \mathbf{v}_n^k)$ which indicates the vertices that fibre k goes through. The superscript of $\mathbf{v}_n^k$ identifies the fibre and the subscript identifies the ordered position of the vertex in that fibre.

The initial step of the Fire algorithm is to smooth the image $\mathbb{I}$ with a Gaussian filter G to reduce noise (Fig. 3.41b). The image is then binarised as defined earlier by setting a threshold.

$$BW(\mathbf{e}) = \begin{cases} 1 & \text{if } I(\mathbf{e}) > T \\ 0 & \text{otherwise} \end{cases}$$

Here pixels at position $\mathbf{e}$ are given the value 1 if they belong to the foreground and 0 if they belong to the background. This is seen in (Fig. 3.41c) on the FC image of dataset 2.

A Euclidean distance is then computed for the pixels which would give the minimum distance from each fibre pixel to a zero-background pixel, this is used to locate the central axis of fibres. As defined earlier the $EucD$ assigns each pixel a value that represents the

shortest distance between this pixel and a non-zero pixel as such (Eq. 3.33)

$$EucD(x,y) = \min_{(x_f,y_f):\, BW(x_f,y_f)=1} \sqrt{(x-x_f)^2+(y-y_f)^2}$$

but in this case, it is the smallest distance between the current pixel and a pixel in the background, highlighting the ridges as such (Eq. 3.75).

$$EucD(x,y) = \min_{(x_f,y_f):\, BW(x_b,y_b)=1} \sqrt{(x-x_b)^2+(y-y_b)^2} \tag{3.75}$$

The distance map is smoothed further with a Gaussian filter G (Fig. 3.41d).

The goal now is to trace along the maximal ridges of the fibres of the distance function. Nucleation points are identified (Fig. 3.41e) as local maximums of the distance map $EucD(\mathbf{e})$ which represent the thickest part of the fibres. These are the starting points for tracing fibres. Nucleation points can be defined as such (Eq. 3.76).

$$NucP = \{\mathbf{e} \in \mathcal{E} \mid EucD(\mathbf{e}) \geq \max(T_{\text{nuc}}, EucD(\mathbf{v})) \;\forall\; \mathbf{v} \in B(\mathbf{e}, s_{\text{xbox}}) \tag{3.76}$$

Here $B(\mathbf{e}, s_{\text{xbox}})$ is a box of size s_{xbox} around point $\mathbf{e}$ which is the current pixel and T_{nuc} is the threshold for the nucleation points. These parameters are decided by the user. $\mathbf{v}$ is a neighbouring point for pixel $\mathbf{e}$. This is used to compare the Euclidean distance $EucD(\mathbf{e})$ and $EucD(\mathbf{v})$ to determine if pixel $\mathbf{e}$ is a local maximum.

After identifying the nucleation points, the next step is tracing the branches extending from the nucleation points (Fig. 3.41f).

$$\text{LMP}_n = \{\mathbf{v} \in \partial B(\mathbf{u}_i, r_i) \mid EucD(\mathbf{v}) \geq \max\{T_{\text{LMP}}, EucD(\mathbf{q})\},\; \forall \mathbf{q} \in \partial B(\mathbf{e}_n, r_n) \text{ s.t. } ||\mathbf{v}-\mathbf{q}||_\infty = 1\} \tag{3.77}$$

Here LMP_n (Eq. 3.77) represents the set of Local Maxima Points (LMPs) at a specific nucleation point $NucP(\mathbf{e})$. $\partial B(\mathbf{e}_n, r_n)$ is the boundary of the neighbourhood box with radius r.

$$\mathbf{v} \in \partial B(\mathbf{u}_i, r_i) \tag{3.78}$$

(Eq. 3.78) states that $\mathbf{v}$, the candidate point for the LMP is inside the boundary region of the neighbour of radius r and centred at $\mathbf{e}_n$ which is the set of points surrounding $\mathbf{e}_n$.

$$EucD(\mathbf{v}) \geq \max\{T_{\text{LMP}}, EucD(\mathbf{q}) \tag{3.79}$$

(Eq. 3.79) defines the condition for a local maximum. It states that $EucD(\mathbf{v})$ at point $\mathbf{v}$ must be greater than or equal to both the user input threshold of the LMP T_{LMP} and the $EucD(\mathbf{q})$. $\mathbf{q}$ is a neighbouring point of point $\mathbf{v}$. $\mathbf{q}$ must be inside the same boundary box as $\mathbf{e}_n$.

$$||\mathbf{v} - \mathbf{q}||_\infty = 1 \tag{3.80}$$

(Eq. 3.80) restricts $\mathbf{q}$ to only include points that are exactly one unit away from $\mathbf{v}$ in the maximum norm. ∞-norm measures the largest absolute difference between the points $\mathbf{v}$ and $\mathbf{q}$. In conclusion this allows the algorithm to follow the ridges of the distance map, identifying points that are likely to lie along the centre of the fibres. Fibres are then traced iteratively extending from one nucleation point to the other through the LMPS. The direction of the fibre is updated as each new LMP is added. The direction is updated using a weighted average of the previous direction, the new segment direction. This is represented as such

$$\hat{\mathbf{d}}_{n+1} = \frac{\text{dirdecay}\hat{\mathbf{d}}_n + (1 - \lambda_{\text{dirdecay}})\boldsymbol{\delta}_j}{\|\text{dirdecay}\hat{\mathbf{d}}_n + (1 - \lambda_{\text{dirdecay}})\boldsymbol{\delta}_j\|_2} \tag{3.81}$$

In (Eq. 3.81), $\hat{\mathbf{d}}_{n+1}$ represents the updated direction vector of the next iteration. λdirdecay is a decay factor controlling how much the new direction influences the fibre orientation, it typically holds a value between 0 and 1. $\boldsymbol{\delta}_n$ is the direction of the new LMP.

The fibres are extended if there are new LMPS and that they follow the direction of the fibre. If the angle between the fibre's current direction $\hat{d}_n$ and the candidate LMP $\hat{\delta}_n$ exceed a threshold $cos(T_{\text{ext}})$ the fibre extension stops (Fig. 3.41g).

After the fibre extension, dangling fibres or short fibres are removed as they are artefacts and do not contribute to the network. Short fibres are identified as those having a length less than l_{dang}. The dangling fibres are also identified from their angle as they would have a tangent angle that is different from any other incoming fibre by atleast $\theta_{\text{dang-extend}}$.

When two fibres meet at their peripheral end points, or their end points are in proximity, and share similar orientation at the ends, these two fibres are joined together to form one long fibre. Euclidean distance is measured and if the distance is below a threshold T_{linkd} and the θ difference is between 130 and 180, the fibres are linked. The final output of the algorithm is the set of fibres (Eq. 3.82) where each fibre is represented as a set of connected vertices (Fig. 3.39h).

$$\mathcal{F} = \mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n \tag{3.82}$$

As this algorithm only accepts three-dimensional images. Two-dimensional images are converted to three-dimensional images by replicating the image across three layers using the Matlab function "concat". The algorithm requests from the user to set a threshold to binarise the image. To keep a consistent method of thresholding, the code of CT Fire for threshold was adjusted to use an Otsu threshold for automation rather than choosing a new threshold for every image. An image's journey within CT Fire is shown in (Fig. 3.40) using the FC image from Dataset 2.

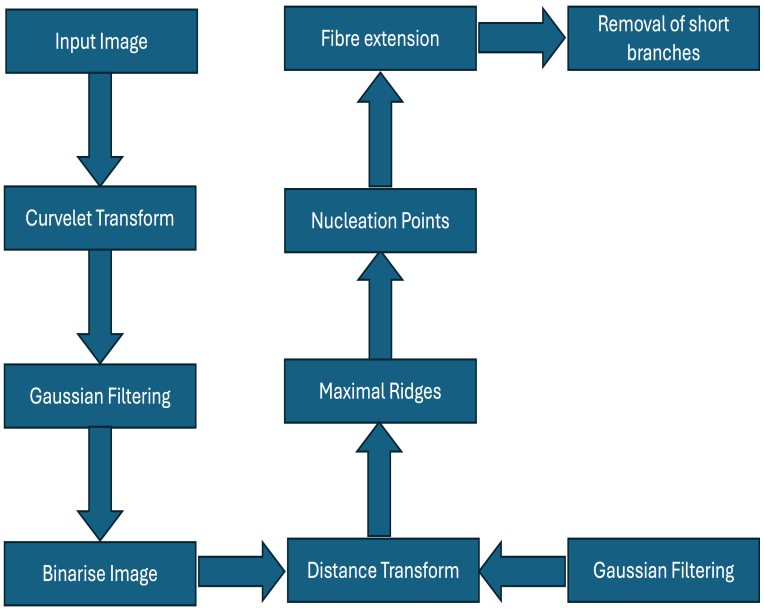


Figure 3.40: CT Fire pipeline follows: The input image is filtered using a curvelet transform followed by Gaussian filtering. The image is then binarised and a Euclidean distance map is extracted which is further smoothed. Maximal ridges and nucleation points are found. The fibres are extended from the nucleation points and short branches are removed.

Graph Based:

The Graph-based algorithm [49, 50] combines fast detect curvilinear anisotropic objects, in combination with a fibre extraction algorithm. The algorithm uses Gabor filters to smooth confocal image fibres and attenuate fibre features. Gabor filters offer flexibility by detecting multiple fibre elements at different frequencies and orientations, avoiding translation and rotation errors. Gabor Filters combine a Gaussian kernel to smooth the image with a sinusodial wave to capture the frequency information to analyse the image at different scales and orientations. The Gabor Kernel can be defined as (Eq. 3.83).

$$\mathbb{K}_{\text{Gab}}(p) = \exp\left(-\frac{1}{2} i^T \Sigma^{-1} p\right) \cos\left(\frac{2\pi \mathbf{s}_{\theta_i}}{\lambda_n} + \phi\right) \tag{3.83}$$

Here $p = (i, j)^T$ is the two-dimensional vector of pixels in Cartesian co-ordinates. $\exp\left(-\frac{1}{2} p^T \Sigma^{-1} p\right)$ is the Gaussian function that defines the shape of the kernel, $\cos\left(\frac{2\pi \mathbf{s}_{\theta_i}}{\lambda_j} + \phi\right)$ is the sinusoidal cosine carrier function that modulates the Gaussian kernel. Σ is the covariance matrix, which determines the Gaussian kernel's shape, spread, and orientation.

$$\mathbf{s}\theta_i = i cos\theta_i + j sin\theta_i \tag{3.84}$$

(Eq. 3.84) and θ_i is the rotation angle (with respect to the horizontal x-axis) of the Gaussian envelope.

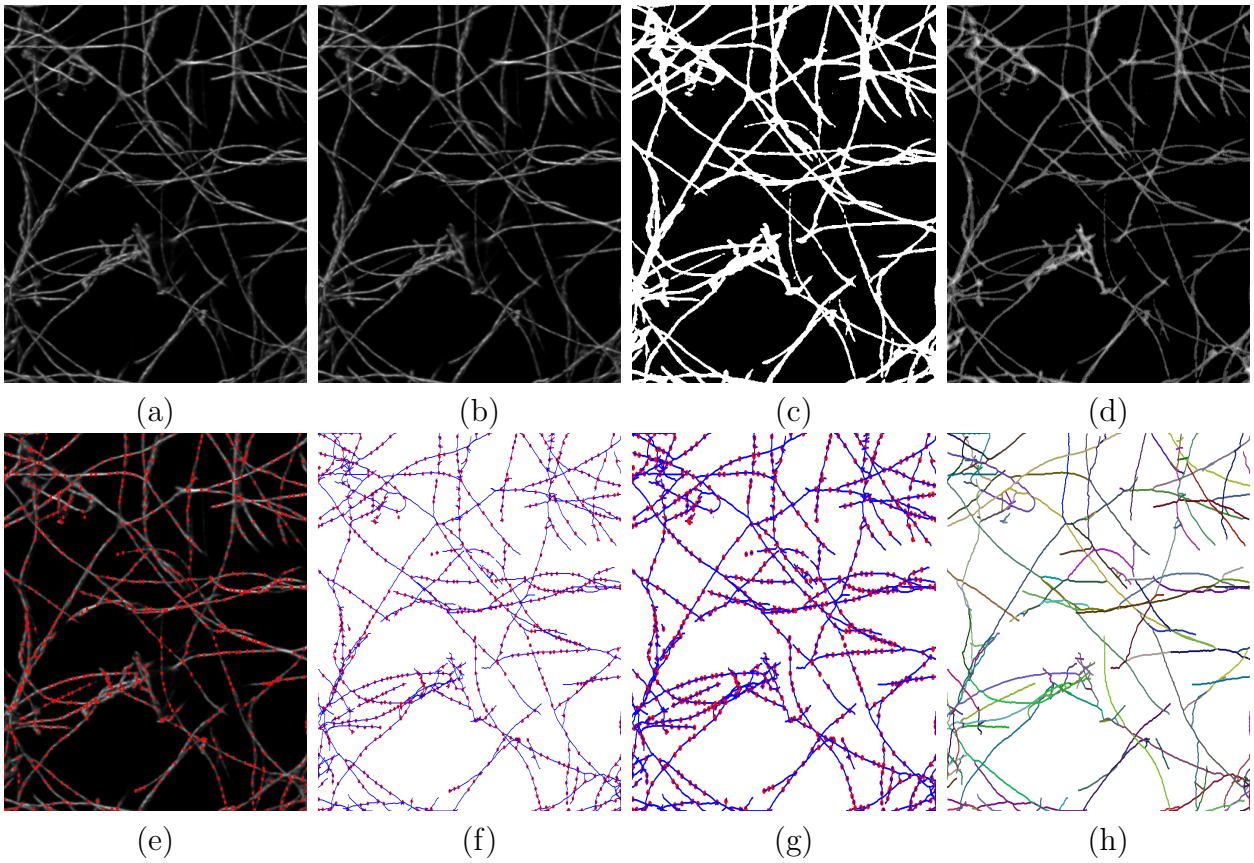


Figure 3.41: CT Fire algorithm applied on the FC image (Fig. 3.14a). (a) FC input image. (b) Curvelet Transform image filtering. (c) Binarisation of the filtered image. (d) Distance transform applied on (c). (e) Nucleation points are identified. (f) Nucleation points are extended. (g) The extension of fibres stops. (h) Final fibres tracing of CT Fire.

Firstly, the algorithm filters the image with the Gabor filter as such (Eq. 3.85).

$$I'^{k}(x,y) = (I * \mathbb{K}_{\text{Gab}})(x,y) \tag{3.85}$$

Here $I^k(x,y)$ is the image intensity at pixel (x,y), I'^k is the filtered image representing the convolution result at pixel (x,y) with the Gabor kernel. $*$ is the convolution operator. This smooths the image and enhances fibre features.

A set of 60 Gabor kernels are applied, each designed with specific parameters that correspond to different orientations and fibre thicknesses.

The fibre orientation is represented by θ_n, this is the angle at which the Gabor kernel is applied, this is used to detect fibres at different angles.

$$\theta_n = \frac{n\pi}{20}, \quad \text{where } n \in \{0,1,2,\ldots,19\} \tag{3.86}$$

(Eq. 3.86) creates 20 different orientations which are evenly distributed from 0 to π

radians, therefore covering a range of angles, allowing the Gabor filters to detect the fibres regardless of the direction the fibres are aligned in.

The thickness of fibres is represented as λ_n which corresponds to the wavelength in pixels of the cosine function in the Gabor filter. The wavelength determines the size of the sinusoidal oscillation within the kernel and is related to the frequency of the detected fibres. The thinnest fibres are detected when $\lambda_n = 6$ pixels, medium thickness fibres when $\lambda_n = 8$ pixels and, the thickest fibres when $\lambda_n = 10$ pixels. These values correspond to different fibre thicknesses, enabling the detection of fibres of varying thicknesses in the image.

The phase parameter ϕ controls the spatial shift of the sinusoidal oscillation, here it is set to 0 which indicates that the filter captures symmetric features in the image, as the cosine wave will start at its maximum value and simplifies the detection of features such as fibres.

Spatial support to the kernel is given by $\sigma_x = 5$ pixels along the x-axis allowing the detection of features that are longer in this direction and $\sigma_y = 3$ pixels along the y-axis which makes the filter anisotropic which is directionally dependent. Anisotropy is extremely suited to detect fibres that have an elongated shape, as they are longer in one direction than in the other.

Once the image is filtered with the set of 60 Gabor kernels, the goal is to retain the best response at each pixel position (x, y), which is the kernel that matches the local structure at each pixel. At each pixel, the best match from the entire Gabor kernel set is kept. The Gabor kernel and its parameters are retained at each pixel location.

$$M_G(x, y) = max_k(I'_k(x, y)) \tag{3.87}$$

In (Eq. 3.87), M_G is the maximum Gabor filtered image and max_k is the maximum kernel co-efficient. The Gabor filtering on the FC image (Fig. 3.43a) is illustrated in (Fig. 3.43b)

The pixel intensity of the detected fibres of M_G relate to the best responding Gabor filter, the parameters can be linked to the physical attributes such as the local fibre orientations and the fibre thickness.

The image is binarised using hysteresis thresholding, where the fibres (foreground pixels) are given the value 1 while the background pixels are given the value 0 (Fig. 3.43c). After which skeletonisation (medial axis) of the foreground pixels is conducted. Skeletonisation is useful as it reduces the amount of data to a representation that holds information of the shape and preserves the connectivity of the original shape. The skeletonisation removes border points of the fibres but preserves the end points of the line segments until no more thinning can be applied (Fig. 3.43d).

This produces a network graph with a collection of nodes which are the fibre ends. Gabor-based features can be combined with the skeleton network graph to reveal details of fibre density and local fibre orientation.

Since the skeletonisation does not manage to represent fibres as required. Post processing is ensued to reconstruct the missing fibres, that were lost due to previous morphological operations and noise in the images.

Potential fibre ends near the fibre extremities along the local orientation of the current pixel p_k are identified. These fibre ends are identified by setting a predefined radius r where

the search will happen and a cone sector (Eq. 3.88).

$$\delta_c \in (0, \pi/2) \tag{3.88}$$

The cone sector is the directional constraint, which is applied within the radial disk. The cone is centred around the local Gabor orientation $\theta_n \in (0, \pi)$.

The selected set of pixels can be defined as (Eq. 3.89).

$$S_p = \left\{ (p_x, p_y) : p_x^2 + p_y^2 \leq r^2,\ \theta_i - \frac{\delta_c}{2} \leq \arctan\left(\frac{p_y}{p_x}\right) \leq \theta_i + \frac{\delta_c}{2} \right\} \tag{3.89}$$

(p_x, p_y) are the relative co-ordinates of pixels in the two-dimensional Cartesian system that considers that $(p_x(p_k), p_y(p_k)) = (0, 0)$.

To make the search of pixels more precise in (S_p), further pixels are filtered. For the pixels that correspond to degree of 1 node (other fibre ends), they will only be considered eligible if they satisfy the condition $p_k \in S_{\text{sym}}$.

The set S_{sym} contains pixels that are symmetric around local orientation θ_p of the candidate pixels. Thus, the condition is (Eq. 3.90).

$$S_{\text{sym}} = \left\{ (p_x^s, p_y^s) : \theta_p - \frac{\delta_c}{2} - \pi \leq \arctan\left(\frac{p_y^s}{p_x^s}\right) \leq \theta_p + \frac{\delta_c}{2} - \pi \right\} \tag{3.90}$$

Only candidate pixels whose orientation is aligned symmetrically with p_k in the direction of θ_p are kept.

To trace the edges, candidate pixels that run along the fibres and are not end points $p_k \in S_{\text{ort}}$ are selected if they are perpendicular to the local orientation θ_p as such (Eq. 3.91).

$$S_{\text{ort}} = \left\{ (p_x^o, p_y^o) : \theta_p - \frac{\delta_c}{2} \pm \frac{\pi}{2} \leq \arctan\left(\frac{p_y^o}{p_x^o}\right) \leq \theta_p + \frac{\delta_c}{2} \pm \frac{\pi}{2} \right\} \tag{3.91}$$

After narrowing down the candidates, the aim is to reconnect the missing fibres. This involves finding the best possible reconnection path between p_k and the selected candidates.

Dijkstra's algorithm is used to find the shortest path, the weights used are based on the image intensity at each pixel, where those pixels were defined as fibres in the original image as their pixels were higher than a certain threshold but might have not been captured during skeletonisation. Amongst all the candidate paths, the path with the minimal weighted length is chosen.

The selected path is then recomputed using the maximum Gabor filtered image M_G as a guideline for the final re-connection this is illustrated in (Fig. 3.43e).

The algorithm affiliates graph networks with fibre morphological skeletons, followed by fibre pruning and post-processing fibre re-connection. Two graphs are used to capture fibre properties: one representing fibre crosslinks or ends, and a skeleton map. These representations capture geometrical and topological features from the confocal images, connecting specific Gabor (example. Orientation) and graph (example. fibre length) characteristics to physical fibre properties. The final fibre graph can be seen in (Fig. 3.43f).

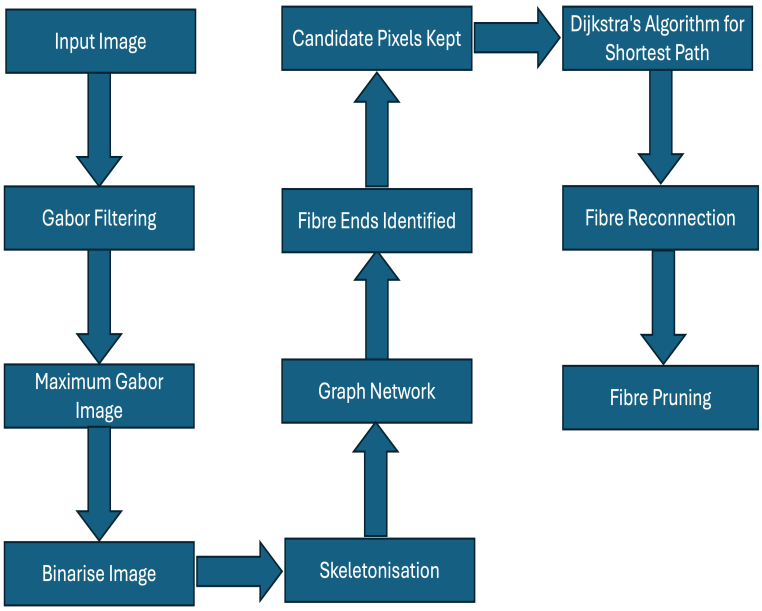


Figure 3.42: Graph Based pipeline follows: The input image is smoothed with a Gabor filter where the maximum Gabor filtered image is extracted. The image is then binarised and skeletonised. A graph network of the fibres is created, the fibre ends are identified, and the candidate pixels are kept. Dijkstra's algorithm is applied to find the shortest path, and the fibres are reconnected. Finally, the fibres are pruned to clean up the delineation.

U-Net:

U-Net [334] is a deep fully convolutional network extensively used in biomedical image segmentation. Its architecture consists of a contracting path which captures detailed features through a series of convolutions, and a symmetric expanding path, which ensures precise localisation. The segmentation map is created by utilising skip connections that link the layers of the contracting and expanding paths. The network's structure is symmetrical to the letter "U", from which it derives its name. U-Net requires minimal training data to perform well. The U-Net used in this thesis (Fig. 3.44) has a depth of 3, and compromises 46 layers, including encoder, bridge, decoder, convolution, softmax and segmentation layers, with skip connections connecting the contracting and expanding paths.

The two main components of the U-Net architecture are the contracting (encoder) and the expanding (decoder) paths.

Encoder Path:

The encoder path is designed to capture contextual information and reduce the spatial resolution of the input image. It is comprised of a series of convolutional layers followed by a pooling operation. The encoder path is like feed forward layers in other convolutional neural networks.

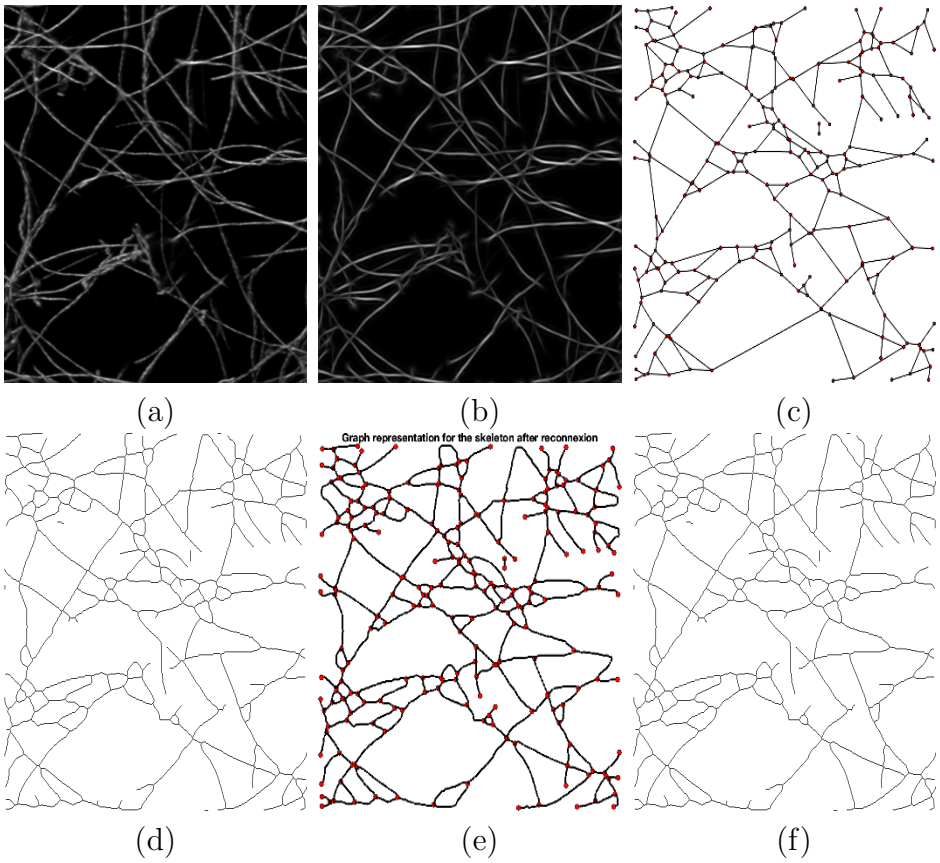


Figure 3.43: Graph Based algorithm applied on the FC image (Fig. 3.14a). (a) fluorescently labelled collagen image. (b) Curvelet transform application on image (a). (c) Binarisation of (b). (d) Skeletonisation of (c). (e) Final reconnection of the selected path from Dijkstra's algorithm. (f) Final fibre graph which acts as the fibre trace.

The contracting path utilises 3×3 convolutional layers with ReLU activation. The convolutions are responsible for extracting more complex data the deeper they are in the network.

Pooling, which is another important feature of the contracting path occurs after every convolution block. In U-Net a 2×2 max-pooling operation is applied. The max-pooling reduced the spatial resolution of the feature maps. The downsampling assists the network in capturing contextual information. The max-pooling operation doubles the number of feature channels, which allows the network to capture more complex information. Feature channels represent the various characteristics or features of an image. Each layer represents a different extracted feature or information from the image, for example, a feature map could look at just the edges and another feature map could look at circular objects. In the contracting path, after each convolutional block and max-pooling layer the image halves in size and the features channels double in size.

The convolutional layers use a stride of 1. Stride is the step size that the convolutional

or pooling layers take when they are moving across the input feature map. Stride indicates how the filter shifts in each operation, stride can affect the spatial dimensions of the output feature maps at each layer. A stride of 1 means that the filter kernel moves one pixel at a time across the feature map. In the pooling layers the filters use a stride of two which reduces the spatial dimensions of the feature maps by half in the downsampling and double its size in the upsampling. In U-Net zero padding is applied. Padding is the process of adding extra pixels around the edges of an input feature map before applying convolution. Padding's main purpose is to control the spatial dimensions of the output feature maps, and it affects how much information is retained at the borders of images. In zero padding extra pixels are added around the border so that the output feature map could have the same spatial dimension as the input image. This is important as the U-Net retains the spatial resolution across the layers to localise features for the segmentation task. For example, for a 5×5 input with a 3×3 filter, a row, and a column of padding around the pixels will yield a 5×5 output. For same padding, the height and width of the input image must be multiples of $2^{\text{EncoderDepth}}$. Padding is applied in the convolution layers of the U-Net.

Activation function:

Rectified Linear Unit (ReLU) is one of the most widely used activation functions in deep learning. It allows the network to be non-linear, enabling it to model complex patterns and relationships in the data. ReLU can be defined as (Eq. 3.92).

$$f(x) = max(0, x) \tag{3.92}$$

where the ReLu outputs x if $x > 0$ and outputs 0 if $x \leq 0$. ReLU helps avoid vanishing gradient problems as it does not confine the value into a small range. ReLU affects the weight updates of the network by influencing gradients during backpropagation. ReLU's condition makes it sparse, therefore neurons in the network will output zero for certain inputs, which in turn will turn off some neurons in each layer. When neurons are turned off, they do not contribute to the gradient calculation thus reducing the amount of weight updates in the network. This sparsity makes the network more efficient by reducing the complexity of the learned observations. Without the linearity of ReLU the model would only be able to model linear relationships regardless of depth.

Decoder Path:

The decoder expanding path is symmetrical to the contracting path and is responsible for creating the segmentation map. It focuses on understanding where the objects to segment are, by localising the objects or structures in the image. The expanding path has convolution layers of the same size as the contracting path, 3×3 convolutions followed by ReLU activation to refine the upsampled features. The convolutional blocks in the expanding path refine the enlarged feature maps by learning new, more detailed representations at higher resolutions. The decoder levels could at times introduce artefacts and unwanted patterns in the upsampled feature maps, the convolution blocks smooth out the artefacts by learning to

filter and refine the feature maps after every upsampling step. Instead of pooling the decoder uses upsampling to increase the spatial resolution, which reverses the effect of pooling.

The skip connections are U-Net's main innovation between the encoder and decoder layers. The connections transfer the feature maps from the contracting path to the expanding path. Thus, combining the higher resolution features from the encoder and the upsampled feature outputs of the decoder, the network can produce more detailed segmentations. The skip connections assist the U-Net to recover the fine details that could be lost due to the downsampling in the encoder levels.

At the end of the expanding path, the final 1×1 convolution is applied to reduce the number of feature maps to the number of output classes. For this thesis's purpose, the number of output classes is 2, as binary segmentation is conducted (foreground and background). The output of the convolution layer is passed to a softmax activation function that produces a probability map, which gives every pixel in the image, a probability of belonging to a particular class.

Classification:

The softmax activation function computes the probability of a pixel belonging to a particular class. Given an output of classification vector in this case 2 classes (Eq. 3.93).

$$\mathbf{z} = \mathbf{z}_1, \mathbf{z}_2 \tag{3.93}$$

The softmax function ensures that the probabilities add up to 1. The softmax function computes the probability $\hat{L}_x$ for each pixel as follows (Eq. 3.94).

$$\hat{L}_x, k = \frac{e^{\mathbf{z}_x, k}}{\sum_{j=1}^{n} e^{\mathbf{z}_x, k}} \tag{3.94}$$

Here $\mathbf{z}_x, n$ are the raw scores for each class k at pixel x and n is the number of classes. The denominator of the fraction $e^{\mathbf{z}_x, k}$ sums the exponentiated logits across all classes, normalising the values.

As a quick example it can be assumed $\mathbf{z} = [2.0, 1.0]$ $e_1^{\mathbf{z}} = e^2.0 \approx 7.39, e_2^{\mathbf{z}} = e^1.0 \approx 2.72$

summing them together would give $7.39+2.72 = 10.11$. Calculating each class probability would compute the following

$$\hat{L}_1 = \frac{7.39}{10.11} \approx 0.73, \hat{L}_2 = \frac{2.72}{10.11} \approx 0.27$$

This means that Class 1 has a probability of 73% while Class 2 has a probability of 27%. Therefore, this pixel will be assigned Class 1.

At the start of training each filter with a kernel size $\mathbb{K} = 3$ is initialised with random values, it is a small matrix of weights. The weights are initialised randomly based on specific initialisation techniques. In the original U-Net manuscript no specific initialisation technique is used, but modern implementation of the U-Net, the He initialisation is commonly used for convolutional layers. The He initialisation is particularly useful when ReLU activation

functions are used. As ReLU only activates where positive values are, part of the weights in each layer do not contribute to the output, which reduces the number of active neurons. The He initialisation is created to keep the variance of the activations consistent across layers in networks, which is a solution to issues such as vanishing and exploding gradients. In He initialisation, the weights are drawn from a normal distribution with a mean of 0 and a standard deviation of (Eq. 3.95).

$$\text{std} = \sqrt{\frac{2}{\mathbb{K} \times \text{Input Channels}}} \tag{3.95}$$

Similarly, the initialisation can be bounded by a limit (Eq. 3.96),

$$\text{limit} = \sqrt{\frac{6}{\mathbb{K} \times \text{Input Channels}}} \tag{3.96}$$

The different filters mean that each filter learns a particular feature, initially low-level patterns such as edges are learned, but as the network progresses the learnable features are refined. As the image passes through the convolutional layers, the filters slide across the image and perform element-wise multiplications between the filter values and image pixels producing the feature maps. The network makes predictions based on the processed feature which are then compared to the ground truth labels and a measurement known as "loss" is calculated. Loss is a measure of how incorrect the predictions are.

For the network to learn from the incorrect prediction, backpropagation is used to calculate the gradient of this loss with respect to each weight in the filters. The gradients indicate the magnitude and direction by which each filter weight should be adjusted to minimise the loss.

The image in U-Net passes through the encoder, and then to the bottleneck, where there is a condensed high-level representation of the image. After the bottleneck it passes through the decoder combined with the skip connections. Finally, the output layer produces a segmentation map that represents the probability for each pixel belonging to the foreground or background for the purpose of this thesis.

Loss:

For segmentation tasks where each pixel is classified into two classes (foreground and background), binary cross-entropy loss calculates how far the prediction was from the actual truth from the pixel label. Calculating the loss of each pixel individually and averaging over all pixels. The loss can be defined in the following way (Eq. 3.97).

$$J = -\frac{1}{n}\sum_{x=1}^{n}\Big[L_x \log(\hat{L}_x) + (1 - L_x)\log(1 - \hat{L}_n)\Big] \tag{3.97}$$

n is the number of pixels in the image; L_x is the true label of pixel x where 1 is for the foreground and 0 is for the background. $\hat{L}_n$ is the predicted probability that pixel x belongs to the foreground.

The goal of the backpropagation is to calculate how much each weight in U-Net contributes to the binary cross-entropy loss and update these weights to reduce the loss. The backward pass of the network training computes gradients with respect to each parameter (weights and biases) in the network. The gradient is first calculated at the output layer and moves backwards hence the name. The gradient of the binary cross-entropy loss at the output layer is computed with respect to the output probabilities $\hat{L}_x$ for each pixel x. The gradient can be defined as (Eq. 3.98).

$$\frac{\partial J}{\partial \hat{L}_x} = -\frac{L_x}{\hat{y}_i} + \frac{1 - L_x}{1 - \hat{L}_x} \tag{3.98}$$

The gradient indicates how much the loss would change if the prediction for each pixel were adjusted. For example, if $\hat{L}_x$ is far from the true label L_x, the gradient will be large, indicating a large error that needs correcting. If $\hat{L}_x$ is close to the true label L_x, the gradient will be small, which will only require a minor adjustment. In backpropagation, the gradient is passed backward through the network. Each layer in the U-Net contributes to producing $\hat{L}_x$ at the output, so backpropagation distributes the gradient through all the network's layers. Adjustments to the weights are necessary to minimise the overall loss and improve the model's segmentation accuracy.

The adjustment is applied using an optimisation algorithm. For this thesis, the "Adam" optimiser is used. As the training advances, the filter weights are continuously updated which allows each filter to become specialised in detecting particular features.

The Adam optimiser is based on adaptive moment estimation, which gives it its name. The Adam optimiser is one of the most popular optimisation algorithms used in training deep neural networks. Adam modifies the learning rate of each parameter individually based on the estimation of both the first moment (mean) and the second moment (variance) of the gradients, which makes it particularly useful in handling sparse gradients and improving convergence rates.

The first moment (mean) μ of the gradients helps to estimate the direction of the gradients, the second moment (variance) σ of the gradients helps to estimate the magnitude of the gradients. The averages are updated regularly at each time step t using exponential decay rates which are controlled by two hyperparameters β_1 and β_2. The parameters of the model which are the weights and biases are updated based on these two estimates, this leads to adaptive step sizes where the learning rate is adjusted based on the direction and the scale of past gradients.

The first moment estimate can be computed as (Eq. 3.99).

$$\mu_t = \beta_1 \mu_t - 1 + (1 - \beta_1) g_t \tag{3.99}$$

g_t is the gradient of the cost function with respect to the parameter at time step t

The second biased moment estimate (variance) can be computed as such (Eq. 3.100).

$$\sigma_t = \beta_2 \sigma_t - 1 + (1 - \beta_2) g_t \tag{3.100}$$

Bias correction is also initialised to account for the initialisation bias at the start of training. At the start of training, the moving averages, the mean and variance of the gradients are initialised at zero. This can introduce a bias, especially in the early iterations. To correct the initialisation bias, a bias correction factor to each moving average is applied. This correction ensures that the moving averages are unbiased estimates of the first and second moments.

For the mean and variance, the bias correction could be represented as such (Eq. 3.101).

$$\hat{\mu}_t = \frac{\mu_t}{1-\beta_1^t}, \quad \hat{\sigma}_t = \frac{\sigma_t}{1-\beta_2^t} \tag{3.101}$$

The parameter which are the weights of the network can be updated as such (Eq. 3.102).

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{\hat{\sigma}_t} + \epsilon}\hat{\mu}_t \tag{3.102}$$

α is the learning rate which is usually initialised at 0.001 and ϵ is a small constant with a value of 10^-8 that is added to prevent division by zero. The decay rates β_1 and β_2 are usually initialised to 0.9 and 0.999 respectively and they control the decay rates for the moving averages of the gradient.

The updated parameter w_{t+1} is calculated by modifying w_t in the direction that minimises the loss, the incorrect predictions. It achieves this by using the corrected gradient estimate and adjusting the step size.

The advantages of using Adam as an optimiser is in its ability to adjust the learning rate for each parameter individually which makes it robust for models with high variance in gradient magnitudes. It is particularly useful for computer vision tasks as it works well even when gradients are sparse. The downside of Adam is that at times it converges to suboptimal solutions if the learning rate is not finely tuned, where it could get stuck on the local minima on highly non-convex surfaces.

U-Net Training:

For the purpose of this thesis, the U-Net trained had a depth of 3, where there was 3 encoder levels and 3 decoder levels (Fig. 3.44).

Each image from (Fig. 3.14a-d) was split into halves for training and testing. BCB Collagen image (Fig. 3.14c) was complemented to highlight fibre features in white, and background in black. After splitting the images into halves, the top half was used for training, patches of [32 32] pixels were used to create the patches of the images. Similarly for the labels the ground truth was loaded and split into halves and then the training half for each image was split into patches of [32 32] pixels. To increase the number of images for training, augmentation was employed. Augmentation techniques (Fig. 3.45) included horizontal and vertical flipping, rotation by 90 degrees and addition of Gaussian noise, this was also done on the ground truth labels to match the images except the Gaussian noise. A total of 8910 images of size 32x32 were used for training.

Model	Learning Rate	Batch Size	Training Images	Testing Images	Accuracy
1	0.0001	16	8910	1782	61.4%
2	0.001	16	8910	1782	53.4%
3	0.0001	32	8910	1782	59.4%
4	0.001	32	8910	1782	61.3%

Table 3.2: U-Net Model Training. 4 models with varying learning rates and batch size were tested. The best performing model is model 1 with a learning rate of 0.0001 and a batch size of 16. This model was used for the rest of this thesis.

Initially, various training parameters were tested, including batch size (16 and 32), and learning rate (0.0001 and 0.001) with 'Adam' as the optimiser, to optimise training parameters and select the best-performing network. Models were trained with 8910 images. The best performing model of the U-Net was trained with a batch size of 16, learning rate of 0.0001 and the "Adam" optimisation algorithm, for 15 epochs as shown in Table 3.2 when tested on 1782 image patches. Training duration was 15 minutes. One of the limitations of this set up is lack of a validation set which could lead to data leakage and bias. This was the final network used for analysis in the rest of this thesis. The patch generation for U-Net is outlined in `github.com/youssefarafat/Trace_Ridges/blob/main/UNET_PatchMaking_general.m` (Fig. 3.46a), training `github.com/youssefarafat/Trace_Ridges/blob/main/UNET_Training_general.m` (Fig. 3.46b) and testing `github.com/youssefarafat/Trace_Ridges/blob/main/UNET_testing_general.m` (Fig. 3.46c).

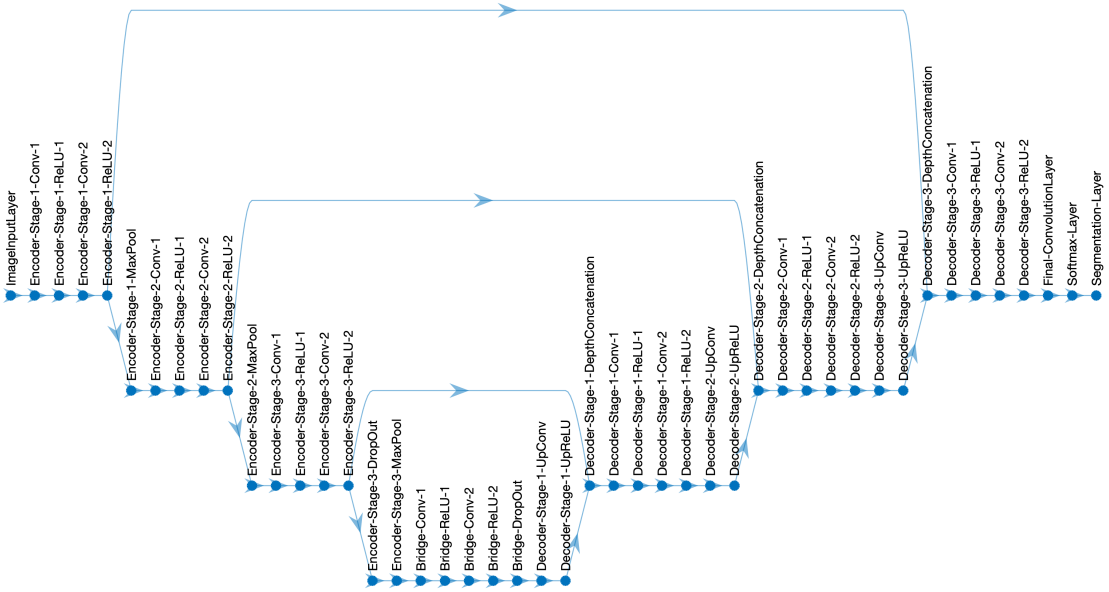


Figure 3.44: Illustration of the U-Net architecture used to segment the fibres. The architecture consists of three encoder levels and 3 decoder levels. Image input layer followed by encoder, bridge, decoder, final convolution, softmax and segmentation layers.

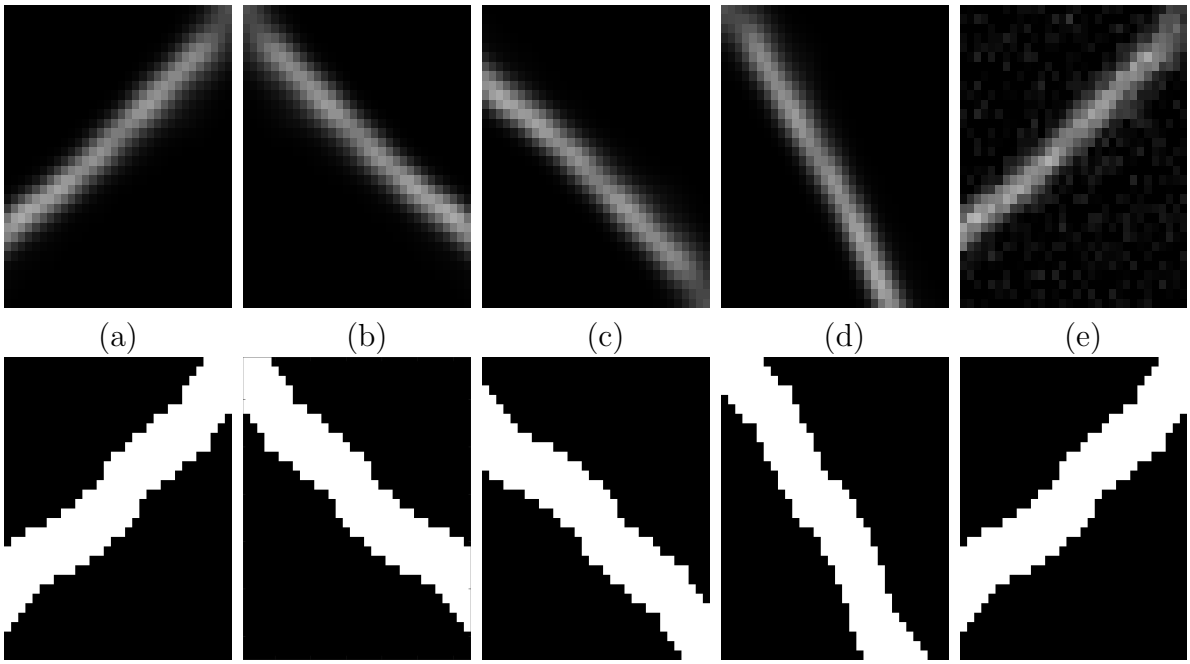


Figure 3.45: Illustration of the data augmentation methods used to train the U-Net. (a) Original patch (b) Horizontally flipped (c) Vertically flipped (d) Rotation by 90 degrees (e) Gaussian blurred, with zero mean and a standard deviation of 0.05. (f-j) labels of the augmented patches. Label (j) is the same as (f) as the Gaussian blur does not affect the label.

Code Blame 290 lines (229 loc) · 9.95 KB Code 55% faster with GitHub Copilot

```
% Define image directory and filename
imageDir = '/MATLAB/UNET/';
imageName = 'A+_N.tif';
imagePath = fullfile(imageDir, imageName);

% Read the image
image = imread(imagePath);
%image = rgb2gray(image); %colour to gray image
% Define patch size
patchSize = [32, 32];

% Create folder to save patches
saveDir = fullfile(imageDir, 'AN_patches');
mkdir(saveDir);

% Extract and save patches
patchIndex = 1;
for r = 1:patchSize(1):size(image, 1)-patchSize(1)+1
    for c = 1:patchSize(2):size(image, 2)-patchSize(2)+1
        % Print loop bounds for debugging
        fprintf('r: %d - %d, c: %d - %d\n', r, r+patchSize(1)-1, c, c+patchSize(2)-1);

        % Calculate end row and column indices for the patch
        endRow = r + patchSize(1) - 1;
        endCol = c + patchSize(2) - 1;

        % Extract patch
        patch = image(r:endRow, c:endCol, :);

```

(a)

Code Blame 28 lines (25 loc) · 783 Bytes Code 55% faster with GitHub Copilot

```
imageSize = [32, 32];
dataDir = '\MATLAB\UNET\trainImages\';
labelDir = '\MATLAB\UNET\trainLabels\';

%%
class_Names = ["BACK","FIB"];

labelIDs = (1:numel(class_Names));
imds = imageDatastore(dataDir);
pxds = pixelLabelDatastore(labelDir,class_Names,labelIDs);

pximds = pixelLabelImageDatastore(imds,pxds)
%%
encoderDepth = 3;
numClasses = numel(class_Names);
[lgraph, OutputSize] = unetLayers(imageSize, numClasses, 'EncoderDepth', encoderDepth);
%%
options = trainingOptions('adam', ...
    'InitialLearnRate', 0.0001, ...
    'MaxEpochs', 15, ...
    'VerboseFrequency', 5, ...
    'MiniBatchSize', 16, ...
    'Shuffle', 'every-epoch', ...
    'Verbose', true);
%%
Unet = trainNetwork(pximds, lgraph, options)
%%
save('UNET_top_Fibre_1.mat', 'Unet');
```

(b)

Code Blame 18 lines (15 loc) · 574 Bytes Code 55% faster with GitHub Copilot

```
classNames = ["BACK", "FIB"];
labelIDs = (1:numel(classNames));

% Create image and label datastores for testing
imdsTest = imageDatastore("testingImages_Dilated");

pxdsTest = pixelLabelDatastore("testingLabels_Dilated", classNames, labelIDs);
%%
load('UNET_Dilated_top_Fibre_4.mat');
%%
predictedLabels = semanticseg(imdsTest, Unet, ...
    'MiniBatchSize', 16, ...
    'Verbose', false);
%%
metrics = evaluateSemanticSegmentation(predictedLabels, pxdsTest);

% Display overall dataset-level metrics (accuracy, mean IoU, etc.)
disp(metrics.DataSetMetrics)
```

(c)

Figure 3.46: Illustration of the workflow of training and testing the U-Net models available on GitHub. (a) U-Net Patch making and dataset splitting. (b) An example of the training workflow of one of the U-Net models. (c) Workflow for testing a U-Net model on unseen data.

3.6 Classification Approaches:

Classification [338] is a machine learning task, where a set of features or image belonging to a sample is labelled based on its category or "class". A classification model is then trained on a set of samples to recognise and learn patterns of the categories or classes. This trained model can then be used to correctly identify unseen samples. Classification [339] can be binary with two categories or multi class with three or more classes. Classic machine learning models include Decision Trees, Naive Bayes, Support Vector Machine, Logistic Regression, and K Nearest Neighbours. Ensemble methods include Random Forest as they combine many decision trees. Deep learning models include Multi-Layer Perceptron, VGG-16, and Vision Transformers. Some models such as the VGG16 and Vision Transformers require the input to be in the form of images rather than a set of features. In this thesis, classification methods were trained on extracted morphological features of extracellular matrix fibres to build a pipeline to classify images based on their extracted features.

3.6.1 VGG-16:

VGG-16 [340] is a deep convolutional neural network that consists of 16 weight layers for classification tasks. Like U-Net it employs multiple convolutional layers stacked together and uses $\mathbb{K} = 3 \times 3$ filters to capture the spatial information while maintaining a manageable number of parameters. Each convolution layer is followed by a ReLU activation function to ensure non-linearity. After several convolutional layers, VGG-16 also includes 2×2 max pooling layers with a stride of two, which downsample the feature maps to reduce their spatial dimension to maintain the relevant information. At the end of the network, VGG-16 has three fully connected layers, but first the output feature map must be flattened into a vector of size 25,088.

Flattening is the process which transforms the output from the convolutional and pooling layers which are in a multi-dimensional format, to a single dimensional vector. Convolutional and pooling layer output feature maps which are three-dimensional $N_r \times N_c \times 3$. The fully connected layer expects a one-dimensional input. For example, if the final pooling layer output has dimensions of $7 \times 7 \times 512$ and by multiplying these values with each other $7 \times 7 \times 512 = 25,088$.

The fully connected layers in VGG-16 act similarly to neurons in traditional neural networks to make the final high-level predictions. They transform the feature maps into a format suitable for classification by assigning class probabilities. They have dense connections where each neuron in a fully connected layer is connected to every neuron in the previous layer, this allows the network to learn complex combinations of high-level features.

The first two fully connected layers have 4096 neurons each, while the final layer has 1000 neurons as this model was designed for the ImageNet Large Scale Visual Recognition Challenge, where the database contained 1000 neurons. For this thesis, 2 neurons are available in the last fully connected layer. After the fully connected layer, there is a softmax layer that converts the outputs (logits) into class probabilities, allowing the model to assign a probability to each class.

The loss function used in VGG-16 for multi-class classification is the categorical cross-entropy loss. The loss function measures how well the predicted class probabilities match the true class labels and this can be defined in the following manner (Eq. 3.103).

$$\text{Loss} = -\sum_{x=1}^{n} L_x \log(\hat{L}_x) \tag{3.103}$$

L_x is the true label, where the value is 1 if the class is correct and 0 otherwise. $\hat{L}_x$ is the predicted probability of class x. n is the number of classes, for the purpose of the work done in this thesis it is 2 classes (Normal, Tumour).

The loss function provides information to the network to update weights during back propagation as done similarly in U-Net. This aims to minimise the difference between the predicted and true labels by iteratively adjusting the network parameters.

Stochastic gradient descent momentum (sgdm) is the optimiser used with VGG16 in this thesis. sgdm is an update of sgd where the gradient descent updates the parameter in the direction of the negative gradient to minimise the lose function. sgdm introduces a momentum term to smooth the updates by keeping information on the previous velocity. This assists the update to be influenced by past gradients rather than relying on the current gradient. This leads to faster convergence and less oscillation by smoothing them.

For an input vector (Eq. 3.93) $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_n]$,
the softmax function computes the probability $\hat{L}_x$ for each pixel as follows, (Eq. 3.104).

$$\hat{L}_x = \frac{e^{\mathbf{z}_x}}{\sum_{j=1}^{n} e^{\mathbf{z}_j}} \tag{3.104}$$

Here $\mathbf{z}_x$ and $\mathbf{z}_j$ are the raw scores for each class and n is the number of classes.

Softmax probabilities are combined with the categorical cross-entropy loss during training to compute how far the predicted probabilities are from the actual class labels which guide the network's weight updates. As the image size is $(512 \times 512 \times 3)$ after every max pooling layer the image size $N_r \times N_c$ gets halved while in the first convolution layer the third channel becomes 64 feature channel and after every max pooling layer it gets doubled. The architecture of VGG-16 can be seen in (Fig. 3.47).

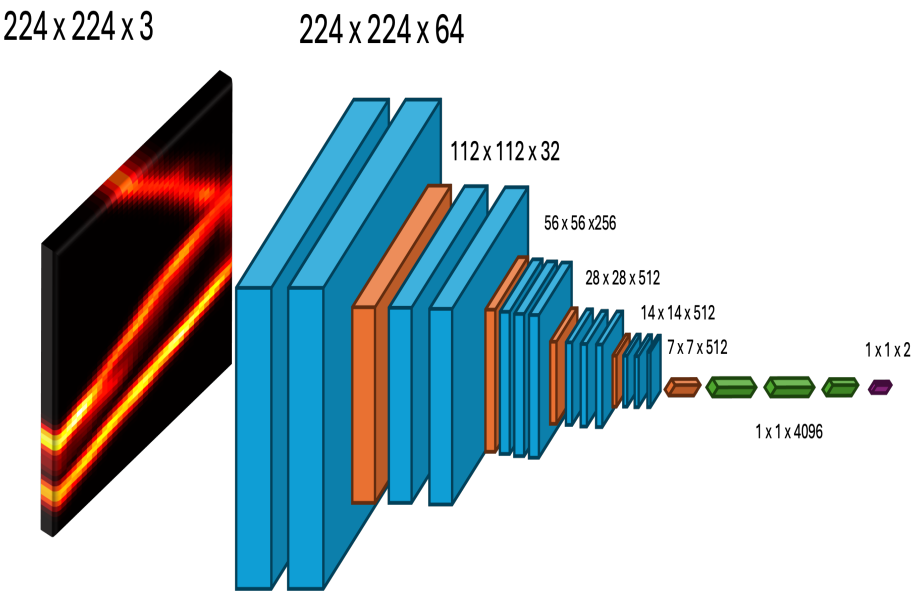


Figure 3.47: Architecture of VGG16. Convolution + ReLu blocks are in blue, max pooling in orange, fully connected layers + ReLu in green and the Softmax layer in purple.

3.6.2 Tiny Vision Transformer (ViT):

The tiny Vision Transformer [341] is a smaller efficient version of the widely used Vision Transformer, designed for image processing tasks. Transformers were originally applied to natural language processing tasks [342] but lately they have shown improved results in image classification tasks. In ViT images are split into tokens or patches of 16×16, to be treated as tokens in natural language processing. These tokens are the input to transformer blocks that learn high level dependencies.

The tiny sized model has 5.7 million parameters. As mentioned, the first task of the Vision Transformer is to split the input images into patches of 16×16, each patch is subsequently flattened and transformed into a fixed linear vector which is mapped into the embedding. This is done similarly to a fully connected network. This is to reduce the dimensionality and sequence length and to prepare the transformer block's input. In addition to this, transformers do not understand positions and spatial structure. A classification token is used in combination with the sequence of patch embeddings. This token carries information across all the patches and represents the image globally as it flows through the transformer. The Vision Transformers use positional encodings to provide information on the spatial positions of the patches; this helps the model understand the spatial relationships in the image.

The transformer includes many encoder layers, with each layer consisting of multi-head-self-attention (MHSA) and MLP

Self-attention captures information from the image on how distant parts in the image relate to each other, this is done by introducing every patch in the image to other patches

in the image, where they interact and learn relationships over long distances. The MHSA, allows the model to learn different relationships between the patches. Each head in the MHSA is an independent self-attention mechanism. Where each head calculated the dot-product attention, which is an operation that measures the degree to which each patch should interact with other patches in the image. It determines this using three vectors, Query, Key and Value. Query is the area in the patch requesting information on other patches. Key represents the information in patches that could be important. Value is the information that will be cascaded and carried forward based on the attention scores. These vectors are produced by multiplying the learned weight matrices to the input. The output of the MHSA is a weighted combination of all the patches, which gives each patch a new representation based on context from other patches.

After every self-attention layer, the MLP network processes each patch embedding individually. The feed forward network processes the input from the self-attention by using two fully connected layers combined with a non-linear activation function GELU. This operation does not alter the patches but refines the representation of each patch. This process helps the model extract the transformations and non-linear relationships of each individual patch.

Between every self-attention and feed-forward sub layers, Layer Normalisation is applied. Layer Normalisation balances the input of the self-attention and the feed forward layer. This helps prevent fluctuations in training. By scaling and shifting the input values of each patch to ensure a mean of zero and variance of 1. This ensures the feature distribution is equalised so that the gradients are stable.

In the final layer the classification token goes through a fully connected layer to produce class probabilities through a softmax activation in the head layer to transform the logits into probabilities. Cross-entropy loss is used as the loss function to assist the updating of weights in backpropagation. The architecture of the Vision Transformer can be seen in (Fig. 3.48).

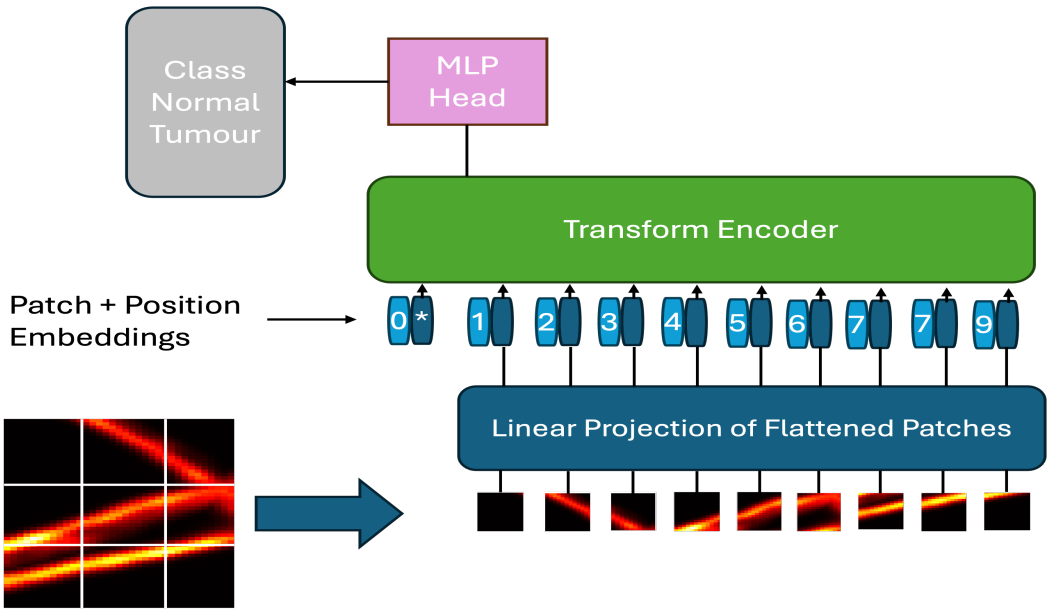


Figure 3.48: Vision Transformer architecture. Patches are extracted and flattened from the input image. They are passed through the linear projection layer, after which the patches and their positional embeddings are matched and pass through the Transformer encoder. The output of the transformer encoder passes through the MLP head and finally the patches are classified.

3.6.3 Decision Trees:

Decision Trees algorithm [343] is a widely used machine learning algorithm. The decision tree has a flowchart-like structure "tree" (Fig. 3.49). Decision trees define a set of rules to split the data features into predefined classes. Each node in the tree is where the rules are applied and a decision on a feature is made. Each branch of the tree is the decision result, defining which data samples have passed through. The leaf nodes are the class labels. Decision trees aim to separate the data into different subsets based on the features. They also aim to make the subsets as pure as possible by minimising the impurity, where each subset contains only data points from a single class.

To choose the best separation of the subsets at each node, impurity is measured using the Gini impurity [344]. Impurity is a measure of homogeneity in the data sample. For example, if we have 3 tanks with stars in them seen in (Fig. 3.50) and a star from each tank is randomly drawn from the tank, if tank 1 has 4 stars of the same colour, tank 2 has 1 star with a different colour, tank 3 had 2 stars of one colour and 2 with another colour. Tank 1 required the least information passage on the colour of the stars in that tank as all the colours of the stars were the same, tank 2 required more information, while tank 3 required the most information. The amount of information passed is an indication of purity. Where tank 1 is the purest and tank 3 is the least pure.

Gini impurity looks at the likelihood of an incorrect classification of a star that was

randomly chosen from a tank if it was labelled according to set of rules of that node. A Gini impurity of 0 is an indication that this node is pure, while if the value is closer to 1 it indicates the presence of impurity.

The Gini impurity can be defined as (Eq. 3.105).

$$Gini(k) = 1 - \sum_{j=1}^{n} P_j^2 \tag{3.105}$$

Here n is the number of classes, for a node k, the Gini impurity looks at sum of the square of the probabilities of each class, which is the proportion of instances of class j in that node. For example, if we have 10 stars in the tank, 6 green stars and 4 blue stars, $\frac{6}{10} = 0.6$, $\frac{4}{10} = 0.4$, the Gini impurity can be formulated as such, $Gini(k) = 1-(0.6^2+0.4^2) = 1-(0.36+0.16) = 0.48$, a Gini score of 0.48 means that the classes are mixed and that it is neither completely pure nor completely impure. The node decides on the split, and 2 branches will come out of the node, which will go to two or more child nodes. At child node 1 where it is assumed that the proportions are 0.8 and 0.2, the Gini impurity can be computed as such $Gini(k) = 1 - (0.8^2 + 0.2^2) = 1 - (0.64 + 0.04) = 0.32$. It can be assumed that child node 2 has the proportions 0.3 and 0.7, $Gini(k) = 1 - (0.3^2 + 0.7^2) = 1 - (0.09 + 0.49) = 0.42$. To check if the separation is sufficient, a weighted average of the Gini impurities is calculated with respect to the child nodes. 60 % of the instances from the parent node went to child node 1 and 40 % of the instance went to child node 2, the weighted average can be computed as such $Gini(split) = 0.6 \times 0.32 + 0.4 \times 0.42 = 0.192 + 0.168 = 0.36$. The Gini split value is compared with the original parent value of 0.48, $0.48 - 0.36 = 0.12$, therefore there was a reduction in the mixing of classes at the node and the rule applied at the parent node was sufficient. Different rules are applied to choose the rule that will reduce the Gini impurity.

Alternatively, to choose the best separation, entropy could be used as the splitting criterion. Entropy is similar to Gini; it measures the disorder or uncertainty in the data. The main aim of entropy in decision trees is to reduce entropy at each split thus achieving a purer subset with less uncertainty. Entropy is outlined in (Eq. 3.106).

$$Entropy(k) = -\sum_{j=1}^{n} log_2(P_j) \tag{3.106}$$

n is the number of classes; P_j is the proportion of elements in class j. $log_2(n)$ represents the maximum disorder, if all the classes are equally distributed. A value of 0 would indicate perfect purity where all the data points belong to a single class.

For example, at child node 1 where it is assumed the proportions are 0.8 for class 1 and 0.2 for class 2 the entropy can be computed as such $Entropy(k) = -(0.8log_2(0.8)+0.2log_2(0.2) = -(-0.258 - 0.464) = 0.722$. Entropy provides a more theoretical measure of impurity in comparison to Gini as it based on information theory.

For the classification task in this thesis, a grid search was conducted for the splitting criterion, maximum depth, minimum samples to split and minimum samples in one leaf. The search was conducted for the following parameters: 'criterion': ['gini', 'entropy'], 'maximum

depth': [3, 5, 10, None], 'minimum samples to split': [2, 5, 10], and 'minimum samples in a leaf': [1, 2, 4, 10]. Additionally, to validate the results k-fold validation was used with 5 folds to ensure the highest performing parameters are used.

When classifying fibre features with Decision Trees, the quality of the split was decided by Entropy, there was a maximum depth of 5until all the leaves had the minimum number of samples to split which was 2. In addition, the minimum number to be at a leaf node was set at 10. While for the classification of the fibres and cells features the criterion was entropy, maximum depth was 3 the minimum number of samples to split was 2 and the minimum number at a given leaf was 10.

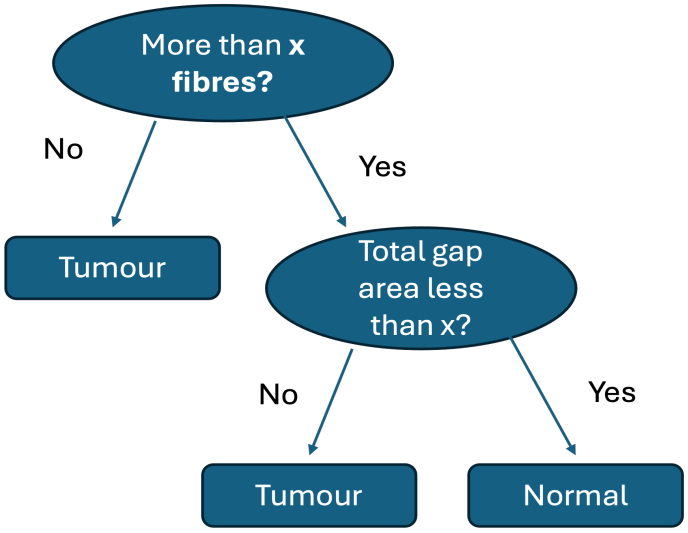


Figure 3.49: Decision Trees representation. A query is first asked to partition the data. Samples that belonged to "Yes" in the first query must be asked another query in order to assign them into the classes. While samples that belonged to "no" were assigned the tumour class.

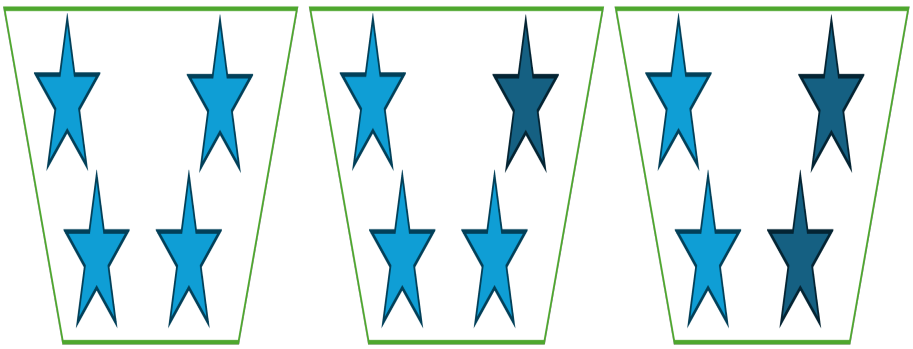


Figure 3.50: Gini Impurity: Three tanks with stars of different colours to show the impurity levels. Tank on the left-hand side has 4 stars of the same colour meaning it is fully pure. Middle tank has 4 stars where 1 star is of a different colour therefore it is less pure than the left-hand side tank. The tank in the right-hand side has 4 stars, 2 stars are light blue, and 2 stars are dark blue therefore the tank is impure.

3.6.4 Naive Bayes:

Naive Bayes methods are supervised learning algorithms that follow the Bayes theorem combined with the assumption of conditional independence where features are independent given the class, in simple terms every pair of features being classified is independent of each other.

The Bayes algorithm can be modelled as such (Eq. 3.107) for class variable L and the feature vector $\mathbf{z_1}, ..., \mathbf{z_n}$.

$$P(L \mid \mathbf{z_1}, \ldots, \mathbf{z_n}) = \frac{P(L)P(\mathbf{z_1}, \ldots, \mathbf{z_n} \mid L)}{P(\mathbf{z_1}, \ldots, \mathbf{z_n})} \tag{3.107}$$

Here $P(L \mid \mathbf{z_1}, \ldots, \mathbf{z_n})$ is the posterior probability which is the probability of class L where there is evidence of $\mathbf{z_n}$.

$P(\mathbf{z_1}, \ldots, \mathbf{z_n} \mid L)$, is the likelihood which is the probability of observing evidence of $\mathbf{z_n}$ given the class L. $P(L)$ is the prior probability, which is the probability of observing class L with no feature and finally, $P(\mathbf{z_1}, \ldots, \mathbf{z_n})$ is the probability of observing feature $\mathbf{z_n}$, this is a constant value when looking at the probabilities of different classes as it only depends on the features regardless of the class, this is usually ignored in Naive Bayes for classification tasks as it would be the same for all class comparisons, this is generally used to normalise the data and ensures that the value of the posterior probabilities sums to 1 in all the classes.

The aim of naive bayes is to find class L that maximises the posterior probability. By using the naive conditional independence assumption where each feature $\mathbf{z_n}$ is independent of other features given class L, mathematically it can be represented as (Eq. 3.108).

$$P(\mathbf{z_j} \mid L, \mathbf{z_1}, \ldots, \mathbf{z}_{j-1}, \mathbf{z}_{j+1}, \ldots, \mathbf{z}_n) = P(\mathbf{z_j} \mid L) \tag{3.108}$$

This simplifies the formula as there is no need to consider the interactions between the features. Given the assumption, the probability of observing the features with class L can

be simplified (Eq. 3.109).

$$P(L \mid \mathbf{z_1}, \ldots, \mathbf{z_n}) = \frac{P(L) \prod_{j=1}^{n} P(\mathbf{z_j} \mid L)}{P(\mathbf{z_1}, \ldots, \mathbf{z_n})} \tag{3.109}$$

$\prod_{j=1}^{n} P(\mathbf{z_j} \mid L)$ is the product of the likelihood of each feature $\mathbf{z_j}$ given a class. since $P(\mathbf{z_1}, \ldots, \mathbf{z_n}$ is the same across all classes it can be ignored when comparing the probabilities of different classes and the formula can be simplified as such (Eq. 3.110).

$$P(L \mid \mathbf{z_1}, \ldots, \mathbf{z_n}) \propto P(L) \prod_{j=1}^{n} P(\mathbf{z_j} \mid L) \tag{3.110}$$

The $\propto$ is "proportional to", which indicates that the goal is to maximise the product $P(L) \prod_{j=1}^{n} P(\mathbf{z_j} \mid L)$.

The algorithm then determines the class L that maximises the product in the following way (Eq. 3.111).

$$\hat{L} = \arg\max_{L} P(L) \prod_{j=1}^{n} P(\mathbf{z_j} \mid L) \tag{3.111}$$

Here $\hat{L}$ is the predicted class, which is determined by finding class L that provides the highest posterior probability.

Therefore, $P(L) \prod_{j=1}^{n} P(\mathbf{z_j} \mid L)$ is the score or value that is calculated for each class, it is a combination of the prior probability $P(L)$ with the likelihood of each feature being observed given class L, this assists in deciding which class L is suitable to makes the observed features most prominent.

A grid search for the smoothing variable was conducted between $1e^{-}9$ to 1. K-fold validation was used with five folds.

In all cases for Naive Bayes, Gaussian Naive Bayes is used. For variance calculations a smoothing variable of 1 is added.

3.6.5 Random Forest:

Random Forest [345] classification is an ensemble machine learning algorithm, that uses a combination of models and combines their prediction to produce a single output, Random Forest uses multiple Decision Trees to formulate the prediction. The purpose of using multiple decision trees is to improve the classification accuracy and control over-fitting.

Random Forest uses bagging, which is a bootstrapped aggregating, this creates several subsets of training data samples by randomly sampling from the original dataset with replacement, which means that each observation may appear more than once or may not appear at all. Here each subset is used to train a different decision tree. When the Decision Trees are being built, a selection of a random subset of features is chosen to split at each node. Randomness helps to improve generalisation, and each tree is not correlated to another tree.

For each bootstrap sample, a decision tree is built, where a subset of random features is used to split at each node, the best split of the selected features is determined based on Gini, as mentioned in Decision Trees. The process is repeated until the stopping criteria is reached such as minimum samples per leaf. After which the prediction is aggregated by passing the test data through each tree. The final prediction for every instance of data is decided by majority voting, the class label L that gets the highest votes from the trees is chosen as the final prediction (Eq. 3.112).

$$\hat{L} = \text{argmax}_L \sum_{j=1}^{n} Ind(\text{Tree}_j(x) = L) \tag{3.112}$$

The summation$\sum_{j=1}^{n}$ looks over all n trees in the Random Forest, for each tree j, it checks the prediction made by that tree for the instance x, $Ind(\text{Tree}_j(x) = L)$ is an indicator function where Ind is 1 if the j-th tree is predicting class L or the instance x, it is 0 otherwise. The argmax looks at which class has the highest count of votes.

A grid search was conducted on the parameters of the random forest to optimise the performance of the model. The parameters were: number of estimators': [50, 100, 200],'maximum depth': [10, 20, None], 'minimum samples to split': [2, 5, 10], 'minimum samples in a leaf': [1, 2, 4], 'maximum features': ['sqrt', 'log2'].

When classifying with features of the fibres using Random Forest, the number of trees in the forest was 200, the quality of the split was measured by Gini. Maximum depth was set at 10 and the number of features to consider before split was decided by 'sqrt' where it selects the $\sqrt{\text{number of total features}}$, the aim of the aim is to introduce diversity among trees as well as improve generalisation and reduce overfitting. Minimum samples on a leaf were set at 1 and minimum samples to split set at 2. This was decided by 5 K-Fold validation.

Similarly, the hyperparameters used when classifying with features of both fibres and cells, the number of trees in the forest was 100. log_2 was used to decide on the number of features chosen, which gives a smaller number of trees. Maximum depth was chosen as 10, the minimum samples on a leaf was set at 2 and minimum samples to split was 2.

3.6.6 Support Vector Machines:

Support Vector Machines (SVMs) [346] are a popular supervised machine learning algorithm that aims to find an optimal boundary that divides data points of different classes. It aims to maximise the margin between the classes. The boundary that separates data points in the dimensional space is the hyperplane. The support vectors are the data points that are close to the hyperplane, these points are important as they determine the hyperplane. The margin is the distance between the data points in each class that are the closest to the hyperplane. The aim of SVM is to maximise the distance between the hyperplane and the closest points. For a training dataset, SVM aims to find a hyperplane that separates the classes with the highest margin possible. For data points that can be separated linearly, SVM will find the hyperplane that separates those data points. While most data points cannot be separated linearly, SVM uses slack variable or the kernel trick to project the data to a higher dimensional space to be linearly separated.

For a training set that has N points (x_j, L_j) where x_j is the feature vector of the j-th point, and L_j is the class label of the j-th point.

The hyperplane can be defined as (Eq. 3.113).

$$\mathbf{w_L} \cdot x - b = 0 \tag{3.113}$$

where $\mathbf{w_L}$ is the weight vector which is normal (perpendicular) to the hyperplane, b is the bias term which shifts the hyperplane.

As mentioned earlier, one of the aims of SVM is to maximise the margin. To maximise the margin, the distance of the hyperplane and the support vectors is maximised, this can be represented as (Eq. 3.114).

$$Margin = \frac{2}{\|\mathbf{w_L}\|} \tag{3.114}$$

Maximising the margin is the same as minimising the $\|\mathbf{w_L}\|$ which is minimising the length of the weight vector. The norm is calculated using the Euclidean distance transform.

The optimisation problem can then be formulated to minimise $\|\mathbf{w_L}\|$, with the constraint that all the points are classified correctly and are on their respective side of the hyperplane (Eq. 3.115).

$$L_i(\mathbf{w_L} \cdot x_j - b) \geq 1, \forall j \tag{3.115}$$

Thus, the optimisation problem can be written as (Eq. 3.116) subject to $L_i(\mathbf{w_L} \cdot x_j - b) \geq 1, \forall j$.

$$\min_{\mathbf{w_L}, b} \frac{1}{2}\|\mathbf{w_L}\|^2 \tag{3.116}$$

As mentioned most data is not linearly separable therefore SVM introduces slack variables $\zeta \geq 0$, which allows a few misclassifications. The optimisation problem can also be defined as (Eq. 3.117).

$$\min_{\mathbf{w_L},b,\zeta} \frac{1}{2}\|\mathbf{w_L}\|^2 + C\sum_{j=1}^{n}\zeta \tag{3.117}$$

For the correctly classified points, $\zeta = 0$ and for the misclassified points, ζ is the distance of that point to its correct side of the hyperplane.

Another method that can be used when the data points are not linearly separated is to use kernels, as mentioned earlier the kernels allow the data to be represented in higher dimensional feature spaces. Kernels such as the linear kernel, polynomial, and radial basis function (RBF) kernels are used. An RBF kernel is used for the purpose of classification in this thesis.

The RBF kernel has a similar shape to a Gaussian kernel due to its similarity to the Gaussian normal distribution (Eq. 3.118).

$$\mathbb{K}(\mathbf{w_L}j, \mathbf{w_L}l) = \exp(\frac{\|\mathbf{w_L}j - \mathbf{w_L}l\|^2}{\sigma}) \tag{3.118}$$

$\|\mathbf{w_L}j - \mathbf{w_L}l\|$ is the Euclidean distance between the data points w_Lj and z_Ll. σ is the parameter that controls the spread of the kernel. The higher the σ values will result in smaller "bumps" around each data point in the feature space, which allows the SVM to fit better to the training data.

The RBF kernel maps the data points from the original feature space into higher dimensional feature space.

This helps with the formulation of the complex hyperplane; the kernel calculates the similarity between the points. Points that are like each other are in a close distance to each other and will have a higher similarity score of 1. While points that are further away will have a value closer to 0. The similarity assists the SVM in deciding the boundary between the classes in the original feature space.

SVM with RBF kernel can predict the class label of a new vector w_L through the following function (Eq. 3.119),

$$f(\mathbf{w_L}) = \sum_{j=1}^{n}\alpha_j L_i \mathbb{K}(\mathbf{w_L}j, \mathbf{w_L}) + b \tag{3.119}$$

α_j is the learned weights for each support vector $\mathbf{w_L}j$, L_j is the class label of the support vector which would either be +1 and -1 for two classes.

When classifying a data point $\mathbf{w_L}$, the SVM will calculate $f(\mathbf{w_L})$, if $f(\mathbf{w_L}) > 0$ then $\mathbf{w_L}$ is classified as +1 which is on the right side of the hyperplane, if $f(\mathbf{z}) < 0$ then $\mathbf{w_L}$ is classified as -1 which is on the left-hand side of the hyperplane.

For all SVM classifications, rbf was the kernel of choice additionally C-regularisation was set to 1, indicating that the model is generalised prioritising a larger margin by allowing more misclassified points over a perfect classification. The gamma coefficient was set at 0.01. These parameters were chosen after running a grid search with a set of parameters then applying K-Fold validation of 5 folds. The set of parameters were: 'C': [0.1, 1, 10, 100], 'kernel': ['linear', 'rbf', 'poly', 'sigmoid'], 'gamma': ['scale', 'auto', 0.01, 0.1, 1].

3.6.7 Logistic Regression:

Logistic regression [347] is a statistical method that is popularly used in binary classification, in the case of this thesis Normal and Tumour. Logistic regression applies the logistic function to predict the probability of a binary outcome.

The logistic function is used to map any feature to a value between 0 and 1. The logistic function can be represented as (Eq. 3.120). $\mathbf{Z}$ is a linear combination of the input features. The probability of an input x belongs to class 1 can be predicted as such (Eq. 3.121).

$$\sigma(\mathbf{Z}) = \frac{1}{1 + e^{-\mathbf{Z}}} \tag{3.120}$$

$$P(Y = 1 \mid x) = \sigma(\mathbf{Z}) = \sigma(w \cdot x + b) \tag{3.121}$$

$P(Y = 1 \mid x)$ is the probability given class x of the positive class (class 1). w is the weight vector for the input features and, b is the bias co-efficient.

The logistic function output can be modelled in terms of odds (Eq. 3.122) and by using the logistic function it can be written as (Eq. 3.123).

$$\text{Odds} = \frac{P(Y = 1 \mid x)}{P(Y = 0 \mid x)} \tag{3.122}$$

$$\text{Odds} = \frac{\sigma(\mathbf{Z})}{1 - \sigma(\mathbf{Z})} = e^{\mathbf{Z}} \tag{3.123}$$

This equation could also mean (Eq. 3.124).

$$\mathbf{Z} = \log\left(\frac{P(Y = 1 \mid x)}{P(Y = 0 \mid x)}\right) \tag{3.124}$$

If the odds are greater than 1, the event is more likely to happen than not, if it is less than 1, the event is less likely to happen.

When training the logistic regression model, the weights w and bias co-efficient b must be optimised to minimise the error. Commonly the maximum likelihood estimation is used to give the cost function (Eq. 3.125).

$$J(w, b) = -\frac{1}{n}\sum_{i=1}^{n}\left[L_i \log(P(Y = 1 \mid x_i)) + (1 - L_i \log(1 - P(Y = 1 \mid x_i)\right] \tag{3.125}$$

Here $P(Y = 1 \mid x_i) = \sigma(\mathbf{Z}_i) = \sigma(w \cdot x_i + b)$ is the predicted probability of the positive class for sample i. n is the total number of training samples. L_i is actual label of sample i. If L_i is 1 and the probability prediction $P(Y = 1 \mid x_i)$ is low, then the logarithmic terms in the cost function $L_i \log(1 - P(Y = 1 \mid x_i)$ will exhibit a large negative number that will increase the cost value. The cost function is averaged to give a single measure of the performance of the algorithm.

The weights and the bias are further improved using the gradient descent method, which updates the weights in the opposite direction of the gradient cost function. The weight and bias values are iteratively updated to lower the loss.

The updates for the weight and bias can be represented as such (Eq. 3.126) and (Eq. 3.127), respectively. Here α is the learning rate.

$$w \leftarrow w - \alpha \frac{\partial J(w,b)}{\partial w}$$

$$w \leftarrow w - \alpha \frac{\partial J(w,b)}{\partial w} \tag{3.126}$$

$$b \leftarrow b - \alpha \frac{\partial J(w,b)}{\partial b} \tag{3.127}$$

The gradient calculations could be computed as outlined in (Eq. 3.128) and (Eq. 3.129).

$$\frac{\partial J(w,b)}{\partial w} = -\frac{1}{n} \sum_{i=1}^{n} \left[L_i - \sigma(\mathbf{Z_i}) \right] x_i \tag{3.128}$$

$$\frac{\partial J(w,b)}{\partial b} = -\frac{1}{n} \sum_{i=1}^{n} \left[L_i - \sigma(\mathbf{Z_i}) \right] \tag{3.129}$$

The main goal of this is to find the best parameters that would give the optimal fit to the training data.

Finally, predictions of the logistic regression on a new data point x can be made as such (Eq. 3.130).

$$P(Y = 1 \mid x) = \sigma(w \cdot x + b) \tag{3.130}$$

The predicted class $\hat{y}$ is decided by setting a threshold.

Thus

$$\hat{y} = \begin{cases} 1 & \text{if } P(Y = 1 \mid \mathbf{x}) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

For all cases of logistic regression, L1 regularisation is used which encourages sparse solutions where some feature coefficients become exactly zero therefore irrelevant features are ignored. liblinear as the optimisation solver which supports L1 regularisation and is used with binary classification or small datasets [348], and C-regularisation which is the regularisation strength set at 0.01 with the fibre features but 0.001 with the fibres cells features. This was decided by running a grid search on a set of parameters and five folds of K-Fold validation. The set of parameters were: 'C': [0.001, 0.01, 0.1, 1, 10, 100], 'penalty': ['l1', 'l2', 'elasticnet', 'none'], 'solver': ['liblinear', 'lbfgs', 'saga', 'newton-cg'].

3.6.8 K Nearest Neighbour:

The K Nearest Neighbour (KNN) [349] is a simpler algorithm for classification. It uses the concept of the nearest neighbour to assign a data point a label based on the labels of its neighbours in the feature space. It is an instance-based learning algorithm therefore does not necessarily require training.

In KNN the aim is to predict a class label of a data point based on the majority class of its nearest neighbours "**K**".

Usually, the "**K**" value is set by the user, and the value of "**K**" is of vital importance as it can affect the performance of the model fitting. If a small value of "**K**" is used, the model will become extremely sensitive as it only looks at those neighbours, a larger value of "**K**" can smooth the decision boundary which may lead to better generalisation, but it can also misclassify some of the labels as it is less sensitive.

The Euclidean distance transform is used to find the "**K**" nearest neighbours of a test point x_{test} and each point of the training data x_i

The Euclidean distance can be represented as such (Eq. 3.131).

$$EucD(x_{\text{test}}, \mathbf{x}_i) = \sqrt{\sum_{j=1}^{n} \left(x_{\text{test},j} - x_{i,j}\right)^2} \tag{3.131}$$

Here the training point is x_i, j and x_{test}, j are the values of the jth feature for the test and ith training point, n is the number of features. With the Euclidean distance each difference is squared and summed across all features as well as the square root is applied to get the total distance.

After extracting the combined distance for every training point, the KNN identifies the closest "**K**" neighbours with the smallest distance value between the test point and the training set. KNN uses majority vote on the nearest neighbours to determine the class of the test points.

This can be represented as such (Eq. 3.132).

$$\hat{L} = \arg\max_{L} \sum_{i \in N_{\mathbf{K}}} Ind(L_i = L) \tag{3.132}$$

$\hat{L}$ is the predicted label, L is the classes, $Ind(L_i = L)$ is the indicator function that is equal to 1 if the neighbours class L_i matches the class L otherwise it is 0.

For all the KNN classifications, the minimum number of neighbours was set at 9 as well as Euclidean distance is used, the weighting function was set at uniform therefore all neighbours were given equal consideration when making predictions. These parameters were chosen after running a grid search on set of parameters and using K-Fold validation of 5 folds for each combination. The set of parameters are number of neighbors': [3, 5, 7, 9, 11], 'weights': ['uniform', 'distance'], 'metric': ['euclidean', 'manhattan', 'minkowski'].

3.6.9 Multi-Layer Perceptron:

Multi-layer Perceptron (MLP) [350] is a supervised learning algorithm that uses a neural network to perform classification. MLP's architecture has an input layer, one or more hidden layers and an output layer. In the hidden layers ReLU is used to act as a non-linear activation function to extract complex patterns in the data.

The input layer is where the features are represented, each feature of the data is represented by one neuron of the input layer. The hidden layers are fully connected layers of neurons, every hidden layer performs transformations on the data by applying weights and biases, the result of that layer is passed onto the next layer with the activation of ReLU. The output layer contains 2 neurons for this thesis, as there are 2 classes to be classified.

A simple neuron with two feature inputs x_1 and x_2 with their respective weights w_1 and w_2 can be represented as (Eq. 3.133) where b is the bias. In MLP forward propagation passes information through the neural network to the output. Every neuron in the hidden layer calculates the weighted sum of the input it gets from the previous layer; it also applies a bias and then the result is passed through the activation function to ensure non-linearity.

$$\mathbf{Z} = w_1 x_1 + w_2 x_2 + b \tag{3.133}$$

The weighted sum considers the weights and bias given in the connections between the previous layer's neuron and the current neuron. The weighted sum $[\mathbf{Z}_j^{(\updownarrow)}]$ can be represented as (Eq. 3.134).

$$\mathbf{Z}_j^{(\updownarrow)} = \sum_{i=1}^{n_{\updownarrow-1}} w_{ji}^{(\updownarrow)} a_i^{(\updownarrow-1)} + b_j^{(\updownarrow)} \tag{3.134}$$

The current layer is $\updownarrow$ and the current neuron is j, the weight $w_j i$ connecting neuron i in layer $\updownarrow - 1$ to neuron j in the current layer $\updownarrow$. $a_i^{(\updownarrow-1)}$ is the output or the activation of neuron i from the previous layer $\updownarrow - 1$. $b_j^{(\updownarrow)}$ is the bias for the current neuron and current layer. $n_{\updownarrow-1}$ is the number of neurons in the previous layer.

The weighted sum, sums over all the neurons in the previous layer to compute the weighted input for neuron j in the current layer l

After the weighted sum $\mathbf{Z}_j^{(\updownarrow)}$ is computed it is passed through to an activation function, for the purpose of this thesis ReLU was used, this can be represented as (Eq. 3.135), where $f(\mathbf{Z}) = max(0, x)$

$$a_i^{(\updownarrow-1)} f(\mathbf{Z}_j^{(\updownarrow)}) \tag{3.135}$$

Cross-entropy loss is used to calculate the loss like (Eq. 3.97)

$$J = -\frac{1}{n} \sum_{i=1}^{n} \left(L^{(i)} \log(\hat{L}^{(i)}) + (1 - L^{(i)}) \log(1 - \hat{L}^{(i)}) \right)$$

n is the number of training samples or data points. Using back propagation, the gradients of the cost function with respect to every weight and bias in the network is extracted (Eq. 3.136).

$$g_t = \frac{\partial J}{\partial w_{ji}^{(\updownarrow)}} \tag{3.136}$$

The gradients are updated for every layer using the 'Adam' optimiser The first moment (mean of gradients), second moment (variance of gradients) and the bias correction are adjusted at each layer and the weights are updated (Eq. 3.137), the bias is also updated similarly and can be represented as (Eq. 3.138).

$$w_{ji}^{(\updownarrow)} \leftarrow w_{ji}^{(\updownarrow)} - \alpha \frac{\mu_t}{\sqrt{\hat{\sigma}_t} + \epsilon} \tag{3.137}$$

$$b_{j}^{(\updownarrow)} \leftarrow b_{j}^{(\updownarrow)} - \alpha \frac{\mu_t^{(b)}}{\sqrt{\hat{\sigma}_t^{(b)}} + \epsilon} \tag{3.138}$$

The weights and bias are updated in the direction that lowers the cost function to a minimum, α is the learning rate.

In the output layer there are two neurons for the two classes where a Softmax function is applied. After the application of the Softmax function the predicted label is the class with the highest probability.

For all MLP classifications there were 50 neurons in the hidden layer with the fibre features and 100 with the fibres and cells features. ReLu activation was used as well as Adam as the optimiser. In addition, initial learning rate was set at 0.0001 and beta 1=0.9, beta 2=0.999 and, epsilon=$1e^{-}08$. These values were decided after running a grid search on a set of parameters. The optimised parameters were chosen after validating the results using K-Fold validation with 5 folds. The set of parameters were: 'hidden layer sizes': [(50,), (100,), (50, 50), (100, 50)], 'activation': ['relu', 'tanh', 'logistic'], 'solver': ['Adam', 'sgd', 'lbfgs'], 'alpha': [0.0001, 0.001, 0.01, 0.1].

3.7 Visualisation Technique:

3.7.1 UMAP:

Uniform Manifold Approximation and Projection (UMAP) [225] is a popular dimensionality reduction technique. UMAP clusters similar data points to each other to highlight groupings of data classes based on their features. It is used on high dimensional data, and it preserves the global and local structures of the data. For example, if we have data points that are represented by 40 features, UMAP reduces the feature representation into two features. UMAP was used in this thesis to visualise and explore the separability between two populations of data based on their extracted feature sets.

3.8 Metrics:

3.8.1 Cell Morphological Metrics:

Cell morphological metrics relate to the physical properties of the detected objects classified as cells. Properties include area, count, and various intensities. These properties can then be used to compare between different cell populations as done in this thesis.

Measurements included:

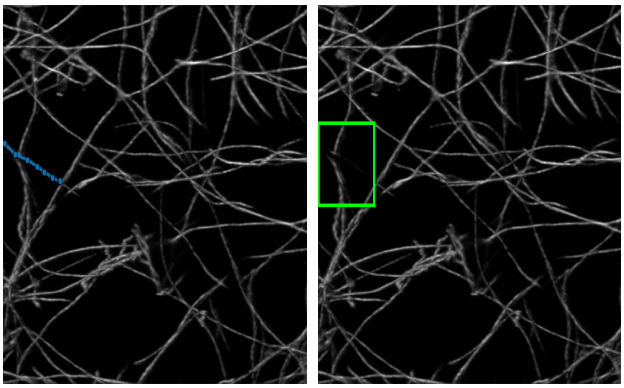
- The number of cells detected, is just how many cells are identified in the image.
- Area of the cells, which is how many pixels is the cell.
- Cell orientation, the angle between the horizontal x-axis of the image and major axis of the cell's ellipse.
- Solidity, which is the number of pixels in the convex hull that are also part of the cell structure.
- Mean intensity, which is the average intensity of the pixel in the images.
- Minimum intensity of the pixels of images, which is the minimum value of intensity of a pixel of that particular cell.
- Maximum pixel intensity of the images, which is the maximum value of intensity of any pixel in that particular cell.
- Maximum Feret Diameter, which is the maximum value of length between the two furthest points on the cell structure.
- Maximum Feret Angle which is the maximum angle value between the horizontal x-axis of the image and the Maximum Feret Diameter of the cells.
- Major Axis Length, which is the length of that cell.
- Minimum Axis Length, which is the width of that cell.
- Aspect ratio is the ratio of the minimum axis length and the major axis length.
- Bright regions are cells or regions of cells that are exhibiting an intensity, one standard deviation higher than the mean intensity.

3.8.2 Ground Truth:

Fibres of FC, FF, BCB and, DME were manually delineated by hand on MATLAB to be used later to compare tracing algorithms (Fig. 3.52).

For the ground truth the fibres were hand delineated in MATLAB (Fig. 3.51). Firstly, an array of zeroes which has the same size as the image is generated to store the drawn fibres as binary masks. An annotation index is set to 0, then a counter is initiated where the index is increased by increments of 1. The MATLAB function "drawpolyline" was used to prompt the user to draw a polyline over the fibres of the image. After the line is drawn, the function "createMask" is used to create a binary mask of the line. The binary mask is then saved as a stack in the initialised array, this was repeated for every separate line. Each fibre was located by eye by zooming in the image to carefully identify every fibre. Each individual drawn line forms a layer of the stack and is labelled. The stack of lines is then flattened into a 2D image where all the drawn lines can be seen. Finally, the drawn lines are dilated to match the width of the actual fibres and removed from the image, so that the user is able to identify which fibres are left to trace. It is important to note that the fibres were hand delineated by one observer, but a better practice would be to have more than one user delineate the fibres. The workflow for generating the ground truth can be found in *https* : *//github.com/youssefarafat/Trace_Ridges/blob/main/drawGroundTruth_general.m.*

Similarly, the DAPI cells of the 10 images were manually segmented by hand to quantify the accurateness of the Cell segmentation. It is important to note that the ground truth was delineated by one observer.



(a) (b)

Figure 3.51: Ground Truth generation. (a) Using the 'drawpolyline" function to draw over the fibre. (b) After the delineation, the fibre is removed from the image.

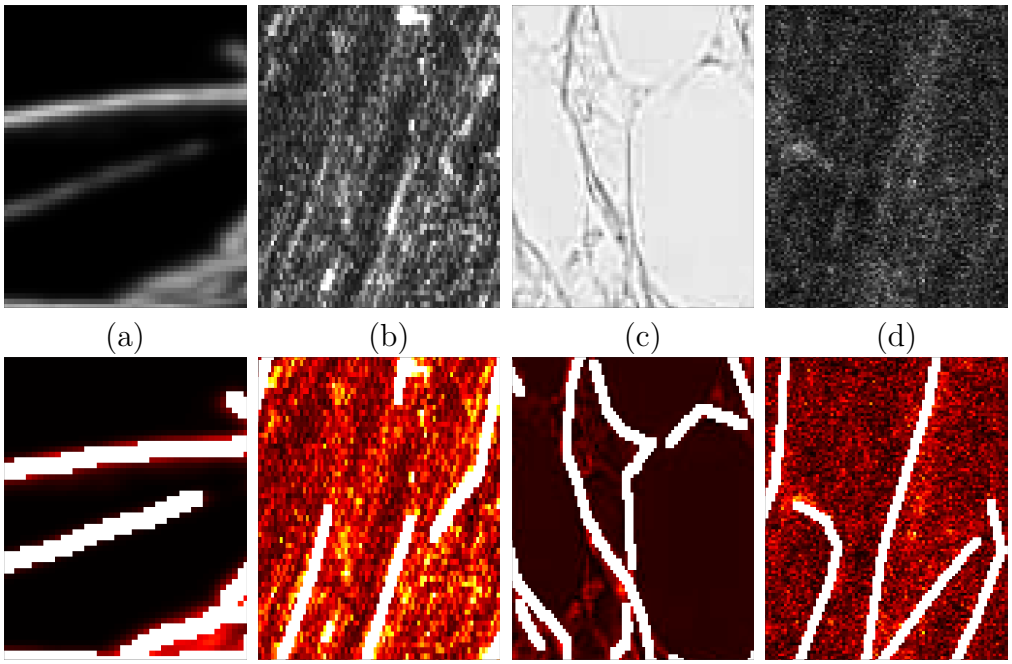


Figure 3.52: Ground Truth of dataset 2 (Fig. 3.14): Region of interest of original images and hand delineated tracing. The algorithms performance with these images were used and compared with the ground truth. ROI: (a) Fluorescently labeled Collagen (FC). (b) Fluorescent Fibronectin (FF), Image courtesy of Dr Esther Castellano, and Dr Cristina Cuesta-Apausa, Universidad de Salamanca. (c) Breast Cancer Biopsy slide (BCB). (d) Disease Mimicking Extra Cellular Matrix (DME). Images (a), (c) and (d) are open source. (e)-(h) Manually delineated ground truth (GT) for the ROIs of images in (a-d). White lines are those of the GT and intensities are displayed in a colourmap with hot colours (black-red-orange-yellow-white).

3.8.3 Pixel-wise Metrics:

Pixel-wise metrics are measurements to quantify accurate segmentations based on whether a pixel was correctly identified to belong to a particular class or not.

Accuracy:

The first metric that was used to compare the segmentation is Accuracy. Accuracy is a measurement of how correct the algorithm is, it takes into account the correctly identified True Positives and the True Negatives so that means that it takes into account the correctly identified cell regions (True Positives) as well as the correctly identified non-cell regions (True Negatives) and compares them with the image as a whole and that includes the correctly identified cell and non-cell regions as well as the incorrectly identified cell and non-cell regions. The scale is from 0-1. With 0 being not accurate and 1 being accurate.

This means that accuracy is measured in the following way (Eq. 3.139).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.139}$$

Jaccard Index:

The second form of measurement was the Jaccard Index. The Jaccard index measures how accurate the segmentation was based on the correctly identified positive regions (True Positives) which in this case is the correctly identified cell regions and it compares the correctly identified regions with the correctly identified positive regions as well as all the incorrectly identified regions, leaving the correctly identified background regions out of the measurement. This makes it a stricter measurement of accuracy and in return a stricter measurement of how well the segmentation is. Therefore, it is expected that the numbers would be smaller in value. The scale is from 0-1. With 0 being not accurate and 1 being accurate.

The Jaccard index is measured in the following way (Eq. 3.140).

$$\text{Jaccard Index} = \frac{TP}{TP + FP + FN} \tag{3.140}$$

Dice Score:

The third form of measurement that was used is the Dice Score. The Dice score is like the Jaccard index in that it only considers the correctly identified positive regions which are the identified cells from the segmentation and compares it with the accurately identified positive regions as well as the all the incorrectly identified regions. It does not consider the correctly identified background regions. The difference between Dice and Jaccard is that it accounts twice for the correctly identified positive regions. In the Dice score the incorrectly identified regions are penalised.

Therefore, the Dice Score is measured in the following way (Eq. 3.141).

$$\text{Dice Score} = \frac{2TP}{2TP + FP + FN} \tag{3.141}$$

The final metric was the comparison of the number of cells detected by the segmentation as compared to the ground truth. Labelled objects were counted and compared.

3.8.4 Fibre Metrics:

This section outlines two fibre-related metrics, firstly the distance calculation which is used to calculate the distance between an algorithm's trace of a fibre and the ground truth, this was used in the objective comparison seen later in the thesis. The second set of metrics are the morphological metrics which once the fibres are traced, the morphological metrics quantify the properties of the fibres.

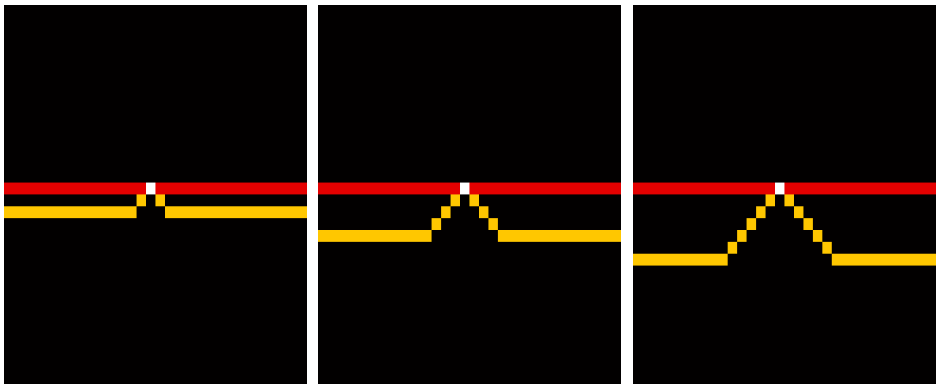
Distance Calculation:

Quantitative comparisons between the GT and the tracing algorithms results were computed using distance metrics, as previously reported in [211]. Conventional pixel-based metrics such as accuracy, true/false positive rate, specificity, and sensitivity are not suited for trace comparisons calculations against the ground truth. This limitation is shown in (Fig. 3.53), where there are three cases that display lines that overlap in just a single pixel but extend from that point in quite different conditions. Although these cases differ, pixel-based metrics would give an identical count of true positives as one in each instance. Similarly, true negatives (black pixels) may appear consistent across cases, providing little useful information for distinguishing between them. Pixel-based metrics are suitable for larger structures, such as cells or nuclei, where comparisons are area-based rather than distance-based between traces. Computing of the distance metrics was previously outlined in [211] but is revisited here. The distance calculation is illustrated with a clear example in (Fig. 3.54) and the Second Harmonic Generation image in (Fig. 3.55) To begin, a distance map was generated (Fig. 3.54b, e) and (Fig. 3.55b, e) for each trace being compared. Then, each map from one set was multiplied by the map from the opposing set. The traces that are further away from the traces from the opposite set will be in areas of higher distance while traces closer to their opposites are in areas of short (low) distance (Fig. 3.54c, f) and (Fig. 3.55c, f). The distance maps can yield several metrics, such as average distance error per trace, maximum distance error per trace and total distance error.

The algorithm result is multiplied (element wise multiplication) by the distance map of the ground truth. The multiplication results in a matrix that holds information on only the pixels present in the algorithm trace i.e., part of a trace that are in the same position as pixels in the ground truth distance map that are far away from a non-zero pixel which is the delineated traces. This is an indication of the mismatch between the algorithm trace and the ground truth. Similarly, the distance map of the algorithm result is multiplied by the ground truth delineation. This is done as some traces might be in the ground truth but not in the algorithm result and vice versa, thus the distance must be measured in both directions. The mean intensity for every labelled object is computed. Intensity values are a measure of distance. The mean intensities are used to calculate the average distance error. The maximum intensities per labelled object are obtained, contributing to the maximum distance error per trace. The intensity values are summed to give the total distance error per image. The average distance error is the average distance between the two fibres of the algorithm trace and the ground truth represented in green (Fig. 3.56). The maximum distance error is the biggest distance between a pixel of the ground truth fibre and a pixel of the algorithm trace represented in red (Fig. 3.56). The total distance error is the total distance between the pixels of the ground truth and the algorithm trace represented in blue (Fig. 3.56).

This method has its own set of limitations, if an algorithm traces many fibres or oversegments inaccurately, then its average error per trace could be significantly less than other algorithms that may have not oversegmented. This is due to the fact that if an algorithm traces many fibres, its traces will not be far from an actual fibre therefore the errors will

be less. Another limitation is the maximum distance error, once again if an algorithm oversegments, its maximum distance error will be smaller in comparison to methods that do not oversegment as its a singular error value of the maximum distance between an actual and algorithm fibre. Code for the distance calculation can be found on GitHub at $https://github.com/youssefarafat/Trace_Ridges/blob/main/calculateErrorDistance.m$.



(a) (b) (c)

Figure 3.53: Visualisation of three cases of two lines where the pixel-wise metrics of true positives, true negatives, false positives, false negatives are identical, but the distance between the lines is different.

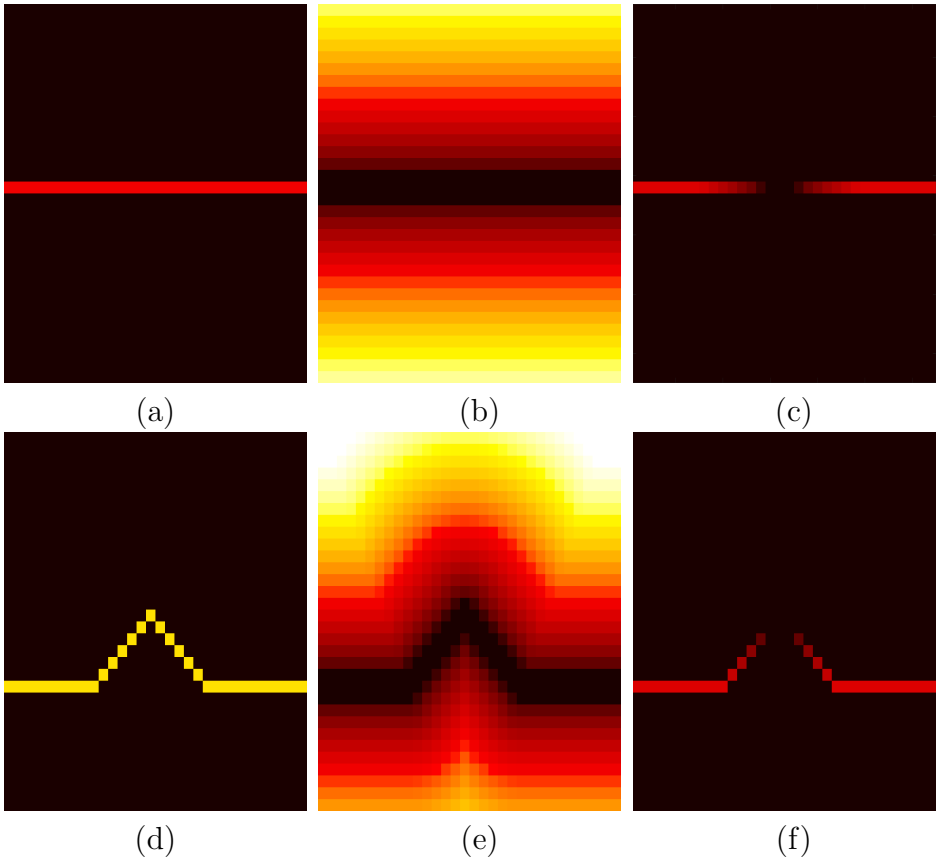


Figure 3.54: Visualisation of the distance error measurement between the two lines of (Fig. 3.53c). (a,d) Two lines that overlap only at one point. (b) Distance map extracted from the red line in (a). Distances follow a hot colour map: black-red-orange-yellow-white. (c) Product of the distance map shown in (e) with the line shown in (a). (e) Distance map extracted from the yellow line in (d). (f) Product of the distance map shown in (b) with the yellow line shown in (d).

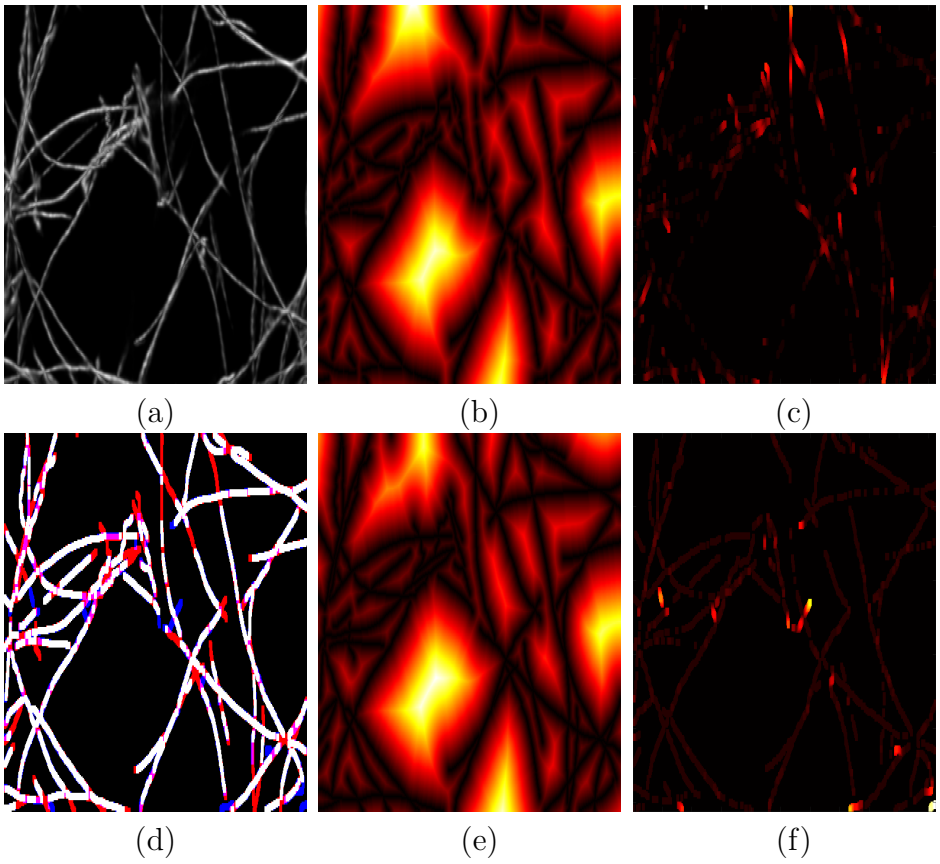
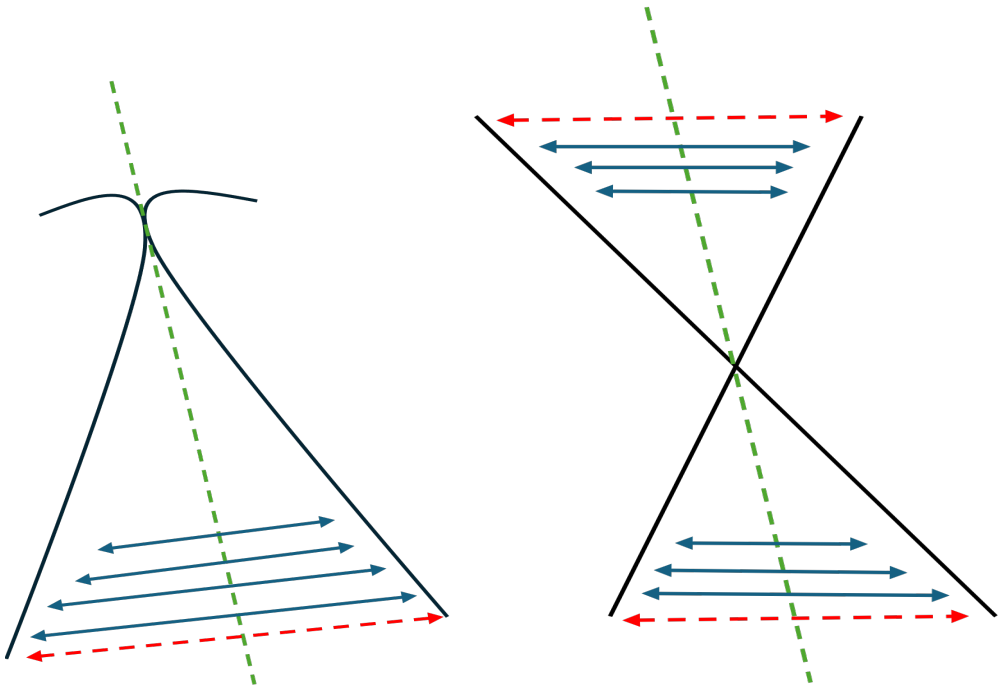


Figure 3.55: Illustration of the calculation of the error distance maps on the fluorescently labeled collagen image shown in (Fig. 3.14a). (a) FC image with a variety of fibres. (b) Distance map calculated from the traces calculated with Trace Ridges. (c) Product of the ground truth and the distance map of (b). (d) Comparison between the results of Trace Ridges and the ground truth. White lines are the pixels from GT and the algorithm trace. Blue lines correspond to Trace Ridges' trace. Red lines are from the GT. (e) Distance map calculated from the ground truth. (f) Product of the traces from Trace Ridges and the distance map of (e). The cases where the lines overlap (white) in (d) do not appear in any of the distance maps (c, f) but the lines that do not overlap (blue and red) appear in the distance maps with increased brightness as they are at a greater distance away from the lines generated by the opposite method.



(a) (b)

Figure 3.56: Illustration of the Distance Metrics. The average distance error per trace is represented by the green line between the two fibres where it takes the average distance between the trace of the algorithm (one line) and the ground truth (the other line). The maximum distance error is represented by the red line which the is the biggest distance between the two fibres. The total distance error accounts for all the distances between pixels of the trace and the ground truth. (a) Representation of the distance error metrics between two fibres (black), one representing the algorithm trace and the other representing the ground truth. The fibres have a region where they are touching and are near each other and a region where they are far away from each other. (b) Representation of the distance error metrics between two fibres (black), one representing the algorithm trace and the other representing the ground truth. The fibres have a region in the middle of the lines where they are intersecting.

Morphological Metrics:

Morphological metrics of fibres relate to the properties of the objects in images detected as fibres. These properties include orientation, alignment, length, and width. Properties also include gap metrics which measures the gap area size between the fibres. These properties can be used to compare populations of fibres.

- Orientation: Orientation is the angle between the x-axis of the image and the major axis of the object/fibre. The value of the orientation is in degrees, and it is between the range of -90 and 90.
- Alignment Score: This is a measure of how aligned the fibres in the image are. As the orientation angles are between [-90, 90] degrees, the angles are shifted by $+90^o$. The angles are then converted to radians, to look at the orientations as positions in a circle. The mean sine and cosine magnitude in radians is calculated. The resultant vector magnitude in the polar co-ordinate system is then found, this gives a score between 0 and 1. 0 means the fibres are not aligned, while 1 means that the fibres in the image are aligned.
- Major Axis Length: This is the length of object/fibre in pixels. Measured between the furthest points of the fibre.
- Minor Axis Length: This is the width of the object/fibre in pixels. Measured between the minor axis of the object.
- Aspect Ratio: The ratio of the major axis length and the minor axis length.
- Area: This is the total count of pixels of objects or fibres in an image.
- Area taken up by fibres: The area the fibres take up of the extracellular matrix.
- Intensity: This is the intensity value of the pixels in an image. Maximum, Mean and, Minimum values of intensities can be calculated.
- Eccentricity: This is a measure of how straight a fibre is.
- Circularity: A measure of how close in shape a fibre is to a circle.
- Curvature: A measure of how much a shape of a fibre deviates from a straight line.
- Number of regions/ Bright Regions: This is the number of regions in the image with an intensity one degree of standard deviation higher than the mean.
- Number of edges/fibres: This is the total count of fibres detected in an image after the delineation.
- Average gap area: Based on the Euclidean distance transform, this is the average gap area in pixels.

- Gap Area: A Euclidean distance transform of the detected fibres is computed where it assigns every pixel a value of measurement of distance to its closest non-zero pixel. Pixels that are of high value correspond to the gaps between fibres. The sum of those pixels that are above the average distance to an edge give the total gap area.
- Largest Gap: Based on the Euclidean distance transform, the largest gap is the value of the largest connected region of low value pixels, measured in pixels.

3.8.5 Object-wise Metrics:

Object-wise metrics are like the pixel-wise metrics with the difference that pixel measurements are not performed but rather a true positive is an accurately predicted label, and a True Negative is an accurately labelled opposite class in the case of this thesis. This has been used to quantify the performance of the classification tasks.

Accuracy:

Accuracy is a measurement of the proportion of correctly classified labels whether they are from class 1 (positive) or class 2 (negative). It can be represented as (Eq. 3.142).

$$\text{Accuracy} = \frac{\text{Correct Classifications}}{\text{Total Classification}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.142}$$

If the dataset is imbalanced, accuracy does not add much value to the analysis, as it could show a higher accuracy percentage if there is one class that is dominating the dataset.

Recall:

Also known as the true positive rate, recall is a measurement of the proportion of the actual class that was predicted correctly out of all the actual positives. False negatives are actual positives that were misclassified. Recall can be represented as such (Eq. 3.143).

$$\text{Recall} = \frac{\text{Correctly Classified Positives}}{\text{All Positives}} = \frac{TP}{TP + FN} \tag{3.143}$$

In simple terms how many of the samples were identified correct from the samples that should have been identified correct.

Recall will add no value if the class samples are imbalanced.

Precision:

Precision is a measurement of the proportion of actual positives from the positive classes identified and can be represented as (Eq. 3.144).

$$\text{Precision} = \frac{\text{Correctly Classified Actual Positives}}{\text{Everything Classified as Positive}} = \frac{TP}{TP + FP} \tag{3.144}$$

In simpler terms, how many classes were actually correct from the samples identified as belonging to that particular class.

If the dataset is imbalanced, precision does not add much value similarly to accuracy, recall and precision.

Recall and Precision are opposites, precision increases when the number of false positives decrease while recall increases when the number of false negatives decrease. Precision is a harsher classifier that questions the actual positives from the classification, while recall is more of a relaxed classification metric that would accept any samples that resemble a positive class.

F1 Score:

The F1 Score is computed from the Recall and Precision, where it is the harmonic mean of the precision and recall score. The harmonic mean pushes for similar values for precision and recall. The F1 Score implies that both recall, and precision are maximised

F1 Score can be represented as (Eq. 3.145).

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{3.145}$$

A score of 1 in F1 Score indicates that the scores of Precision and Recall are high.

F1 Score is the best measurement for imbalanced datasets.

3.9 Summary:

This chapter explored the various techniques and performance metrics used throughout this thesis. Acquisition methods such as microscopical imaging techniques used in this thesis are explained and visualised to give an overview on the appearance of images used throughout this research. In this chapter different datasets were explored and outlined. Dataset 1 contained 3 sets of images representing the cell nuclei, actin filaments and fibronectin for the mutant and wild-type populations. Dataset 2 contained four images of fibre-like structures from different backgrounds. Dataset 3 contains three different sets of images, with each set containing two specific populations. The first set had images of collagen fibres at different concentrations (1.7 and 6.6 mg/ml), the second set contained images of collagen control samples and collagen with protease inhibitors. The final set contained collagen control samples and images of collagen exhibiting Marangoni Flow. Dataset 4 contained two populations of Tissue Micro-Arrays representing normal breast tissue and tumour breast tissue. The images contained both collagen and cell nuclei. An image is defined as well as different image colour spaces and dimensions such as grayscale, RGB, HSV and binary images are introduced. The role of filtering in image processing is explained followed by the introduction of traditional image processing techniques that are building blocks to many computer vision techniques, these techniques were explored with a sample image. Segmentation and in specific cell segmentation is outlined and CellPose a cell nuclei segmentation method that

was used in an investigation in (Section 5) is introduced. Tracing algorithms are then outlined and explained. Tracing algorithms are used to delineate fibre like structures; in this thesis they are used to explore images that contain fibres of the extracellular matrix such as fibronectin and collagen. This section explored tracing algorithms by applying the same image across each algorithm. This chapter outlines classification methods that are used in Chapter 6 to investigate which classification method performs best in classifying normal and tumour features that have been extracted using Trace Ridges. They are also used to investigate whether Trace Ridges assists in the classification or if deep learning methods can classify populations from the images alone. UMAP, a dimensionality reduction visualisation is outlined and then used in Chapter 6 to show the relationship between the two populations of normal and tumour. The ground truth is then explored and outlined; the ground truth is used to analyse the performance of cell segmentation as well as the performance of tracing algorithms. The chapter then explores the performance metrics used throughout this thesis as well as morphological metrics. Firstly, it outlines metrics extracted for cells then outlines how the performance of cell segmentation is measured using pixel-wise metrics. It then moves on to the fibres where it explores how the distance calculation is conducted which is deemed useful when measuring the accuracy of tracing. It then outlines the morphological metrics and features that are extracted by Trace Ridges. Finally, object wise metrics are discussed which assist in analysing the performance of the classification tasks.

Chapter 4

Exploration of Data:

4.1 Overview:

This chapter explores images in Dataset 1 by applying traditional image processing techniques to extract and quantify meaningful insights from the cell nuclei, actin filaments and fibronectin fibres. This is conducted to explore and understand the data as well as experiment with different image processing techniques. In this chapter, other pre-processing pipelines developed for the purpose of investigations of this thesis are presented.

4.2 DAPI Stained Images:

4.2.1 Segmentation:

WT images are images that contain normal cell nuclei while the RBD images, the cells were mutated, therefore it is hypothesised that they have differences between them. The first approach to investigate the cells focused on the nuclei as it was hypothesised that the differences of a WT and a mutant population would manifest on the morphological characteristics of the images. For instance, the size of nuclei, orientations, distributions within the region observed, etc. In order to quantify meaningful insights from the cell nuclei, they first must be segmented and extracted.

The nuclei were stained with DAPI and appeared blue (Fig. 4.1a). The image was passed through a low pass filter with a Gaussian kernel $\mathbb{K} = 5$ Otsu threshold was applied to segment objects in the image and separate objects into foreground and background (Fig. 4.1d). Otsu is used instead of setting a fixed value for segmentation so that it is specific for each image. Objects segmented are then labelled (Fig. 4.1e).

Labelled objects with an area greater than 48 pixels were considered large cells, which were then "closed". As previously explained closing is when an object goes through dilation followed by erosion. In this case the "large" cells are filled through dilation and then they are eroded to ensure that they do not exceed their original size and no two objects merge with each other. Through testing it was found that objects with a circularity value larger

than 0.7 are single cells. Merged cells were detected by using the inverse parameter or the single cells (Fig. 4.1d). To distinguish whether the object is that of 2 or 3 cells that are merged, the area was used. Objects with an area less than 4000, were of two different cells merged. While objects with an area greater than or equal to 4000 are three merged cells.

Objects of three merged cells are separated firstly by erosion with a linear structuring element that had a radius size of the minor axis length of that object to ensure it acts on the cell boundaries. The number of the structuring element was 90 minus the orientation of that object; this ensures that the structuring element aligns with the orientation of the cell's long axis which helps when dealing with elongated or stretched cells. The cells were then subjected to dilation with a kernel size of 15×15, this can be seen in (Fig. 4.1e). Watershed was used to distinguish the components of the merger cells into three distinctive parts based on the inverted image where the foreground pixels became the background pixels and vice versa, erosion ($\mathbb{K} = 5$) was used to clean up the image. Watershed can be seen in [351] [352], then objects of two merged cells were separated

Firstly, the Euclidean distance transform was computed [304] and as suggested in [353, 354]. Using the inverse of the distance transform, the Watershed method as previously explained was used to split the two cells into distinctive cells. Erosion with a structuring element of 6×6 of ones was applied to remove some of the edges of the cell to ensure that they do not overlap anymore.

Cells on the edges of the image were removed as they were often fragments and their characteristics would have been misleading. Cells were removed using their Extrema property. Cells with the minimum Extrema value that is less than one and cells with the maximum Extrema value that is greater than the value of the rows were removed (Fig. 4.2a).

The detected cells with their edges can be seen in (Fig. 4.2b). The image was partitioned to form a Voronoi diagram as suggested in [307]. The partitions of the Voronoi diagram are based on how close the originating point in each partition is to the points outside the partition. The more crowded the area in terms of cells, the smaller the partition of the originating point. The Delaunay Triangulation between the detected cells [308] and as used in [309] can be seen in (Fig. 4.2c).

(Fig. 4.2d) illustrates the detected cells from (Fig. 4.2b) with the edges laid on top of them for visual representation. (Fig. 4.2e) shows the Delaunay Triangulation between the detected cells [308] and as used in [309]. Firstly, the image was partitioned to form a Voronoi diagram as suggested in [307], such as that each centroid of each detected cell had its own partition. The partitions are based on how close the originating point in the partition is to the points outside the partition. The more crowded the area in terms of cells, the smaller the partition of the originating point.

4.2.2 Ground Truth Segmentation Comparison:

To assess the quality of the segmentation process, the results were compared against a manually labelled ground truth and the metrics of Accuracy, Jaccard, Dice and number of cells were calculated.

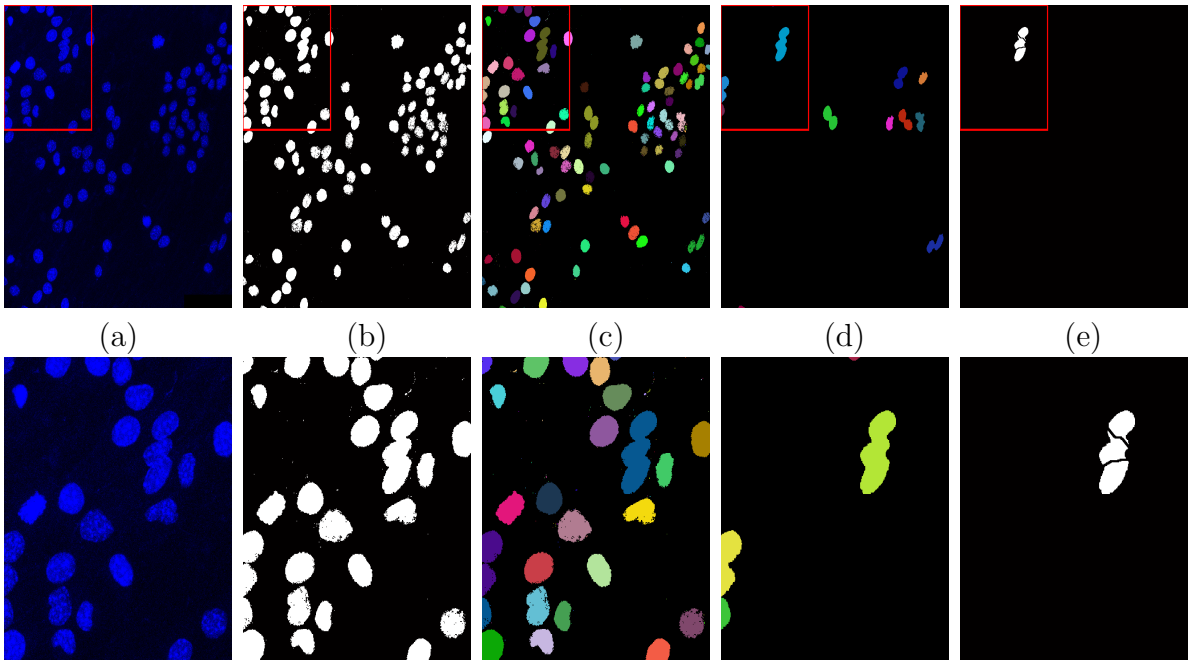


Figure 4.1: Processing of the DAPI channel to segment the cell nuclei. Top images: illustrate the full field of view. (a) Input Image. (b) Otsu threshold applied to the input image. (c) Labelling of objects from (b) where each segmented object is lablled with a different colour. (d) Detection of all merged cells from image b. (e) Detection and Separation of objects of 3 merged cells using Watershed. Bottom images (f)-(j): Region of interest indicated by the bounding box of the top row.

To calculate the metrics previously mentioned, it was necessary to determine the TP, TN. FP, and FN, which are illustrated in (Fig. 4.3).

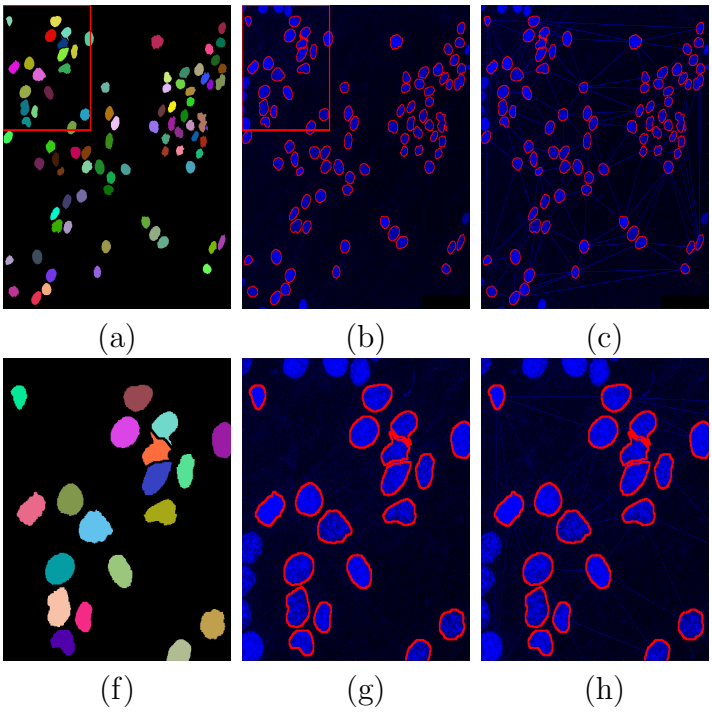


Figure 4.2: Processing of the DAPI channel to segment the cell nuclei continued. (a) Cells and fragments of cells on the border of the image a removed. (b) Original segmented cells. (c) Delaunay triangulation applied on cells from (b).

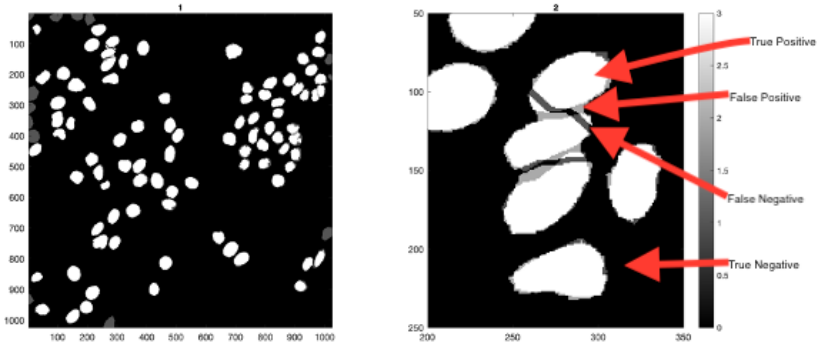


Figure 4.3: Pixel-Pixel Segmentation Comparison. White regions or pixels are the True Positive (TP). Light grey regions are False Positive (FP). Dark grey regions are False Negative (FN). The black pixels are True Negative (TN).

4.3 Phalloidin Stained Images:

The first approach to investigate the actin filaments focused on the segmenting the filaments as whole objects or as "islands" as it was hypothesised that the differences of a WT and a mutant population would manifest on the morphological characteristics of the images. For instance, the size of the islands, orientations, number of islands, etc.

Segmentation of the actin filaments is the primary goal for the Phalloidin images, the raw image is seen in (Fig. 4.4a). The input image was filter with a Gaussian filter of size 5×5 [355]. Binary image was obtained by setting a threshold of 120 (Fig. 4.4b). The threshold cleaned the image further. The thresholded components were labelled (Fig. 4.4c).

The image was then filled, dark pixels that were surrounded by light pixels were converted into light pixels [356]. The image was cleaned by removing small artefacts that do not belong to any of the island spaces. The image was cleaned by removing objects of connected pixels that had less than 150 connected pixels, this is also known as area opening [303], followed by closing which used a disk-shaped kernel with a size value of 35×35.

The distance map of the islands is computed (Fig. 4.4e). The closed image was dilated to enhance the shape of the islands, a disk-shaped kernel with a size value of 15×15 is used.

(Fig. 4.5a), illustrates the skeleton map of islands. The skeleton map is useful as it shows the points where the islands are joined together [305] [306], which could be used to separate the islands.

The break-off points were computed by selecting the points of the skeleton map which had a value less than 30 in the computed distance map, the dilated break-off points (Fig. 4.5b) [301].

Edges of the islands were detected and edges with an area less than 100 pixels are removed (Fig. 4.5c). The break-off points were removed from the segmented islands, resulting in the separated islands. Some of these islands were very small (Fig. 4.5d) as a result of the multiple dilations which were removed, and the rest were labelled (Fig. 4.5e).

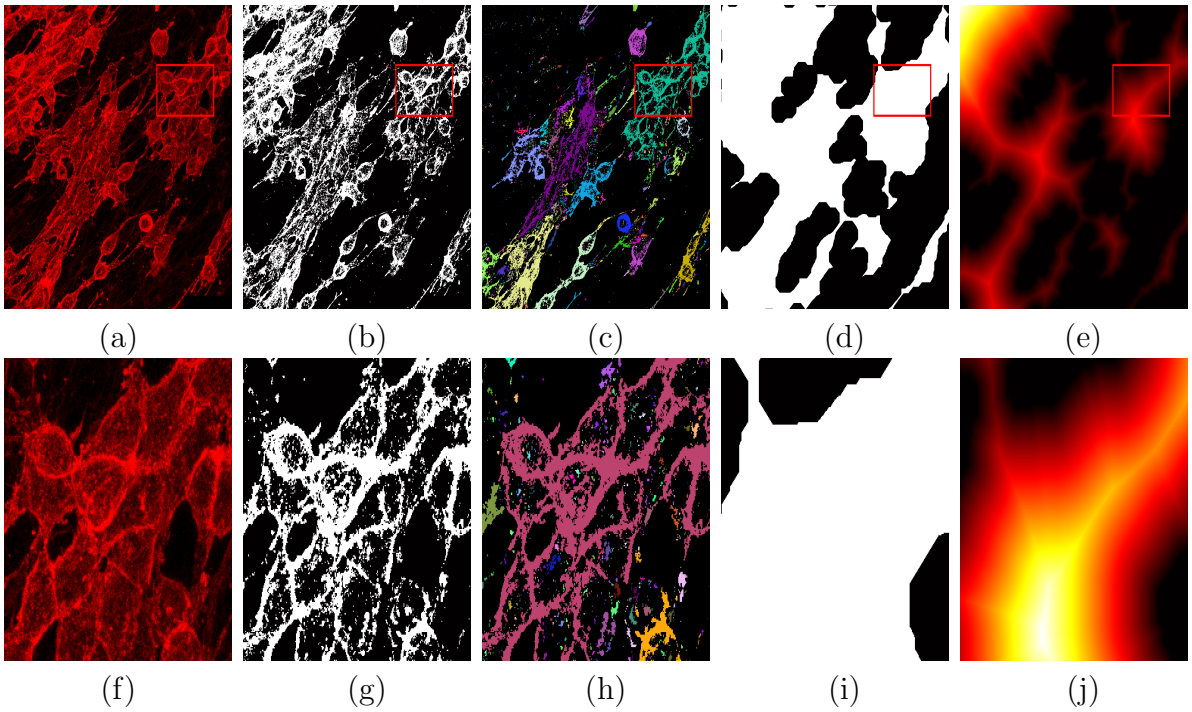


Figure 4.4: Phalloidin actin filaments island segmentation. Top images: full field of view. Red bounding box showing a region of interest. (a) Input Image. (b) Binary image obtained with a threshold of 120. (c) Objects from (b) labelled. (d) Expansion of island regions followed by the erosion of outer regions of the islands. (e) Distance map of islands from (c). Bottom images (f)-(j): Region of interest of (a)-(e) indicated by the red region of the top row.

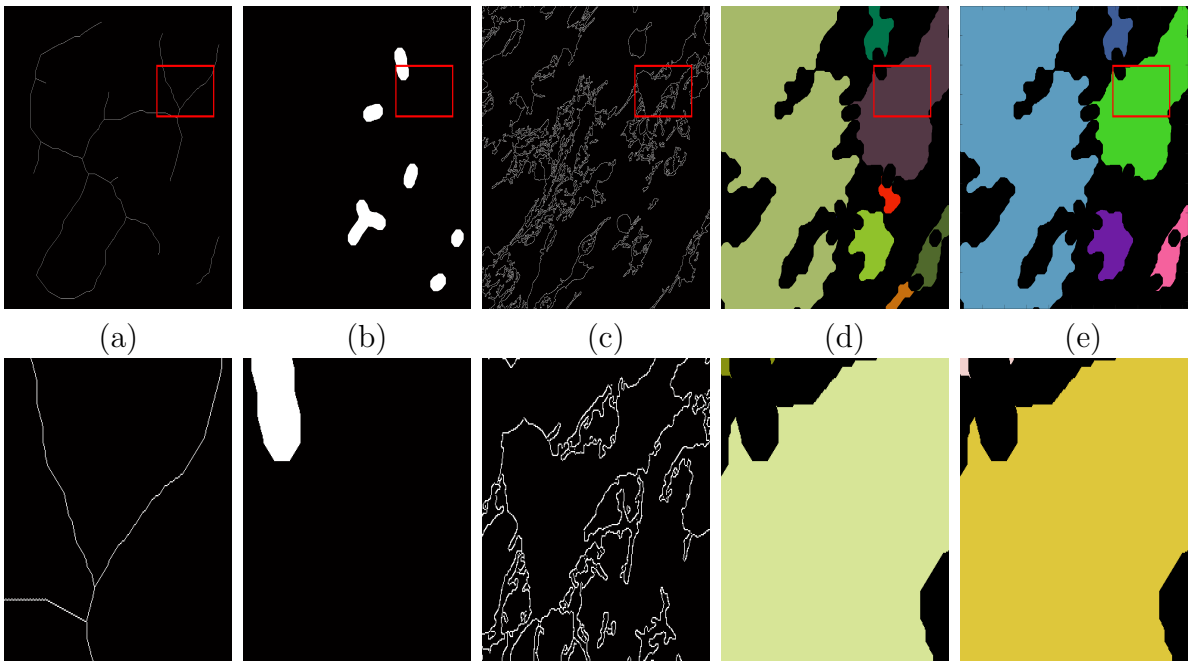
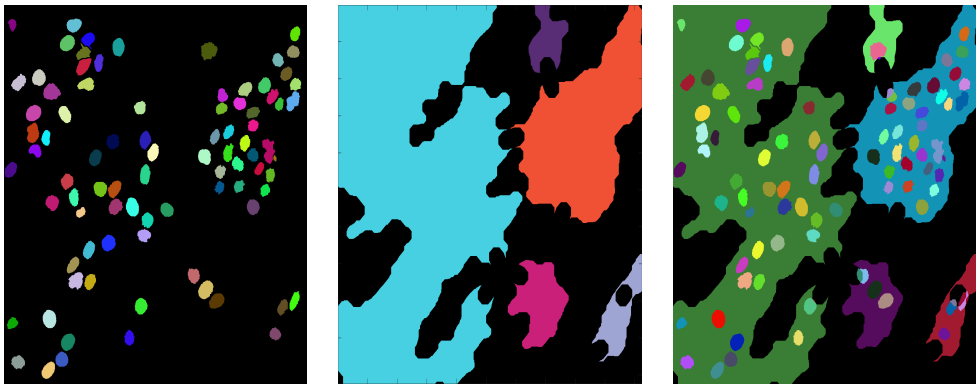


Figure 4.5: Phalloidin actin filaments island segmentation continued. (a) Skeleton Map of islands. ((b) Dilation of break-off points identified. (c) Removal of small edges from (Fig. 4.5e). (d) Islands after removal of smaller edges and the identification of small islands. (e) Final islands labelled. (f)-(j) represent the region of interest of (a)-(e) as indicated by the red bounding box.

4.4 DAPI + Phalloidin Combinatory Properties:

After the segmentation of the cell nuclei from the DAPI channel and the actin filaments from the Phalloidin channel. It was hypothesised that the differences of a WT and a mutant population would manifest on the morphological characteristics of the images. For instance, the ratio of number cell nuclei relative to area of islands, orientations, number of unique nuclei per island, etc. To test the hypothesis the segmented cell nuclei (Fig. 4.6a) were placed on the segmented actin filaments islands (Fig. 4.6b) to give one image (Fig. 4.6c) where comparisons of the extracted metrics were made.



(a) (b) (c)

Figure 4.6: DAPI and Phalloidin combination. (a) DAPI Segmented Cell Image. Each cell has been labelled with a different colour as each cell has been identified as a different object. (b) Phalloidin Segmented Island Image. Each island segmented is labelled in a distinctive colour. (c) DAPI and Phalloidin channels combined. It can be seen that the cells fall on the islands. The cells that belong to one island all are labelled with the same colour.

(Fig. 4.7), illustrates the combination of the DAPI and Phalloidin layers for the ten images from each channel, which results in ten images of DAPI and Phalloidin combination. These images are used for the comparison of properties.

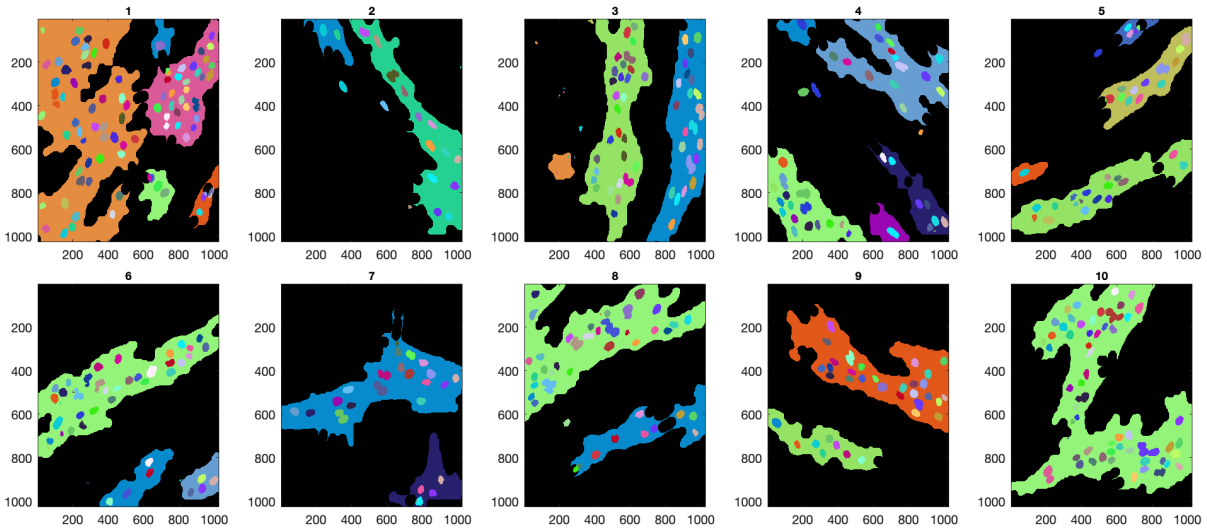


Figure 4.7: DAPI and Phalloidin combination of all images. The result of the DAPI segmented cells and Phalloidin segmented actin filaments islands of the same cell culture were merged together. The top five images are those of the RBD images. The five images below are those of the WT.

4.5 Fibronectin Images:

The first approach to investigate the fibronectin fibres focused on tracing the fibres as it was hypothesised that the differences of a WT and a mutant population would manifest on the morphological characteristics of the images. For instance, the number of fibres, orientations, gap areas, etc.

Visually looking at the Fibronectin images the structures are fine hairs with certain gaps which are the darker regions of the images. To be able to extract metrics from the images, the fibres were traced using Canny Edge Detection with a filter of size 2×2. This can be seen through the middle images in (Fig. 4.8). After which, the edges were labelled.

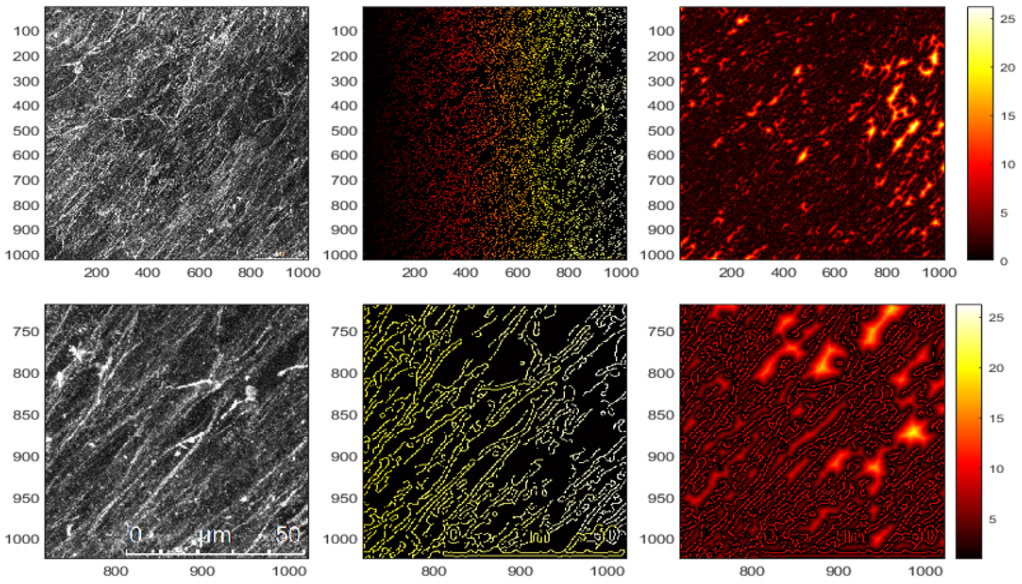


Figure 4.8: Fibronectin Analysis. Left image: Input image RBD_LKR13_1_Fibronectin from the repository. Middle Image: Edges of Fibronectin fibres using canny method. Right image: Gaps between Fibronectin fibres, brighter red areas show bigger gaps. Bottom row: Region of interest.

Furthermore, in the method to obtain more metrics the gaps between the edges were obtained, this was done by using a Euclidean distance transform and computing the distance map. It then calculates the number of pixels that are further away from the average distance to an edge. This can be seen in the right-hand images of (Fig. 4.8). The bigger the gap the further away a point is from an edge of the fibronectin hairs, the higher the intensity and the more yellow it is. The components of the edges and the gap were particularly useful in obtaining interesting metrics.

4.6 Segmentation Performance Results & Morphological Metrics Comparisons:

4.6.1 DAPI Images:

To verify if the segmentation of the DAPI cells was sufficient, comparisons of the segmentation and the ground truth were made. The ground truth included all the cells except those on the borders of the images.

Segmentation:

DAPI Directory	Accuracy	Jaccard Index	Dice Score
1	0.9726	0.7843	0.8791
2	0.9891	0.6886	0.8156
3	0.9876	0.8438	0.9153
4	0.9780	0.7224	0.8388
5	0.9904	0.8102	0.8951
6	0.9880	0.8089	0.8943
7	0.9936	0.8440	0.9154
8	0.9874	0.8120	0.8963
9	0.9898	0.8085	0.8941
10	0.9821	0.8010	0.8895

Table 4.1: DAPI segmentation measurements of Accuracy, Jaccard Index and Dice Score. The number in the DAPI directory refers to the images from the DAPI directory.

The measurements from Table. 4.1 are of the central cells which are only the whole cells and do not include partial cells on the borders of images. The mean accuracy was 0.9859. The mean Jaccard index was 0.7924. The mean Dice Score was 0.8834. (Fig. 4.9) illustrates the pixel-pixel segmentation comparisons with all measurements of Accuracy, Jaccard Index, and the Dice Score. A fourth measurement was added, the Cells/GT, and this is the number of cells detected by the segmentation compared to the number of cells detected in the ground truth.

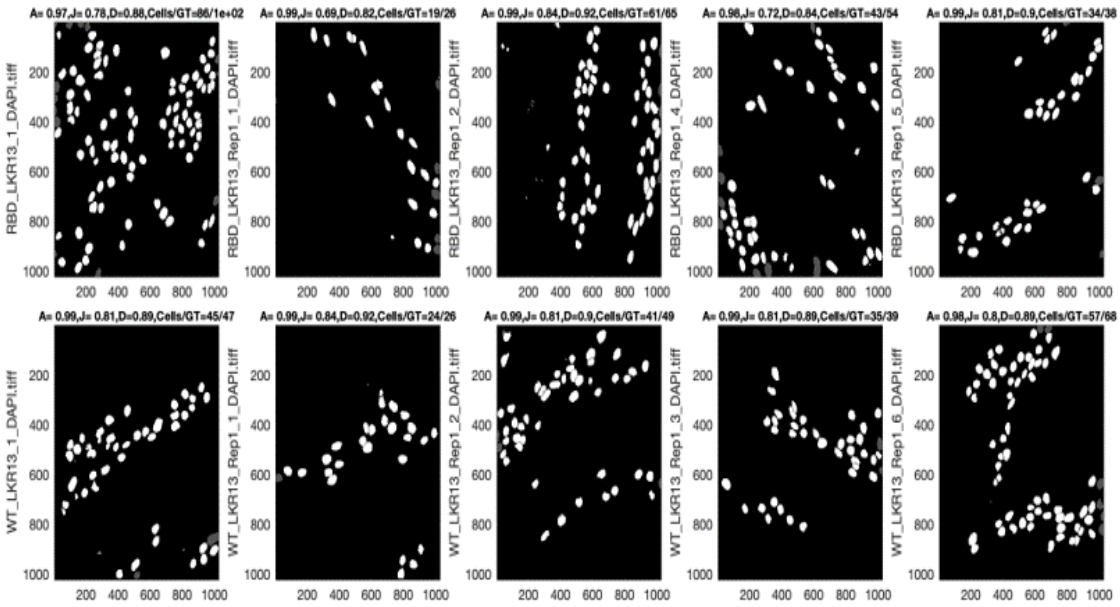


Figure 4.9: Pixel-Pixel Segmentation of the Central Cells. The top five images are those of the RBD and bottom five images are those of the WT. Each image contains the values of Accuracy (A), Jaccard Index (J), Dice Score (D) of the segmentation as well as Cells/GT, which is the number of cells segmented compared to the number of cells in the ground truth (GT).

Cell Morphological Features:

Since it was hypothesised that morphological properties of the cell nuclei could show differences between the mutant and the WT population several properties were extracted, the number of cells detected, Area of the cells, Solidity, Mean intensity, Minimum intensity, Maximum pixel intensity, Maximum Feret Diameter, Maximum Feret Angle were extracted for comparison.

Both the mean and the standard deviation were calculated for all the metrics per image. The population results are then compared using boxplots to observe any differences. The boxplots for the mean values can be seen in (Fig. 4.10) and for the standard deviation (Fig. 4.11).

By observing the boxplots only one measurement observed a difference when comparing populations of the RBD and WT populations which was the standard deviation of the Maximum Feret Angle (Fig. 4.11).

As seen from (Fig. 4.11), on average RBD cells have a lower Max Feret Angle value, as well as a larger variation in angle sizes than the WT cells.

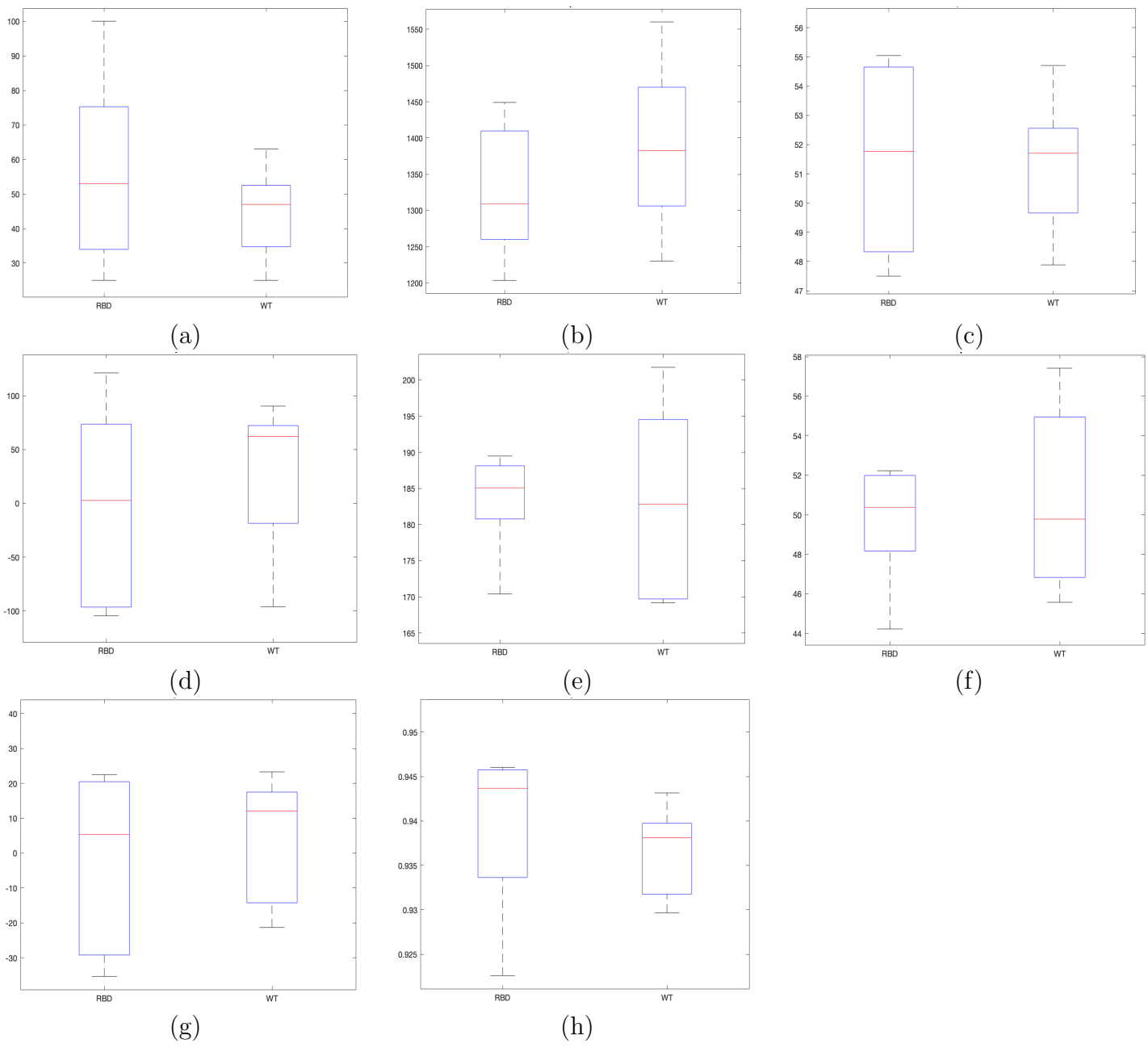


Figure 4.10: Boxplot illustration when comparing mean cell properties of RBD and WT. (a) Number of cells. (b) Mean Area. (c) Mean Maximum Feret Diameter. (d) Mean Maximum Feret Angle. (e) Mean of the Mean Intensity. (f) Mean Minimum Intensity. (g) Mean Orientation. (h) Mean Solidity. None of the boxplots are observing any differences when comparing the populations.

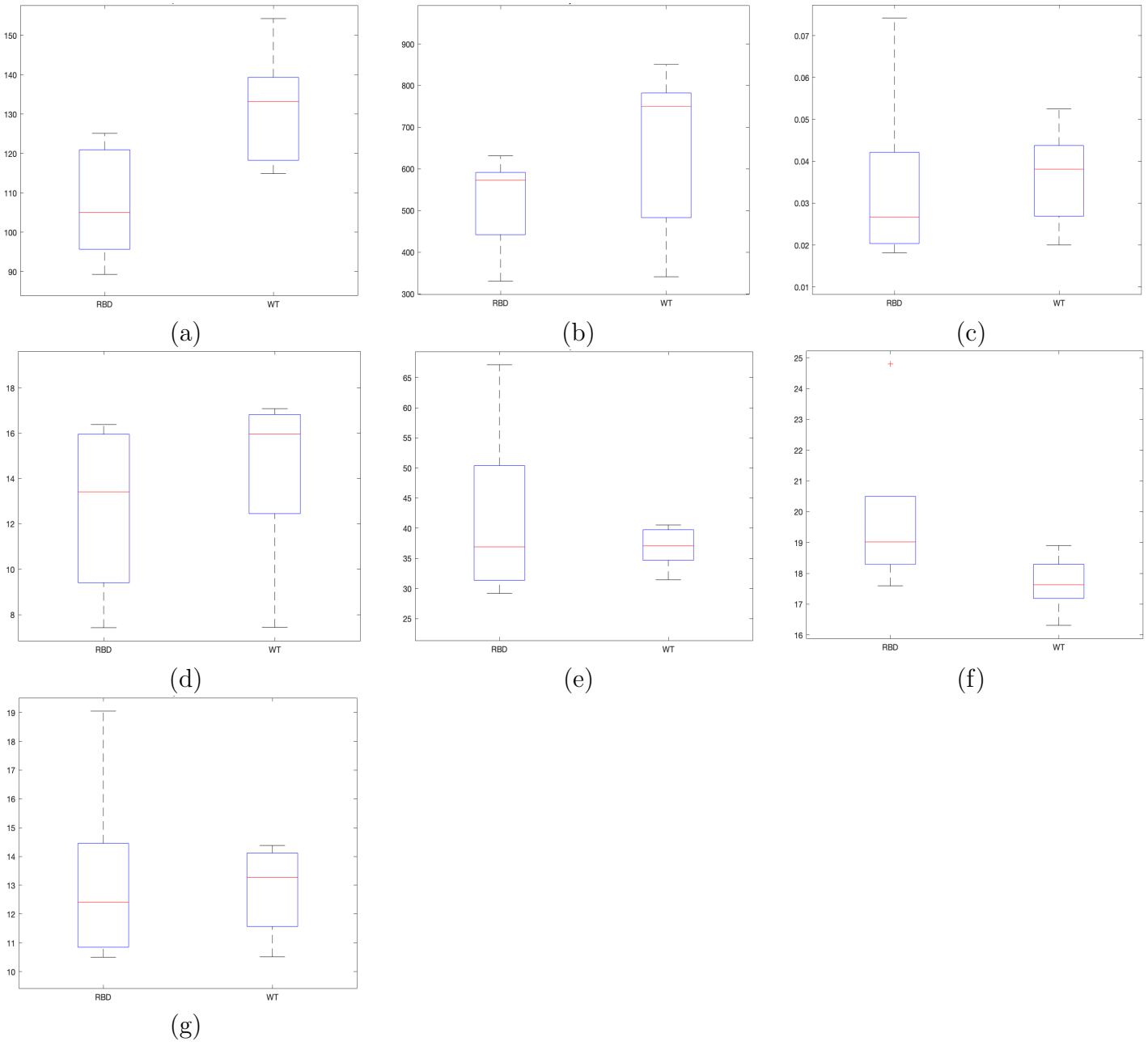
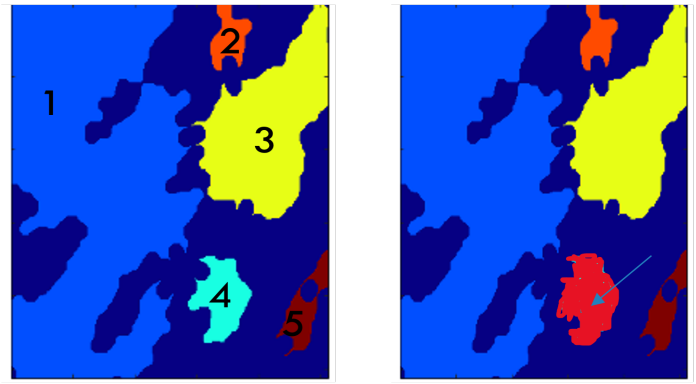


Figure 4.11: Boxplot illustration when comparing the standard deviation of the cell properties of RBD and WT. (a) Std Maximum Feret Angle. (b) Std Area. (c) Std Solidity. (d) Std Maximum Feret Diameter. (e) Std Orientation. (f) Std Mean Intensity. (g) Std Minimum Intensity. Only when comparing the standard deviation of the maximum Feret angle of the RBD and WT populations the boxplots show a difference.

4.6.2 Phalloidin Images:

The actin filaments of the Phalloidin channel were successfully segmented into islands. It was hypothesised that comparing morphological properties of the mutant and WT populations might lead to the discovery of differences between them. Two metrics were of interest to compare and those were the number of islands and the area size of the actin filament islands.

The number of islands (Fig. 4.12a) is how many islands there are per image. The area of islands (Fig. 4.12b) measures the average area of islands in an image. The measurement with differences was the number of islands.



(a) (b)

Figure 4.12: Phalloidin island morphological properties. (a) Number of Islands measurement. (b) Area of Islands measurement. The island covered in red to illustrate the area of the island. Which is a measurement of the number of pixels of that island.

From (Fig. 4.13) on average RBD images seem to have more islands in the Phalloidin channel compared to the WT images (Fig. 4.14).

From (Fig. 4.14a) it is evident that there is a difference when comparing the number of islands of the RBD and WT populations. the RBD images have a wide variation in the number of islands while the WT images have a small variation. The highest number of islands in the RBD images is 5 islands while the highest number of islands in the WT images is 3. The lowest number of islands in any RBD image is 2 islands while for the WT images the lowest number is 1 island. When comparing the area of the islands (Fig. 4.14b), there appears to be no differences. There is a higher variation in the area sizes of islands from WT images compared to the islands from the RBD images.

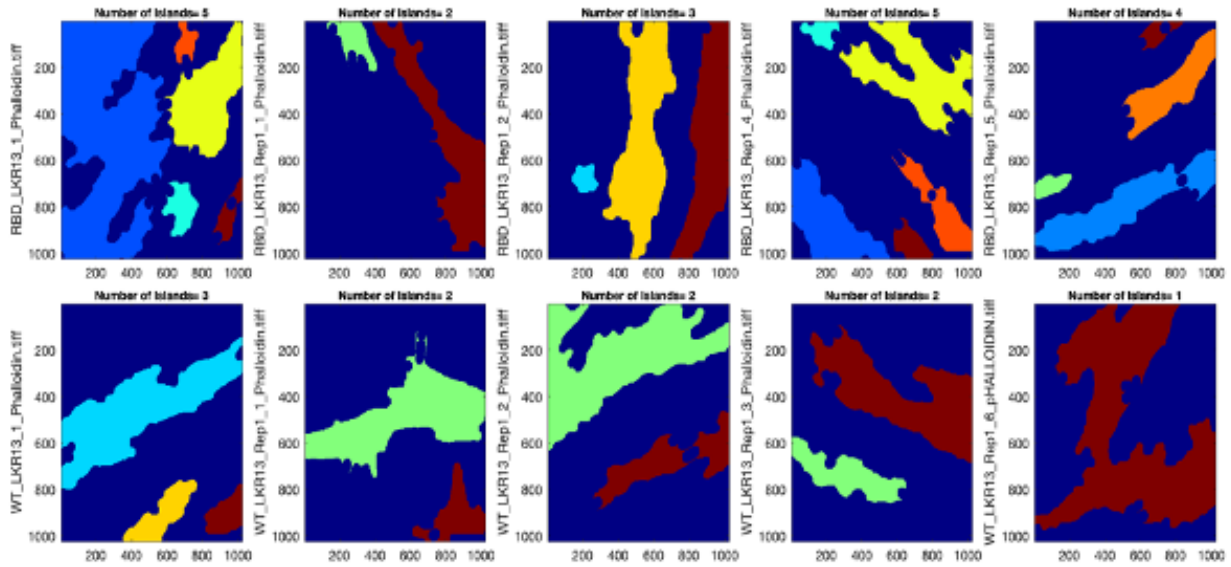
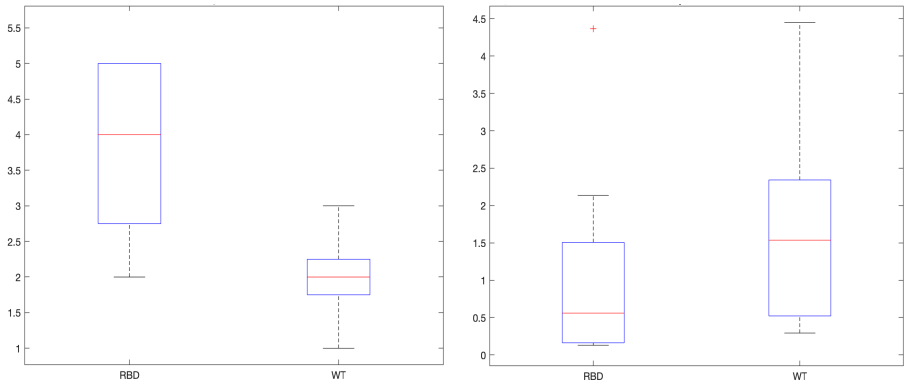


Figure 4.13: Number of Phalloidin Islands. Each image has the number of islands segmented at the top. The top five images are those of the RBD and the bottom five images are those of the WT.



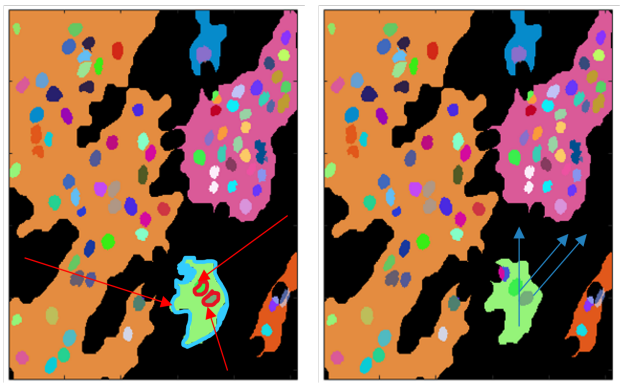
(a) (b)

Figure 4.14: Boxplot illustration of the comparison between RBD and WT when comparing the Phalloidin islands extracted metrics. (a) Boxplot of the number of islands. (b) Boxplot of the Area of islands ($X10^5$).

4.6.3 DAPI & Phalloidin Combination:

After successfully combining the cell nuclei and the actin filament island. Two metrics were extracted to compare between the mutant and WT populations. The two measurements were, area comparison of the DAPI cell nuclei and actin filament islands and finally, cell nuclei and actin filaments orientation comparison.

The total area of all the cells within an island is measured and that is compared to the area of their corresponding island seen in (Fig. 4.15a). The orientation of the cells from



(a) (b)

Figure 4.15: DAPI and Phalloidin combination morphological properties. (a) Area comparison of DAPI cells and Phalloidin islands. Area taken up by the cells of a particular island compared to the area value of that island. (b) Orientation comparison of DAPI cells & Phalloidin islands. The point at which the cells of a particular island are pointing to, compared to the orientation of that island.

the DAPI channel that are on one island are compared to the orientation of that island illustrated in (Fig. 4.15b).

Orientation comparison of DAPI cells and Phalloidin islands showed statistical significance when comparing the RBD and WT population (Fig. 4.16a). While, when the ratio of cells per area and the mean unique cells per area of actin filament islands and the area comparison of the cells per area of actin filaments islands showed no differences (Fig. 4.16b, c, d).

As it can be seen in (Fig. 4.16a), that there are differences between the boxplots indicating a difference between the populations when comparing the DAPI cells and their corresponding actin filaments island orientation. There is a higher variation in the WT images, and the average orientation is slightly higher than zero, while for the RBD images the orientations are averaging at around zero and there is a smaller variation. There also seems to be many outliers for the RBD orientations. The WT cells and islands seem to have a more positive orientation compared to the RBD orientation. When observing the other comparisons (Fig. 4.16b, c, d), there are no differences between the boxplots.

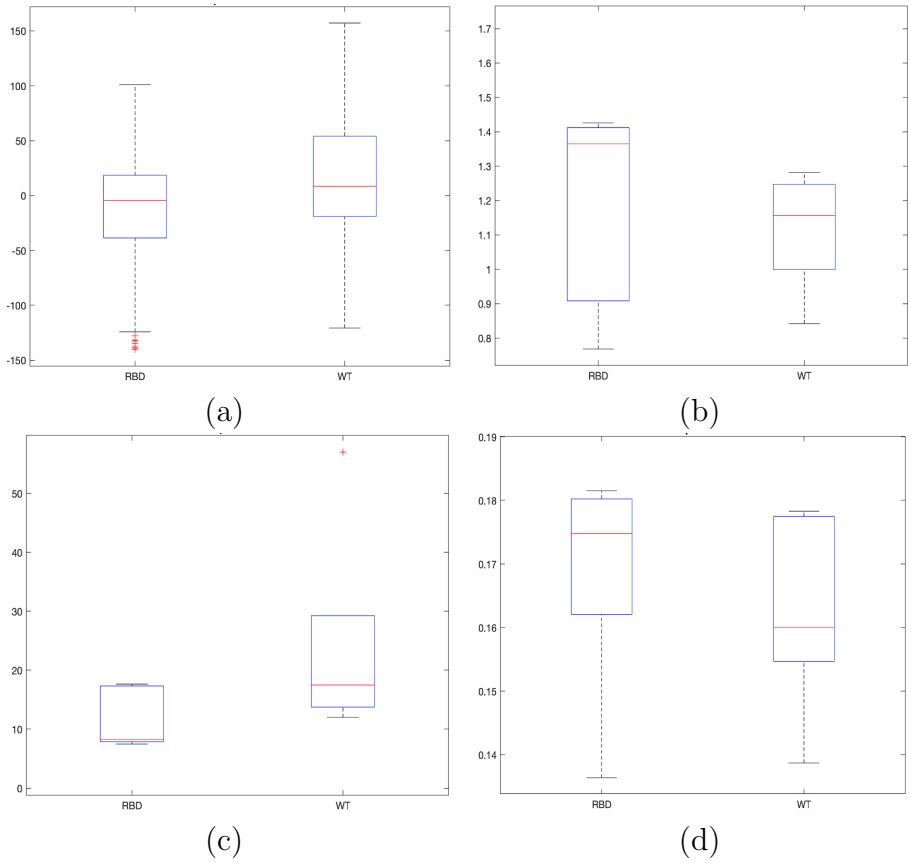


Figure 4.16: Boxplot illustrations of the comparison between RBD and WT when observing the combined cell and Phalloidin islands metrics. (a) Differences in the orientation of the DAPI cells and Phalloidin islands. (b) Ratio of number of cells per area of island Boxplot ($X10^{-}4$). (c) The number of unique cells per islands. (d) Area comparison, where the total area of cells occupying an island is compared to the area of that island.

4.6.4 Fibronectin Images:

Excerpts of this section have been published as a conference abstract in MIUA 2023, Aberdeen, UK.

After successfully tracing the fibres of the fibronectin using Canny Edge detection two morphological properties were extracted to test the hypothesis to investigate whether there were differences between the mutant and RBD populations. Edges were extracted and the gaps between the fibres were computed. The two measurements were the average major axis length relative to the size of the image and the relative area of gaps.

As it can be seen in (Fig. 4.17), images of the Fibronectin dataset have been processed, and their edges have been identified and labelled. The major axis length for each edge was measured. The major axis length is the length in pixels of the major axis of the edges. The values of all the major axes of all the edges were then added and the mean was found. The mean was then divided by the image's size (row value) to give the relative mean value of

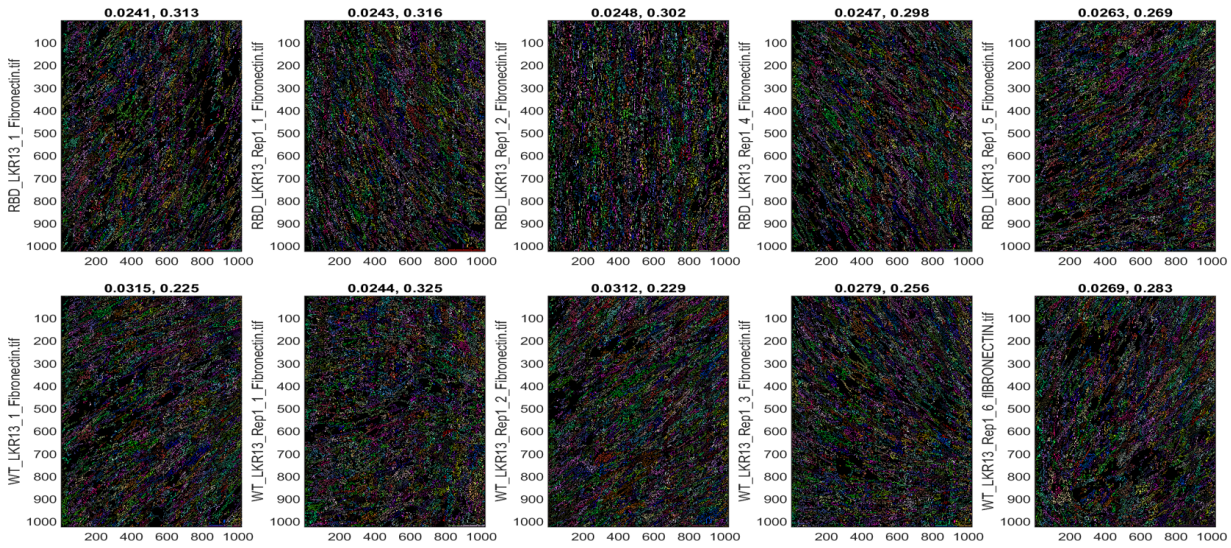


Figure 4.17: Fibronectin Edges Measurement & Major Axis Length. Each edge is labelled with a random colour to distinguish between the objects. The top left value is the relative mean value of the major axis length. The top five images are that of the RBD, while the bottom five images belong to the WT.

the major axis length for that image. This is the first number in the title of each image in (Fig. 4.17).

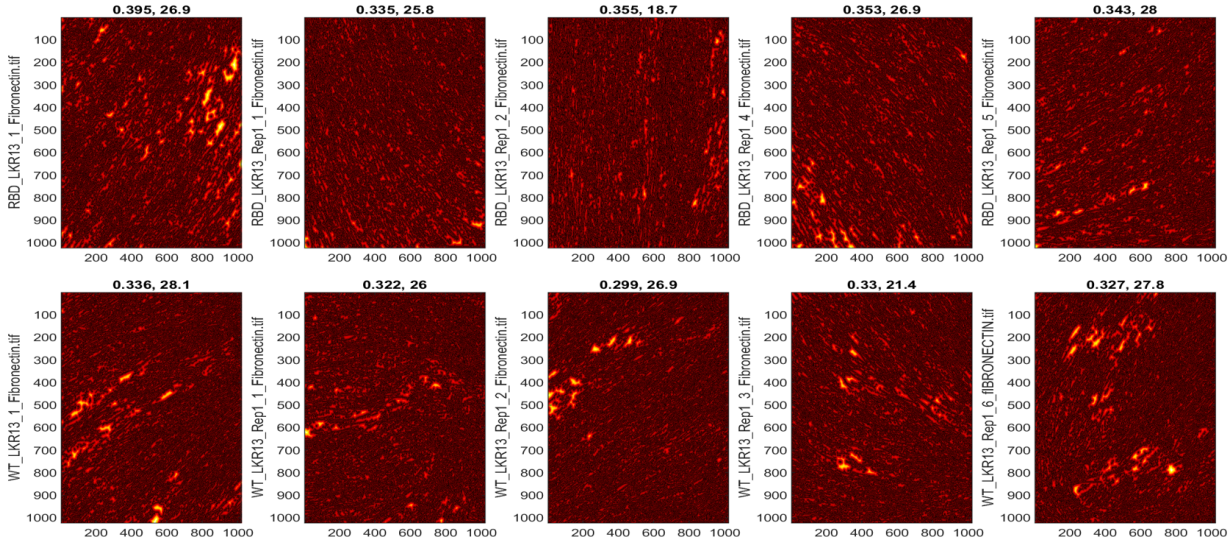


Figure 4.18: Fibronectin Relative Gap Area Measurement. The left value is that of the mean relative value of the gaps. The top five images are that of the RBD. The bottom five images are that of the WT.

It can be viewed from (Fig. 4.18), that images of Fibronectin have been processed by

applying a distance transform on them. The gap area was found by calculating the sum of the pixels in the distance transformation that is greater than the mean pixel intensity of the distance transformation. Therefore, there are regions of different intensities, the further away a pixel is from an edge the brighter and more yellow the colour. The closer a pixel is to an edge the dimmer its intensity and the darker red it is as evident in (Fig. 4.18). This measurement is then divided by the size of the image in this case the rows and the columns to give the relative value of the area of the gap compared to the image size, after which the gaps area (distance) are added and a mean for the image is calculated this can be found in (Fig. 4.18) and it is the first number in the title.

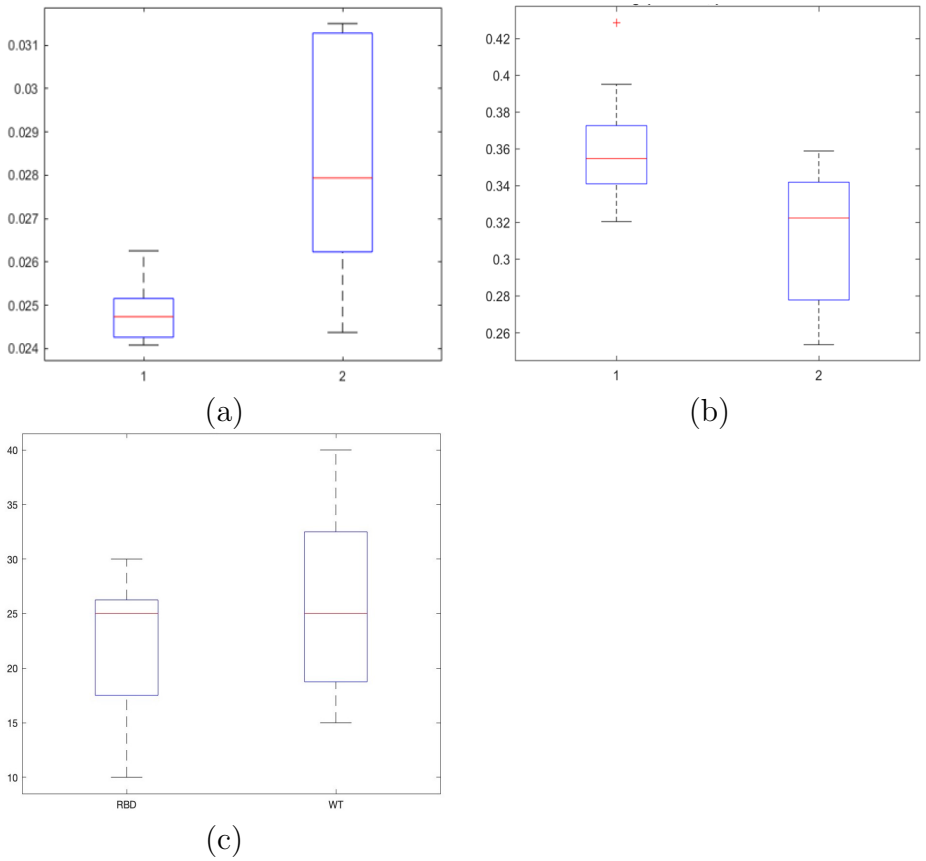


Figure 4.19: Boxplot illustration of the comparison between RBD and WT when observing the differences in the (a) Fibronectin Relative Major Axis Length of Edges. (b) Fibronectin Relative Area of Gaps. (c) Orientation of the fibres. When comparing the populations, the boxplots of the relative major axis length and the relative gaps between the fibres observed a difference.

When comparing the relative major axis length of the RBD and WT fibres, the boxplots show a difference. As can be seen from the boxplots in (Fig. 4.19a), RBD images are denoted here as 1 and WT images are denoted here as 2. It can also be seen that there is a much higher variation in the WT major axis lengths, while the RBD major axis length is more

orderly, and its range is small. The average relative major axis length in RBD images is 0.0248 while for the WT images it is 0.0280.

From (Fig. 4.19b), the boxplots show a difference when comparing the relative gap area of the RBD and WT. RBD is denoted here as 1 and WT is denoted as 2 in the boxplot. WT gap areas have a higher variation than the RBD, on average RBD had a higher average relative area of the gaps at a value of 0.358 while the WT was at a value of 0.322. The differences can also be seen visually from (Fig. 4.18), by looking at the bright yellow regions. While when comparing the orientation of the fibronectin fibres (Fig. 4.19c), no differences can be seen.

Although the samples analysed are small in quantity, the investigation shows that the Edge Detection method is not an ideal tracing method due to its affinity of over-segmenting. The results are encouraging and promote further exploration of the mutant and wild type extracellular matrix fibronectin fibres.

The work done in this chapter is primarily to experiment with different image processing techniques to develop a processing pipeline. The different shapes of objects for each channel analysed allowed development in the understanding of image processing techniques as well as conducting comparisons. Cells and fibres were easier to segment and trace, while the actin filaments proved challenging due to their structure extending beyond the borders of images. There were also differences observed between the samples which promotes further exploration with a developed algorithm and in particular the extracellular matrix fibres.

4.7 Image Preprocessing Pipelines:

Pre-processing is an important task in image processing with the main goal of preparing an image to be passed through an algorithm, whether for tracing or segmentation. The main objective of this thesis is to trace images of fibronectin and collagen. In materials, we discussed several image datasets that use different acquisition combinations, such as the type of microscope and staining agent used. Images vary in noise levels and must be processed to prepare for segmentation or tracing. The TMA cores of Dataset 4 contain both cells and collagen fibres that must be separated and passed to their respective processing methods (segmentation and tracing). In this section we outline the different pre-processing pipelines that were used throughout this thesis to prepare particular datasets for tracing.

4.7.1 No Preprocessing:

No individual pre-processing was conducted on images of Dataset 2. These images were used to compare tracing algorithms. In this experiment, the images were filtered with three filtering methods, no filtering, Gaussian filtering, and DnCnn filtering. Subsequently, these images were passed onto 7 tracing algorithms for comparison. This is explained in detail in (Section 5.3.5).

4.7.2 Fluorescent Fibronectin Images:

From the comparison of tracing algorithms (Section 5.3.5), Dataset 1 images exhibited the lowest distance errors when pre-processed with DnCnn. For the experiment in (Section 5.4) all the images in Dataset 2 were passed through DnCnn before tracing.

4.7.3 Increasing Collagen, Protease Inhibited, Marangoni Flow images:

Images of Dataset 3 were used to investigate whether the accumulation of collagen fibres leads to the remodelling of the cytoskeleton. The images were pre-processed with a simple averaging kernel of ones of size $\mathbb{K} = 3$ to slightly smooth the image before tracing.

4.7.4 TMA Cores:

The TMA cores of dataset 4 are fully unprocessed, the images exhibit both cells and collagen fibres. The two components must be isolated so that segmentation can be conducted on the cells and tracing could be done to the collagen fibres to extract useful information on the features of normal and tumour components.

The input image is a three channel RGB image (Eq. 3.3),

$$\mathbb{I} = N_r \times N_c \times 3$$

The RGB image was converted to HSV seen in (Fig. 4.20a), as fibres and cells have different hues and saturation values from the staining agents.

The first step of this conversion is to normalise the values of the Red, Green, Blue colour channels to be in the range of [0, 1], this is typically done by dividing the intensity values by 255. Value of the HSV as mentioned earlier is the brightness of the colour and therefore the maximum of the normalised RBG values for every (x, y) pixel (Eq. 4.1). The value channel of the HSV image can be seen in (Fig. 4.20d).

$$V(x, y) = max(R(x, y), G(x, y), B(x, y)) \tag{4.1}$$

Saturation of the HSV is the colour intensity and for every pixel (x, y) can be represented as (Eq. 4.2).

$$S(x, y) = \frac{V(x, y) - min(R(x, y), G(x, y), B(x, y))}{V(x, y)} \tag{4.2}$$

The minimum is extracted here as the difference between the maximum and minimum is an indication of the spread of colour, with higher values indicating a purer colour while lower values indicate less colourfulness. The saturation channel of the HSV image can be seen in (Fig. 4.20c).

To calculate the Hue, the channel that determines the maximum value determines the colour, if R is the maximum the hue is calculated as (Eq. 4.3).

$$H = 60 \times \frac{G - B}{V - min(RGB)} \tag{4.3}$$

If G is the maximum, then (Eq. 4.4) applies.

$$H = 120 + 60 \times \frac{B - R}{V - min(RGB)} \tag{4.4}$$

Finally if B gives the maximum, then (Eq. 4.5) applies.

$$H = 240 + 60 \times \frac{B - R}{V - min(RGB)} \tag{4.5}$$

Normally the HUE values are in the range of [0 360], but here they have been normalised to be in the range of [0 1]. The Hue channel of the breast core can be seen in (Fig. 4.20b).

To extract the cells two thresholds on two different channels of the HSV image were set, in the hue channel, any pixels that were below 0.75 were kept seen in (Fig. 4.20e), and in the saturation channel only pixels with a value below 0.3 were retained. Combining both thresholds results in the extraction of cells seen in (Fig. 4.20f).

To extract the fibres, the hue channel of the HSV image was looked at where only pixels with a value higher than 0.5 are kept, but less than 0.75 which is seen in (Fig. 4.20j), fibres appear in green. Secondly a second threshold is set where only pixels that have a value larger than 0.2 in the saturation channel are kept. Those two thresholds combined extract

the fibres seen in (Fig. 4.20k). Finally random Gaussian noise is added to make sure that the center of the fibres exhibits the highest intensity, preparing it to be traced (Fig. 4.20l).

To ensure that the gap analysis shown in (Fig. 5.34) is only within the each core's boundaries and its contents and not outside the core, a mask of the core boundaries was made, by using a large disk kernel of size $\mathbb{K} = 28$ in the dilation operation (Fig. 4.20g), holes inside the core are filled (Fig. 4.20h), then the core is eroded with a kernel of size $\mathbb{K} = 17$ (Fig. 4.20i). After which the objects in the image are labelled. As there was some noise that was on the outside of the core, dilating resulted in the noise forming objects. These small objects were removed by keeping only the object with the largest area.

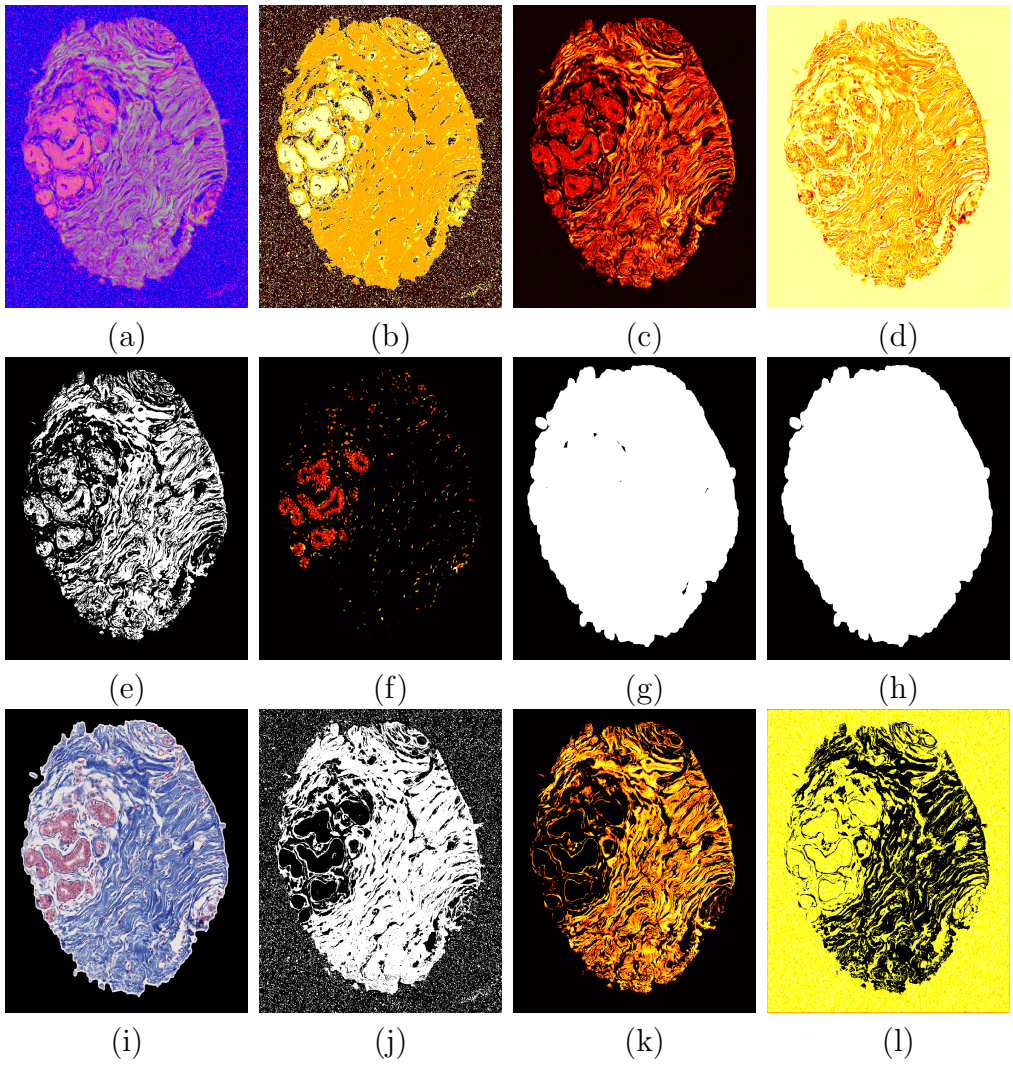


Figure 4.20: Illustration of the TMA cores preprocessing step to separate the cells from the fibres to be traced and segmented. (a) TMA core RGB image is converted to the HSV colour space. (b) The HUE channel of the HSV image. (c) Saturation channel of the HSV image. (d) Value channel of the HSV image. (e) Pixels that are above the values of 0.75 in the HUE channel for cell nuclei detection. (f) Combination of (e) and pixels with value below 0.3 in the saturation channel to separate the nuclei. (g) To only detect the contents of the actual cores and not other impurities in the image a dilation operation is ensued with a kernel of size 28 to form a mask of only the core. (h) Cores are filled to fill any holes inside. (i) An erosion operation is ensued with a kernel size of 17 to ensure the size of the mask matches only the contents of the mask. (j) Pixels higher than 0.5 but less than 0.75 in the hue channel belong to the collagen fibres. (k) A second threshold where only pixels higher than 0.2 in the saturation channel belong to the fibres, finally separating the fibres. (l) Noise is added to assist in the delineation of the fibres with Trace Ridges.

4.8 Summary:

This chapter explores image processing techniques to segment and delineate DAPI stained cell nuclei, Phalloidin stained actin filaments and fluorescently labelled fibronectin. This chapter also investigates the relationship between the cell nuclei and the actin filaments. Based on extracted morphological features this chapter also aims to investigate differences between the mutant and wild type populations. The cell segmentation algorithm achieved an accuracy of 0.9859. The mean Jaccard index was 0.7924 and the mean Dice Score 0.8834. When comparing the cell nuclei of the mutant and wild type only the standard deviation of the maximum Feret angle had differences. After successfully segmenting the actin filaments islands and comparing the mutant and wild type populations, two properties showed differences those were number of actin filaments islands and the average area of the actin filaments islands. When combining the segmented cell nuclei and actin filaments islands and comparing the two populations there were differences in the relationship of the cells and the actin filaments orientation. After the successful delineation of the fibronectin, the mutant and wild type populations were compared with each other. Two properties showed differences, and these were the relative major axis length of the fibres and the relative total area between the fibres. The analysis on the fibronectin proved promising and prompted further study. This leads on directly to the next chapter where it was identified that Edge Detection is not the optimum algorithm to trace fibre like structures therefore an objective comparison is needed to determine which algorithm traces fibre-like structures best. Other image pre-processing pipelines are outlined, including which images had no pre-processing applied to them. One of the experiments outlined in (Section 5) uses the FF images, the preprocessing of these images was outlined here. Additionally, another investigation requires pre-processing of dataset 3 which was just filtered with a simple averaging kernel. Images of dataset 4 are made up of TMA cores, this chapter outlines how each core was isolated and how the fibres and cell nuclei in those images were distinguished from each other. This exploration of data also set the tone for the work done in the thesis and the contributions.

Chapter 5

Results:

5.1 Overview:

This chapter contains the main contributions of this thesis. Trace Ridges, a novel approach is presented in this chapter. It is followed by the comparison of the tracing algorithms presented in (Section. 5.3.5) which has been published in PLOS ONE [357], the analysis of the modified cell line was published as a conference article [37], the analysis of the Marangoni flows, protease inhibitors and increasing collagen concentration contributed to the analysis published by Maiques [358] and the TMA analysis is expected to be submitted as a journal paper by the end of April.

5.2 Trace Ridges:

In this thesis, a novel tracing method *Trace Ridges* is proposed, a method combining Watershed for ridge identification and Edge Detection which *breaks* any ridges that run from a main ridge towards the sides of the basins, separating them from the minor ones. It prioritises brighter, longer ridges for cleaner results. Trace Ridges accepts images of any dimensionality, converting 3D images to grayscale and the maximum intensity projection is taken. Trace Ridges requires minimal input from the user.

5.2.1 Algorithm Steps:

A detailed view of the algorithm follows:

The image $\mathbb{I}$ is read, and the size of the image is checked $N_r \times N_c \times 1$ for grayscale images and $N_r \times N_c \times 3$ for RGB images. If the image is grayscale, the image is kept as is and follows through the rest of the pipeline. If the image is RGB, the image is converted into grayscale using a weighted sum of each colour channel based on how each colour's brightness is perceived by the human eye, with green being the most sensitive. The coefficients used to calculate the grayscale values are the same as those used to calculate luminance.

The new image $\mathbb{I}'$ can be derived as such (Eq. 5.1).

$$\mathbb{I}'(x,y) = 0.2989 \cdot I_R(x,y) + 0.5870 \cdot I_B(x,y) + 0.1140 \cdot I_G(x,y) \tag{5.1}$$

Each pixel value intensity $\mathbb{I}'(x,y)$ is derived by multiplying the coefficients with the intensities present in each of the colour channels for the same pixel location (x,y)

For images that have more than three channels, the maximum intensity projection (MIP) is obtained (Fig. 5.5a). The maximum intensity projection is the maximum value of intensity I of the pixel at each (x,y) co-ordinate across all channels of the image and returns the highest value. Therefore, the image shape changes from $N_r \times N_c \times 3$ to $N_r \times N_c \times 1$ and the value at each pixel (x,y) is the value returned by the MIP. This can be represented as such (Eq. 5.2), where $x \in 1,2,...,N_r$, $y \in 1,2,...,N_y$, and $k \in 1,2,...,k$.

$$\mathbb{I}'(x,y) = \max_k \mathbb{I}(x,y,k) \tag{5.2}$$

A threshold, distGap is set at 1.5 pixels below which ridges are considered low level or dark and therefore are ignored. The Gaussian filter kernel $\mathbb{K}$ size is set at 3.

The image is filtered with a Gaussian kernel as such (Eq. 3.8)

$$\mathbb{I}'(x,y) = \sum_{i=-a}^{a} \sum_{j=-a}^{a} \mathbb{I}(x+i, y+j) \cdot G(i,j)$$

and since it is a 3x3 kernel, the values of the kernel are

$$\begin{bmatrix} (-1,-1) & (0,-1) & (1,-1) \\ (-1,0) & (0,0) & (1,0) \\ (-1,1) & (0,1) & (1,1) \end{bmatrix}$$

With a σ value of 1, the Gaussian filter function can be simplified to (Eq. 3.7).

$$G(i,j) = \frac{1}{2\pi} \exp\left(-\frac{i^2+j^2}{2\sigma^2}\right)$$

The Gaussian filtering on the image can be seen in (Fig. 5.5b), and the image and its contents became smoother and brighter.

5.2.2 Watershed Transform:

Watershed Transform is applied to the Gaussian filtered image where it views the image as a topographic surface. Where higher intensity pixels are peaks or ridges and form the foreground while low intensity pixels are those of background and looked at as basins (Fig. 5.5c).

Firstly, the filtered image $\mathbb{I}(x,y)$ is converted to a gradient magnitude image as such (Eq. 3.44),

$$M(x,y) = \sqrt{\nabla_x^2(x,y) + \nabla_y^2(x,y)}$$

∇_x and ∇_y are the partial derivatives of the image with respect to x and y. Partial derivatives highlight areas of intensity changes which correspond to the ridges or edges.

A set of markers is set where each marker is for a different edge or region in the image. This is done by using the inverse of the Euclidean distance transform to highlight regions where objects exist not the background, it could also be found using thresholding. Markers are set as such (Eq. 3.50)$\mathbb{M}(x, y) = \mathbf{R}$ where pixel (x, y) belongs to the region $\mathbf{R}$.

The image is then "flooded" as mentioned earlier. When "water" from different regions or markers meet which is where the basins are, a boundary is placed to separate the regions. The process continues until all regions have been "flooded" and all boundaries have been identified. The watershed lines or ridges represent the boundaries between regions. In this case the watershed lines trace the centre line of the fibres, and they are called ridges. Watershed Transform has an affinity to over-segment if an image is noisy.

5.2.3 Edge Detection:

Edge Detection is applied to the Gaussian filtered image (Fig. 5.5d) to extract all the regions where sudden intensity changes have occurred, indicating the presence of fibres where they appear as two lines, ideally one on each side of the centre line of the ridges. The intersection between the edges of Edge Detection and the ridges of Watershed are used to break the minor ridges that run down from a main ridge towards the sides of the basins, to separate them from the major ridges.

The first step in Edge Detection is Gaussian filtering, as the image has been previously convoluted with a Gaussian kernel of size 3. Minimal filtering in Edge Detection is applied, a kernel $\mathbb{K}$ of size 2 is used. Matlab was used to process the Canny Edge Detection, with $\sigma = \sqrt{2}$ as the default and was kept as is. The image is filtered as such (Eq. 3.39).

$$\mathbb{I}'(x, y) = \sum_{i=-a}^{a} \sum_{j=-a}^{a} \mathbb{I}(x + i, y + j) \cdot G(i, j)$$

The spatial derivatives of the Gaussian filtered image in the x and y directions are extracted and could be represented as such in the x direction (Eq. 3.42) and for the y direction (Eq. 3.43).

$$\nabla_x(x, y) \approx \sum_{i=-a}^{a} \sum_{j=-a}^{a} \mathbb{K}_x(i, j) \cdot \mathbb{I}(x + i, y + j)$$

$$\nabla_y(x, y) \approx \sum_{i=-a}^{a} \sum_{j=-a}^{a} \mathbb{K}_y(i, j) \cdot \mathbb{I}(x + i, y + j)$$

These convolutions provide approximations of the partial derivatives ∇_x, ∇_y. $\mathbb{I}(x+i, y+j)$ represents the intensity of the neighbouring pixels around (x, y) and $\mathbb{K}_x(i, j)$ and $\mathbb{K}_y(i, j)$ are the values of the Sobel kernel at the position (i,j).

Sudden changes in intensity present in high derivative values indicate the presence of edges. The magnitude of the gradient which is the strength of the change of intensity at each pixel can be computed as such (Eq. 3.44) and the gradient direction as such (Eq. 3.45).

$$M(x, y) = \sqrt{\nabla_x^2 + \nabla_y^2}$$

$$\theta(x, y) = \tan^{-1}\left(\frac{\nabla_y}{\nabla_x}\right)$$

The magnitude informs which edges are strong and which edges are weak, and direction reveals the orientation of the edge at that pixel measured in degrees or radians.

Non-maximum suppression is applied to thin the edges by keeping only the local maxima along the gradient direction. This is done by checking the magnitude of the current pixel (x, y) in the gradient direction for example if $\theta(x, y)$ is $0°$, a comparison of $M(x, y)$ to $M(x-1, y)$ and $M(x+1, y)$, if $M(x, y)$ is not a higher value than the pixels in its direction, it is set to 0.

Double thresholding is then applied to identify the strong and weak edges. Strong edges are kept while weak edges are discarded if they are not connected to a strong edge. The threshold values are obtained heuristically and automatically by Matlab depending on the image.

The high and low thresholds can be represented as such, T_h and T_l respectively.

Strong edges can be represented as (Eq. 3.46), weak edges (Eq. 3.47), and non-edges (Eq. 3.48).

Hysteresis is then applied to determine which edges to retain or remove, if a weak edge is connected to a strong edge pixel, it is kept as part of an edge, if not, the edge is discarded.

5.2.4 Watershed Combined with Edge Detection:

The intersection between the edges and Watershed lines is used to break the minor ridges that come from a main ridge towards the sides of the basins; therefore, they separate from the major ridges. This solves and counteracts the well-known problem of the affinity of Watershed to over-segment. This method prioritises brighter regions of the image and longer ridges to give a clearer and cleaner trace.

To accomplish this, we perform point element-wise multiplication between the Watershed ridges image and the output of the edge detection image, specifically where no edges are present. This ensures that any ridges identified by the Watershed algorithm are constrained within the boundaries defined by the detected edges, preventing ridges from extending beyond the Edge Detection lines. This image can be referred to as allRidgesNoEdges (Fig. 5.5e).

The combination of these two steps of Trace Ridges is shown in (Fig. 5.1a). The region has three strong and bright ridges with a small, distorted region highlighted with a green rectangle. The results of Watershed and Edge Detection are illustrated separately in (Fig. 5.1b, c) and the overlap in (Fig. 5.1d) where the edges are shown in red, and the

watershed is in yellow. The intersections where the two overlap are shown as white. These points are discarded thus breaking the small watersheds and maintaining the main ridges. Edge detection traces two fibres for the same fibre as also seen in (Fig. 5.2). In (Fig. 5.2), a fibre image with a green rectangle indicating an ROI for inspection. Intensity changes occur in the green rectangle; this is illustrated by observing the intensity profile using "improfile". The intensity is at first very low as it is in the background. Once it moves onto the yellow fibre there is a sharp increase in intensity after which it stays stable until there is a sharp decrease in intensity as it went from an area of high intensity (foreground) to an area of low intensity (background). The derivative of the changes in intensity is shown in red. Edge Detection detects these changes in intensity and as there are two derivatives, Edge Detection traces two lines per fibre. Watershed (Fig. 5.3a) treats the image or the objects of an image as a topographic surface whereby objects of an image that are elevated will be treated as foreground pixels around which the Watershed lines occur. Traces running down from the peaks to the basins must be removed therefore Edge Detection is used to break off these Watershed lines. Thresholding can then be applied (Fig. 5.3b), where any ridges below the intensity of the horizontal line threshold are removed.

5.2.5 Tracing clean-up:

A distance transform is applied to the output binary image BW of the Edge Detection. The distance transform assigns a value to each pixel in the image that is a measure of distance to its closest non-zero pixel. Therefore, pixels that are further away from the edges will have a high value while pixels near or on edges will either be 0 or have low values. In (Fig. 5.5f) pixels from the fibre to the furthest point from the fibre follow the colours red-orange-yellow-white. The distance transformation can be represented as outlined in (Eq. 3.33).

$$EucD(x,y) = \min_{(x_f,y_f):\, BW(x_f,y_f)=1} \sqrt{(x-x_f)^2 + (y-y_f)^2}$$

Here (x, y) represent the pixel co-ordinates of the binary image BW. (x_f, y_f) are the co-ordinates of a foreground pixel in BW where $BW(x_f, y_f) = 1$. The Euclidean distance transform assigns a value of the shortest distance to each pixel (x, y) to its closest (x_f, y_f).

A point element-wise multiplication of the allRidgesNoEdges image and the distance transform image with the condition of only keeping pixels that have a distance less than 5 pixels, creating a binary mask. This ensures that only the main ridges are kept, and the Watershed lines of the basins are removed. This image can be called allEdgesNoRidges2 (Fig. 5.5g). This can be represented as

$$\mathbf{m}(x,y) = \begin{cases} 1 & \text{if } EucD(x,y) < 5 \\ 0 & \text{if } EucD(x,y) \geq 5 \end{cases}$$

Here the mask $\mathbf{m}$ and the image product can be represented as (Eq. 5.3).

$$\mathbb{I}' = \mathbb{I}(x,y) \cdot \mathbf{m}(x,y) \tag{5.3}$$

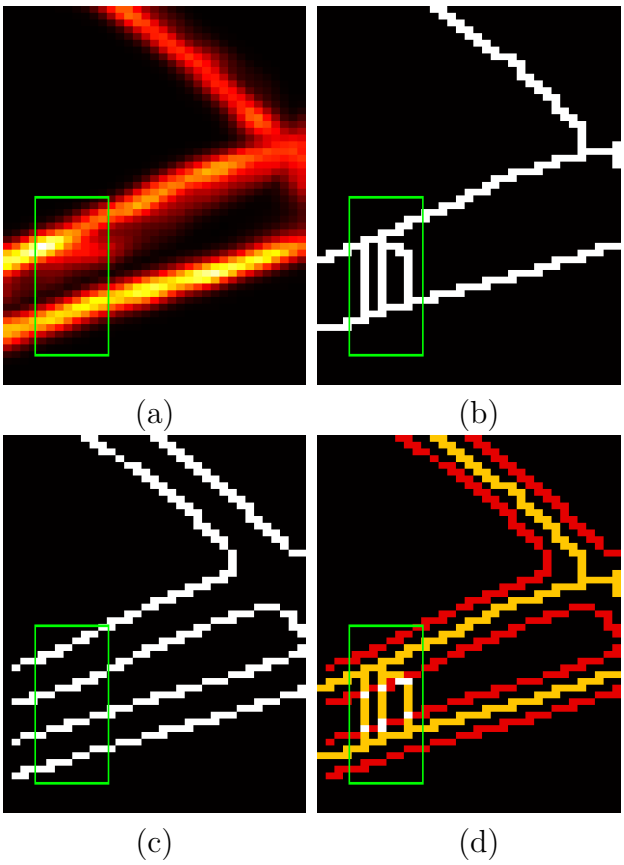


Figure 5.1: Illustration of the Trace Ridges methodology. (a) Region of interest of a region of interest of the FC image that has strong ridges and a region with distortion highlighted with a green rectangle. (b) The result of a Watershed transform, that traces the strong ridges but various smaller ones that are not of interest. (c) Edges detected with Canny's algorithm showing the sides of the strong ridges. (d) The combination of Edge Detection (red lines) with Watershed (yellow lines). White pixels correspond to the intersection of both techniques. The Watershed will be broken by removing these intersecting pixels.

The next step is to remove the darker regions of the Watershed, as Trace Ridges prioritises longer and brighter ridges for a cleaner delineation. The ridges in the dark regions are not of interest as they are hardly visible.

To remove the darker regions of the watershed a global Otsu thresholding is applied which aims to maximise the between class variance represented in (Eq. 3.10).

$$\sigma_B^2(T) = w_1(T) \cdot w_2(T) \cdot [\mu_1(T) - \mu_2(T)]^2$$

$w_1(T)$ is the weight or probability of class 1 where pixels value range between 0 and T, similarly, for $w_2(T)$, where pixel values range between T+1 and 255. $mu_1(T)$ and $mu_2(T)$ are the mean intensities of class 1 and class 2, respectively.

To ensure that the Otsu threshold is scaled to the intensity range of an 8-bit grayscale

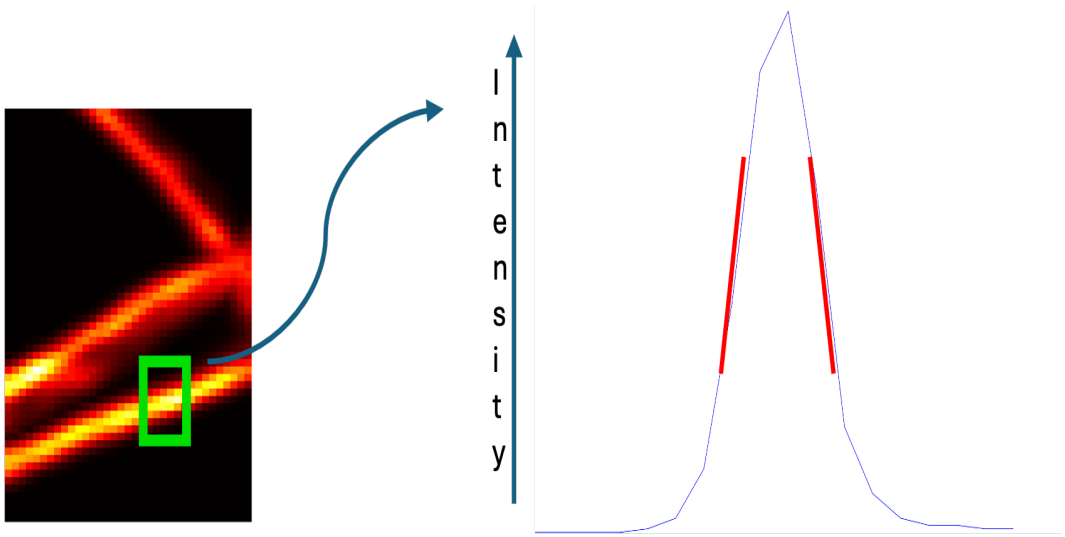


Figure 5.2: Edge Detection intensity profile and derivatives. (a) A sample fibre image, with a green rectangle box indicating a region of interest of a section of the fibre and its surrounding. Intensity profile using Matlab "improfile" illustrating the changes from top to bottom of the green box. The intensity changes as it goes from the surrounding background into the fibre and again out of the fibre into the background. The red lines indicate the derivative which are the edges Edge Detection detects.

image, the scale is multiplied by 255. This multiplication converts the normalised threshold value (0, 1) into a corresponding threshold in the range of pixel intensities. The threshold value is further scaled by 0.9 to make it more sensitive to lower-intensity pixel values to detect more ridges that could have been overlooked with the Otsu threshold. The scaled threshold can be represented as T.

From allRidgesNoEdges2 we only keep the features and the ridges that are present in the input grayscale image that are above the threshold.

$$\mathbb{I}' = \mathbb{I}_1 \cdot (\mathbb{I}_2 > T) \tag{5.4}$$

In (Eq. 5.4), $\mathbb{I}_1$ is the allRidgesNoEdges2 image and $\mathbb{I}_2$ is the grayscale input image. The product image is referred to as allRidgesNoEdges3.

To clean up the delineation further, spurious ridges from Watershed are removed. Skeletonisation reduces the ridges into 1-pixel wide lines. Skeletonisation preserves the overall shape and connectivity of the objects and helps reduce the data complexity. Skeletonisation of the binarised allRidgesNoEdges3 is conducted and can be represented as such (Eq. 3.34).

$$\text{skel}(BW) = \{(x, y) \mid EucD(x, y) = \max(EucD(x_f, y_f)) \text{ for neighbours } (x_f, y_f)\}$$

(x, y) belongs to the skeleton if the $EucD(x, y)$ is equal to the maximum distance $\max(EucD(x_f, y_f))$ among the pixel neighbours (x_f, y_f).

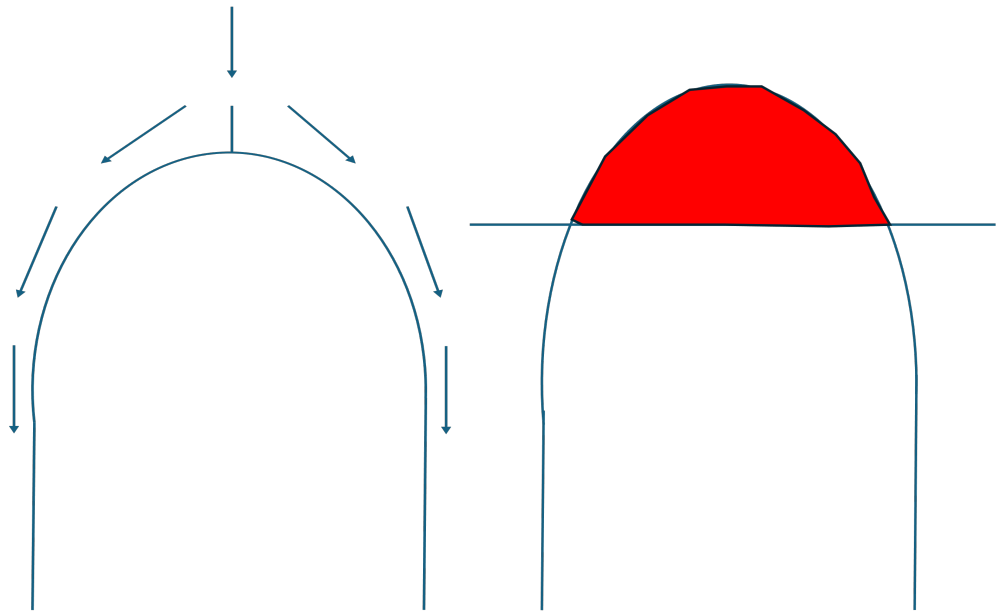


Figure 5.3: Illustration of the Watershed and Thresholding methods. (a) The Watershed method traces lines of ridges, and the traces run off to the basins at the bottom of the ridges. The arrows represent the flow from the top of a peak to the basins. (b) Thresholding shown by the horizontal line drawn across the object. Only portions of the object that are above the line have been kept.

Only branches that are longer than 5 pixels are kept ensuring the retention of the skeletal structure, while removing artefacts from the image (Fig. 5.5h).

Large ridges are identified as ridges that have a value larger than 5 in the major axis length. In some cases, a thicker ridge may produce two parallel watershed lines with a small hole in between those lines. These cases are detected by their Euler characteristic. When there are no holes, the Euler number is 1, if there are holes, the Euler number will be less than 1. These holes are filled in using the Matlab function "imfill". Which floods the holes in the binary image. The function defines holes as a set of background pixels that would not be reached by filling the background around the edges of the ridges. The filling operation could also be done by "closing", whereby dilation occurs to fill the holes followed by erosion to retain the original structure of the ridges. Closing can be represented as (Eq. 3.20).

$$\text{closing}(I, SE) = (I \oplus SE) \ominus SE$$

Where dilation $D(x, y) = (I \oplus SE)$ increases the foreground pixels followed by, $E(x, y) = (D \ominus SE)$ which removes some pixels of the foreground. Dilation (Eq. 3.19) is $D(x, y) = \max\{I(x+x_i, y+y_i) \mid (x_i, y_i) \in SE\}$ and Erosion (Eq. 3.17) is $E(x, y) = \min\{I(x+x_i, y+y_i) \mid (x_i, y_i) \in SE\}$.

To ensure that parallel lines with a hole between them are represented accurately. The holes were filled as mentioned, which indicates that they are between those two, these lines

are subsequently filled. This filled object is now much larger than a ridge therefore these objects must be thinned to leave a trace that runs through the position of that small hole that represents the thicker ridges (Fig. 5.5i).

The filled holes are identified and separated from the rest of the ridges (Fig. 5.5j). They are thinned using erosion. All ridges are recombined together (Fig. 5.5k) and spurious objects that could have been generated by the morphological operations are cleaned (Fig. 5.5l). Finally, the large ridges are labelled satisfying the following condition, if $BW(x, y) = 1$. Initially, labels are set at 0 where they are iteratively increased as more objects in the binary image are labelled (Eq. 3.27).

To further clean the delineation, Trace Ridges discards small and short fibres by setting three different thresholds of combinations of the major axis length len and the minor axis length (2wid). The ridges that are kept satisfy the following thresholds (Eq. 5.5), (Eq. 5.6) and (Eq. 5.7).

$$T_1 = \{L_k \mid len(L_k) > 10 \text{ and } 2wid(L_k) < 3 \tag{5.5}$$

$$T_2 = \{L_k \mid len(L_k) > 20 \text{ and } 2wid(L_k) < 10\} \tag{5.6}$$

$$T_3 = \{L_k \mid len(L_k) > 20 \text{ and } 2wid(L_k) \geq 10\} \tag{5.7}$$

L_k is the labelled object. For ridges that satisfy T_3 since they are long and wider than other ridges and have branches, their branches are identified and removed (Fig. 5.6a). Only ridges that are longer than four pixels are kept to further produce a cleaner delineation (Fig. 5.6b). The result of the three thresholds can be seen in (Fig. 5.5m), (Fig. 5.5n) and (Fig. 5.5o). All the ridges from the three thresholds are recombined together (Fig. 5.6c).

5.2.6 Missed ridges

There are instances where Watershed fails to identify and miss a ridge, particularly when both sides of the ridge contribute to a single surrounding basin. For these cases, ridge detection is accomplished through Edge Detection. This involves looking at the edges present through Edge Detection but not in the cleaned final form ridges from Watershed (Fig. 5.6d).

This can be represented by (Eq. 5.8).

$$\mathbb{I}'(x, y) = \mathbb{I}_1(x, y) \cdot (1 - \mathbb{I}_2(x, y)) \tag{5.8}$$

Where $\mathbb{I}_1$ is the Edge Detection image and $\mathbb{I}_2$ is the image of the final form of ridges from Watershed.

The ridges detected solely through Edge Detection are filled (Fig. 5.6e) and thinned resulting in one singular line representing the ridge, as typically Edge Detection results in two ridges for the same fibre (Fig. 5.6f). If the Edge Detection results in only one edge for a ridge, these are discarded and are treated as low level ridges or noise.

Moreover, Trace Ridges eliminates small and curved fibres, focusing on retaining longer and straighter ridges from Edge detection (Fig. 5.6g). If Trace Ridges identifies fewer than 100 missed ridges it only retains those that have a *len* above 15 pixels and a $2wid$ less than 10 pixels. If there are more missed ridges, a stricter rule is applied where the ridges must have a *len* of at least 25 pixels and a $2wid$ less than 10 pixels (Fig. 5.6h).

The ridges detected from Watershed are further looked at and their branch points are removed (Fig. 5.6i), in some instances the ridge structures are curled and can be broken into two separated ridges at the junctions. Those junctions can be found by looking at the corner points of the ridges. The corner points are detected using the Matlab function "detectHarrisFeatures" and the minimum accepted quality of corners was set at 0.45 the range is between $[0, 1]$. The minimum accepted quality of corners is a metric based on the strongest corner in the image, in this case 0.45. This means that only corners kept whose values is at least 45 percent of the strongest corner metric in the image are kept.

5.2.7 Ridges Combination:

Once all the ridges are detected, Trace Ridges merges the ridges detected with the combination of Watershed and Edge Detection as well as the missed ridges detected by Edge Detection, prioritising longer and brighter fibres (Fig. 5.6j). Many morphological metrics can be extracted from the traced fibres which are further explained in (Section 3.8.4).

5.2.8 Trace Ridges Implementation:

Trace Ridges' implementation is straight forward as it is available as a function. The input image is chosen and is the only parameter required, $[fibronectinOut, fibronectinOut2, dataOut] = Trace_Ridges(\mathbb{I}, cannySize, distGap)$. The canny size for the Edge Detection can be set but if it is not set, the default is 2 as earlier mentioned, as well as the distGap can be set but if it is not set then the default 1.5 is used. The steps taken by Trace Ridges are illustrated in (Fig. 5.4).

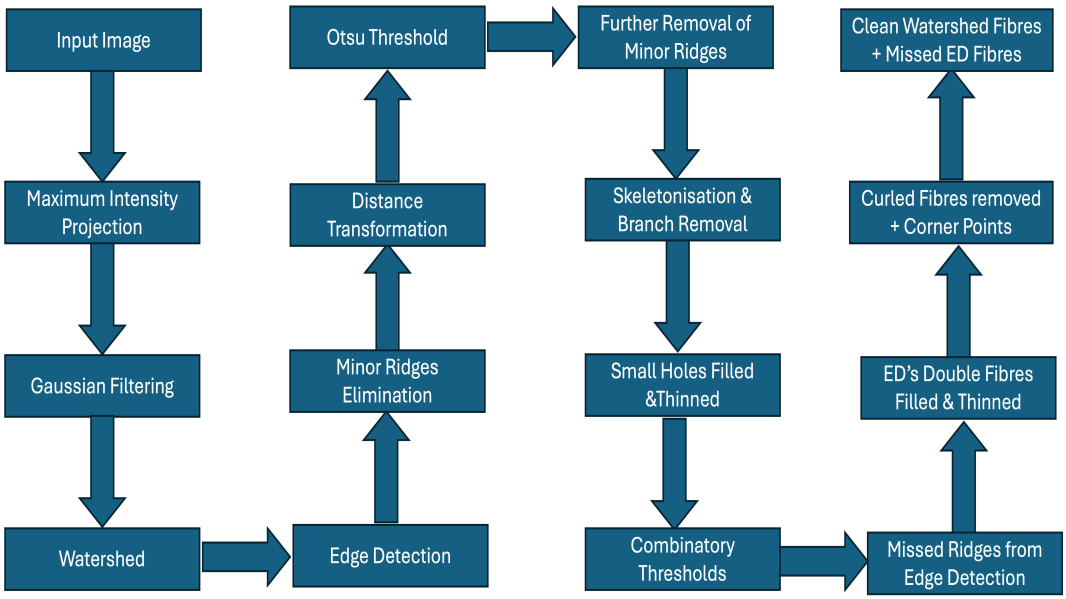


Figure 5.4: Trace Ridges pipeline follows: The maximum intensity projection of the input image is extracted. The MIP image is then smoothed. Watershed lines are extracted as well as Edge Detection. Edge Detection lines break the Watershed lines that are running towards the basins and thus eliminates the minor ridges. A distance transform is applied followed by Otsu thresholding which results in further removal of minor ridges. The fibre structures are skeletonised and the branches are removed. When two lines that belong to the same fibre are in close proximity, they form a small hole between them which is filled and thinned resulting in one line to represent the fibre. Three levels of threshold are applied to remove minor ridges and keep only fibres that are long and bright. Ridges missed by Watershed are detected by Edge Detection, where the proxies of fibres from Edge Detection are filled and thinned to leave only one ridge to represent the fibre. Curled fibres are then removed, and corner points are identified. The delineation is cleaned further to remove any minor ridges to give the final fibre delineation.

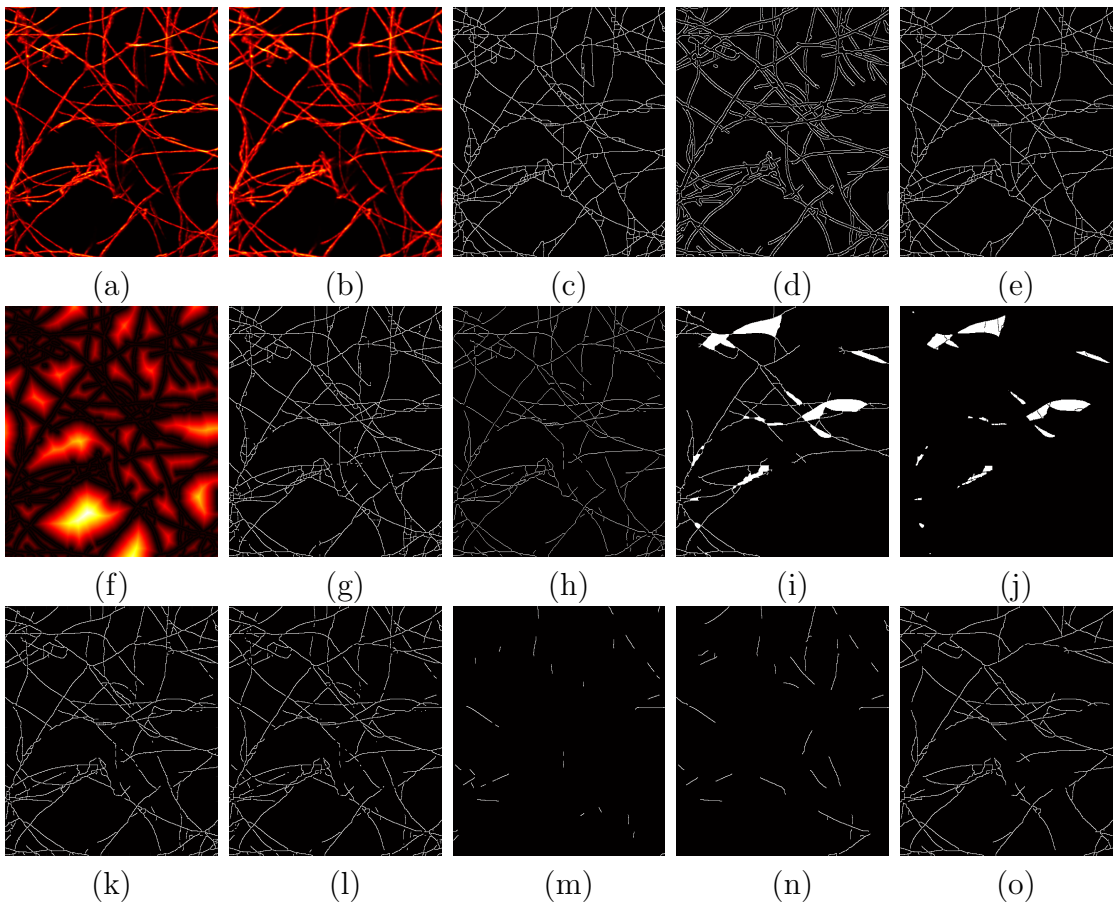


Figure 5.5: Trace Ridges algorithm steps to trace an image of fibre-like structures. Image used is the FC image from Dataset 2. (a) Maximum intensity projection of the input image. (b) Gaussian filtering of the MIP. (c) Watershed lines of the smoothed image. (d) Edge Detection application on the smoothed image. (e) Breaking of the Watershed lines by the Edge Detection lines to separate minor ridges. (f) The application of the distance transform on the binary image of the Edge Detection. (g) Only pixels kept are those that have a distance less than 5 pixels from the fibres. (h) Skeletonisation of the image in (g) to remove further minor ridges. (i) Small holes between traces of the same fibres are filled. (j) Filled holes are identified and separated. (k) Filled holes are thinned which result in one trace per fibre and they are recombined with the other ridges. (l) Spurious objects are removed, and the delineation is further cleaned. (m) T_1 threshold applied on Trace Ridges. (n) T_2 threshold applied on Trace Ridges. (o) T_3 threshold applied on Trace Ridges.

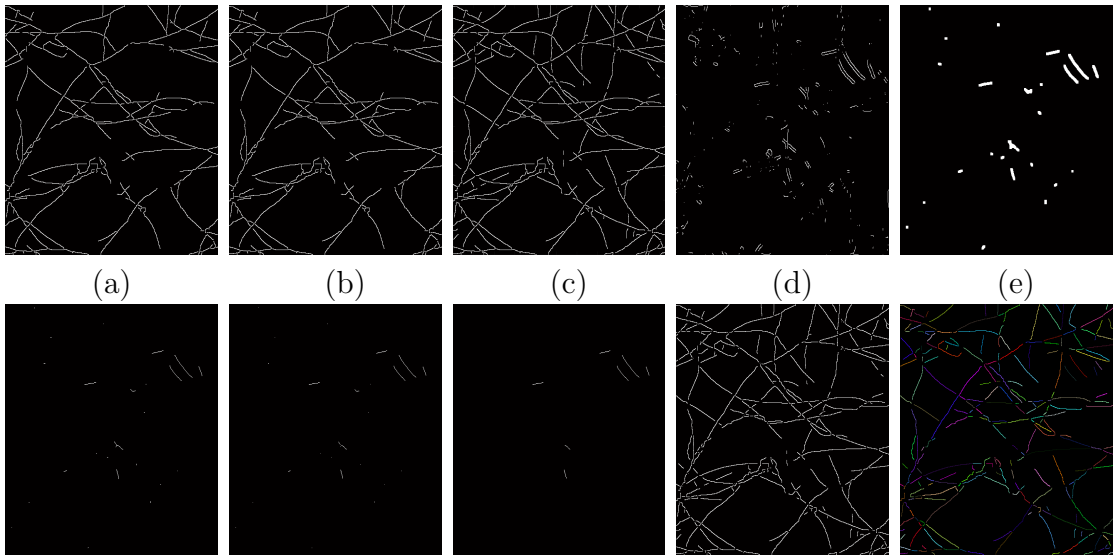


Figure 5.6: Trace Ridges algorithm steps to trace an image of fibre-like structures continued. (a) Branches from T_3 are identified and removed. (b) Only ridges that are longer than four pixels are kept. (c) The ridges from the three thresholds are recombined together. (d) Edges from Edge Detection not detected by the Watershed. (e) Edges from Edge Detection are filled. (f) Filled edges are thinned to give one trace for each fibre. (g) Curved fibres are removed. (h) Thresholding of missed ridges based on their count. (i) Branch points of ridges from Watershed are identified and removed. (j) Final ridges detected by Trace Ridges.

5.3 Tracing Algorithms Comparison:

Excerpts of this section have been published to appear as a journal article in PLOS One [357]

5.3.1 Introduction:

The structural characteristics of cells and their surrounding environment have been studied with interest in cancer research and other diseases [25]. However, progressively attention is being given to the microenvironment that surrounds these cells and its major influence on the cellular behaviour [359]. A major component of the microenvironment is the Extracellular matrix (ECM), which holds an important role in cancer progression [27, 28, 29, 30] and other conditions [26]. The ECM provides structural support [31] to tissue cells. The ECM may appear in different shapes and structures; each organ has a particular ECM structure that supports its purpose. The ECM is able to modify its shape and remodel with the help of ECM proteinases [360, 361]. Despite its importance, the ECM has received less focus and attention [33] compared to other microenvironmental such as macrophages, soluble factors, or blood vessels. This could be attributed to its complex three-dimensional networks [34, 35], which is made up of many elements such as enzymes, glycoproteins, and collagen.

While visually analysing the ECM structure can reveal important insights, simply observing the images with the eye can make it challenging to compare different populations with each other. Structural properties such as fibre length, distribution, shape and orientation could provide valuable information about the microenvironment and treatment effects if it were objectively analysed. Therefore, accurately identifying and quantifying these fibre-like structures requires robust tracing algorithms that are able to detect singular fibres.

Several methods, both manual and automatic, have been developed for tracing elongated fibrial structures. Some are general methods that can be applied beyond tracing, for instance, skeletonisation [313, 314, 315], Watershed [316, 317], medial axis transform [318, 319, 320], Scale Space [45, 46] and Edge Detection [44]. However, many existing techniques have been optimised for specific applications, such as neuron tracing [321, 322, 323], retinal imaging [324, 325, 326], angiography [327, 328, 329], intravital vasculature observed with confocal, fluorescent, and light microscopy [330, 331, 332], and extracellular matrix [333, 47]. In recent years, deep learning models and in particular U-Net [334] have been widely used in image analysis. U-Net has achieved excellent results without requiring hand-crafted features or algorithm. Instead, these models learn using pairs of data and labels to perform the segmentation tasks, which in this case, involves detecting fibre-like structures.

This study presents an objective comparison of various fibre tracing algorithms as it is essential to ensure fair evaluation, facilitate method selection, and promote reproducibility in fibre analysis research. To ensure variability and diversity, four types of images with varying noise levels were selected, they have been mentioned before but for clarity the images are replicated here (Fig. 5.7).

The images from dataset 2 have very distinctive characteristics. Images such as (Fig. 5.7b) and (Fig. 5.7d) contain many fibres and are noisier by nature. The fibres in those images

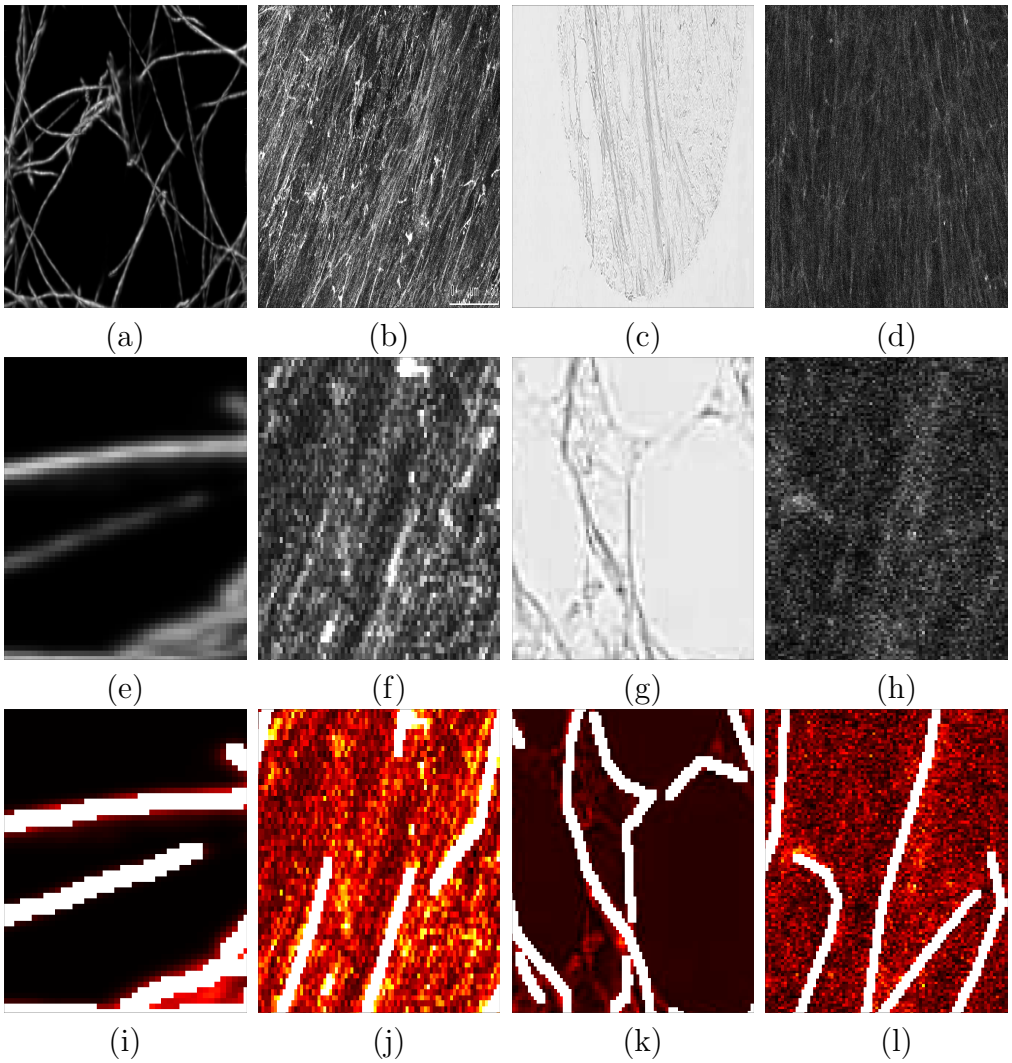


Figure 5.7: Images used in objective comparison, Dataset 2: Various images of extracellular matrix fibres from different microscopical methods. These images are used to compare the tracing algorithms (a) Fluorescently labeled Collagen (FC). (b) Fluorescent Fibronectin (FF), image courtesy of Dr Esther Castellano, Dr Cristina Cuesta-Apausa Universidad de Salamanca. (c) Breast Cancer Biopsy slide (BCB). (d) Disease Mimicking Extra Cellular Matrix (DME). Images (a), (c) and (d) are open source. (e-h) Regions of interest (ROI) of the images in (a-d). (i-l) Manually delineated ground truth (GT) for the ROIs of images in (e-h). White lines are those of the GT and intensities are displayed in a colormap with hot colours (black-red-orange-yellow-white).

may be difficult to differentiate with the eye. While image (Fig. 5.7a) is much clearer to see and simpler as it does not contain many fibres.

The ground truth data mentioned earlier was used for the objective evaluation between the algorithms performance and to serve as training data for the U-Net (Fig. 5.7i-l). Six

algorithms including our novel approach *Trace Ridges* that are used for the application of fibrial tracing are evaluated: Edge Detection [44], CT Fire [48], Scale Space [45, 46], Twombli [47] and Graph Based [49, 50]. Additionally, a U-Net [334] was trained using patches of data and labels.

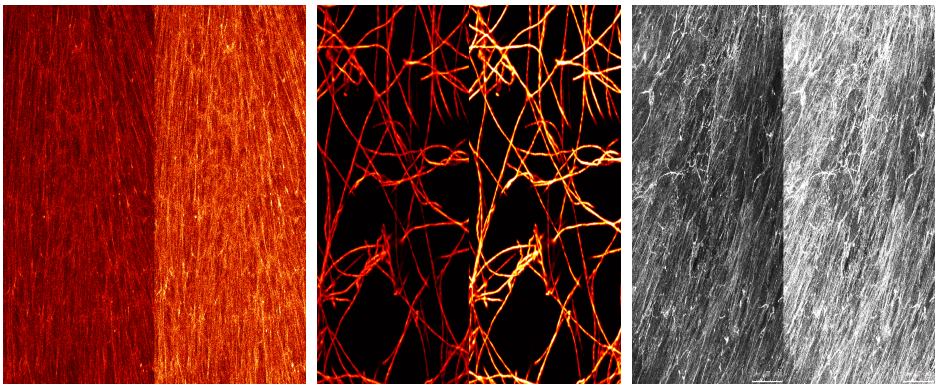
No filtering indicates that the images were not pre-processed with any filter before being subjected to the algorithms. The Gaussian filter implemented here used a kernel size of $\mathbb{K} = 3$ with a $\sigma = 1$. DnCnn was employed as described in (Section. 3.5.1).

A secondary investigation explores the effect of High Pass filtering on algorithm performance and was implemented as mentioned in (Section. 3.5.1).

Another filtering option is the use of a high pass filter which intends to sharpen an image and enhance its features by removing low frequency components.

Generally high pass filters add noise to images therefore they would affect the tracing of fibrial structures and cause the algorithms to incorrectly trace. In order to justify the reasoning of only comparing the algorithms with no filtering, low pass Gaussian filtering and DnCnn filtering. A small study is conducted to compare 4 of the mentioned algorithms (Scale Space, Edge Detection, Trace Ridges and U-Net) with four different filtering methods (No filtering, Gaussian filtering, DnCnn and High Pass Filtering (HPF) across three of the images (DME, FC and FF) to observe the effect of high pass filtering on the tracing of algorithms when it is compared to the ground truth.

The three images used for the investigation with the high pass filtering can be seen in (Fig. 5.8). Here the image features which are the fibres appear to be sharpened.



(a) (b) (c)

Figure 5.8: The effect of applying a high pass filter to the DME, FC, FF images. The images are split in half where the left hand side of the images is the original unfiltered image and the right hand side is the high pass filtered image. The images features appear to be brighter and sharper. (a) DME image under high pass filter, fibres in this image appear brighter. (b) FC image under high pass filtering, fibres appear brighter and sharper hence the yellow colour. (c) FF image under high pass filtering, fibres appear whiter indicating they are brighter and sharpened.

5.3.2 Quantitative comparison:

To determine which method traces fibres more accurately quantitative metrics comparison between the ground truth and the results of the algorithms were calculated. Four quantitative metrics were assessed: total distance error, average distance error, maximum distance error, and computational time (Section. 3.8.4). The distance errors reflect how closely each algorithm's trace results matched the manually delineated ground truth.

Total Distance Error:

The initial comparison examines total distance errors, the cumulative distance errors per trace from the GT to the algorithm trace, generated by the distance maps as explained. A summary of the quantitative results is provided in Table 5.1, comparing algorithms with and without filtering. Filtering significantly improves the performance of Edge Detection, Scale Space, Twombli, Graph Based and Trace Ridges algorithms. Both U-Net and CT Fire show increased errors after filtering, due to U-Net training on non-filtered images, and CT Fire applies its own internal filtering. Notably, U-Net performs better on BCB and FC images when filtered with DnCnn, though using DnCnn filtering for other images increases U-Net's errors, impacting its overall effectiveness of tracing. Considering filtering-algorithm pair combination as separate distinct methods, U-Net without filtering performs best, followed by Graph Based with DnCnn, Trace Ridges with Gaussian filtering, Trace Ridges with DnCnn, and Trace Ridges without filtering. Averaging total distance errors across all images, regardless of filtering, reveals that Trace Ridges performs best overall, followed by Graph Based, U-Net, Twombli, Scale Space, CT Fire, and Edge Detection. Yet, Trace Ridges does not always perform the best individually, as seen with some of the more complex images such as the BCB image where U-Net outperformed all other methods. Another limitation of Trace Ridges, is if there is a fibre where both sides lead to the same basin, the Watershed algorithm would not identify it as a fibre.

Method		Images							
	Filter	DME	FC	FF	BCB	Average	Rank	avg/Method	Rank
ED	No Filt	947,550	35,905	754,120	224,120	490,424	21	429,173	7
	Gaus Filt	834,660	36,248	714,530	208,090	448,382	20		
	DnCnn	620,970	32,591	612,600	128,690	348,713	19		
CTF	No Filt	311,580	13,594	232,140	83,653	160,242	16	171,529	6
	Gaus Filt	342,290	22,687	269,290	79,694	178,490	18		
	DnCnn	330,190	29,437	266,950	76,848	175,856	17		
SS	No Filt	206,620	15,654	318,770	72,760	153,451	14	146,592	5
	Gaus Filt	209,020	15,493	326,510	75,720	156,686	15		
	DnCnn	215,930	15,334	212,400	74,890	129,639	7		
TW	No Filt	218,870	14,160	286,120	60,969	145,030	11	141,988	4
	Gaus Filt	186,810	14,328	280,020	49,700	132,715	8		
	DnCnn	248,420	14,087	280,280	50,091	148,220	12		
UNET	No Filt	180,610	12,938	129,040	47,940	92,632	1	129,457	3
	Gaus Filt	268,560	12,119	248,620	45,840	143,785	9		
	DnCnn	292,440	12,220	260,070	43,090	151,955	13		
GB	No Filt	247,040	14,953	209,990	103,750	143,933	10	123,648	2
	Gaus Filt	191,540	14,693	203,640	68,019	119,473	6		
	DnCnn	154,756	19,502	188,520	67,372	107,538	2		
TR	No Filt	208,580	9,141	168,170	70,320	114,053	5	112,120	1
	Gaus Filt	198,810	9,537	166,690	63,520	109,639	3		
	DnCnn	195,270	13,727	162,760	78,910	112,667	4		

Table 5.1: Quantitative Comparison by total error of distance in pixels of the following Tracing methods: Edge Detection (ED), CT Fire (CTF), Scale Space (SS), Twombli (TW), U-Net, Graph based (GB) and Trace Ridges (TR), with three filtering options on the following images: Disease Mimicking ECM (DME), Fluorescently labelled (FC), Fluorescent Fibronectin (FF) and Breast Cancer Biopsy (BCB) images. Along with the errors per image, average per filtering option, average of the methodology and rank are presented. Results are sorted by rank with the worst results at the top (7) and best results at the bottom (1).

Average Distance Error:

The average distance error calculates the average error by dividing the total error by the number of traces, with the results displayed in Table 5.2. Filters improve the accuracy and performance of several methods, including Edge Detection, CT Fire, Scale Space, Twombli, Graph Based, and Trace Ridges. However, U-Net does not benefit from filtering, as it was trained on unfiltered images. Specifically, DnCnn filtering improves U-Net's accuracy for the DME and BCB images, while Gaussian filtering enhances the performance only for the BCB image. Ranking the filter-method combination pairs, it is found that Edge Detection with DnCnn performs best, followed by Edge Detection without filtering, Trace Ridges with Gaussian, Edge Detection with Gaussian, and Twombli with Gaussian. When averaging

results per method, Edge Detection ranks the highest (best performing), closely followed by Trace Ridges and Twombli. Edge Detection performs the best due to the proximity of its traces to the GT, although it has an affinity to over-segment, as it produces for one fibre, two edges. By contrast, Trace Ridges, which achieves the lowest total distance error and second-best average distance error, is seen as a more robust and accurate tracing method. Yet, Trace Ridges does not always perform the best individually, for example with the DME image, Twombli with Gaussian filtering, CT Fire with no filtering and DnCnn outperforms Trace Ridges in the average distance error as well as with the BCB images, Twombli with Gaussian and DnCnn filtering outperforms Trace Ridges.

Method		Images							
	Filter	DME	FC	FF	BCB	Average	Rank	avg/Method	Rank
GB	No Filt	19.3	15.0	17.7	10.5	15.6	21	12.0	7
	Gaus Filt	15.3	4.1	17.3	6.6	10.8	18		
	DnCnn	11.0	7.4	13.0	7.0	9.6	13		
UNET	No Filt	10.7	12.9	10.7	8.2	10.6	16	11.0	6
	Gaus Filt	11.8	13.6	14.4	7.5	11.8	20		
	DnCnn	12.7	9.7	14.3	6.1	10.7	17		
CTF	No Filt	8.6	15.7	14.3	8.4	11.7	19	9.6	5
	Gaus Filt	11.4	3.8	11.4	7.9	8.6	11		
	DnCnn	9.5	6.4	10.1	7.5	8.4	9		
SS	No Filt	10.7	5.2	16.1	8.6	10.1	15	9.5	4
	Gaus Filt	10.4	4.7	15.8	8.2	9.8	14		
	DnCnn	10.7	5.0	10.4	8.2	8.6	10		
TW	No Filt	10.0	5.0	10.6	6.9	8.1	8	8.2	3
	Gaus Filt	9.7	4.7	10.3	5.7	7.6	5		
	DnCnn	14.5	5.5	10.3	5.4	8.9	12		
TR	No Filt	11.5	3.1	9.8	7.6	8.0	6	7.8	2
	Gaus Filt	10.2	3.4	8.7	6.7	7.3	3		
	DnCnn	10.5	4.3	9.3	8.3	8.1	7		
ED	No Filt	8.4	4.8	9.4	6.4	7.3	2	7.2	1
	Gaus Filt	8.5	4.9	9.1	6.6	7.3	4		
	DnCnn	8.3	5.1	8.8	5.7	7.0	1		

Table 5.2: Quantitative Comparison by average error of distance in pixels of the following Tracing algorithms; Edge Detection (ED), CT Fire (CTF), Scale Space (SS), Twombli (TW), U-Net, Graph based (GB) and Trace Ridges (TR) with three filtering options on Disease Mimicking ECM (DME), Fluorescently labelled (FC), Fluorescent Fibronectin (FF) and Breast Cancer Biopsy (BCB) images. Along with the errors per image, average per filtering option, average of the methodology and rank are presented. Results are sorted by rank with the worst results at the top (7) and best results at the bottom (1).

Maximum Distance Error:

The maximum distance error assesses the maximum individual error of a trace within each image; this is detailed in Table 5.3. Filtering improves the performance of several methods, including Edge Detection, CT Fire, Scale Space, U-Net, Graph Based, and Trace Ridges, while it increases the maximum error of Twombli and affects its performance negatively. Interestingly, Gaussian filtering benefits U-Net, while DnCnn filtering negatively influences its performance, particularly on the FC image, which impacts the overall maximum error. Among filter-method combinations, Graph Based with Gaussian filtering performs the best, followed by U-Net with Gaussian, U-Net without filtering, Edge Detection with DnCnn, and U-Net with DnCnn. When looking at the average results solely based on the methods, U-Net ranks the highest by achieving the lowest maximum distance error, followed by Graph Based and Edge Detection. Trace Ridges underperforms on this metric due to its tendency to favour brighter and longer fibres, leading to higher singular maximum errors. While this metric offers useful insights, total and average distance errors are more reliable indicators of an algorithm's robustness and accuracy.

Method		Images							
	Filter	DME	FC	FF	BCB	Average	Rank	avg/Method	Rank
SS	No Filt	65.2	50.4	69.6	202.4	96.9	21	94.9	7
	Gaus Filt	63.6	55.6	66.4	191.8	94.4	20		
	DnCnn	61.4	57.4	59.2	195.5	93.4	19		
CTF	No Filt	68.9	89.0	81.0	73.6	78.1	16	77.5	6
	Gaus Filt	57.5	98.5	75.5	82.7	78.5	17		
	DnCnn	58.4	96.5	75.7	73.1	75.9	15		
TR	No Filt	68.5	55.0	95.2	102.1	80.2	18	76.1	5
	Gaus Filt	84.3	50.4	91.9	70.4	74.3	14		
	DnCnn	76.3	48.8	88.2	81.7	73.7	13		
TW	No Filt	62.4	65.8	57.8	68.7	63.7	8	66.3	4
	Gaus Filt	69.8	66.9	60.5	60.5	64.4	9		
	DnCnn	79.2	89.0	62.7	51.8	70.7	12		
ED	No Filt	45.4	59.4	52.3	119.7	69.2	10	66.3	3
	Gaus Filt	46.1	59.4	52.3	119.7	69.4	11		
	DnCnn	50.3	86.0	58.3	46.1	60.2	4		
GB	No Filt	58.9	72.1	65.8	56.2	63.2	7	60.7	2
	Gaus Filt	59.6	53.8	62.9	47.8	56.0	1		
	DnCnn	71.6	55.8	70.6	53.2	62.8	6		
UNET	No Filt	58.8	63.4	66.4	50.9	59.9	3	59.8	1
	Gaus Filt	56.1	63.4	57.6	50.4	56.9	2		
	DnCnn	57.2	89.0	56.0	47.8	62.5	5		

Table 5.3: Quantitative Comparison by maximum error of distance in pixels of the following Tracing algorithms; Edge Detection (ED), CT Fire (CTF), Scale Space (SS), Twombli (TW), U-Net, Graph based (GB) and Trace Ridges (TR) with three filtering options on Disease Mimicking ECM (DME), Fluorescently labelled (FC), Fluorescent Fibronectin (FF) and Breast Cancer Biopsy (BCB) images. Along with the errors per image, average per filtering option, average of the methodology and rank are presented. Results are sorted by rank with the worst results at the top (7) and best results at the bottom (1).

Time:

The computational processing time for each algorithm, combined with the various filtering approaches on each image, was recorded and is shown in Table 5.4. Interestingly, filtering accelerated each algorithm, with DnCnn consistently delivering the quickest results across all combinations. Each subsequent algorithm filter pairs in Table 5.4 saw improvements in both performance and speed, except for Graph Based with DnCnn, which outperformed Twombli.Overall, Edge Detection emerged as the fastest algorithm, followed closely by Trace Ridges. U-Net was approximately three times slower than Trace Ridges, followed by Scale Space and CT Fire. Computational processing time was noticeably affected by image noise levels, with noisier images (FF and DME) taking longer to process than those with less noise.

Method		Images							
	Filter	DME	FC	FF	BCB	Average	Rank	avg/Method	Rank
GB	No Filt	1387.0	88.0	550.0	146.0	542.75	21	433.08	7
	Gaus Filt	836.0	55.0	664.0	153.0	427.00	20		
	DnCnn	527.0	84.0	573.0	134.0	329.50	16		
TW	No Filt	383.0	211.0	777.0	306.0	419.25	19	385.92	6
	Gaus Filt	681.0	340.0	438.0	152.0	402.75	18		
	DnCnn	520.0	303.0	380.0	140.0	335.75	17		
CTF	No Filt	180.0	60.0	145.0	77.0	115.50	15	102.42	5
	Gaus Filt	167.0	57.0	110.0	71.0	101.25	14		
	DnCnn	135.0	49.0	113.0	65.0	90.50	13		
SS	No Filt	98.6	4.8	72.7	45.4	55.40	11	55.78	4
	Gaus Filt	86.0	5.2	71.8	82.5	61.36	12		
	DnCnn	87.0	5.1	68.0	42.2	50.57	10		
UNET	No Filt	23.0	5.2	21.1	20.1	17.33	9	12.23	3
	Gaus Filt	20.1	5.0	18.9	23.7	16.92	8		
	DnCnn	2.5	2.2	2.4	2.6	2.44	5		
TR	No Filt	8.0	0.5	4.0	2.7	3.81	7	3.00	2
	Gaus Filt	4.5	0.8	3.1	3.0	2.87	6		
	DnCnn	4.0	0.6	2.9	1.8	2.32	4		
ED	No Filt	1.2	0.1	0.5	0.5	0.58	2	0.53	1
	Gaus Filt	0.7	0.5	0.5	0.6	0.58	3		
	DnCnn	0.8	0.1	0.4	0.4	0.43	1		

Table 5.4: Quantitative Comparison by time taken in seconds of the following Tracing algorithms; Edge Detection (ED), CT Fire (CTF), Scale Space (SS), Twombli (TW), U-Net, Graph based (GB) and Trace Ridges (TR) with three filtering options on Disease Mimicking ECM (DME), Fluorescently labelled (FC), Fluorescent Fibronectin (FF) and Breast Cancer Biopsy (BCB) images. Along with the errors per image, average per filtering option, average of the methodology and rank are presented. Results are sorted by rank with the worst results at the top (7) and best results at the bottom (1).

5.3.3 Visual Assessment:

Visual comparison of the algorithms performance could also bring insight into the efficacy of those algorithms. Figures (Fig. 5.9), (Fig. 5.11), (Fig. 5.13), (Fig. 5.15), (Fig. 5.10), (Fig. 5.12), (Fig. 5.14) and (Fig. 5.16) present a comparison between the algorithm results and the ground truth (GT), with the GT shown in red and the results in blue. White lines are the pixels where the algorithm's trace is aligned with the GT, representing accurate traces. In some instances, a purple colour may appear; this is an artefact caused when a blue line is extremely close to a red line.

Fluorescent Collagen Image:

The FC image is clear, making it easy to identify fibres, and is a simple image that is easily traced by the algorithms. Edge detection (Fig. 5.9) often produces two edges for each fibre. As shown in (Fig. 5.10), applying DnCnn enhances tracing by increasing the regions of white overlap with the ground truth. The CT Fire algorithm performs well on the original image (Fig. 5.9) but struggles with the filtered approaches, as seen in the reduced white regions in (Fig. 5.10), suggesting filtering may hinder its accuracy. Scale Space remains unaffected by filtering and produces consistent, though suboptimal results (Fig. 5.10). Twombli shows irregular traces and misses bright easily seen fibres (Fig. 5.9), with DnCnn further worsening its performance (Fig. 5.10). U-Net performs well with many white overlap regions (Fig. 5.9), and filtering does not appear to impact its segmentation. The Graph Based method misses several fibres, this is highlighted in red. Graph Based also identified extraneous elements when DnCnn is applied (Fig. 5.9) Trace Ridges demonstrates the best visual performance, achieving the most accurate fibre delineations (Fig. 5.9) . However, Trace Ridges with DnCnn performs the least effective among filtered methods on Trace Ridges (Fig. 5.10).

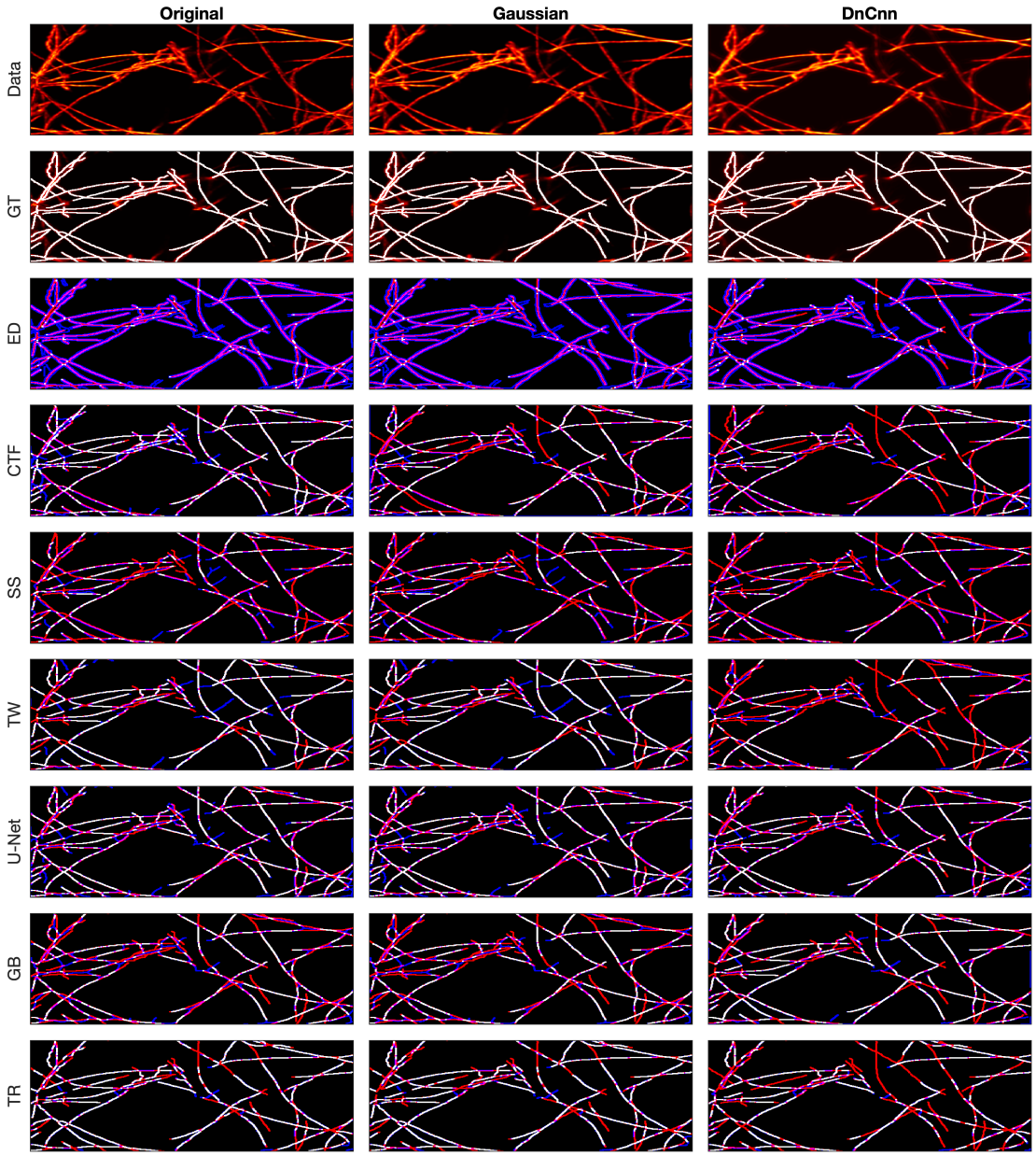


Figure 5.9: Comparison of tracing algorithms on FC image (full view). The seven algorithms were applied on the image, three different filters: Original/no filtering, Gaussian, and DnCnn were used. Data corresponds to the tracing images; GT is the manually drawn ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is shown in blue and areas where they overlap is white.

Fluorescent Fibronectin Image:

For the fluorescent fibronectin image (Fig. 5.11), several key observations are notable. Although this image contains more noise than the previous FC example, similar patterns are

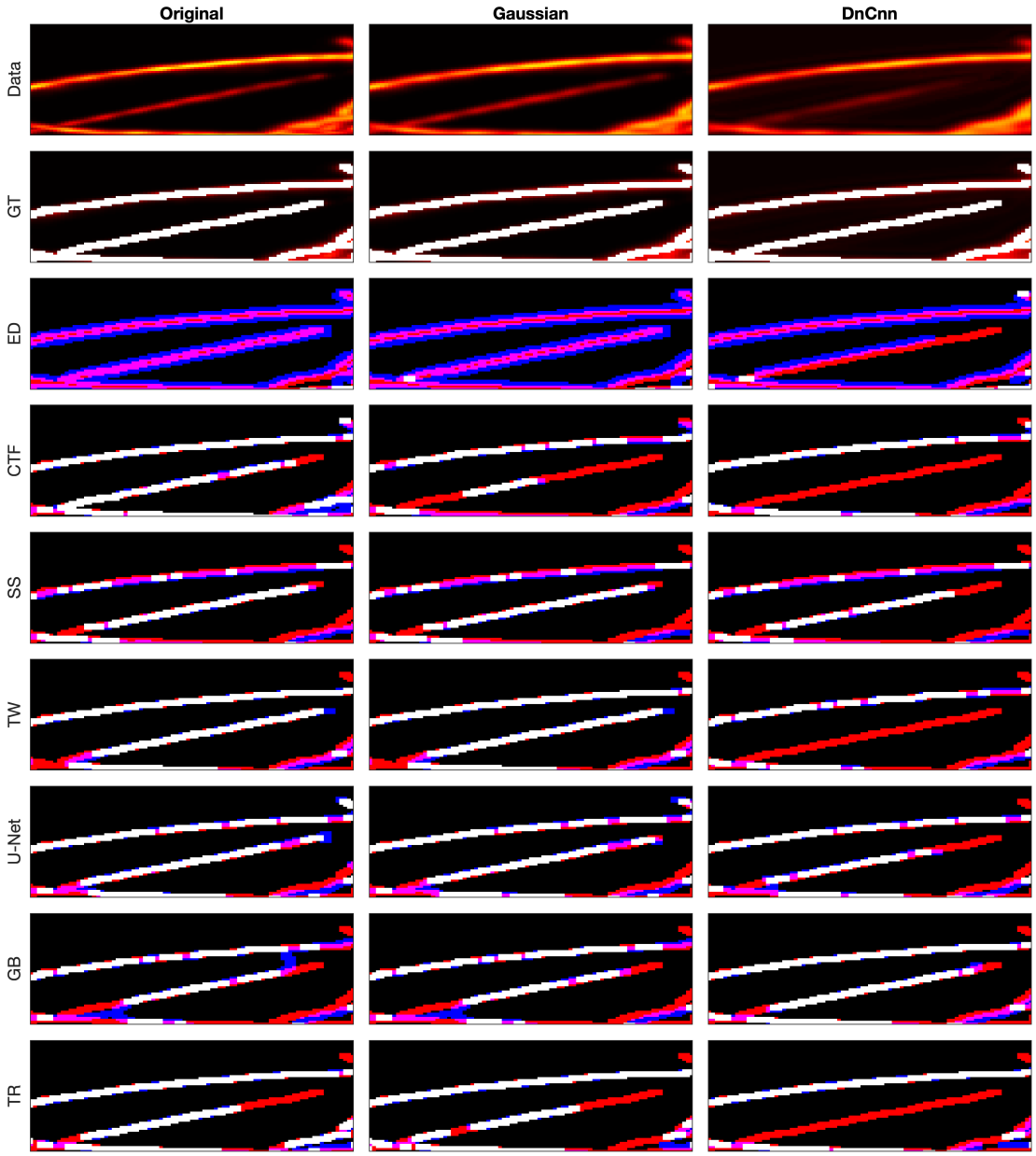


Figure 5.10: Comparison of tracing algorithms on the FC image (region of interest). he seven algorithms were applied on the image, three different filters: Original/no filtering, Gaussian, and DnCnn. Data corresponds to the images traced; GT is the manually delineated ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is shown in blue and areas where they overlap is white.

observed: both CT Fire and Edge Detection produce a high number of traces, exceeding the actual number of fibres present. Filtering improves the performance of Edge Detection (Fig. 5.12) but worsens the CT Fire's output. Scale Space and Trace Ridges generate

fewer results overall, with Scale Space identifying darker fibres and Trace Ridges emphasising the brighter ones. Trace Ridges also shows fewer false blue traces and larger white regions, reflecting its stricter criteria (Fig. 5.12), and it shows improved performance when filtered (Fig. 5.11) and (Fig. 5.12). The Graph based method and Twombli display oversegmentation, noticeable with the excess of blue lines (Fig. 5.12), but their performance improves slightly with filtering. In contrast, U-Net struggles with filtered images, as it was trained on unfiltered images and performs best with the original image and when no filtering is applied, as seen in (Figs. 5.11, 5.12).

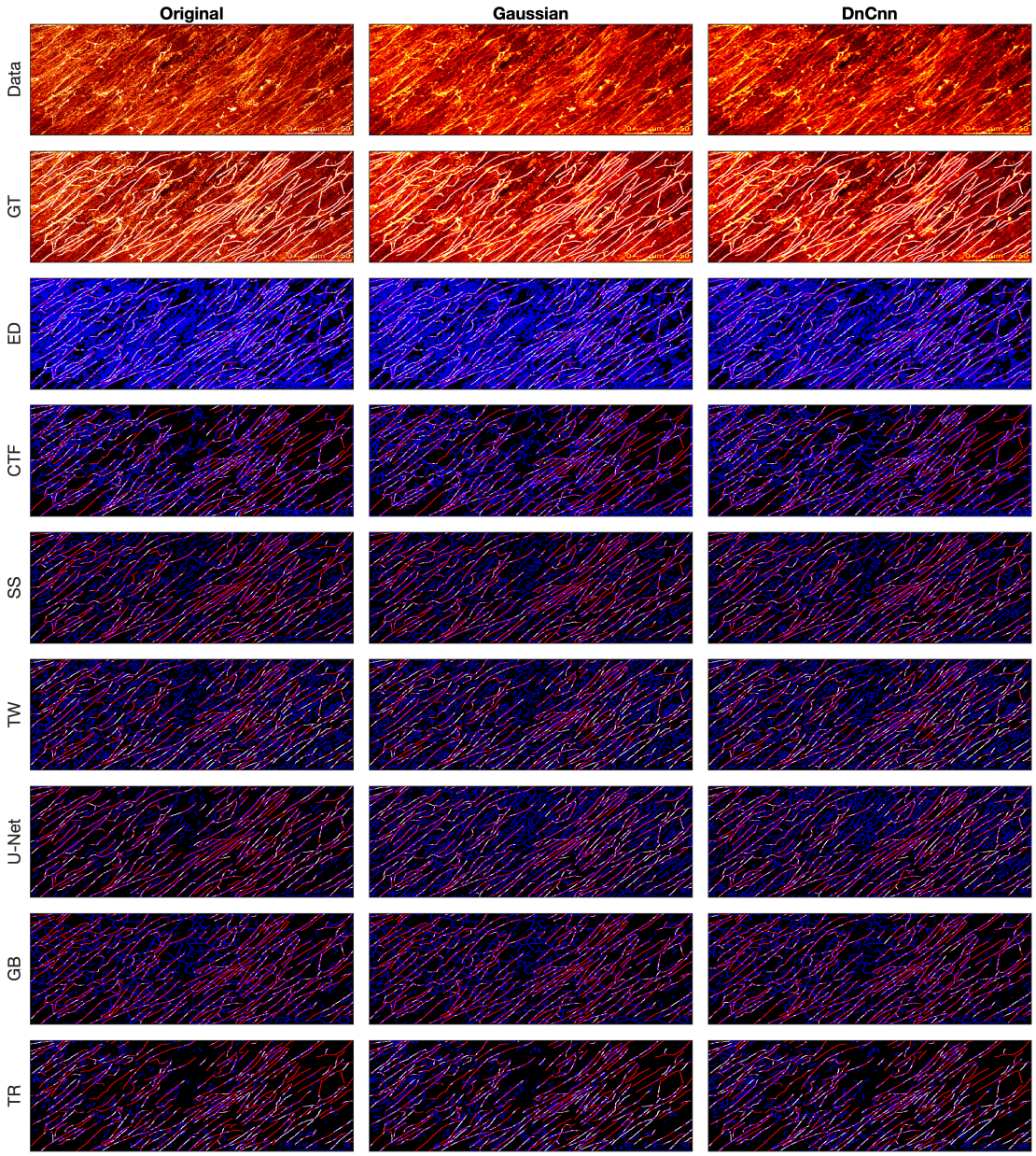


Figure 5.11: Comparison of tracing algorithms on FF image (full view). The seven algorithms were applied on the image, three different filters: Original/no filtering, Gaussian, and DnCnn. Data corresponds to the images traced; GT is the manually delineated ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is shown in blue and areas where they overlap is white.

Breast Cancer Biopsy Image:

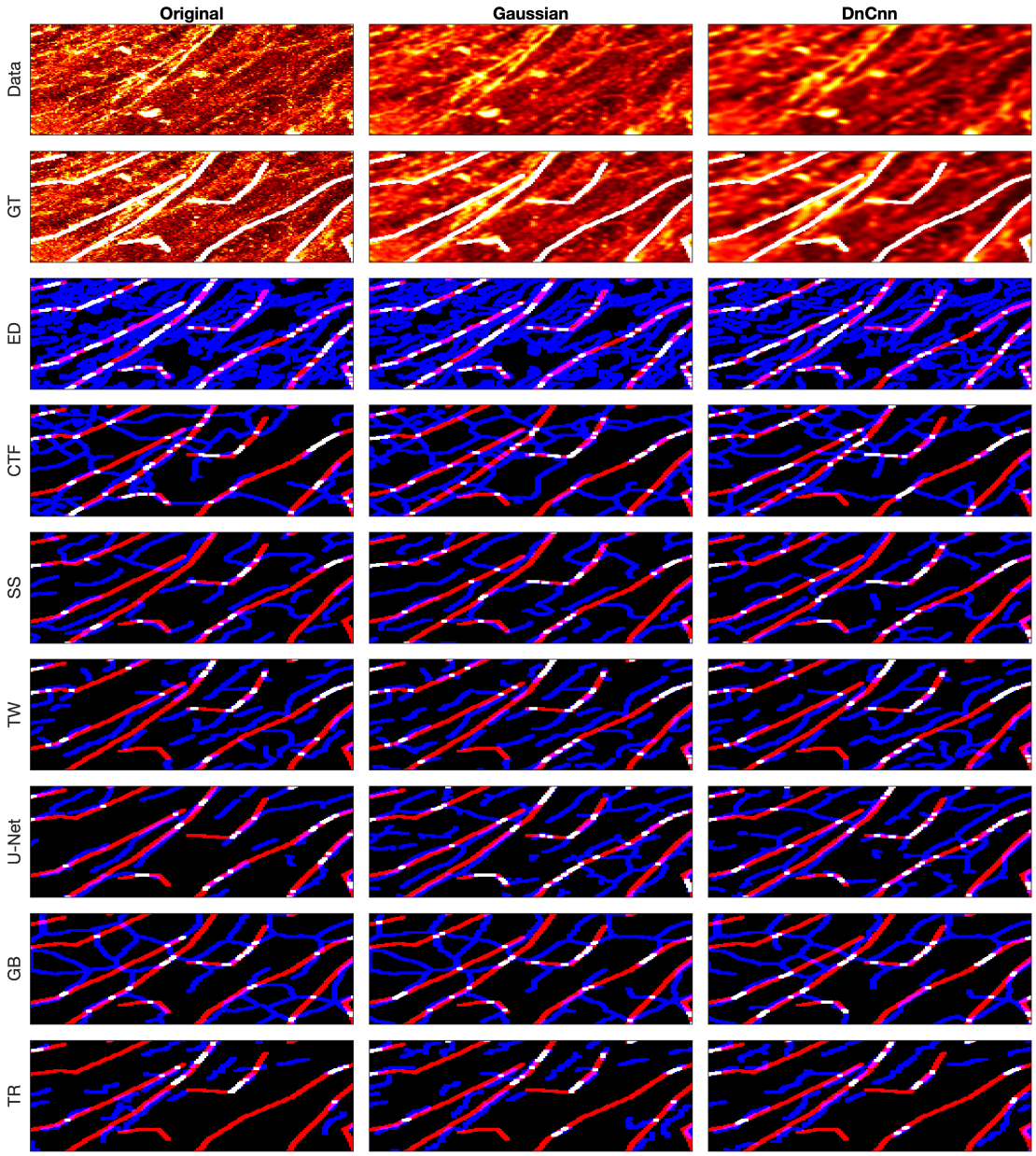


Figure 5.12: Comparison of tracing algorithms on FF image (region of interest). The seven algorithms were applied on the image, three different filters: Original/no filtering, Gaussian, and DnCnn. Data corresponds to images traced; GT is the manually delineated ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is shown in blue and areas where they overlap is white.

For the BCB image, similar patterns emerge: Edge Detection tends to over-segment resulting in inaccuracies (Fig. 5.13) and (Fig. 5.14). However, pre-filtering approaches help produce cleaner images and reduce errors. CT Fire fails to capture major fibres, instead

identifies noise as false objects, as shown in. (Fig. 5.13) and (Fig. 5.14) Filtering improves CT Fire's accuracy (Fig. 5.14). Scale Space, unusually, detects noise and spurious objects, deviating from its usual behaviour (Fig. 5.14), with filtering having minimal impact on its performance. Twombli performs well, particularly in identifying fibres in the upper right corner (Fig. 5.14). Gaussian filtering enhances Twombli's performance while DnCnn hinders its effectiveness, resulting in the degradation of many white line regions. U-Net's results are notable, with extensive white regions indicating precise delineation (Fig. 5.13), and filtering further improves U-Net by reducing spurious objects, contrary to its response on other images. Trace Ridges maintains its strict approach, producing cleaner delineations (Fig. 5.14) but misses some key fibres. The Graph-based algorithm behaves similarly to Trace Ridges, with filtering effectively reducing spurious traces (Fig. 5.14).

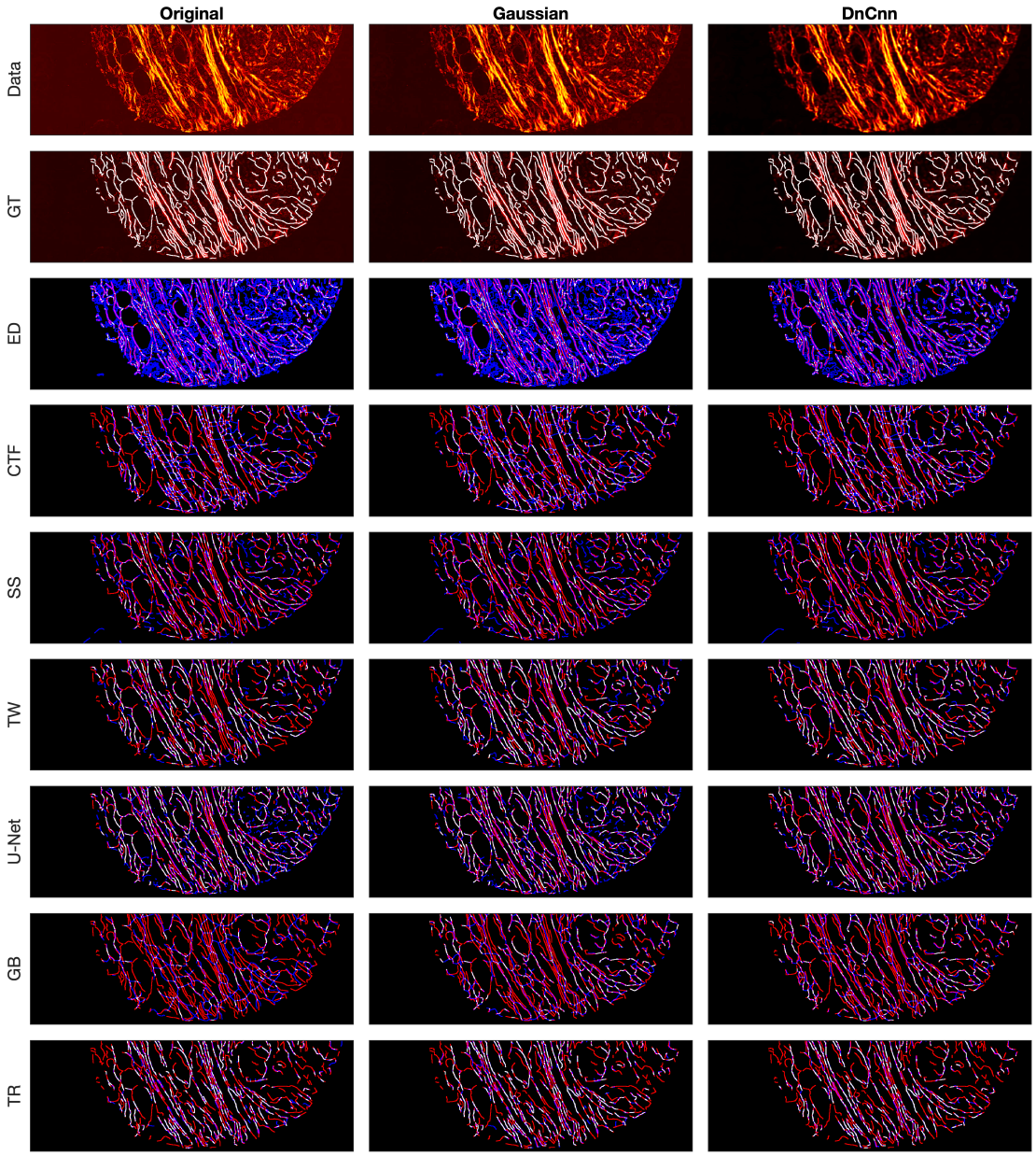


Figure 5.13: Comparison of tracing algorithms on Picrosirius red Breast Cancer Biopsy image (full view). The seven algorithms were applied on the image, three different filters: Original/no filtering, Gaussian and DnCnn. Data corresponds to the images traced; GT is the manually delineated ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is shown in blue and areas where they overlap is white.

Disease Mimicking ECM Image:

The visual inspection of the disease mimicking ECM (DME) image (Fig. 5.15) provides important insights. Both Gaussian and DnCnn filtering noticeably improves image clarity.

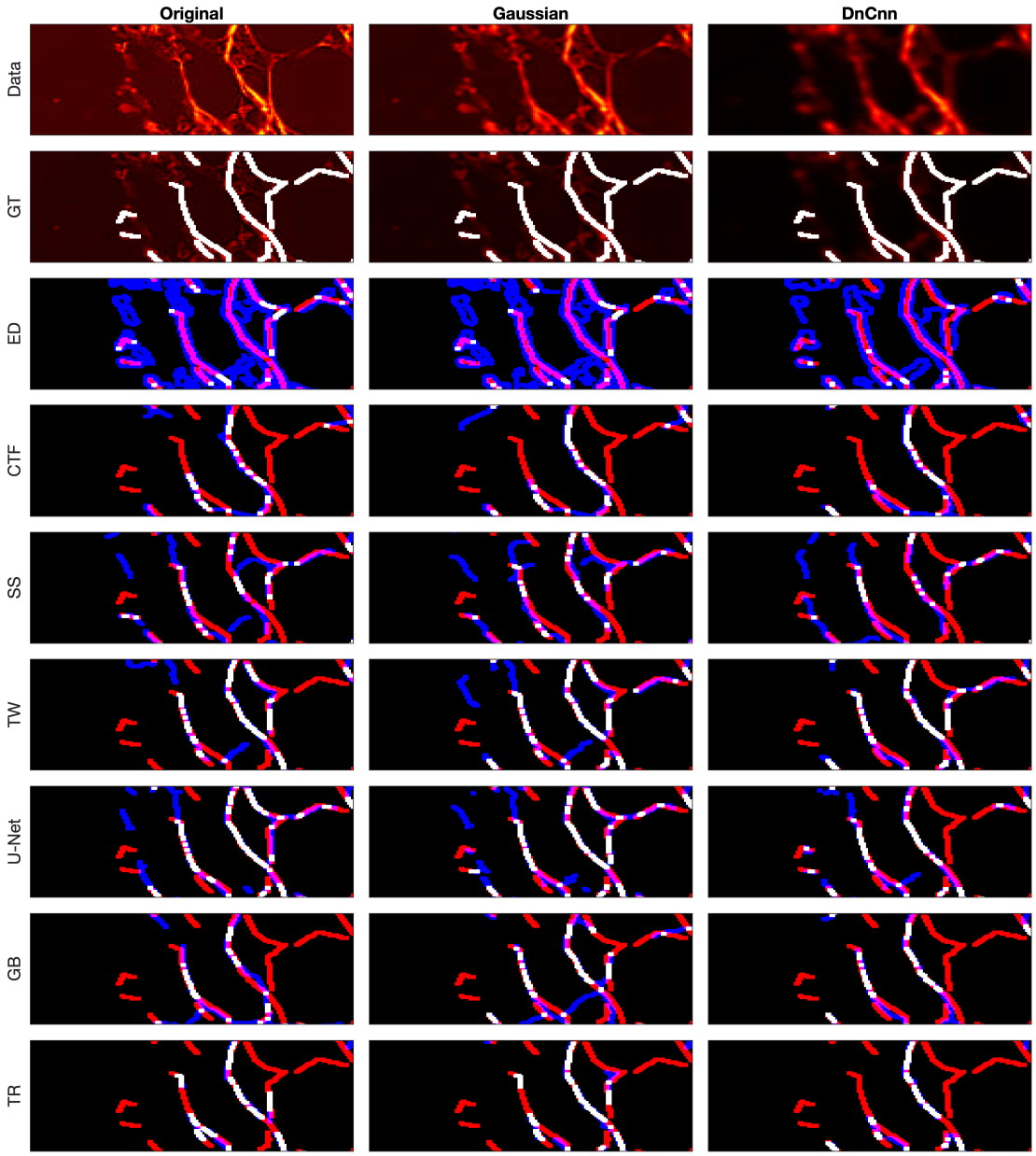


Figure 5.14: Comparison of tracing algorithms on Picrosirius red Breast Cancer Biopsy image (region of interest). The seven algorithms were applied on the image, three different filters: Original/no filtering, Gaussian and DnCnn. Data corresponds to images traced; GT is the manually delineated ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is shown in blue and areas where they overlap is white.

Edge Detection still performs poorly, as it continuously produces two traces for one fibre. Filtering significantly improves Edge Detection, shown in (Fig. 5.16). Here DnCnn provides

the better tracing. CT Fire outperforms Edge Detection but still introduces some spurious elements, with Gaussian and DnCnn filtering leading to additional traces that are not present in the original image, as seen in the ROI (Fig. 5.16). Scale Space has fewer traces due to its multi-resolution capabilities, although it may occasionally merge distinct fibres; Gaussian and DnCnn filtering increase fibre traces, though the latter has caused longer fibres to separate. Twombli yields moderate results, tracing some non-existent fibres but still outperforming certain other methods. Filtering improves Twombli's delineation, with Gaussuan filtering enhancing its performance, evident by the increased number of white lines in (Fig. 5.16). DnCnn filtering however, introduces some false traces. U-Net performs well on the unfiltered image but shows reduced accuracy on filtered versions due to its training on original images. The Graph based algorithm shows strong results, with filtering and in particular DnCnn boosting its accuracy and clarity, as seen in (Fig. 5.16). Trace Ridges, similar to Scale Space, produces cleaner traces, effectively separating intersecting fibres for accuracy. Filtering further improves Trace Ridges, notably seen in (Fig. 5.16).

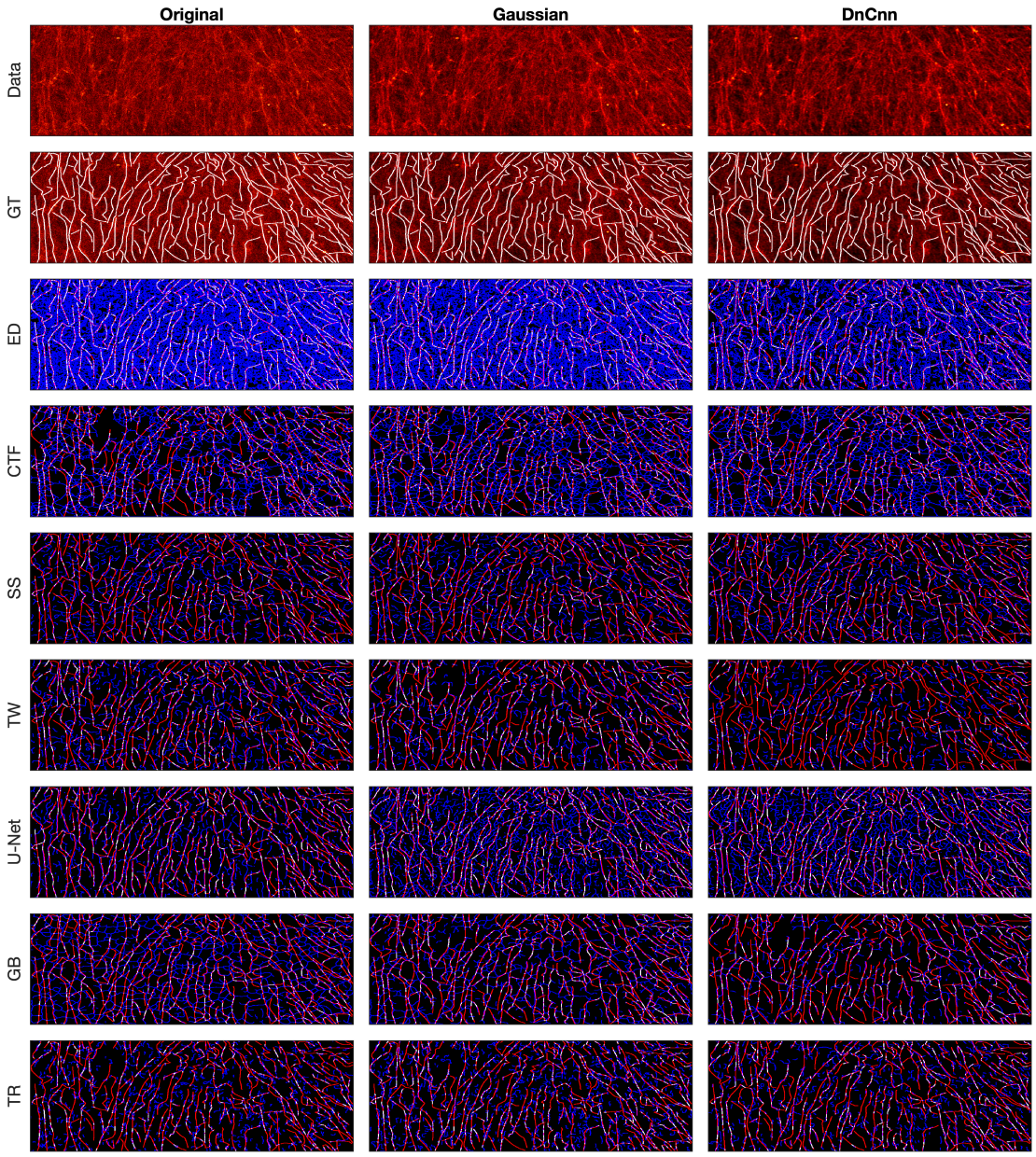


Figure 5.15: Comparison of tracing algorithms on DME image (full view). The seven algorithms were applied on the image, three different filters: Original/no filtering, Gaussian and DnCnn. Data corresponds to images traced; GT is the manually delineated ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is shown in blue and areas where they overlap is white.

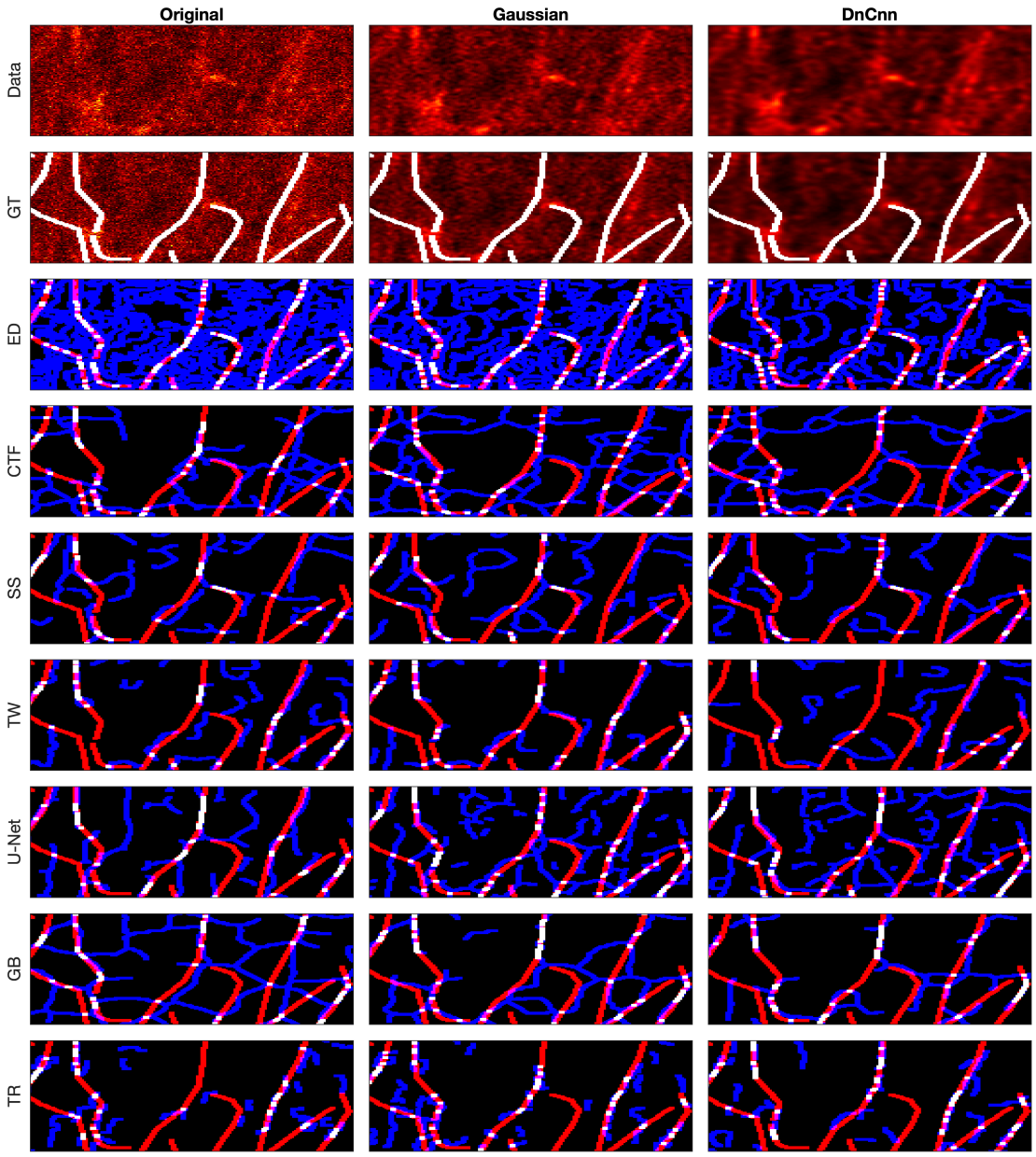


Figure 5.16: Comparison of tracing algorithms on DME image (region of interest). The seven algorithms were applied on the image, three different filters: Original/no filtering, Gaussian and DnCnn. Data corresponds to images traced; GT is the manually delineated ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is shown in blue and areas where they overlap is white.

5.3.4 High Pass Filter investigation:

Similarly, to the first investigation, the objective comparison with the HPF observes the total distance error, average distance error, maximum distance error and computational time.

Total Distance Error:

Firstly, looking at the total distance errors in Table 5.5. When analysing the average value for filter and method pair, it can be seen that the use of high pass filters hinders the performance of the algorithms this is seen in Scale Space and Trace Ridges where it doubles the second highest total distance errors for this method and U-Net where it once again doubles the second highest total distance errors for this method. Edge Detection is the only method where high pass filtering improves the total distance errors this is due to the fact that Edge Detection detects two edges for each fibre. The effect of a high pass filter is that it constricts the width of the fibres making it easier for the algorithm to trace two edges for each fibre. When observing the rank per method, Scale Space is the best performing algorithm when considering a high pass filter as the total distance errors achieved with a high pass filter are higher than other filtering options, but they do not deviate away significantly from the values of those filtering options. Trace Ridges comes out second due to it achieving a great performance with other filtering options followed by U-Net. Both algorithms performance was hindered significantly due to the application of a high pass filter. Finally, Edge Detection is the worst performing algorithm due its performance with other filtering options.

METHOD		DME	FC	FF	Average	Rank	avg/Method	Rank
ED	No Filt	947,550	35,910	754,760	579,407	16	462,370	4
	Gaus Filt	834,660	36,250	714,400	528,437	15		
	DnCnn	583,550	32,260	604,050	406,620	14		
	HPF	417,240	24,990	562,820	335,017	12		
UNET	No Filt	**180,600**	12,900	**129,100**	**107,533**	1	213,392	3
	Gaus Filt	268,600	12,100	247,700	176,133	9		
	DnCnn	280,000	11,100	242,900	178,000	10		
	HPF	941,900	26,500	207,300	391,900	13		
TR	No Filt	209,950	9,820	164,910	128,227	2	164,584	2
	Gaus Filt	214,980	**9,790**	167,240	130,670	4		
	DnCnn	194,600	16,030	176,150	128,927	3		
	HPF	585,030	31,260	195,250	270,513	11		
SS	No Filt	206,620	15,650	198,440	140,237	5	145,201	1
	Gaus Filt	209,020	15,490	201,820	142,110	6		
	DnCnn	208,790	15,780	206,640	143,737	7		
	HPF	225,580	19,250	219,330	154,720	8		

Table 5.5: Quantitative Comparison by total error of distance in pixels of the following Tracing algorithms; Edge Detection (ED), Scale Space (SS), U-Net and Trace Ridges (TR) with four filtering options (no filtering, Gaussian filtering, DnCnn, High pass filtering) on Disease Mimicking ECM (DME), Fluorescently labelled (FC) and Fluorescent Fibronectin (FF) images. Along with the errors per image, average per filtering option, average of the methodology and rank are presented. Results are sorted by rank with the worst results at the top (4) and best results at the bottom (1).

Average Distance Errors:

The results of the average distance errors can be seen in Table 5.6. Observing the average distance errors for the filter and method pair, it can be seen that high pass filters have worsened all the algorithms performances, although for Scale Space and Edge Detection the values do not deviate far from the other filtering options. While for Trace Ridges and U-Net the error values with a high pass filter have doubled those of other filtering options. Observing the average per method, Edge Detection performs the best followed by Scale Space then Trace Ridges and U-Net where their average values were affected due to the large errors with the high pass filter.

METHOD		DME	FC	FF	Average	Rank	avg/Method	Rank
UNET	No Filt	10.704	12.942	10.706	11.451	13	14.217	4
	Gaus Filt	11.756	13.565	14.425	13.249	14		
	DnCnn	12.032	7.737	12.900	10.889	12		
	HPF	48.876	4.940	10.021	21.279	16		
TR	No Filt	11.155	**3.416**	10.124	8.231	6	10.664	3
	Gaus Filt	9.851	3.577	**8.936**	**7.455**	1		
	DnCnn	10.130	5.116	9.493	8.246	7		
	HPF	34.361	8.840	12.963	18.722	15		
SS	No Filt	10.718	5.168	10.520	8.802	10	8.891	2
	Gaus Filt	10.352	4.733	10.559	8.548	9		
	DnCnn	10.333	4.957	10.301	8.530	8		
	HPF	12.288	5.791	10.972	9.684	11		
ED	No Filt	8.284	4.813	9.374	7.490	2	7.529	1
	Gaus Filt	8.498	4.875	9.126	7.499	3		
	DnCnn	8.421	5.269	8.936	7.542	4		
	HPF	**8.169**	5.289	9.299	7.586	5		

Table 5.6: Quantitative Comparison by average error of distance in pixels of the following Tracing algorithms; Edge Detection (ED), Scale Space (SS), U-Net and Trace Ridges (TR) with four filtering options (no filtering, Gaussian filtering, DnCnn, High pass filtering) on Disease Mimicking ECM (DME), Fluorescently labelled (FC) and Fluorescent Fibronectin (FF) images. Along with the errors per image, average per filtering option, average of the methodology and rank are presented. Results are sorted by rank with the worst results at the top (4) and best results at the bottom (1).

Maximum Distance Error:

Similarly, observing the maximum distance errors in Table 5.7. It can be seen from the average maximum errors per filtering and method pair that only Edge Detection benefited from filtering with a high pass where Edge Detection with DnCnn performed worse overall from those pairs. The application of a high pass filter with Scale Space, Trace Ridges and U-Net performed worse than any other combination of each respective method and filter combinations. Overall, by observing average maximum error per method Edge Detection performs the best followed by Scale Space, U-Net and finally Trace Ridges. Trace Ridges performed poorly in this metric as it favours brighter and longer fibres to trace.

METHOD		DME	FC	FF	Average	Rank	avg/Method	Rank
TR	No Filt	69.806	54.974	91.239	72.006	12	86.001	4
	Gaus Filt	87.881	**49.602**	91.948	76.477	13		
	DnCnn	62.482	54.849	123.075	80.135	14		
	HPF	166.255	79.438	100.463	115.385	15		
UNET	No Filt	58.801	63.450	66.398	62.883	9	75.868	3
	Gaus Filt	56.104	63.450	57.012	58.855	3		
	DnCnn	55.325	65.616	60.904	60.615	6		
	HPF	193.755	96.811	72.788	121.118	16		
SS	No Filt	65.249	50.398	62.438	59.362	4	61.556	2
	Gaus Filt	63.625	55.628	64.350	61.201	7		
	DnCnn	62.831	56.000	61.751	60.194	5		
	HPF	66.295	65.155	64.955	65.468	10		
ED	No Filt	**45.403**	59.413	**52.266**	**52.361**	1	58.384	1
	Gaus Filt	46.071	59.403	**52.266**	52.580	2		
	DnCnn	56.000	86.009	56.558	66.189	11		
	HPF	54.464	72.245	60.511	62.407	8		

Table 5.7: Quantitative Comparison by maximum error of distance in pixels of the following Tracing algorithms; Edge Detection (ED), Scale Space (SS), U-Net and Trace Ridges (TR) with four filtering options (no filtering, Gaussian filtering, DnCnn, High pass filtering) on Disease Mimicking ECM (DME), Fluorescently labelled (FC) and Fluorescent Fibronectin (FF) images. Along with the errors per image, average per filtering option, average of the methodology and rank are presented. Results are sorted by rank with the worst results at the top (4) and best results at the bottom (1).

Time:

Observing the computational time taken to trace the fibres in Table 5.8, it can be seen that when comparing the average time taken per filter and method pair that high pass filtering hindered the performance of Scale Space while DnCnn filtering was the fastest to process with Edge Detection followed by Gaussian filtering then high pass filtering and no filtering. For Trace Ridges high pass filtering improves the processing time of Trace Ridges due to less fibres being detected followed by DnCnn, Gaussian and no filtering. For U-Net, Gaussian filtering was the fastest followed by high pass filtering, no filtering and then DnCnn filtering. Overall, when observing the average time taken in seconds per method Edge Detection was the fastest followed by Trace Ridges which was 3 times faster than the next best algorithm U-Net which was followed by Scale Space.

METHOD		DME	FC	FF	Average	Rank	avg/Method	Rank
SS	No Filt	96.443	4.297	70.589	57.109	15	60.575	4
	Gaus Filt	82.277	3.831	84.875	56.994	14		
	DnCnn	75.698	3.632	70.245	49.858	13		
	HPF	95.555	5.193	134.269	78.339	16		
UNET	No Filt	21.054	4.363	18.724	14.713	11	14.788	3
	Gaus Filt	20.084	4.273	18.977	14.445	9		
	DnCnn	19.735	4.582	21.796	15.371	12		
	HPF	18.218	4.334	21.319	14.624	10		
TR	No Filt	7.708	0.820	4.106	4.211	8	3.077	2
	Gaus Filt	5.275	0.565	3.890	3.243	7		
	DnCnn	4.392	0.525	4.070	2.996	6		
	HPF	2.300	0.571	2.705	1.858	5		
ED	No Filt	0.955	0.113	**0.407**	0.492	4	0.411	1
	Gaus Filt	0.596	**0.090**	0.443	**0.376**	2		
	DnCnn	0.553	0.105	0.434	0.364	1		
	HPF	**0.522**	0.091	0.624	0.412	3		

Table 5.8: Quantitative Comparison by time taken in seconds of the following Tracing algorithms; Edge Detection (ED), Scale Space (SS), U-Net and Trace Ridges (TR) with four filtering options (no filtering, Gaussian filtering, DnCnn, High pass filtering) on Disease Mimicking ECM (DME), Fluorescently labelled (FC) and Fluorescent Fibronectin (FF) images. Along with the time taken to process each image, average per filtering option, average of the methodology and rank are presented. Results are sorted by rank with the worst results at the top (4) and best results at the bottom (1).

5.3.5 Comparison Summary:

Trace Ridges the novel method was compared with six-existing tracing algorithms. Three filtering options were applied to each image before being passed to each algorithm. Four biological images were used to compare the algorithms. For the four images, a manual ground truth delineation was created for each image and the distance maps were used to quantify the errors of the distance between the ground truth and an algorithm's delineation. It was found that Trace Ridges performs the best overall having achieved the lowest total distance error, followed by Graph Based and U-Net, yet, Trace Ridges does not always perform the best individually as seen with BCB image. Edge Detection had the lowest average distance error, followed by Trace Ridges and Twombli. Edge Detection's low average is attributed to its affinity to over-segment thus it is usually not that far from a complement ground truth trace. U-Net achieved the lowest maximum distance error followed by Graph Based and Edge Detection. Trace Ridges performed poorly in this metric as it favours longer and brighter fibres, so it is prone to have a higher singular maximum error.

Filtering enhanced the performance of all algorithms except for U-Net. This is because U-Net was trained on non-filtered image patches. Edge Detection was the fastest algorithm,

followed by Trace Ridges and U-Net. The introduction of filtering made the algorithms faster except for U-Net. Overall Trace Ridges emerges as the superior algorithm that can be used to trace fibre-like images. Many morphological features can be extracted using Trace Ridges which could be used to compare populations of data. This chapter will further explore some of the applications of Trace Ridges.

When experimenting with a high pass filter, the algorithms performances are hindered where only Edge Detection benefits. Looking at the total distance errors, Scale Space, Trace Ridges, and U-Net had increased errors when the images were filtered with a high pass filter and the tracing accuracy deteriorated. The same behaviour was observed when analysing the maximum distance error. Observing the average distance errors revealed that the implementation of a high pass filter worsened algorithm performance for this metric. High pass filtering assisted in faster image processing for the algorithms in comparison to other filtering options.

Manual assessment of the ECM such as its orientation and density is subjective and at times inconsistent. Tools which provide quantitative descriptors such as fibre alignment, length and curvature which are known to correlate with cancer aggressiveness and tumour micro-environment remodelling can provide information to aid in biomarker discovery and tumour micro-environment remodelling. [156], in this study authors find that increased collagen density and in particular aligned fibres correlate strongly with poor prognosis and higher metastasis rates. Authors in [362] applied CT-Fire to images of non-small cell lung cancer and quantified the collagen fibre descriptors which correlated with poor prognosis. Similarly, authors in [363] utilised Twombli to quantify collagen fibres of triple-negative breast cancer tissue, where they found that higher fibre density and length correlated with aggressive cancer phenotypes and metastasis risk.

Methods such as Trace Ridges can automate complex analyses which are typically carried out by pathologists and produce reproducible results. This could provide scalable solutions for digital pathology and consistent results can be produced at a faster rate.

Algorithms may also reveal hidden ECM biomarkers, Twombli and deep learning approaches have been used to uncover collagen alignment and curvature features which correlated with tumour invasiveness [364]. This shows that computational tools can open doors for new diagnostic or prognostic biomarkers based on ECM morphology and heterogeneity.

As observed through the various examples, computational tools such as Trace Ridges can be used to quantify and extract meaningful properties of the extracellular matrix which can be used to investigate and diagnose conditions such as cancer.

5.4 Modified Cell Line Investigation:

Excerpts of this section have been published as a conference article in CGVC 2024, London, UK. [37].

In vitro experiments have demonstrated that certain alterations of the DNA sequence of cells [365] i.e., the creation of mutant cells can lead to distinct cellular characteristics and interactions. This experiment explores how mutant fibroblasts differ from wild type fibroblasts in their role in forming the extracellular matrix.

Whilst visual assessments of the ECM structure can reveal intriguing features, they may not be sufficient to compare different cell populations accurately. Analysing specific ECM structural elements, such as the shape, distribution, and count of fibrial glycoproteins, can provide valuable insights into the micro-environment and effects of treatments. Quantifying these fibres will open the door to revealing their properties. Comparing the properties of the RBD and the WT can reveal characteristic differences that can be used to distinguish between the two populations.

Various tracing techniques have been developed, including Edge Detection, Scale Space, Watershed algorithms, CT Fire algorithm, Twombli, Graph-Based methods and popular deep learning architectures such as U-Net. In a recent study, the Trace Ridges algorithm was found to be the most robust and accurate for tracing fibre-like structures. The recent study also revealed that DnCnn filtering was the most suitable filtering method for Images of fibronectin in Dataset 1, this experiment also served as an opportunity to demonstrate the application of trace ridges in tracing fibre-like structures and extracting morphological properties.

From the objective comparison, for the fluorescent fibronectin image, Trace Ridges achieved the lowest total distance error, with 165,850 pixels, in comparison to the second-best performing method, Graph-based, which observed 200,716 pixels. Trace Ridges had the second lowest average distance error per trace of 9.27 pixels, close to Edge Detection's 9.1 pixels. However, Trace Ridges is considered more reliable for this measure, as Edge Detection has an affinity to over-segment.

In terms of speed, Trace Ridges had a processing time of 3.33 seconds, second only to Edge Detection's 0.467 seconds, while the next fastest algorithm, U-Net was significantly slower with 14.13 seconds. The application of DnCnn lowered the distance errors for the fluorescent fibronectin images. The total distance error was 162,760 pixels with DnCnn in comparison to 166,690 of Gaussian filtering and 168,170 of no filtering. The image was then processed faster with 2.9 seconds compared to 3.1 and 4.0 seconds for Gaussian and no filtering, respectively.

Images of the extracellular matrix for two populations (mutant, n=5, and wild type, n=5) are shown in (Fig 3.12c) and (Fig 3.13c) respectively and repeated for clarity here (Fig. 5.19a) for the RBD and (Fig. 5.20a) for the wild type. The images were processed using Trace Ridges to trace the fibres (Fig. 5.17). Once the images were traced morphological features were extracted, two of those metrics showed a statistical significance when comparing images of the populations. First, the number of detected fibres, which represents the total edges identified by the algorithm (Fig. 5.18c), (Fig. 5.19b) for the RBD and (Fig. 5.20b) for

the Wild type. Secondly, the relative gap area is deemed significant. Brighter areas in the images reflect greater distances from fibres. The relative gap area represents the size of gaps relative to the entire image area (Fig. 5.18d), (Fig. 5.19c) for the RBD and (Fig. 5.20c) for the Wild type.

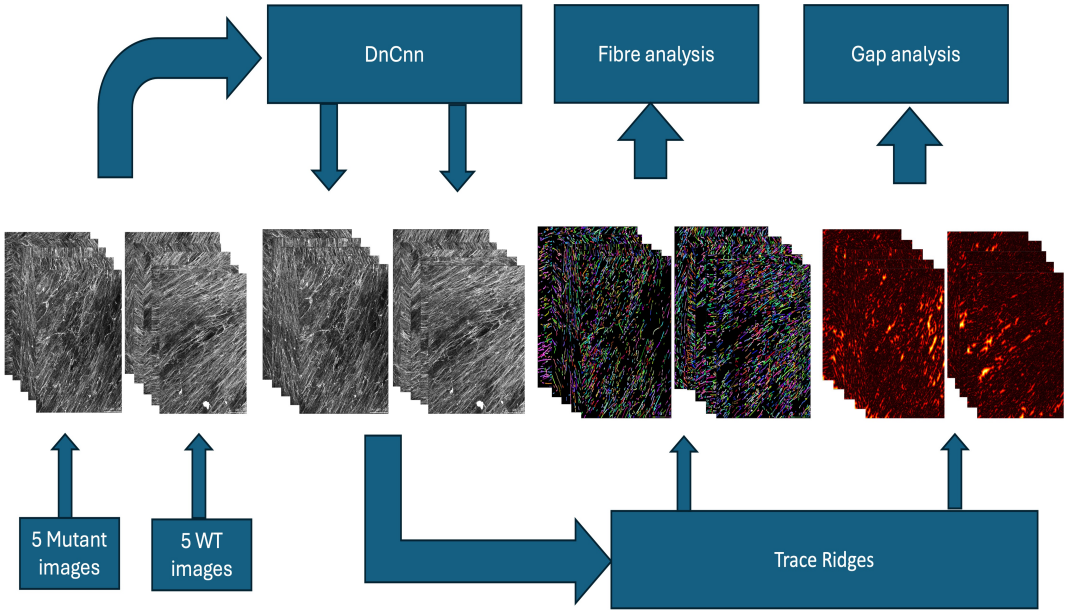


Figure 5.17: Implementation pipeline of Trace Ridges. Fluorescent fibronectin images are passed through DnCnn for filtering which are then processed by Trace Ridges. Labelled fibres and distance maps are produced where fibre and gap analysis are carried out respectively.

Trace Ridges was implemented on the images from dataset 1 as seen in (Fig. 5.17). Boxplots for the two metrics are displayed in (Fig. 5.21), showing statistically significant differences, with p-values of 0.00091 and 0.014.

Wild type extracellular matrix had a higher fibre count than the mutant extra-cellular matrix, as shown in Table A.1. For WT, the maximum fibre count was 2,363 compared to 1,851 for the mutant group. The minimum fibre count for mutants was 1,503, while the WT was 1,964, resulting in a higher median count for the WT. The WT group also had a wider range (Fig. 5.21a), with its lowest fibre count exceeding the highest count observed in mutants.

Visually, in (Fig. 5.19b) and (Fig. 5.20b), images of the mutant culture (Fig. 5.19b) show fibres that are more dispersed, spread out, and fewer in number. In contrast, the WT (Fig. 5.20b), fibres appear densely packed and abundant. This higher fibre count in the WT extracellular matrix suggests that, biologically, it contains more fibres on average than the mutant fibroblasts.

The wild type displayed a wider range of fibre counts across the images, suggesting greater variability. In contrast, the mutant fibroblasts had a more consistent and predictable fibre count, falling within a narrower range.

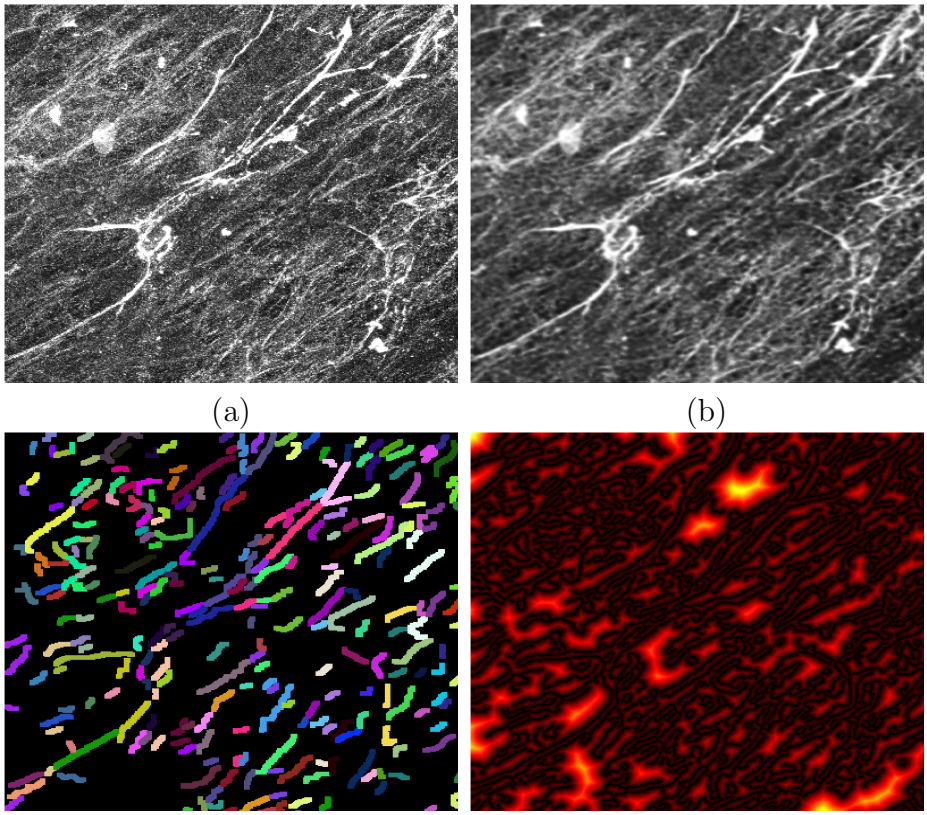
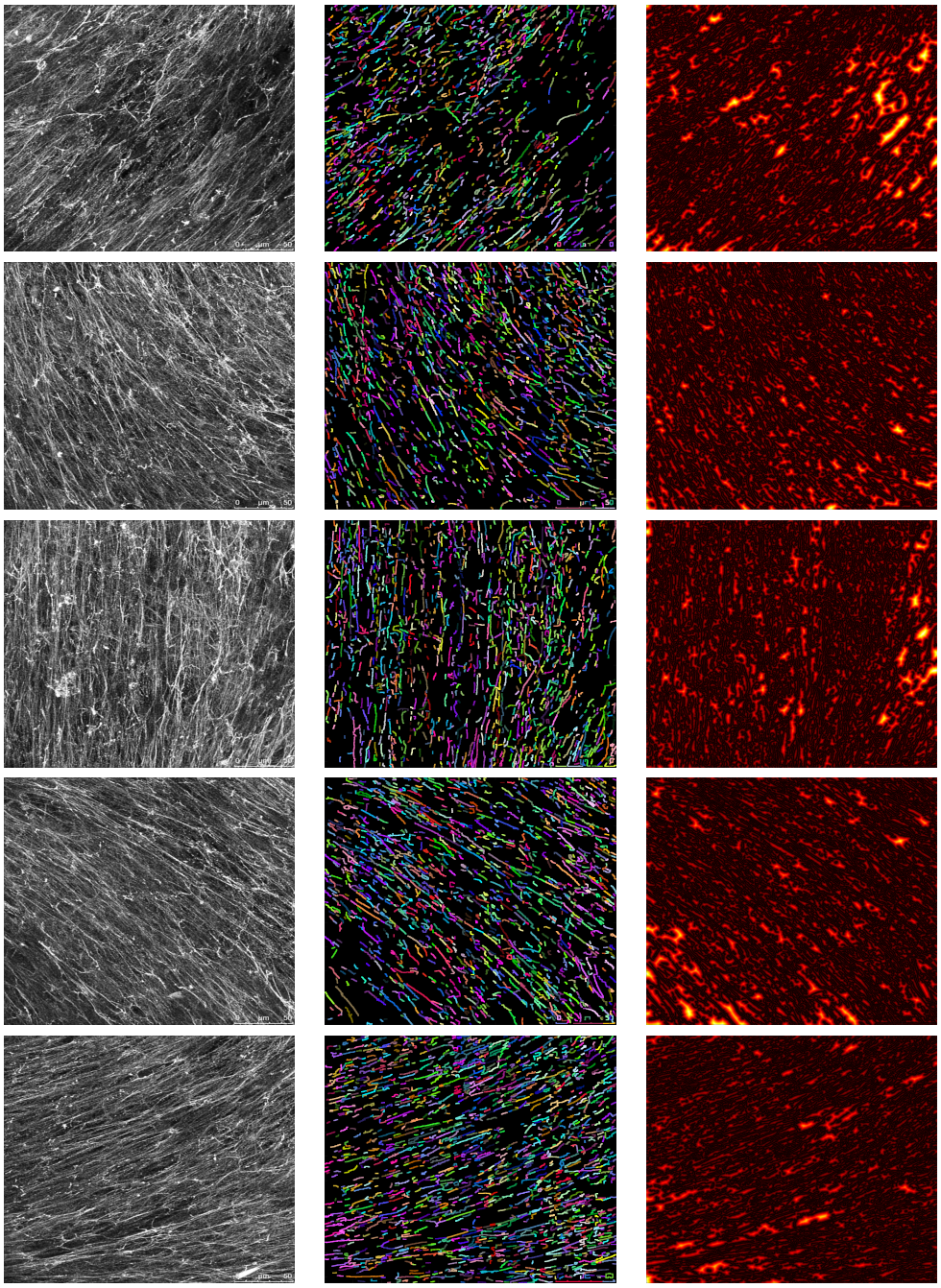
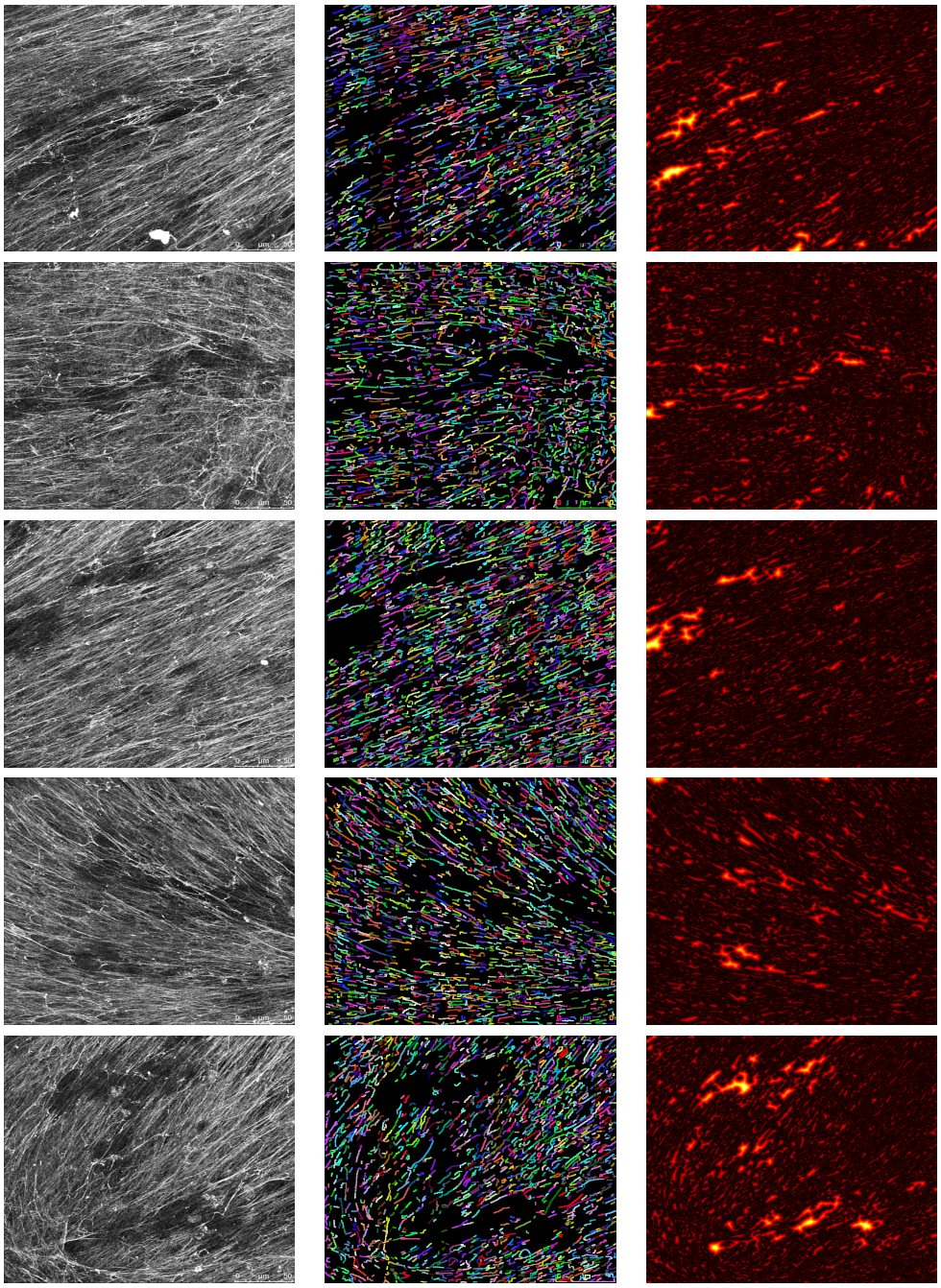


Figure 5.18: Processing the FF images through the Trace Ridges pipeline. (a) Fluorescent Fibronectin region of interest (ROI). (b) DnCnn application on Fluorescent Fibronectin (ROI). (c) Number of edges (ROI). Colours are randomly assigned to distinguish between the fibres. (d) Gaps between fibres (ROI). Brighter yellow means pixel is further away from an edge.



(a) (b) (c)

Figure 5.19: Mutant Fluorescent Fibronectin Fibres and Gaps detected by Trace Ridges. (a) Images representing RBD mutant. (b) Fibres delineated with Trace Ridges. Colours of fibres are assigned randomly to differentiate between the edges. Black regions in the images are areas of gaps between the fibres. The mutant fibres appear spread out, with the images exhibiting significantly large gaps between them. (c) Gap areas of fibres. Brighter yellow regions are pixels that are further away from fibres. Black regions represent fibres. The mutant images exhibit large, more prominent bright yellow regions, suggesting greater and more frequent separation between the fibres.



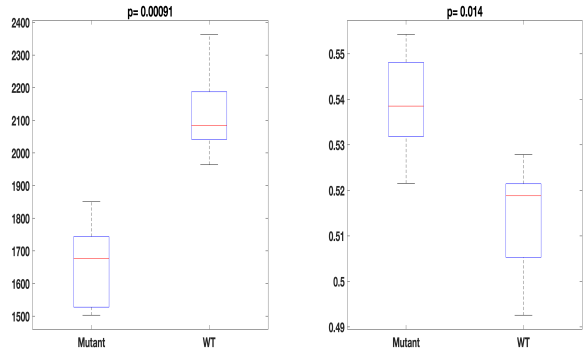
(a) (b) (c)

Figure 5.20: WT fluorescent Fibronectin Fibres and Gaps detected by Trace Ridges. (a) Images representing WT. (b) Images representing WT. Colours of fibres are assigned randomly to differentiate between the edges. Black regions in the images are areas of gaps between the fibres. The wild type extracellular matrix show a high fibre density with minimal gaps between the fibres. (c) Gap areas of fibres. Brighter yellow regions are pixels that are further away from fibres. Black regions in the image represent fibres. Visually, the wild type extracellular matrix exhibits small gaps and more frequent dark regions, indicating a denser network of fibres

Mutants showed a higher average relative gap size compared to wild type images as seen in Table A.1. The largest gap size for mutants was 0.5542, compared to 0.5278 for WT. While the smallest gap for mutants was 0.5215, versus 0.4926 for the WT. This indicates that the WT had greater variability (Fig. 5.21b), with the WT mean gap size being close to its maximum value. Visual inspection of the gap distribution in (Fig. 5.19c) and (Fig. 5.20c) reveals that mutant culture images (Fig. 5.19c) feature more regions with high intensities (yellow), indicating a greater number of large gaps, while WT images (Fig. 5.20c) display more dark regions (black), suggesting fewer large gaps.

The larger gaps in the mutant extracellular matrix imply increased spacing between the fibres, reflecting a lower fibre density and a less complex structural organisation than in WT cultures.

These findings show an inverse relationship: as fibre count in the extracellular matrix increases, gap size decreases, consistent with [366]. In that study, fibronectin growth with and without Ficoll was examined, and it was found that the addition of Ficoll promoted higher fibronectin fibre counts, resulting in smaller gaps. The inverse relationship can be seen using the scatter plot in (Fig. 5.22). The regression line indicates the trend that as the gap size decreases the number of fibres increases.



(a) (b)

Figure 5.21: Boxplots of RBD and WT morphological properties. (a) Number of edges with a $p - value$ of 0.00091 (b) Gap Area with a $p - value$ of 0.014.

Using the Bonferroni correction, a statistical method that adjusts to reduce the risk of false positives when performing multiple hypotheses tests.

The Bonferroni correction makes the significance threshold stricter. The new p-value threshold is computed by dividing the current overall significance level (0.05) by the number of hypotheses tests. For example, in this investigation there were 20 metrics tested therefore the p-value can be computed as such (Eq. 5.9). The new p-value suggests that only the number of edges has a p-value below the adjusted p-value.

$$\text{p-value} = \frac{0.05}{20} = 0.0025 \tag{5.9}$$

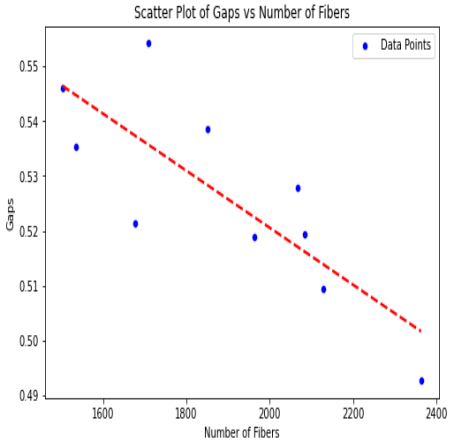


Figure 5.22: Scatter Plot of Gap size and the number of fibres. The scatter plot indicates that as the gap size decreases the number of fibres increases.

5.5 Matrix Mechano-sensing:

Excerpts of this section have been published as a journal paper in Nature Communications [358].

The ECM is made up of fibrillar collagens, fibronectin, elastin, and laminins [31]. Abnormal ECM deposition assists processes of tumour cell proliferation, invasion, and metastasis [367]. A stable ECM environment is affected during cancer progression, which allows the tumours to become exposed to collagen-rich stroma. The extracellular matrix is organised differently in pathological conditions such as cancer and ageing [368, 369, 370, 371]. Aligned ECM fibres create pathways that facilitate tumour cell invasion [372, 373], and the organisation of collagen fibres is associated with the onset and progression of the disease [106]. Collagen structures are different depending on their location on the tumour body. Those at the tumour edge experience higher levels of alignment, allowing for cell migration guided by contact signals [372, 374, 375]. However, it remains unclear how dispersing cancer cells detect ECM characteristics at the invasive front and convert them into signals that promote metastasis.

Fibre alignment is observed with increasing tissue stiffness as well as cell invasion [372, 376]. Actomyosin machinery determines the overall cell stiffness and oversees the adaptation of migratory modes in response to the changes in the ECM stiffness [377]. Actomyosin machinery is a contractile mechanism in animal cells that induces mechanical stress and is in charge of cell migration where a leader cell pulls on the matrix and the follower cells push the matrix away to create the migration tracks.

In this study, the ECM is broken down into three distinctive areas, each area promotes a set of gene activations within cancer cells that supports their ability to invade surrounding tissue at the primary tumour site and to spread to distant parts of the body, forming metastases. The findings highlight a complex relationship between the ECM's physical structure and the way cancer cells adjust their gene expression in response. This influences their in-

vasive and metastatic behaviour. The three distinctive areas were the centre of the tumour body (TB) and the invasive front (IF), this is the area that is in contact with the matrix. The invasive region can be split into two distinct regions, the proximal invasive front (PIF) which is 50-100 μm of boundary with the extracellular matrix. The distal invasive front (DIF) which is 100-200 μm away from the PIF.

High collagen content and deposition is complementary to poor patients' outcome [378, 379], as increased collagen concentrations can affect cancer cell invasion [28], while aligned matrix promotes invasion [28, 380]. In this investigation, the accumulation of collagen and whether it leads to cytoskeletal remodelling is investigated through the addition of protease inhibitors and increasing collagen concentration. Furthermore, to model if radial alignment enhances invasion, collagen gels are polymerised with Marangoni Flows. The images of collagen are then traced with Trace Ridges and morphological properties of the populations are compared. This investigation provides a use case on the application of image processing techniques to further our understanding of the ECM and its role in the progression of cancer.

5.5.1 Protease Inhibitors:

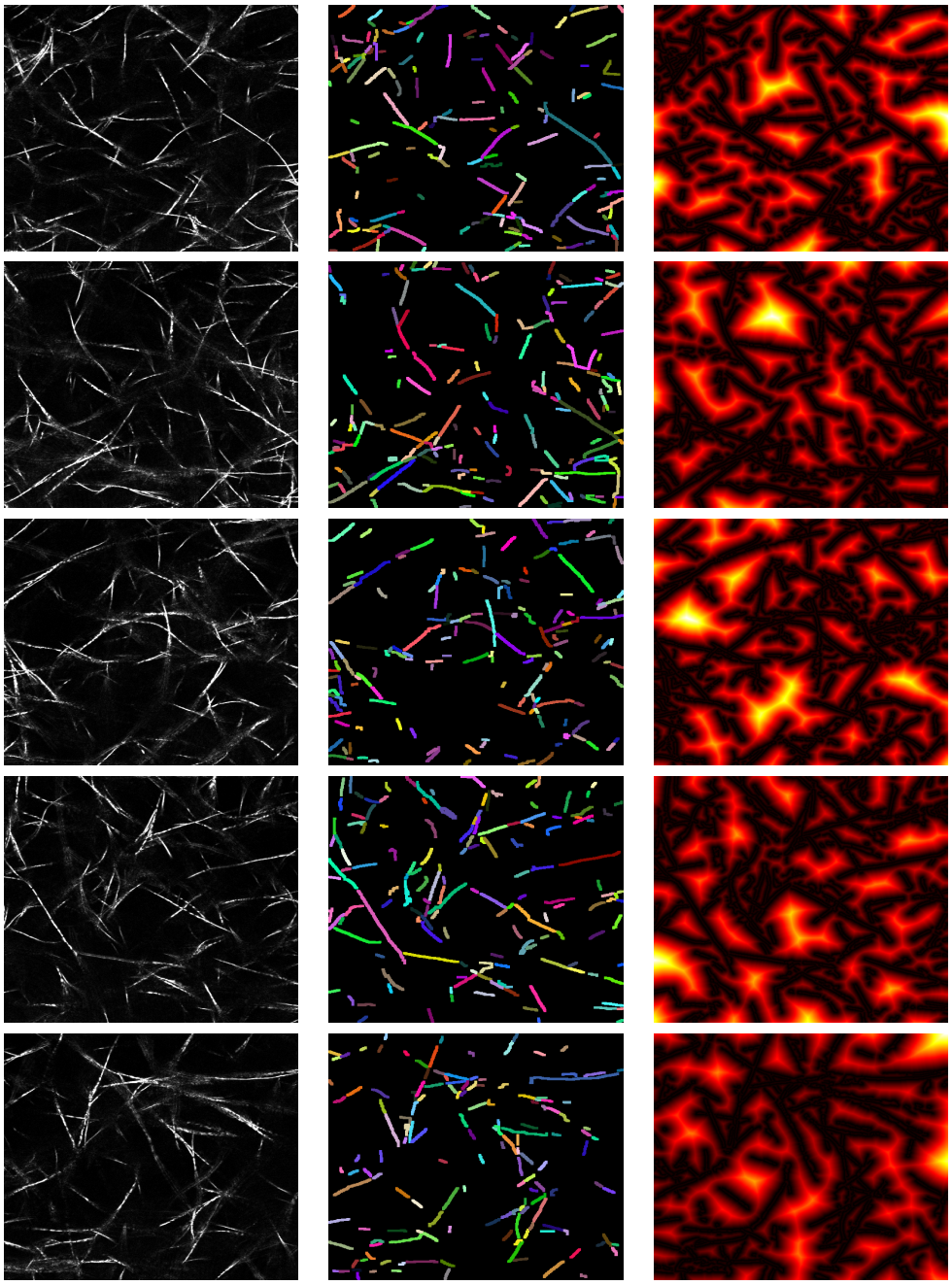
To investigate if the accumulation and abundance of collagen I promotes cytoskeletal remodelling, it was hypothesised that disabling ECM degradation in vitro would result in an increase of the local matrix density as seen in the invasive front of tumours in vivo. Blocking the growth of the ECM degradation can be achieved with the addition of protease inhibitors (PIs) as mentioned earlier in (Section. 3.3.3). PIs were added in to cells grown in the collagen I matrices and images were captured microscopically (CRM) to investigate against the control. The control set is a collection of images where ECM degradation was allowed. Understanding the differences between them could bring insights to how the invasive front of tumours alters the ECM.

Application of Trace Ridges on the control and Protease Inhibited collagen fibres is illustrated in (Fig. 5.23) and (Fig. 5.24) respectively.

Discussion:

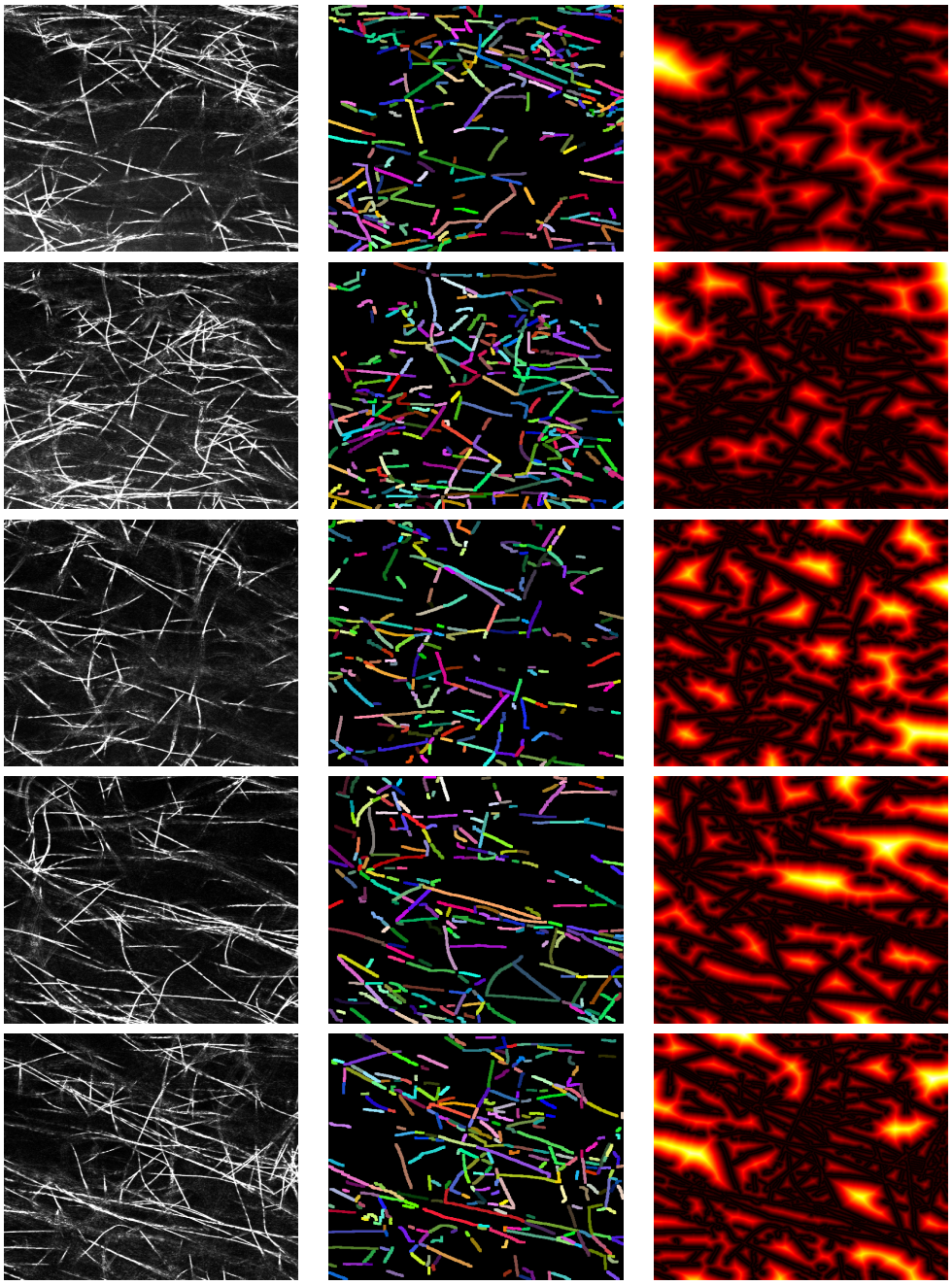
Collagen I fragments activity has declined after the application of PIs, there was an increase of width, straightness, alignment, and aspect ratio (MinAxMajAx) of the fibres. This indicates that adding the PIs resulted in the accumulation of the matrix fibres around cells.

Two-sample t-tests were conducted to compare the populations. This can be seen in (Fig. 5.25). The Protease inhibited samples are represented on the left while the control is represented on the right.



(a) (b) (c)

Figure 5.23: Control samples of Collagen I before the application of Protease Inhibitors. (a) The raw image samples. (b) Collagen fibres detected by Trace Ridges. Fibres have been assigned colours randomly to differentiate between them. (c) Gap analysis maps of the detected fibres. Pixels further away from a fibre form a gap and are shown in yellow and white.



(a) (b) (c)

Figure 5.24: Collagen I fibres after the addition of the Protease Inhibitors. (a) The raw image samples. (b) Collagen fibres detected by Trace Ridges. Fibres have been assigned colours randomly to differentiate between them. (c) Gap analysis maps of the detected fibres. Pixels further away from a fibre form a gap and are shown in yellow and white.

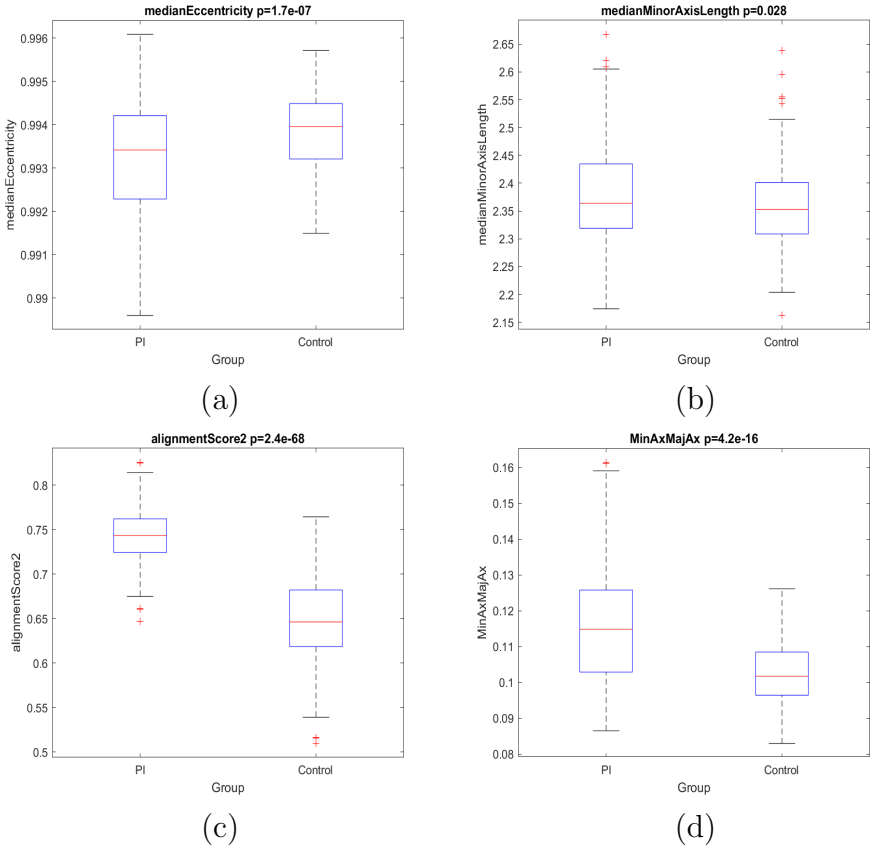


Figure 5.25: Boxplots of comparisons of control samples of Collagen and Collagen with the addition of Protease Inhibitors (PI). (a) The median eccentricity exhibited a $p - value$ of $1.7e^{-}07$. (b) The median minor axis length exhibited a $p - value$ of 0.028. (c) The alignment score which is a measure of how aligned the fibres in each population are exhibited a $p - value$ of $2.4e^{-}68$. (d) The aspect ratio (MinAxMajAx) exhibited a $p - value$ of $2.4e^{-}68$. All these measurements showed a statistical significance when comparing the PI and the control populations.

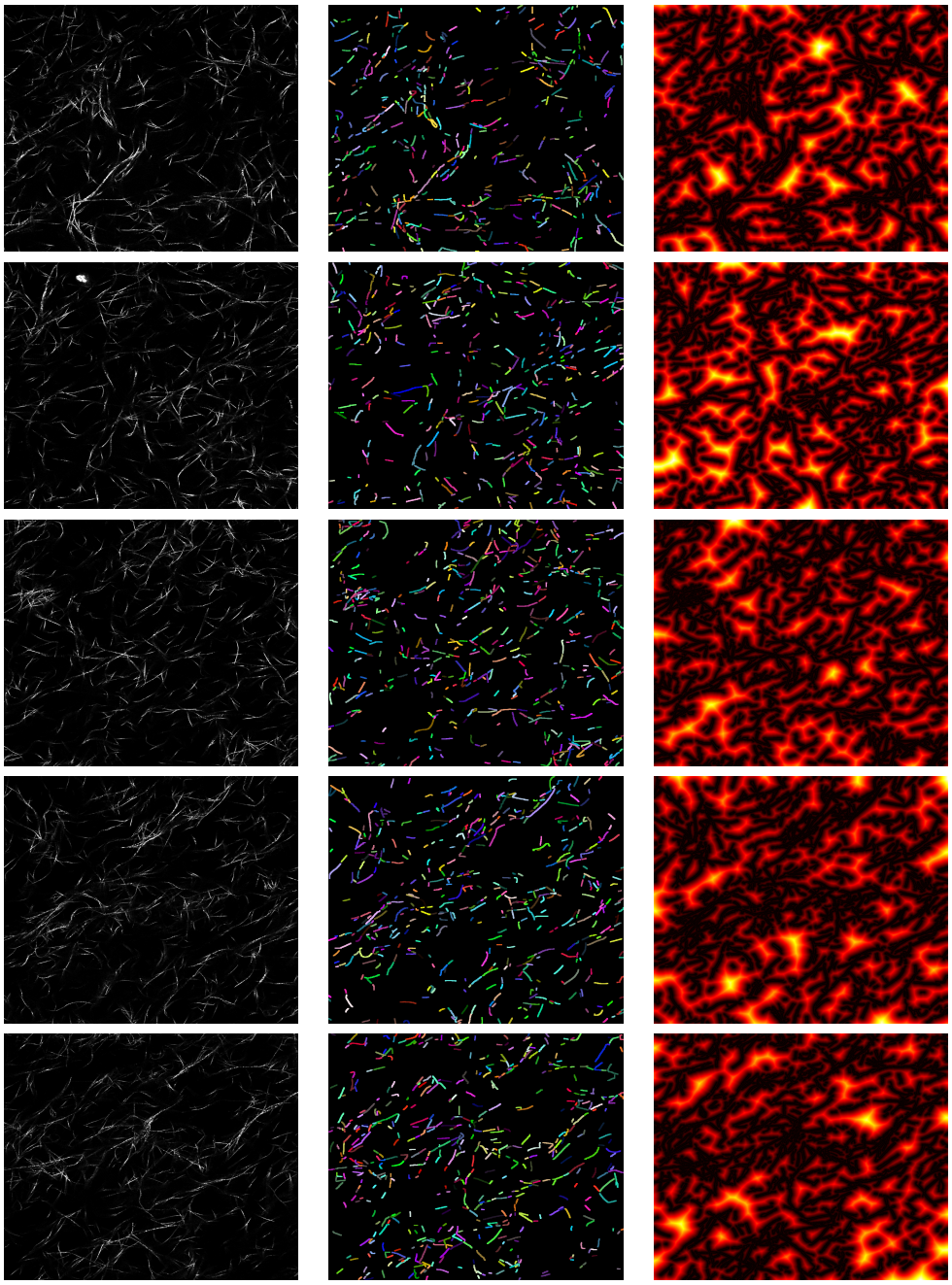
5.5.2 Increasing Collagen Concentration:

To further investigate this hypothesis in vivo of the accumulation and increase of matrix density effects. Cancer cells were grown in matrices containing an incremental increase of collagen I concentration. Increasing amounts of collagen previously have shown increased stiffness [381, 382]. Collagen concentrations were prepared at 1.7 mg/ml and 6.6 mg/ml. Images of the different culture grown collagen at the two concentrations were observed with the microscope (CRM) and passed to Trace Ridges for tracing and for morphological comparisons (Fig. 5.26) and (Fig. 5.27) for 1.7 mg/ml and 6.6 mg/ml.

Increasing concentrations of collagen from 1.7 to 6.6 mg/ml resulted in a number of structural changes of the fibres, mainly increase in the width or thickness of the fibres and the straightness. The length of the fibres has decreased, yet the aspect ratio has increased. The fibres have also become less aligned as previously reported in [383]. This illustrated by the two-sample t-test comparisons of the populations in (Fig. 5.28). This has previously been mentioned for matrices with high collagen concentrations.

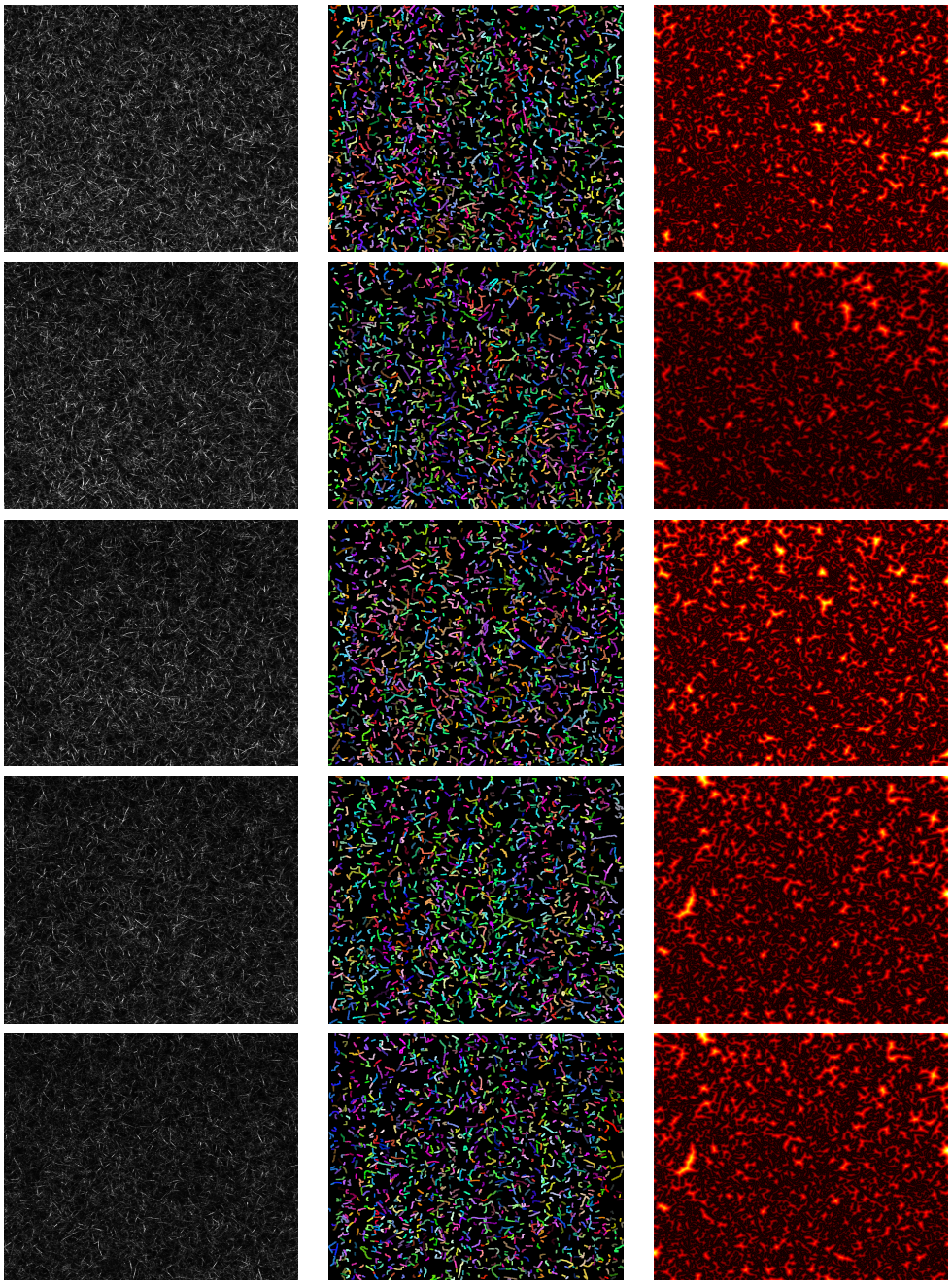
Discussion:

An increase in collagen content after blocking proteases or increasing the concentration results in cell individualisation and cell contraction due to the density. Protease inhibition promotes the migration mode to change to amoeboid strategies while an increase in collagen concentration will lead to overall less invasion, where only individual amoeboid invasion is resulted. These findings show that although matrix-induced confinement increases actomyosin levels, a balanced amount of collagen I, with fibres that are more aligned, best supports cell invasion.



(a) (b) (c)

Figure 5.26: Collagen I fibres at the concentration of 1.7 mg/ml. (a) The raw image samples. (b) Collagen fibres detected by Trace Ridges. Fibres have been assigned colours randomly to differentiate between them. (c) Gap analysis maps of the detected fibres. Pixels further away from a fibre form a gap and are shown in yellow and white.



(a) (b) (c)

Figure 5.27: Collagen I at the concentration of 6.6 mg/ml. (a) The raw image samples. (b) Collagen fibres detected by Trace Ridges. Fibres have been assigned colours randomly to differentiate between them. (c) Gap analysis maps of the detected fibres. Pixels further away from a fibre form a gap and are shown in yellow and white.

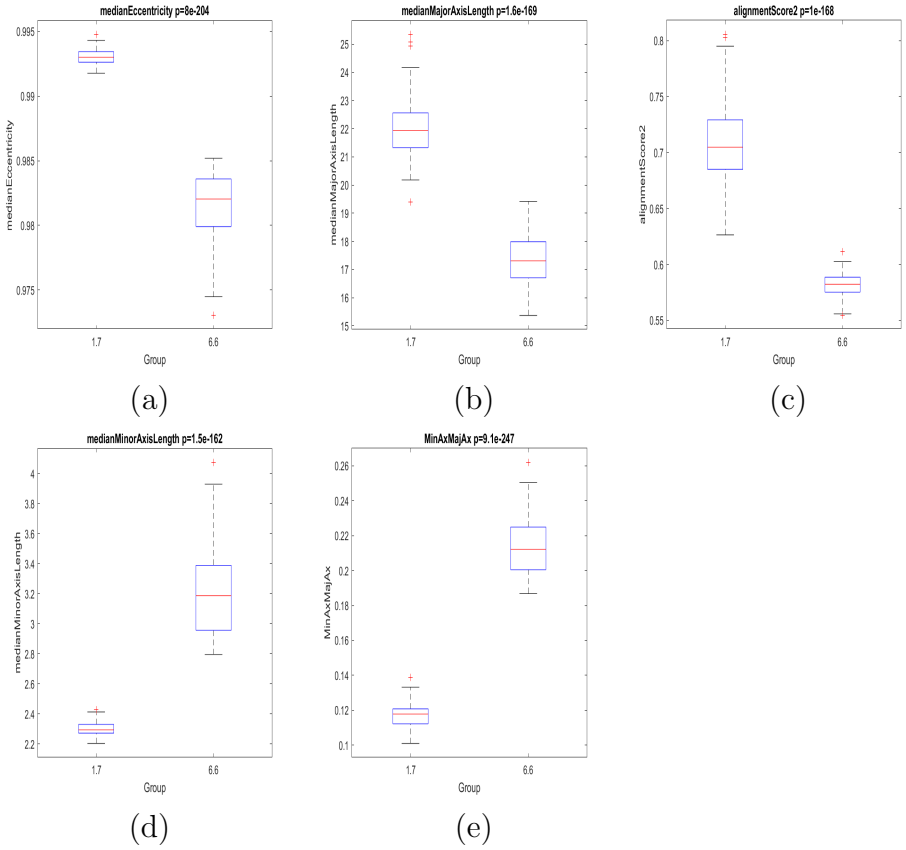


Figure 5.28: Boxplots of comparisons of the collagen morphology when the ECM concentration is increased from 1.7 mg/ml to 6.6 mg/ml. (a) The median eccentricity exhibited a $p-value$ of $8e^{-}204$. (b) The median major axis length exhibited a $p-value$ of $1.6e^{-}169$. (c) The alignment score which is a measure of how aligned the fibres in each population are exhibited a $p-value$ of $1e^{-}168$. (d) The median minor axis length exhibited a $p-value$ of $1.5e^{-}162$. (e) The aspect ratio exhibited a $p-value$ of $9.1e^{-}247$. All these measurements showed a statistical significance when comparing the PI and the control populations.

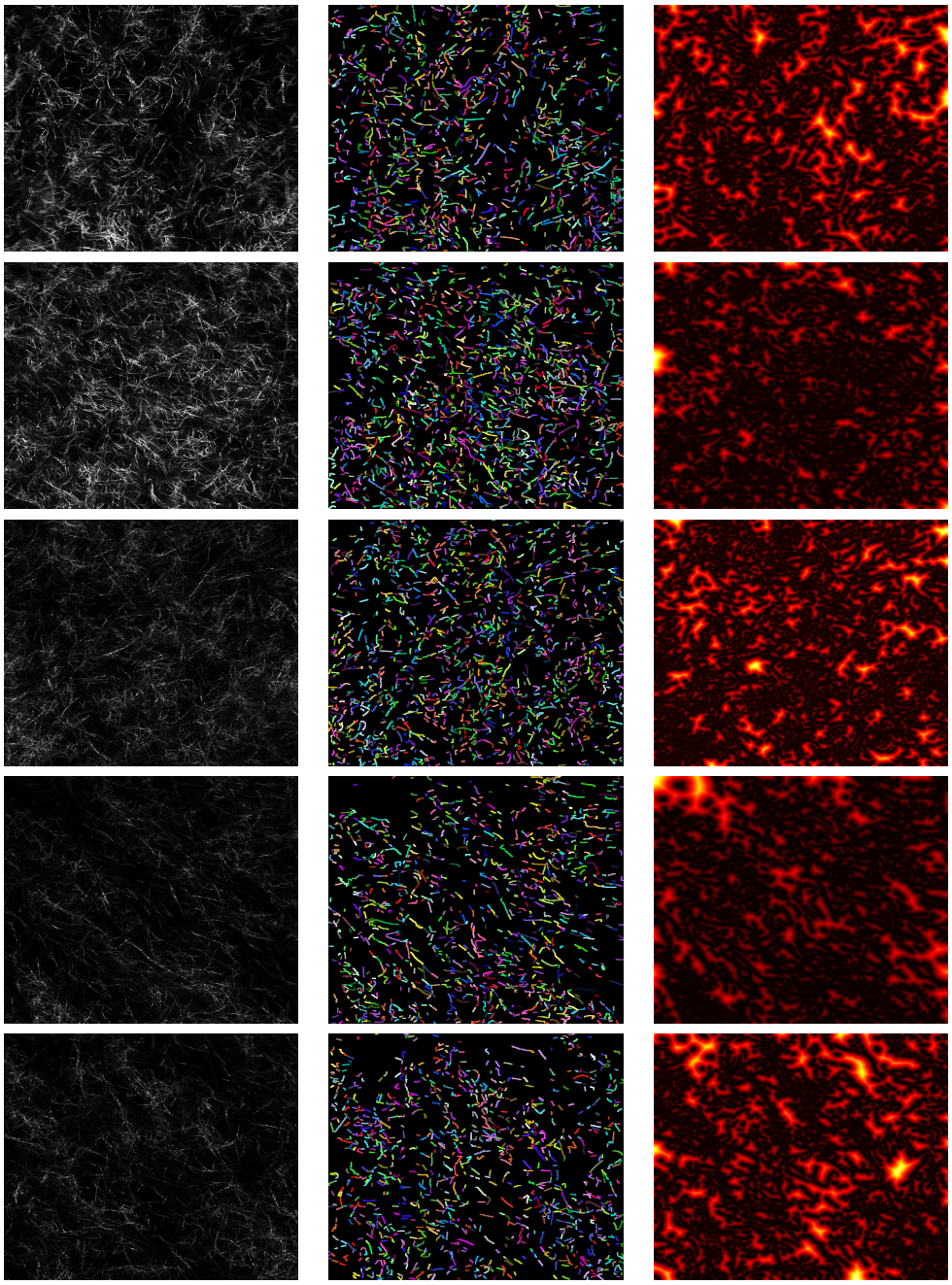
5.5.3 Marangoni Flows:

In the context of breast cancer, it is known that the ECM structure with aligned matrix fibres with additional perpendicular fibres is associated with poor prognosis [372, 106], or outcome. To further test the hypothesis that matrix accumulation affects the cytoskeletal properties, while radial alignment enhances invasion, collagen gels that were polymerised with Marangoni flows were used, as explained previously (Section. 3.3.3). This mimics and models the DIF. The Marangoni flows were compared to a set of controls where no changes have been subjected to the ECM. The images were observed with Trace Ridges, where morphological comparisons were made. This is seen in (Fig. 5.29) and (Fig. 5.30) for the control and Marangoni samples.

Discussion:

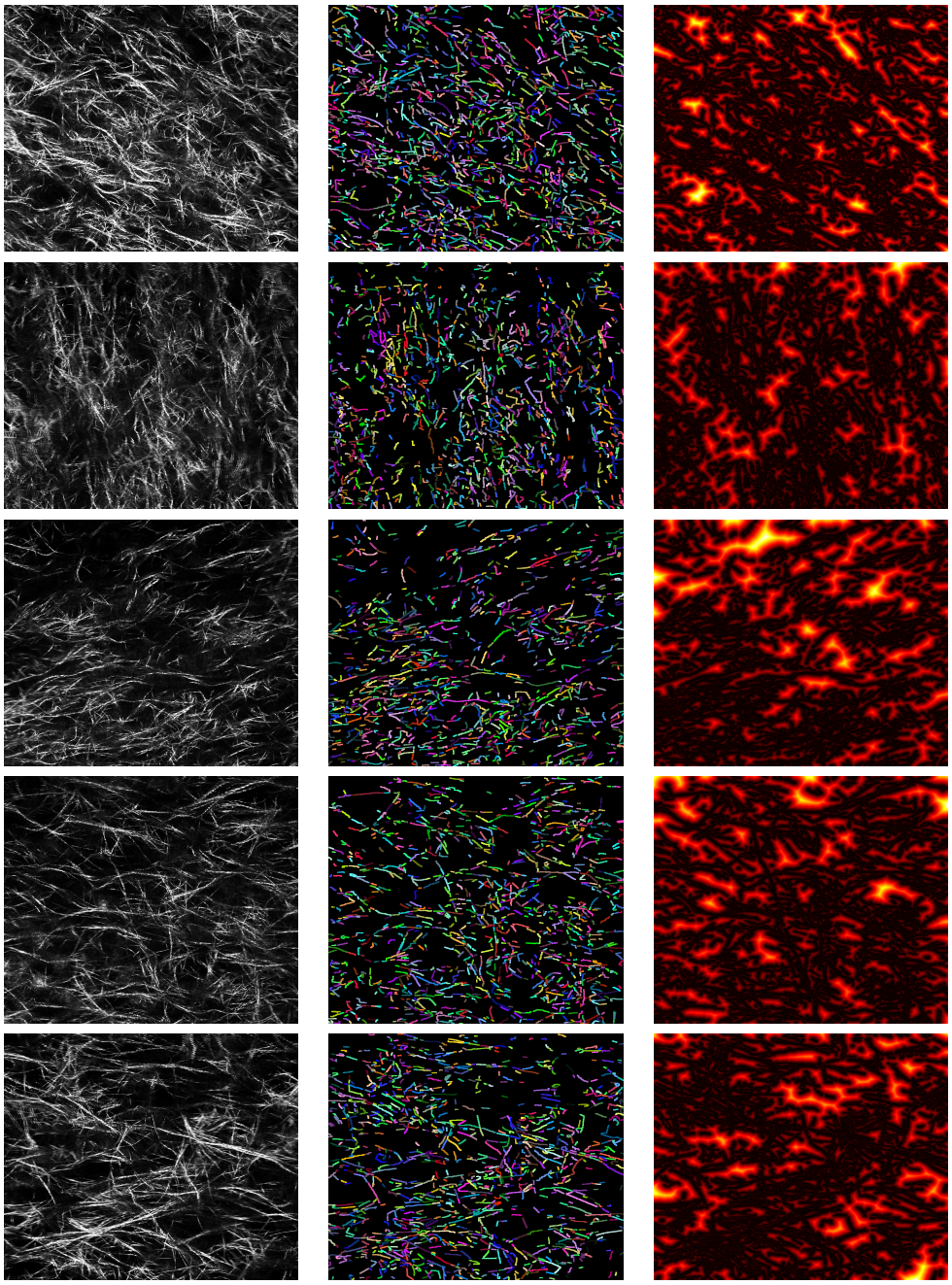
Fibre length, straightness and alignment have increased. While a decrease was observed in the aspect ratio. The Marangoni flows modelled the topological features of the distal invasive front in human and mouse tissue. This is where the fibres were oriented radially towards the centre of the gel [280]. This can be seen in the t-test comparisons in (Fig. 5.31).

These findings support that Marangoni flow-driven matrices promote high actomyosin activity that is driven by abundance and that maximal individual invasion is promoted by long radially oriented fibres.



(a) (b) (c)

Figure 5.29: Control samples of Collagen I before the application of $MgCl_2$. (a) The raw image samples. (b) Collagen fibres detected by Trace Ridges. Fibres have been assigned colours randomly to differentiate between them. (c) Gap analysis maps of the detected fibres. Pixels further away from a fibre form a gap and are shown in yellow and white.



(a) (b) (c)

Figure 5.30: Collagen I fibres after the addition of $MgCl_2$. (a) The raw image samples. (b) Collagen fibres detected by Trace Ridges. Fibres have been assigned colours randomly to differentiate between them. (c) Gap analysis maps of the detected fibres. Pixels further away from a fibre form a gap and are shown in yellow and white.

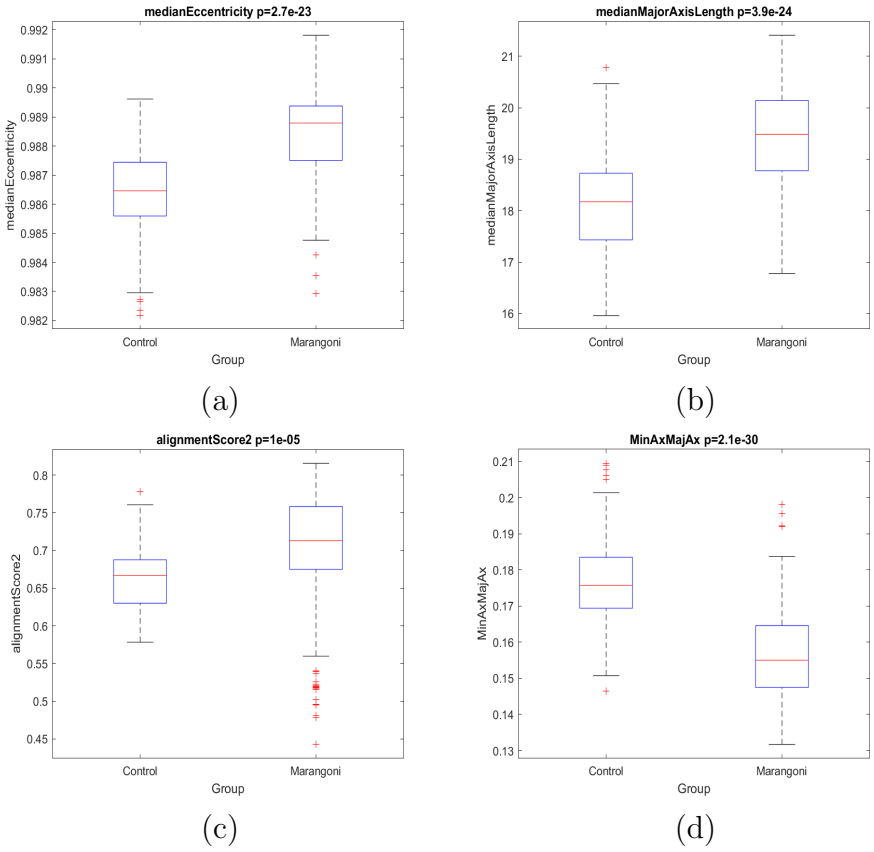


Figure 5.31: Boxplots of comparisons of Collagen fibres architecture and organisation with the addition of $MgCl_2$. (a) The median eccentricity exhibited a $p - value$ of $2.7e^{-}23$. (b) The median major axis length exhibited a $p - value$ of $3.9e^{-}24$. (c) The alignment score which is a measure of how aligned the fibres in each population are exhibited a $p - value$ of $1e^{-}05$. (d) The aspect ratio exhibited a $p - value$ of $2.1e^{-}30$. All these measurements showed a statistical significance when comparing the PI and the control populations.

5.5.4 Conclusion:

Increasing the amount of collagen I around the cells either by adding more collagen or protease inhibition or by Marangoni flows that control the orientation of the collagen fibres. The collagen network becomes denser around the cells. The denser collagen structures confine the cells tightly. As a result, the cells retract their bodies and round up which activates their actomyosin machinery. The increase in the actomyosin contractility allows the cells to generate larger internal forces to handle the pressure from the dense collagen around them. It is the specific arrangement of longer collagen fibres in the radial alignment that actually promote higher levels of cell migration. This structure builds the paths that allow the cells to move more freely through the matrix which supports the migration despite the dense architecture of the ECM. These findings can assist patients with particular tumour architectures to benefit from certain therapies that target the pro-invasive programs associated to collagen architecture.

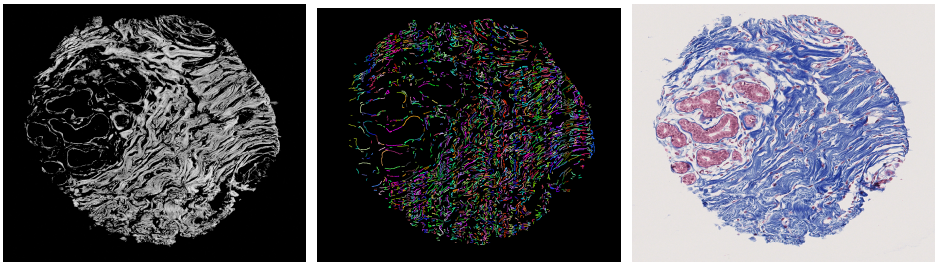
This work sheds light into the importance of the physical conditions in which the matrix organisation play a substantial role in cell migration and cell survival including development, fibrosis, aging, and immune responses. It contributes to the creation of the road map aggressive tumours take before they metastasis. This work exhibits the effective use of Trace Ridges as a computational tool that can be used to assist in the investigation of the behaviour of cancer through quantifying and extracting morphological properties of the extracellular matrix.

5.6 Evaluation of Morphological Properties and Machine Learning Approaches for Tissue Micro-array Image Classification:

Breast Cancer is a disease that affects millions of people [43]. As explained earlier the extracellular matrix plays a significant role in the progression of disease and treatment effects. In the context of cancer, understanding more about the structural changes that occur to the ECM when metastasis occurs could provide insight into how cancer affects the ECM which could lead to better predictions as well as improved treatment plans for patients of particular ECM structures.

Tissue micro-arrays of breast tissue were prepared and supplied by Dr Oscar Maiques (QMUL). Each image of dataset 4 makes up one core. These cores are used for the comparison of the normal and tumour breast tissue. 120 of normal cores (Fig. 5.33a) and 120 of tumour cores (Fig. 5.34a) are compared. The images contain both cells and collagen fibres. The two distinctive properties were separated as explained and shown in (Section. 4.7.4).

Throughout this section, two investigations are constantly mentioned that take place. These investigations use different number of images of dataset 4 for analysis (Section. 3.3.4). Investigation 1 uses only the collagen fibres of images of dataset 4 (Fig. 5.32a, b). While Investigation 2 uses the whole TMA cores of dataset 4 where the images include both the collagen fibres and cell nuclei (Fig. 5.32c). The images have been mentioned before but for clarity they are reproduced here again.



(a) (b) (c)

Figure 5.32: Examples of images used in the investigations of Tissue Micro-Arrays. (a) Investigation 1 uses the isolated collagen fibres from TMA cores. These images are traced by Trace Ridges and later morphological features are extracted. (b) Investigation 1 uses the traced edges of the fibre images of the TMA cores (Fig. 5.32a). (c) Investigation 2 uses the raw TMA cores to include both the fibres and the cell nuclei.

5.6.1 Investigation 1 Metrics:

Firstly, the collagen cores of dataset 4 were assessed and images that contained impurities or artefacts were eliminated from the comparison to provide better and more accurate results. This provided 106 Normal cores and 114 Tumour cores for Investigation 1.

In Investigation 1 the images were passed through Trace Ridges, and 20 morphological metrics were extracted based on the fibres maps in (Fig. 5.33b) and (Fig. 5.34b) for normal and tumour cores respectively as well as the gaps maps in (Fig. 5.33c) and (Fig. 5.34c) for normal and tumour cores respectively. The metrics extracted were either the average or the standard deviation of the morphological features for each image being traced. The values are then grouped based on their population and the populations are compared using a two-sample t-test with a confidence interval of 95%. For a statistical significance to occur between two populations a p-value less than 0.05 must be observed.

There were statistical significances observed when comparing the normal collagen fibres to the tumour collagen fibres (Fig. 5.35)when investigating the average eccentricity, average major axis length, the total area taken up by the gaps, alignment, average mean intensity, average minor axis length, aspect ratio, number of fibres, and std of orientation.

The average eccentricity (Fig. 5.35a) possesses a similar range for both the normal and tumour. The tumour has a positive skewness, therefore, on average there are more tumour samples that possess a higher average eccentricity than that of the normal samples. In addition, the median of the tumour eccentricity is higher than the normal.

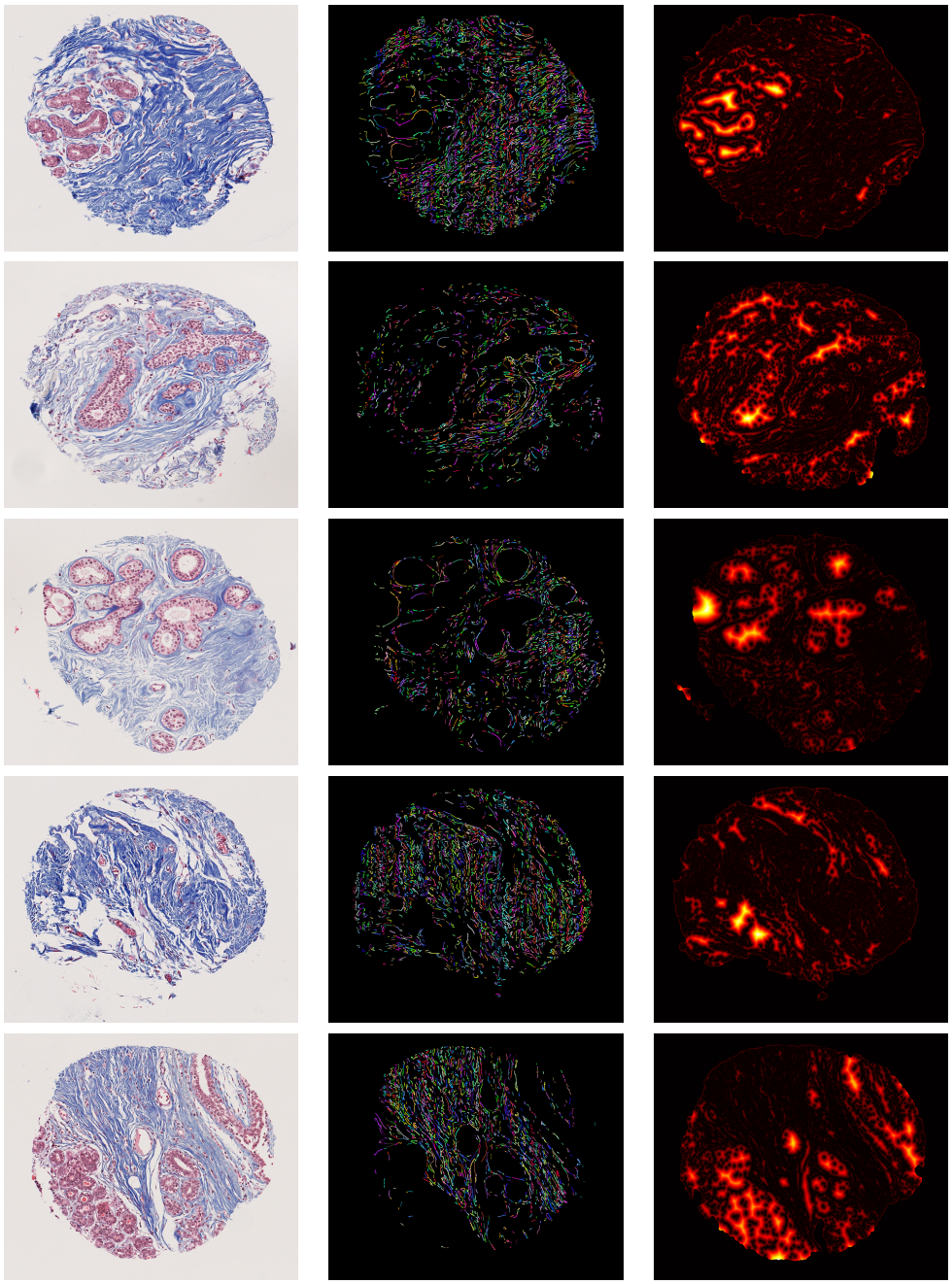
Tumour lengths (Fig. 5.35b) had a wider range than normal, but the normal and tumour samples possessed a comparable median. The tumour samples were positively skewed, having more length values higher than their median. In the interquartile range of the tumour samples, there were more samples that were above the median. This also indicates that those values of the tumour fibre lengths that were below the median were constricted to a small range. The average major axis relative to the image size was the same as the average major axis length, just the length values change relative to the image size.

The total gap area (Fig. 5.35c) increased when comparing normal and tumour samples, the median of the tumour was higher than that of the normal. The normal samples are negatively skewed, therefore there were more values of total gap area below its median, while the tumour had a normal distribution. The interquartile range for the tumour was smaller in comparison to the normal, there were more values below or above the interquartile range of tumour total gap area. The total gap area relative was the same as the total gap area, only the gap area values have changed relative to the image size.

Fibres were more aligned in the tumour samples than the normal samples (Fig. 5.35d). Both populations exhibited a normal distribution and wide ranges.

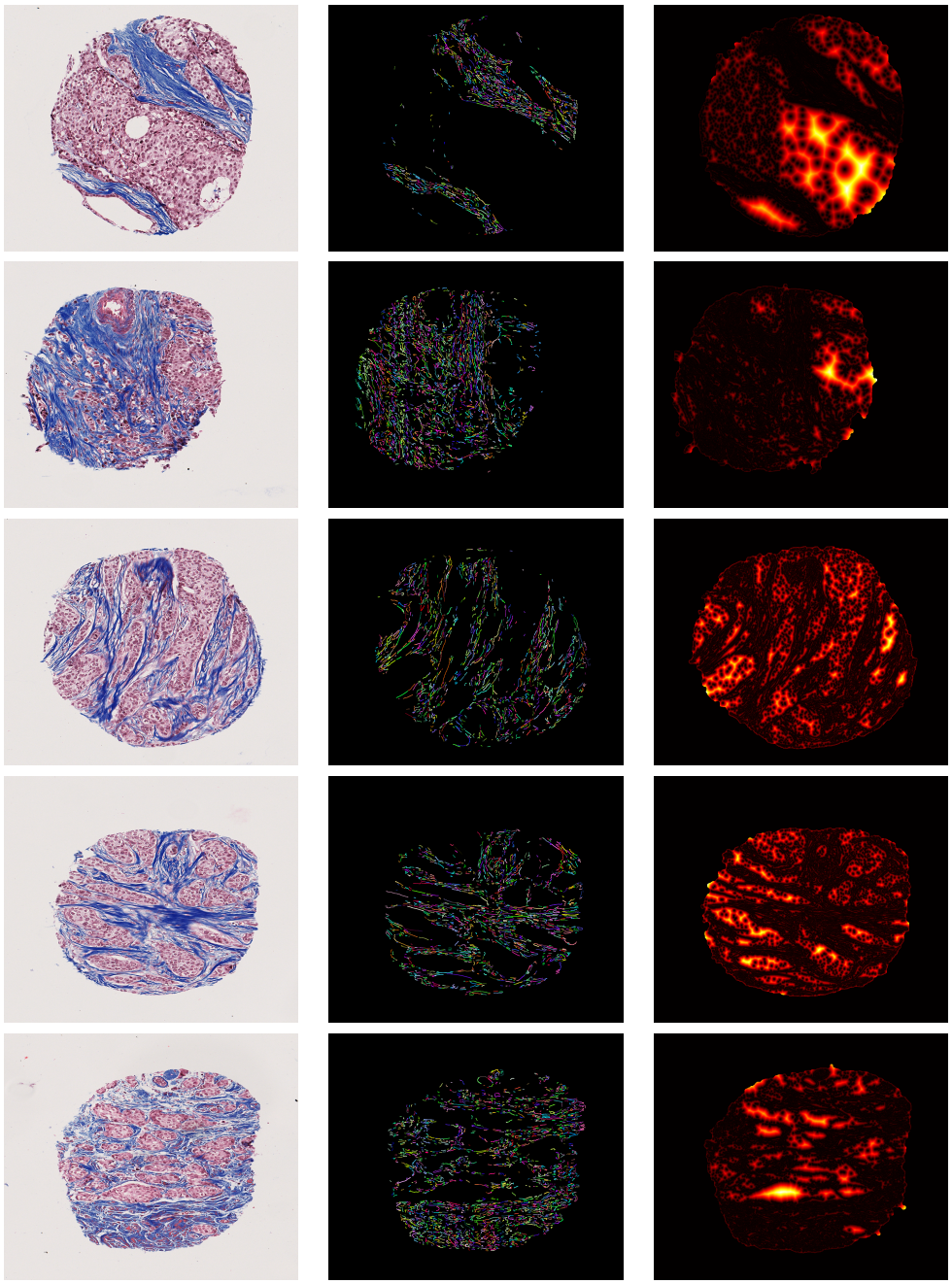
The mean of intensity (Fig. 5.35e) also decreased in the tumour population. Both populations have similar ranges. The normal is negatively skewed therefore the range below the median is larger than the range above the median. The median of normal is higher than that of the tumour.

The average minor axis length (Fig. 5.35f), which is also the width, is lower in the tumour than the normal population. The tumour population has a wider range than that of the normal. The tumour population is negatively skewed, therefore the range of values below the median is larger than that of the range above the median. The median width of the normal population is higher than the tumour population. This result confirms that the longer the fibre the thinner the fibre is.



(a) (b) (c)

Figure 5.33: Fibres and Gaps of the Normal TMA cores detected by Trace Ridges for Investigation 1. (a) Images representing Normal cores. (b) Normal core Fibres detected by Trace Ridges. Colours are assigned randomly to distinguish different edges.Black regions in the images are gaps between the fibres. (c) Normal core Gaps. White-yellow regions represent pixels further away from the fibres.



(a) (b)

Figure 5.34: Fibres and Gaps of the Tumour TMA cores detected by Trace Ridges for Investigation 1. (a) Images representing Tumour cores. (b) Tumour core Fibres detected by Trace Ridges. Colours are assigned randomly to distinguish different edges.Black regions in the images are gaps between the fibres. (c) Tumour core Gaps. White-yellow regions represent pixels further away from the fibres.

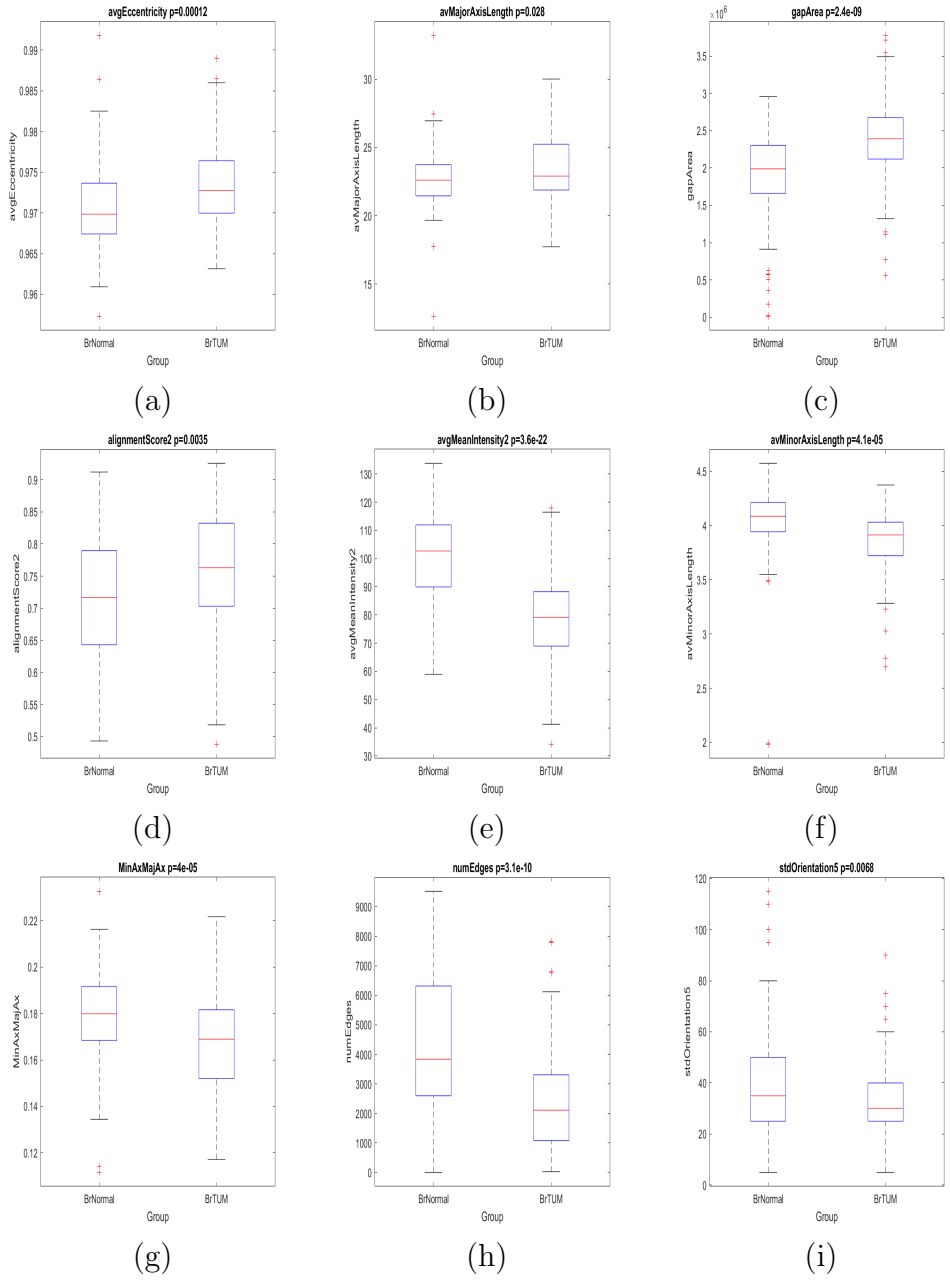


Figure 5.35: Boxplots of morphological metrics that showed statistical significance when comparing the Normal to Tumour. (a) The average eccentricity had a $p-value$ of 0.00012. (b) The average major axis length had a $p-value$ of 0.028. (c) The total gap area had a $p-value$ of $2.4e^{-}09$. (d) The alignment score of the fibres had a $p-value$ of 0.0035. (e) The average mean intensity had a $p-value$ of $3.6e-22$. (f) The average minor axis length had a $p-value$ of $4.1e-05$. (g) The aspect ratio of the fibres had a $p-value$ of $4e^{-}05$. (h) The number of fibres had a $p-value$ of $3.1e-10$ (i) The std of the orientation of the fibres had a $p-value$ of 0.0068.

The aspect ratio (Fig. 5.35g) is the ratio of the minor axis length and major axis length. The tumour aspect ratio median is lower than that of the normal. The tumour population exhibits a larger range of aspect ratio values. The tumour population is negatively skewed. This also confirms that the tumour fibres on average are thinner but longer hence the lower aspect ratio.

Tumour populations exhibit a lower number of fibres given a core sample where their range is smaller therefore the values are more predictable in comparison to the normal cores. The median of the normal population is higher than that of the tumour. The normal and tumour populations are positively skewed (Fig. 5.35h). This metric also supports the theory of bigger total gap areas in tumour populations. The more fibres there are the less gap area the sample would have. This has been confirmed previously in (Section. 5.4).

When the standard deviation of the orientation was compared (Fig. 5.35i). The tumour samples fibre orientations deviated less from the mean of the orientation which could indicate that the fibres are more aligned in comparison to the normal fibres. The normal fibre samples had a wider range as well as a higher median.

Several other metrics have shown a statistical significancewhen comparing the normal to tumour fibre populations (Fig. B.0.1), and (Fig. B.0.2) these metrics are; area taken up by fibres in the ECM, largest gap, average maximum intensity, average minimum intensity, std major axis length, std mean intensity, std minimum intensity, number of regions, std eccentricity, std minor axis length and std maximum intensity.

Using the Bonferroni Correction with 20 hypotheses tested the adjusted p-value becomes 0.0024. Considering the adjusted p-value, the average major axis length, the largest gap, std of the major axis length, std Eccentricity, alignment score are no longer statistically significant.

Investigation 2 Metrics:

To compare the nuclei morphological metrics, dataset 4 was further refined to include only images that are eligible, this reduced the number of images to 174 cores where there was 67 number of Normal cores and 107 number of Tumour cores which make up the images for Investigation 2. The reduction in the number of cores is due to the Normal cores not having sufficient number of nuclei for analysis. In Investigation 2 fibre morphologies were extracted once again through Trace Ridges to check that they are consistent even with a smaller dataset. The cell nuclei were segmented using CellPose (Fig. 5.36a, b) for the normal cores (Fig. 5.36c, d) for the tumour cores. The metrics for the cell morphology were developed in the same way as the Trace Ridges fibre metrics and are outlined in (Section. 3.8.1). Finally, comparisons of normal and tumour populations are conducted.

All of the fibre metrics when comparing Normal cores and Tumour cores maintained their statistical significance. In addition to the fibre metrics further 18 cell metrics were tested.

The statistically significant metrics when comparing the cell nuclei of the normal and tumour populations were the average gap size "distGaps" between the nuclei, the total gap area, the largest gap, area of cells, average mean intensity, average major axis length, average minor axis length, and number of nuclei (Fig. 5.37).

The average gap area (Fig. 5.37a) between the tumour nuclei was smaller than that of the normal nuclei. The range of the average gap area in the normal is slightly wider than that of the tumour. The tumour range is very small, that the maximum and minimum average values are not far from each other. The median of the normal samples is higher than the median of the tumour samples.

The total gap area (Fig. 5.37b) between the nuclei in the tumour samples was smaller than that of the normal samples, both samples follow a normal distribution and exhibit a similar range. The median of the normal samples is higher than the tumour.

The normal population exhibited larger gaps (Fig. 5.37c) between the nuclei than the tumour nuclei. The normal population had a wider range with a higher median. Both populations were positively skewed.

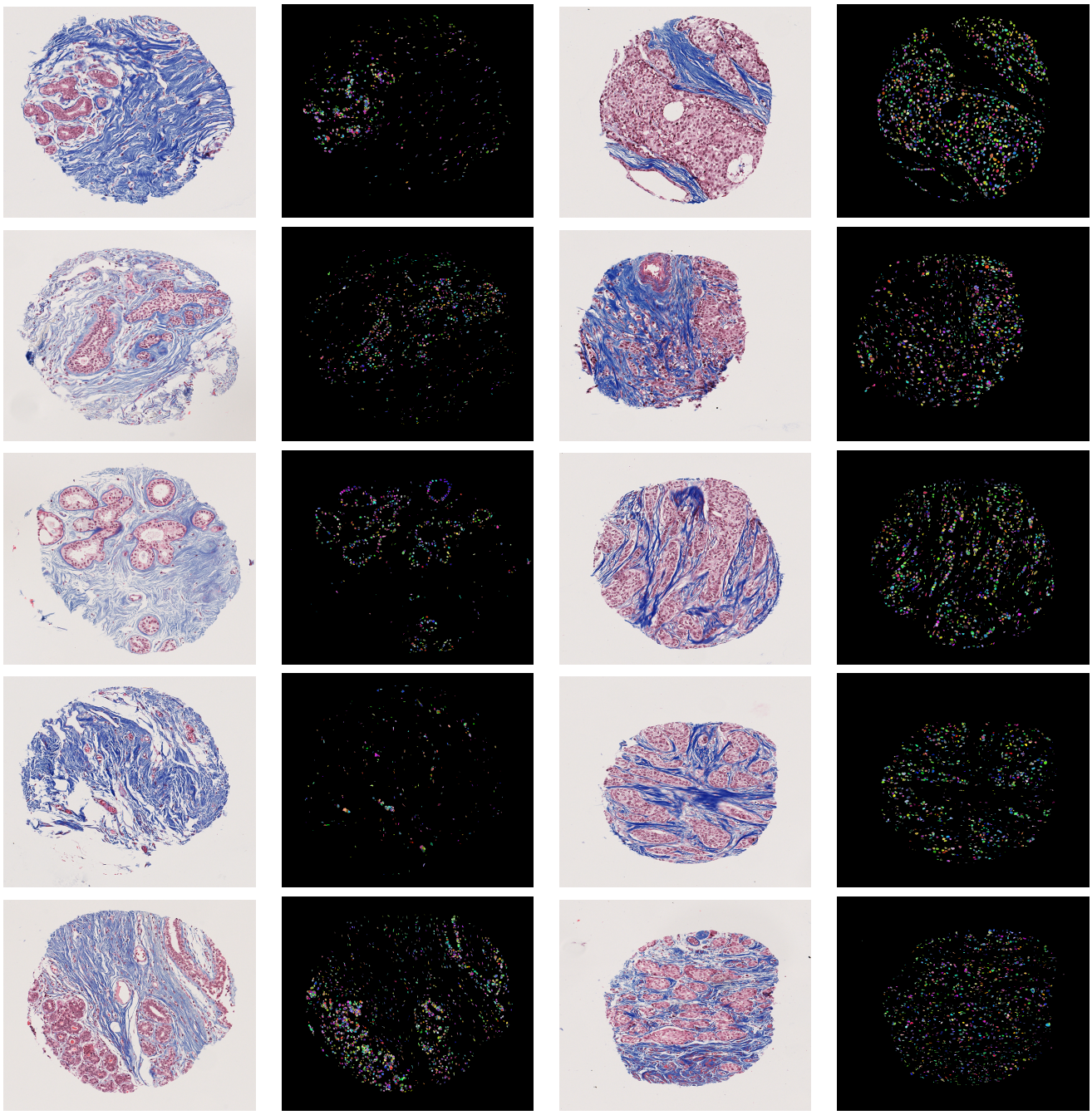
These three metrics are consistent, averaging smaller so there are less gaps and the largest gaps in each sample are smaller than normal.

The area of cells (Fig. 5.37d) is the total area taken up by the nuclei in the core, the tumour nuclei exhibited a larger total area taken up per sample in comparison to the normal samples. The range of the total area of the nuclei is bigger than normal. The normal population has a smaller median than the tumour.

The average mean intensity (Fig. 5.37e) of nuclei in the tumour was higher than the normal, both populations have a normal distribution. Both populations have comparable ranges, but the normal samples are negatively skewed. The median value of the normal is smaller than the tumour.

The intensity values are cohesive where the tumour nuclei exhibit a higher minimum intensity (Fig. C.0.1c) and a higher maximum intensity (Fig. C.0.1d) therefore, they will have a larger mean intensity.

The average major axis length (Fig. 5.37f) is higher in the tumour nuclei. Both popu-



(a) (b) (c) (d)

Figure 5.36: Cells of the Normal TMA cores detected by CellPose. Colours are assigned randomly to distinguish between the different cell nuclei. (a) Images representing Normal cores. (b) Images representing Normal nuclei found in the cores. (c) Images representing Tumour cores. (d) Images representing Tumour nuclei found in the cores. Black regions in the images are gaps between the nuclei.

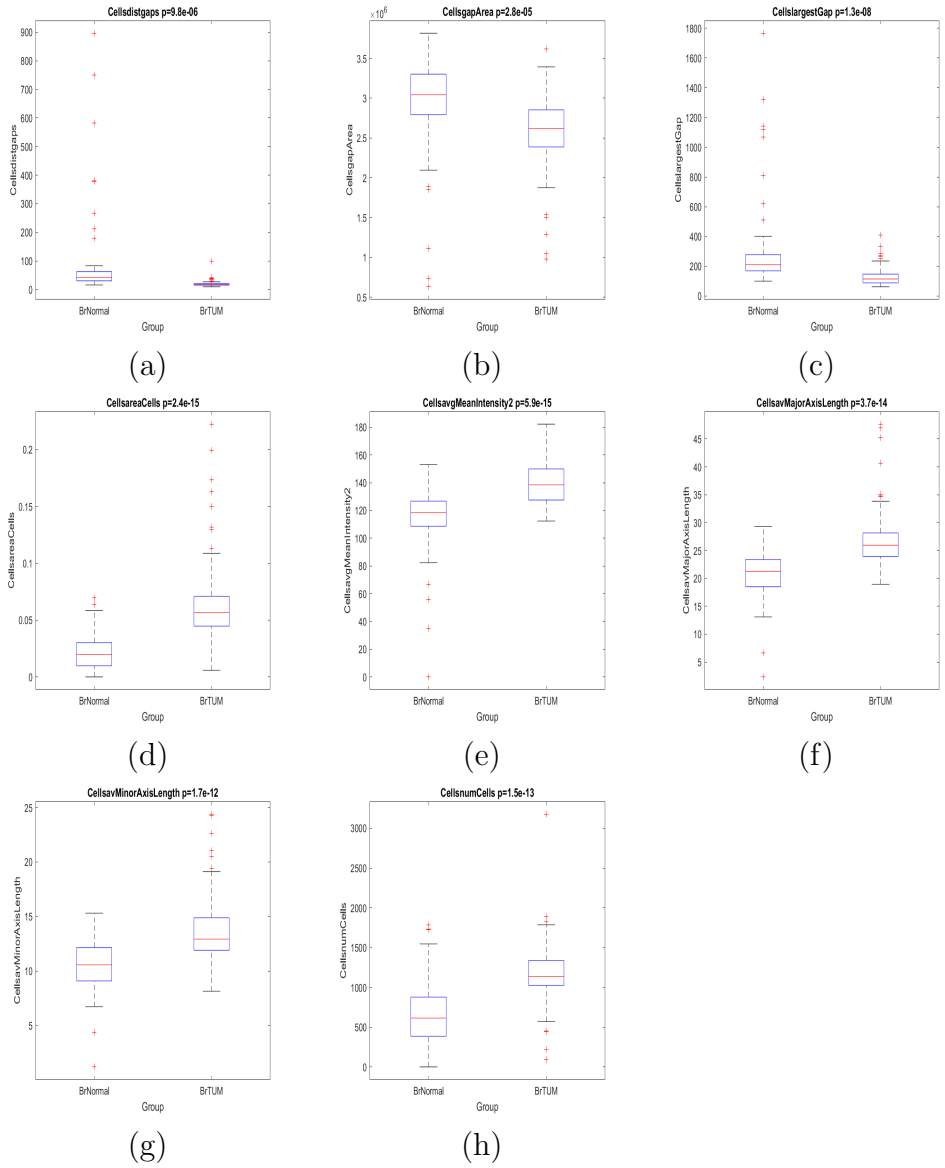


Figure 5.37: Boxplots of cell nuclei metrics when comparing Normal and Tumour cell nuclei. (a) The average gap area between the nuclei had a $p - value$ of $9.8e - 06$. (b) The total gap area between the nuclei had a $p - value$ of $2.8e - 05$. (c) The largest gap area between the nuclei had a $p - value$ of $1.3e - -5$. (d) The total area taken up by the nuclei had a $p - value$ of $2.4e - 15$. (e) The average mean intensity had a $p - value$ of $5.9e - 15$. (f) The average major axis length had a $p - value$ of $3.7e - 14$. (g) The average minor axis length had a $p - value$ of $1.7e - 12$. (h) The number of cells had a $p - value$ of $1.5e - 13$.

lations exhibit a similar range. The median of the average length of the tumour nuclei was higher than that of the normal nuclei.

The average minor axis length (Fig. 5.37g)of the tumour nuclei was higher than that of

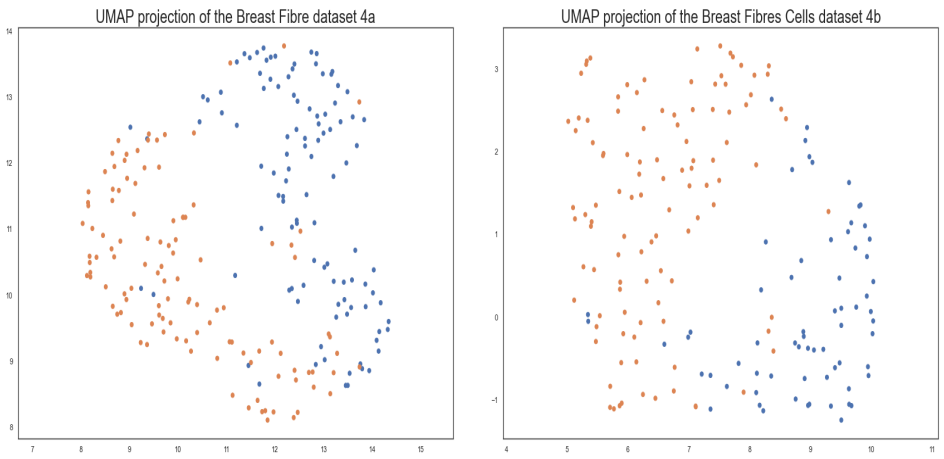
the normal. The tumour had a wider range than the normal as well as a larger median. The tumour population was positively skewed.

The tumour nuclei cells were on average longer and wider, therefore they were bigger in size than nuclei found in the normal cores.

The number of cell nuclei (Fig. 5.37h) in the tumour cores was higher than the normal cores, in addition to this the median of the tumour cell nuclei is slightly smaller than the highest values of cell nuclei count in the normal cores. The tumour population is positively skewed. The normal population exhibits a larger range than the tumour population.

Several other metrics have shown statistical significance when comparing the normal nuclei population to the tumour nuclei population (Fig. C.0.1) These metrics were; std Eccentricity, avg mean intensity, average (avg) minimum intensity, avg maximum intensity, avg major axis length, avg minor axis length, number of cells, number of regions, standard deviation (std) major axis length, std minor axis length, std minimum intensity.

Using the Bonferroni correction the adjusted p-value for 39 metrics of fibres and cells is 0.0013. Therefore, the average minimum intensity of cells, std eccentricity and std mean intensity of the cells are added to the list of the no longer statistically significant metrics.



(a) (b)

Figure 5.38: UMAP representation of fibres and cells of TMA cores. The Normal samples are shown in blue, and the tumour samples are shown in orange. (a) UMAP representation of Investigation 1 features of just fibres. (b) UMAP representation of Investigation 2 features of both fibres and cells. It can be seen that in both cases the normal and tumour populations can be differentiated from each other.

5.6.2 UMAP Representation:

It is clear that the Normal and Tumour fibres and cell nuclei exhibit different characteristics that can distinguish them from each other. To further visualise these characteristics, the metrics extracted for each image were saved and a UMAP was used to visualise the sets of features. The UMAP reduces the number of features from 20 in the fibres to 2 and 38 in fibres and cells into 2 dimensions.

It can clearly be seen from (Fig 5.38a) and (Fig 5.38b) that the populations are easily separable when their dimensionality is reduced, this further confirms the findings that the populations have features which make them distinct.

5.6.3 Investigation 1 classification:

With the metrics extracted by Trace Ridges, machine learning algorithms can be trained for classification. A set of unseen image features can be classified to either belonging to Normal or Tumour. Furthermore, deep learning algorithms can also be trained to classify unseen images as either Normal or Tumour.

Machine learning is a subclass of algorithms that allows machines to understand and learn without programming beforehand. The main function of machine learning is to employ advanced algorithms and statistical techniques to observe the data presented and make a classification, prediction, or cluster the data based on their properties.

Many machine learning and deep learning methods have been utilised to examine medical data. With the advancements of technologies and the retrieval of countless amounts of

Model	Training	Testing	Type
Decision Trees	176	44	Feature set
Nave Bayes	176	44	Feature set
Random Forest	176	44	Feature set
SVM	176	44	Feature set
Logistic Regression	176	44	Feature set
KNN	176	44	Feature set
MLP	176	44	Feature set
VGG16	13869	3467	Images
ViT	5246	1311	Images

Table 5.9: Number of feature sets and images used for training and testing for the classification task of investigation 1. Only VGG16 and ViT used images, the rest of the models were trained and tested on feature sets.

predictor diagnostic features and images, medical professionals are striving for faster and more accurate diagnosis of disease and treatment selection. Machine learning has been used to classify medical conditions in [339]. Commonly used machine learning algorithms on medical data include Decision Trees, Naive Bayes, Random Forest, Support Vector Machine, K-Nearest Neighbours, Multi-Layer Perceptron. These algorithms have been explained in detail with their implementation in (Chapter 3.6.9).

The algorithm was solely trained on the fibres which is just the fibres traced by Trace Ridges and their metrics extracted which are the same images used in Investigation 1 (Fig. 5.32a). 80% of the data was used for training and 20% used for testing. This makes up 44 image feature sets with 20 for normal and 24 for tumour, a summary of the number of feature sets can be found in Table. 5.9. Additionally, K-Fold validation for 5 folds was used to validate the training and reduce bias caused by a single train test split, it is an approach commonly used with small datasets. Precision, Recall, and F1-Score were extracted for each class and finally the overall accuracy.

The algorithms were successfully able to classify unseen image feature sets accurately setting an average accuracy of 87%. Looking at the table of results Table. 5.10, Random Forest achieves the highest precision for the normal class, Random Forest, SVM and MLP for the tumour class. The recall for the normal class all algorithms except Logistic Regression achieve a score above 90%, for the tumour class Random Forest, Logistic Regression and KNN achieve the highest score. F1-Score is used as it combines both precision and recall. An ideal algorithm should have a high precision and recall. The highest F1 score for both the normal and tumour classes was achieved by Random Forest with 93% and 93% respectively. Subsequently Random Forest achieved the highest accuracy of 93%, followed by SVM and KNN with 91% and MLP with an accuracy of 89%. A confusion matrix for each of the algorithms is presented (Fig 5.39). The number indicates the number of feature sets classified correctly for a particular class (TP) and misclassified (FN), classified correctly as the other class (TN), classified incorrectly as this class (FP).

To further test classification techniques, deep learning algorithms were trained on the iso-

	Metrics						
	Precision		*Recall*		*F1-Score*		*Accuracy*
Method	Normal	TUM	Normal	TUM	Normal	TUM	
Decision Trees	0.78	0.90	0.90	0.79	0.84	0.84	0.84
Nave Bayes	0.79	0.95	0.95	0.79	0.86	0.86	0.86
Random Forest	**0.87**	**1.00**	**1.00**	0.88	**0.93**	**0.93**	**0.93**
SVM	0.83	**1.00**	**1.00**	0.83	0.91	0.91	0.91
Logistic Regression	0.81	0.75	0.65	0.88	0.72	0.81	0.77
KNN	0.86	0.95	0.95	0.88	0.90	0.91	0.91
MLP	0.80	**1.00**	**1.00**	0.79	0.89	0.88	0.89
VGG16_1 unseen	0.83	0.88	0.87	0.84	0.85	0.86	0.85
VGG16_2 unseen	0.85	0.88	0.88	0.85	0.86	0.87	0.86
ViT_1 unseen	0.80	0.84	0.83	0.81	0.82	0.82	0.82
ViT_2 unseen	0.84	0.81	0.78	0.86	0.81	0.84	0.82
VGG16_1 post TR	0.68	0.68	0.67	0.67	0.67	0.68	0.68
VGG16_2 post TR	0.81	0.61	0.41	0.91	0.55	0.73	0.66
ViT_1 post TR	0.75	0.55	0.18	0.94	0.29	0.69	0.57
ViT_2 post TR	0.84	0.54	0.11	**0.98**	0.19	0.69	0.55

Table 5.10: Classification metrics of machine learning and deep learning algorithm on images of investigation 1 using Decision Trees, Naive Bayes, Random Forest, SVM, Logistic Regression, KNN, MLP that use features extracted by TR, VGG16_1 on unseen images, VGG16_2 on unseen images, ViT_1 on unseen images, ViT_2 on unseen images (Fig. 5.32a), VGG16_1 post tracing with TR, VGG16_2 post tracing with TR, ViT_1 post tracing with TR and ViT_2 post tracing with TR (Fig. 5.32b) that use only the fibres images of dataset 4, this group of images makes up the images for Investigation 1.

lated fibre images (Fig. 5.32a) as well as the fibre tracing output of Trace Ridges (Fig. 5.32b). VGG-16 and Vision Transformers were used for the classification tasks. Each were trained with 80% of the image patches and tested on 20% of the image patches that were not processed through TR (Fig. 5.32a). VGG-16 had 17,336 patches for training and testing while ViT had 6557 patches for training and testing after removing completely empty images, a summary of the number of images used for training and testing can be found in Table. 5.9. One of the limitations of this experiment is the lack of a training set for the deep learning models due to data size constraints. Additionally, the difference in the number of images used for the VGG16 and ViT models is due to their image input size requirements. For both VGG-16 and ViT algorithms, they were trained with a learning rate of 0.0001 and two batch sizes were used of 16 (VGG16_1 and ViT_1) and 32 (VGG16_2 and ViT_2). Both the algorithms employed stochastic gradient descent momentum (sgdm) as the optimiser. Code for training and testing the machine learning and deep learning pipelines is publicly available in *https* : $//github.com/youssefarafat/Trace_Ridges/tree/main/Machine\%20Learning\%20Models$.

Looking at Table. 5.10 of the classification of fibre images with VGG16 and ViT unseen, VGG16_2 achieved the highest precision score for the normal population with a score 84.73%. For the tumour class VGG16_2 achieves the highest score of 88.11%. For the recall VGG16_2 achieves the highest score for the normal class with 87.63%, while ViT_2 achieves a score of 86.04% for the tumour class. For both the normal and tumour classes VGG16_2 achieved the highest F1-Score of 86.2% and 86.7%, respectively. The highest accuracy of 86.4% was achieved by the VGG16_2.

Fibres were traced with Trace Ridges and the traces were saved as images (Fig. 5.32b) which made up the same number of patches as the unseen images. The images were then used for classification tasks of the VGG16 and ViT networks. The results of the classification performance can be seen in Table. 5.10 and are referred to as VGG16 and ViT post TR. For the precision, ViT_2 achieved the highest score for the normal class with 84.2% and VGG16_1 achieved the highest score of 67.6% for the tumour class. ViT_2 suffers in the recall score for the normal class with a score of 10.5% and VGG16_1 achieves the highest score of 66.5%. For the tumour class, the highest score was achieved by the ViT_2 with a score of 98.13%. Therefore, the F1-Score must be looked at as it balances both the recall and precision. The highest F1-Score for the normal class of 67.1% was achieved by the VGG16_1, the ViT_2 also suffered with a score of 18.7%. For the tumour class both ViT_1 and ViT_2 achieved a score of 69.3% while VGG16_1 had a score of 68.3%. When looking at the overall accuracy, VGG16_1 achieved the highest value of 67.7% followed by VGG16_2 with 66.1%, ViT_1 with 57.2% and 55.5% of ViT_2.

The findings illustrate that the best classification methods to use in the case of this dataset or a similar dataset are the machine learning algorithms and specifically the Random Forest or SVM. Decision trees achieved an accuracy that is 2% less than the highest accuracy achieved by the deep learning architectures and Logistic Regression had an accuracy that was 9% lower. The machine learning algorithms required Trace Ridges, while the deep learning algorithms learned better on their own, without needing the tracing of Trace Ridges

to achieve the highest accuracy between them. Vision Transformers require substantial amounts of data for high classification accuracy. Image data was limited which could be a factor that influenced the performance of the Vision Transformers in this task. Trace Ridges is required to achieve the highest classification accuracy possible, by tracing the collagen fibres, extracting their metrics and providing the set of features to the machine learning algorithms. Confusion matrices were generated for all the classification tasks which are illustrated in (Fig 5.39) and (Fig 5.40). From the confusion matrices the precision, recall, F1-Score for each class and overall accuracy can be derived.

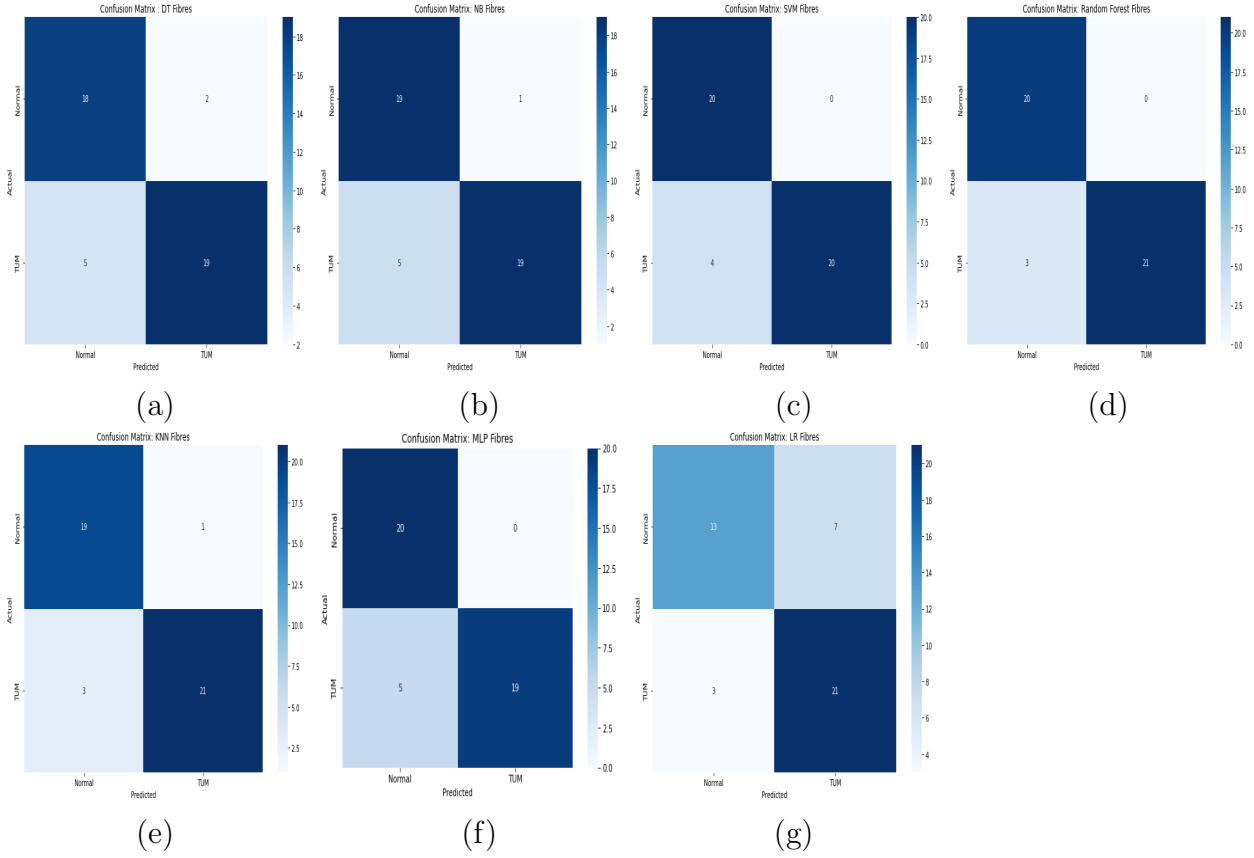


Figure 5.39: Confusion Matrices of the classification tasks on dataset 4a. (a) Decision Trees. (b) Naive Bayes. (c) Support Vector Machines. (d) Random Forest. (e) K Nearest Neighbour. (f) Multi Layer Perceptron. (g) Logistic Regression.

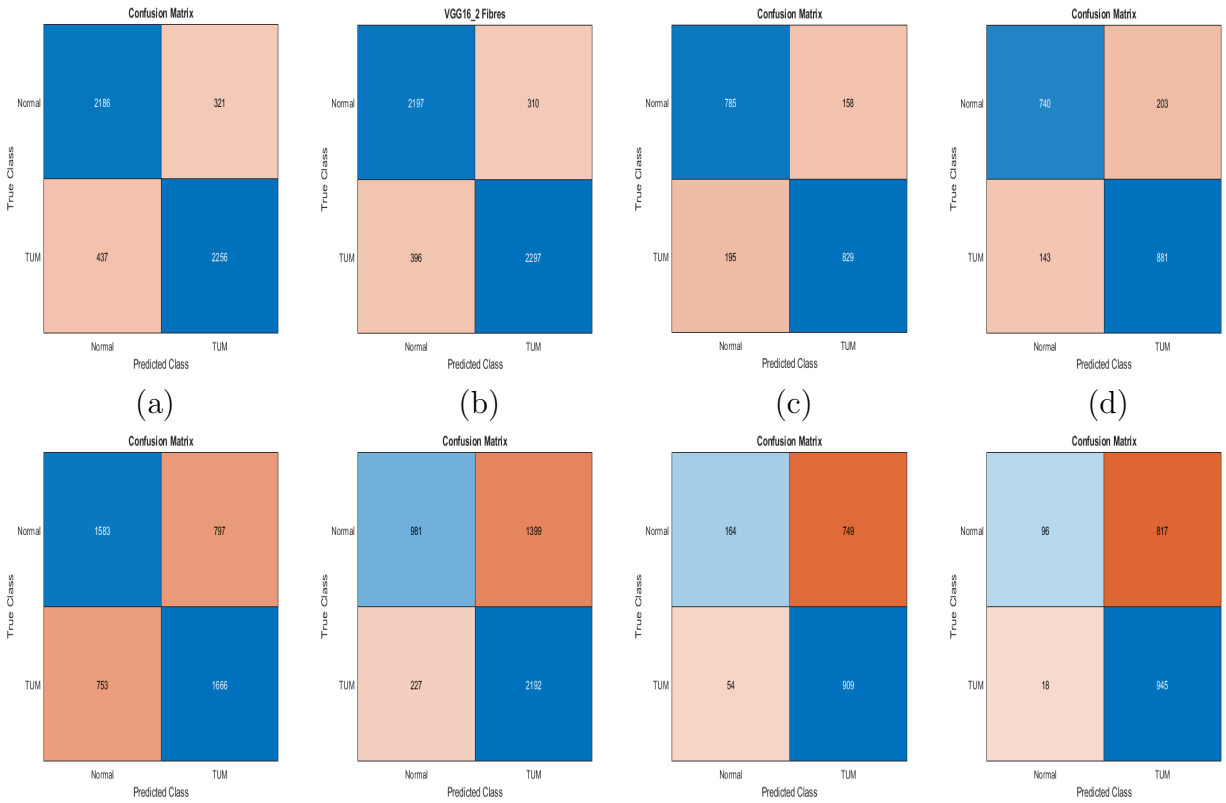


Figure 5.40: Confusion Matrices of the classification tasks on dataset 4a continued. (a) VGG Net_1. (b) VGG Net_2. (c) ViT Net_1. (d) ViT Net_2. (e) VGG16 Net_1 after Trace Ridges. (f) VGG16 Net_2 after Trace Ridges. (g) ViT Net_1 after Trace Ridges. (h) ViT Net_2 after Trace Ridges.

Model	Training	Testing	Type
Decision Trees	139	35	Feature set
Nave Bayes	139	35	Feature set
Random Forest	139	35	Feature set
SVM	139	35	Feature set
Logistic Regression	139	35	Feature set
KNN	139	35	Feature set
MLP	139	35	Feature set
VGG16	14209	3552	Images
ViT	4180	1045	Images

Table 5.11: Number of feature sets and images used for training and testing for the classification task of investigation 2. Only VGG16 and ViT used images, the rest of the models were trained and tested on feature sets.

5.6.4 Investigation 2 classification:

Similarly, for investigation 2, machine learning and deep learning algorithms were also trained to classify the images of TMA cores as a whole (Fig. 5.32c). The images contained both fibres and cells. The fibre features were extracted using Trace Ridges (Fig. 5.33b) and (Fig. 5.34b) while the cell nuclei are segmented using CellPose (Fig. 5.36b, d). The feature sets were also partitioned 80% for training and 20% for testing, this resulted in 35 feature sets with 19 for normal and 16 for the tumour class, a summary of the number of feature sets can be found in Table. 5.11. Additionally, K-Fold validation for 5 folds was used to validate the training and reduce bias caused by a single train test split, it is an approach commonly used with small datasets.

Precision, Recall, F1-Score, and accuracy were calculated. Looking at the results of the machine learning algorithms Table. 5.12, Random Forest, SVM, KNN and MLP achieve the highest precision score for the normal population of 100%, SVM and MLP also achieve the highest score for tumour of 80%, while KNN drops to 73% and Random Forest to 67%. For the recall, SVM and MLP achieved the highest score of 79%. For the tumour class, Random Forest, SVM, KNN and MLP all have achieved a score of 100% followed by Decision Trees and Naive Bayes with 94%. The highest F1-Score for both the normal and tumour was achieved with SVM and MLP with scores of 88% and 89%, respectively for normal and tumour. The highest accuracy was also achieved by SVM and MLP with an accuracy of 89% followed by Decision Trees and KNN with an accuracy of 83%.

The accuracy of classification of the fibres (93% with Random Forest) was marginally higher than the accuracy achieved for the classification of both features of fibres and cell nuclei (89% with SVM and MLP). SVM would be the preferred method for the overall classification of those two datasets. SVM achieved an accuracy of 91% with images of Investigation 1 and 89% with images of Investigation 2 to give an average of 90%. Random Forest on the other hand achieved an accuracy of 93% with images of Investigation 1 and 77% with images of Investigation 2 to give an average of 85%. MLP achieved an accuracy of 89% on images of Investigation 1 and 89% on images of Investigation 2 to give an average

	Metrics						
	Precision		*Recall*		*F1-Score*		*Accuracy*
Method	Normal	TUM	Normal	TUM	Normal	TUM	
Decision Trees	0.93	0.75	0.74	0.94	0.82	0.83	0.83
Nave Bayes	0.93	0.71	0.68	0.94	0.79	0.81	0.80
Random Forest	**1.00**	0.67	0.58	**1.00**	0.73	0.80	0.77
SVM	**1.00**	0.80	0.79	**1.00**	0.88	0.89	0.89
Logistic Regression	0.79	0.62	0.58	0.81	0.67	0.70	0.69
KNN	**1.00**	0.73	0.68	**1.00**	0.81	0.84	0.83
MLP	**1.00**	0.80	0.79	**1.00**	0.88	0.89	0.89
VGG16_1 unseen	0.96	0.95	0.93	0.98	**0.95**	0.96	**0.96**
VGG16_2 unseen	0.96	**0.96**	**0.94**	0.97	**0.95**	**0.97**	**0.96**
ViT_1 unseen	0.93	0.82	0.68	0.96	0.78	0.89	0.85
ViT_2 unseen	0.90	0.87	0.79	0.94	0.84	0.91	0.88

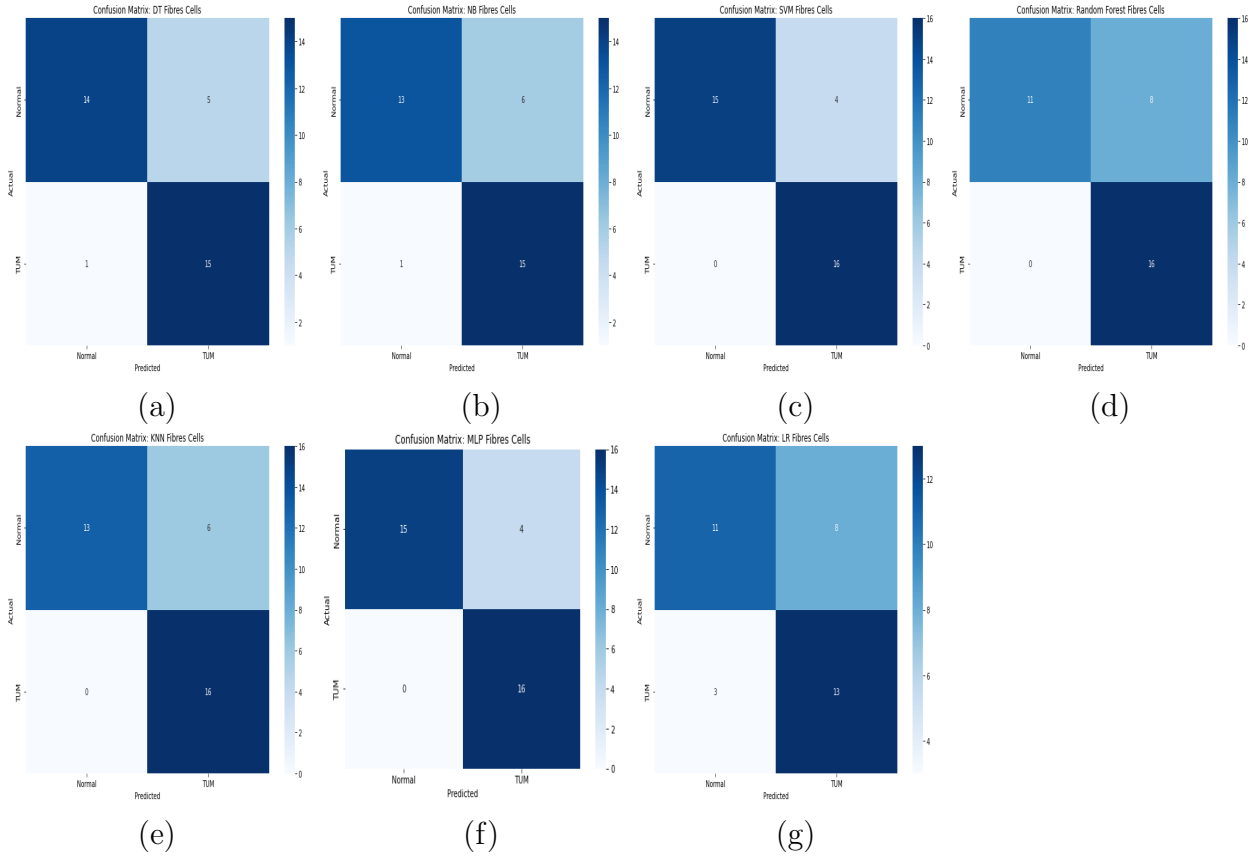
Table 5.12: Classification metrics of machine learning and deep learning algorithm on images of investigation 2 using Decision Trees, Naive Bayes, Random Forest, SVM, Logistic Regression, KNN, MLP that use features extracted by TR, VGG16_1 on unseen images, VGG16_2 on unseen images, ViT_1 on unseen images, ViT_2 on unseen images (Fig. 5.32c), that use both the fibres and cell nuclei images of dataset 4, this group of images makes up the images for Investigation 2 (Fig. 5.32c).

of 89%. SVM and MLP seem to have performed exceedingly well in this classification task. The confusion matrices can be seen in (Fig 5.41).

Similarly, the deep learning techniques VGG16_1, VGG16_2, ViT_1 and ViT_2 were trained and tested on the unseen images of dataset 4 that contain both fibres and cells those make up the images for Investigation 2. VGG-16 had 17,761 patches for training and testing while ViT had 5225 patches after removing completely empty images, a summary of the number of images used for training and testing can be found in Table. 5.11. Observing Table. 5.12, VGG16_2 achieved the highest precision score of 96% for both normal and tumour while VGG16_1 achieved a score of 96% for the normal population and 95% for the tumour population. In terms of precision the ViT networks perform well with the normal population achieving a score of 93% and 90% respectively while for the tumour population their performance was hindered achieving a score of 82% and 87% respectively. In terms of the recall metric, the VGG networks continued their great performance where the VGG16_1 network achieved a score of 93% and 98% for the normal and tumour populations respectively. VGG16_2 achieved scores of 94% and 97% respectively. For the F1-Score VGG16_2 achieves the highest scores of 95% and 97% for the normal and tumour populations followed by VGG16_1 with scores of 95% and 96% respectively. The highest accuracies were achieved by the VGG16 networks, where they both achieved an accuracy of 96% followed by ViT_2 with an accuracy of 88% and ViT_1 with an accuracy 85%. The improvement seen in the performance of the deep learning networks with images in Investigation 2 in comparison to

images in Investigation 1 could be attributed to the increased complexity of the images as in the case of Investigation 2 they contain both fibres and cells while Investigation 1 only contains fibres. This could have led to the networks to learn noise. ViT's require a large quantity of training images but for this investigation the number of images used for training were not sufficient to reach the full potential of the network. Performance metrics can also be derived from the confusion matrices (Fig. 5.42).

VGG16 networks performed better in this classification task in comparison to the machine learning techniques. Only SVM and MLP achieved accuracies higher than the ViT networks. This could be attributed to the image complexities changing.



(a) (b) (c) (d)

(e) (f) (g)

Figure 5.41: Confusion Matrices of the classification tasks on dataset 4b. (a) Decision Trees. (b) Naive Bayes. (c) Support Vector Machines. (d) Random Forest. (e) K Nearest Neighbour. (f) Multi Layer Perceptron. (g) Logistic Regression.

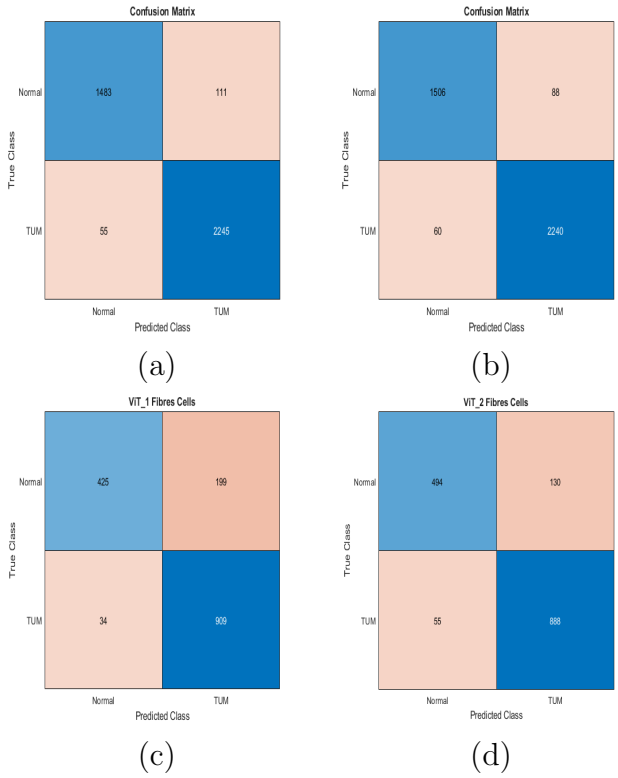


Figure 5.42: Confusion Matrices of the classification tasks on dataset 4b continued. (a) VGG Net_1. (b) VGG Net_2. (c) ViT Net_1. (d) ViT Net_2.

5.6.5 Investigation 1 Clinical Information:

In the process to extract more meaningful connections and a deeper understanding of the dataset. Clinical information was provided for each image of the tumour samples of images used in Investigation 1 (Fig. 5.32a). The clinical information provided covered age, subtype, ER (Estrogen Receptors) (positive or negative), PR (Progestrone Receptors) (positive or negative), CERB2 a protein which plays a role in cell growth and division, the gene that codes CERB2 may be over expressed in breast cancer [384] (when two positives were present FISH a technique used to locate and detect specific DNA sequences on chromosomes [268] was conducted to check whether its truly a positive or a negative), stage of cancer and whether lymph node metastasis occurred. For each category, the images were divided into populations for example. (ER samples positive and negative) and (lymph nodes of 0 and 1).

Trace Ridges was then used to trace the fibres in the images to extract morphological features to compare the populations of data.

Firstly, for the ER, out of 114 tumour images (Fig. 5.32a), 105 were positive and 9 were negative.

Out of number of fibres, average major axis length, average minor axis length, std of major axis length, average eccentricity, gap Area, largest Gap, MinAxMajAx (aspect ratio), std of Orientation and alignment Score.

The average eccentricity, aspect ratio and standard deviation of orientation showed a statistical significance when the populations were compared with one another (Fig. 5.43a-c).

Samples exhibiting a negative ER had a higher aspect ratio than the positive samples but a small range in comparison to the positives. Standard deviation of the orientation of the negative was much higher than the positives therefore fibres of the negative samples largely deviated away from the normal in comparison to the positive samples while having a much larger range than the positives.

For the PR, from 114 tumour images 63 were positive and 51 were negative.

Out of all the metrics analysed, average major axis length, aspect ratio, standard deviation of major axis length showcased statistical significance when the populations were compared (Fig. 5.44)

Positive PR fibres experienced larger lengths in average and a higher median while both ranges of positive and negative samples were comparable. Fibres exhibiting a negative PR status had a higher aspect ratio and a wider range in comparison to the positive fibres. The positive fibres deviated more from the normal of the average major axis length in comparison to the negative, yet the negative samples had a wider range.

For the CERB2 status of positive and negative, 16 samples were positive and 98 were negative.

After tracing the images with Trace Ridges only the total gap area between the fibres showed a statistical significance when both populations were compared (Fig. 5.43d). Images with a negative CERB2 status exhibited a larger total gap area in comparison to the positive images. Negative images experienced a wider range.

For the lymph node involvement, there were 74 images where there was 0 lymph node involvement and 40 images where there was lymph node involvement.

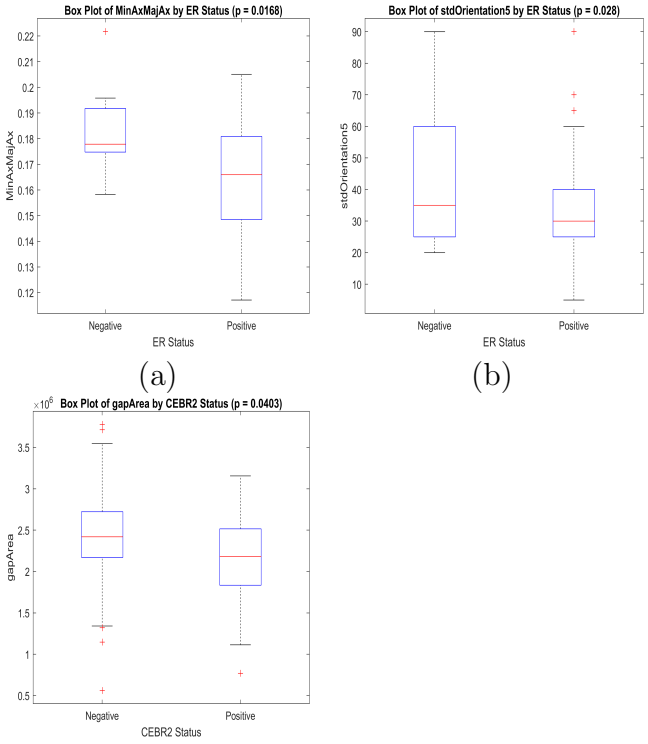


Figure 5.43: Boxplots of statistically different morphological metrics when comparing normal and tumour populations observed with ER and CERB2. With ER the following metrics were statistically significant: (a) Aspect Ratio. (b) Standard deviation Orientation. The following metric was observed with CERB2: (c) Total gap area between fibres.

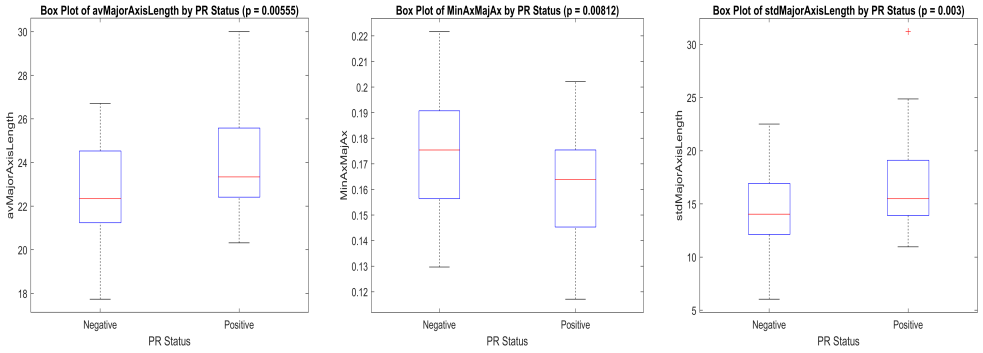
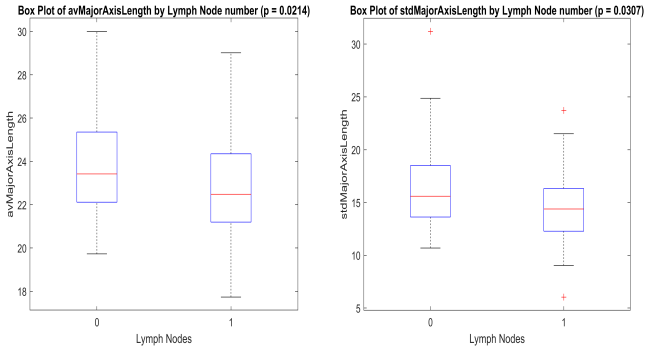


Figure 5.44: Boxplots of statistically different morphological metrics when comparing Normal and tumour populations observed with PR. (a) Average major axis length. (b) Aspect ratio. (c) Std major axis length.



(a) (b)

Figure 5.45: Boxplots of statistically different morphological metrics when comparing Normal and tumour populations observed with lymph node involvement. (a) Average major axis length. (b) Std major axis length.

2 metrics exhibited statistical significance when comparing the populations with each other (Fig. 5.45). Fibres with no lymph node involvement exhibited a higher average major axis length in comparison to the other fibres yet they had a smaller range of values. Fibres with no lymph node involvement deviated more from the normal values of the major axis length when compared to the other population. Yet both populations exhibited similar ranges.

Samples of the data were split into two categories based on the age, the mean of the age of patients was 53.4. The samples were split to either belong above the average age of 53 or below the age of 53. For the patients below the average age, 29 had no lymph node involvement while 22 had lymph node involvement. For the patients above the age 45 had no lymph node involvement while 18 samples had lymph node involvement.

Applying Bonferroni correction would adjust the $p-value$ to 0.0055 for the nine hypotheses tested. The only metrics that would have stayed statistically significant are the average major axis length and std of the major axis length of the PR populations. Although many of the populations are imbalanced the results are promising and promote further investigation.

5.6.6 Conclusions:

This investigation highlights how Trace Ridges and computational tools can be used to investigate the role of the extracellular matrix in cancer progression and provide reproducible results. Trace Ridges was used here to quantify and extract meaningful morphological properties of the extracellular matrix. Next, the morphological features of the normal and tumour populations were compared to find differences. Machine learning techniques were then trained and tested to classify populations of data. It was found that machine learning techniques classify better with known morphological features properties over unseen images of the extracellular matrix. The images provided were accompanied with clinical information in regards to their ER, PR, CERB2 status as well as lymph node involvement. Findings on the differences between the status populations was not fruitful due to the limited unbalanced

dataset. The work conducted here overall solidifies the importance of computational tools for the diagnosis of cancer.

5.7 Summary:

Firstly, in this chapter Trace Ridges our novel approach is introduced which uses methods of Watershed and Edge detection as well as morphological selection to delineate fibre-like structures. Trace Ridges favours longer and brighter fibres. This chapter investigated tracing algorithms, the objective comparison compared 7 algorithms with 3 different filtering options across 4 different images of different backgrounds. For each of the four images a hand delineated ground truth was made to compare the algorithms traces to the ground truth. Distance errors were used to quantify the performance as well as computational time. For the total distance error Trace Ridges performs best by achieving the least errors on average. Edge Detection performed best for the average distance error but that is due to its affinity to over segment therefore it is never far from a ground truth fibre. Trace Ridges achieved the second lowest average. The lowest maximum error was achieved by U-Net while Trace Ridges did not perform that well this is due to its affinity to favour brighter and longer fibres. The fastest average computational time was achieved by Edge Detection followed closely by Trace Ridges which was four times faster than the next algorithm. Filtering improved all the algorithms tracing performance except for U-Net. Overall Trace Ridges emerges as the best performing algorithm. Next Trace Ridges is used to delineate and extract fibres of fluorescently labelled fibronectin fibres. The metrics extracted by Trace Ridges are then used to compare the small populations of mutant and wild type fibres. Two metrics showed statistical significance when compared. There were more fibres in the wild type population than the mutant population. Subsequently, the other metric that showed statistical significance which was the relative gap area size where there was a larger overall gap area in the mutant samples in comparison to the wild type. Trace Ridges was then used to investigate whether certain modifications of the extracellular matrix collagen led to cytoskeletal remodelling. It was first used to extract and quantify fibres of collagen of different concentrations (1.7 and 6.6 mg/ml). It was observed that increasing the collagen resulted in structural changes of the fibre morphology including number of fibres, width, length, aspect ratio and alignment. Then fibres of control collagen were compared to collagen with the addition of protease inhibitors, it was observed that there was an increase in the length of collagen, as well as width, alignment, and the aspect ratio of fibres since both the length and width have increased. Trace Ridges was then used to investigate whether radial alignment leads to the remodelling of the ECM, this is instigated through the addition of $MgCl_2$. Trace Ridges delineated and quantified fibres of the control and the Marangoni Flows equivalent of the control to find differences between the populations. It was observed that there were differences in length, thickness, straightness, and aspect ratio. The last investigation looked at utilising Trace Ridges to understand further the impact of cancer on collagen fibres. Firstly, Trace Ridges extracted and quantified fibres of the normal and tumour breast tissue and statistical analysis was conducted, it was observed that there were 32 metrics that had statistical significance such as straightness and length. In an effort to understand the influence of tumours. The TMA cores also consist of cell nuclei which were segmented using CellPose. Features developed and outlined were using to quantify the properties of segmented cells. Finally, the two populations were compared and a further 18

metrics deemed statistically significant when comparing the cell nuclei of the normal and tumour breast tissue including average gap area and total gap area. A UMAP was used to visualise the clustering of the populations considering their extracted features. Furthermore, we aimed to use the features extracted by Trace Ridges as input into the classification techniques to observe whether the role of Trace Ridges in classification is important and helps in correct identification or not. To investigate these multiple classification methods were trained and tested on features extracted by Trace Ridges and they were compared to state-of-the-art deep learning image classification techniques that require no features to perform the task.

Patient clinical information provided more to the story, where it was observed that when comparing a positive and a negative ER there was a statistical significance when comparing the average eccentricity, aspect ratio and standard deviation of orientation of the fibres. When comparing positive and negative PR status, it was observed that the average circularity, average major axis length and the standard deviation of the minimum intensity had statistical significance. When comparing the CERB2 populations only the total gap area had a statistical significance. In terms of the lymph node involvement 5 metrics observed a statistical significance when the populations were compared. These metrics were the average circularity, average major axis length, standard deviation of circularity, standard deviation of the major axis length and the standard deviation of the minimum intensity.

Chapter 6

Conclusion

6.1 Summary

A fast automatic tracing algorithm was presented for the delineation of fibre-like structures.

The novel algorithm combines methods of Watershed with Edge Detection. Watershed view an image as a topographic surface where pixels of high intensity are elevated, this allows it to identify the ridges. Edge Detection is used to break the ridges that run from a main ridge towards the sides of the basins, separating them from the minor ones. The algorithm further refines the images using morphological operations by removing short branches and dim fibres. Trace Ridges prioritises brighter and longer fibres for a cleaner delineation and accepts any dimensionality image.

Trace Ridges was compared and tested with other tracing methods across 4 biological fibre-like images of different backgrounds with three filtering options; no filtering, Gaussian filtering, and DnCnn filtering. Trace Ridges outperformed state-of-the-art tracing methods in terms of accuracy and speed where accuracy is measured using distance metrics such as total distance error, average distance error and maximum distance error that compute the difference in distance between an algorithm's trace to the ground truth delineation. Trace Ridges achieved the lowest total distance errors as well as the second lowest average distance errors after Edge Detection. Computational time is how long it took an algorithm to process an image. Trace Ridges was three times faster than the next best method.

The algorithm was further used as a tool to quantify fibres of the extracellular matrix to extract meaningful insights. Trace Ridges was used to compare mutant and wild-type fibronectin populations. Trace Ridges was used to investigate whether the accumulation of collagen leads to cytoskeletal remodelling by comparing populations of increasing collagen concentrations, control and protease inhibited collagen populations and control and collagen observing Marangoni Flow populations. Finally, Trace Ridges was used to quantify and compare collagen fibres from Tissue Micro-Array cores of normal and tumour breast tissue.

6.2 Major Contributions:

- A scoping review of Computational Image Analysis Techniques, Programming Languages and Software Platforms Used in Cancer Research was presented. The scoping review uses the publicly available PubMed database. The scoping review concatenated the keywords cancer, pathology, images, or imaging in combination with additional specific keywords in traditional image processing techniques, machine learning, deep learning techniques and computational languages. The scoping review was focused on trends between the years of 1990 and 2022. The scoping review aimed to answer a number of queries including are artificial intelligence techniques now replacing these earlier image processing and computer vision methods in cancer image analysis? As well as which traditional and deep learning computational techniques have been most used in the analysis of cancer-related image analysis?

- After experimenting with images of cell nuclei, actin filaments and extracellular matrix fibronectin fibres with traditional image processing techniques. The extracellular matrix had the most deterministic factors when comparing population of data. Initially, the fibronectin fibres were traced with Edge Detection which has an affinity to oversegment therefore a robust tracing algorithm was needed. The major contribution of this thesis is our novel approach Trace Ridges. The algorithm combines methods of Watershed and Edge Detection with morphological operation to trace fibre-like structure.

- Leading on from the previous contribution, to test whether Trace Ridges performs better than the current state-of-the-art methods, an objective comparison of 7 algorithms across four images and three filtering options was carried out. Trace Ridges had the lowest distance errors and lowest average distance errors when compared to the ground truth as well as the second fastest computational time.

- Trace Ridges was used in a number of investigations with biological significance. Firstly, it was used to trace and quantify fibres of fibronectin for mutant and wild-type cells. Morphological features were extracted and the populations compared. There were differences in the number of fibres and average gap size between the RBD and WT populations. Secondly, Trace Ridges was used to investigate if collagen accumulation would lead to cytoskeletal remodelling. Populations of increasing concentrations, control and protease inhibited collagen fibres as well as control and collagen fibres exhibiting Marangoni flows were traced with Trace Ridges, their morphological properties extracted, and the populations compared. The findings show that increasing collagen the network becomes denser around the cells, which confines them tightly, which as a result activates their actomyosin machinery. The specific arrangement of longer collagen fibres in the radial alignment promotes cell migration. The findings contributed to the creation of the road map aggressive tumours take before they metastasise. Finally, Trace Ridges was used to trace collagen fibres of the Tissue Micro-Arrays of normal

and tumour breast tissue. The population were compared, and a number of findings were presented.

6.3 Major Conclusions:

- From the scoping review machine learning and deep learning techniques are replacing traditional image processing techniques as evident by the number of entries in PubMed. Software platforms such as ImageJ and Fiji are widely used, and QuPath is increasing in popularity.

- The combination of simpler algorithms such as Edge Detection and Watershed outperforms more complex time-consuming methods such as CT Fire, Twombli, Graph Based and U-Net. Improving on their accuracies as well as processing speed.

- Following from the last point, as evident in the objective comparison, not all images are the same. Some images are processed better with different types of filters that attenuate specific features for tracing or smoothing noise. This could also be interpreted as some algorithms perform better under different preprocessing filters. In addition, some images require pre-processing like the case of the Tissue Micro-Arrays cores to be prepared for tracing. In some cases, each image or set of images must be dealt with in a way that only suits them.

- The extracellular matrix is an important biological component that plays a major role in the progression of diseases and in specific cancer. Through the number of investigations Trace Ridges was used to trace and quantify fibres of the extracellular matrix such as the fibronectin and collagen. After tracing with Trace Ridges morphological properties of the quantified fibres were extracted and populations were compared. There were differences observed when comparing mutant and wild-type fibronectin in terms of the number of fibres and total gap size. When investigating whether the accumulation of collagen leads to cytoskeletal remodelling, differences in population of increasing collagen, protease inhibition and Marangoni Flow exhibited differences which led to the creation of the road map that aggressive tumours take before they metastasise. When comparing populations of normal and tumour collagen of the breast tissue, many metrics proved different. These findings can help with early patient diagnosis and tailored treatment plans.

6.4 Discussion & Reflections:

In this section, results of work done throughout the thesis is discussed as well as reflections upon completing the work.

- Image analysis plays a great role in our understanding of biological processes in relation to cells and their organelles, the micro-environment that surrounds these cells and diseases such as cancer. Amongst many uses, image analysis was used to automatically segment and analyse histological tumour sections, with the cytoskeleton, image analysis has been used on the cytoskeleton to segment and analyse how actin stress fibres behave and finally, with the ECM, image analysis was utilised to further our understanding on how the ECM remodels in cancer progression. All these mentioned tasks were generally carried out by pathologists who would spend a great time identifying and labelling the individual cells and fibres to analyse them. With the introduction of image analysis and the automatic algorithms that now exist, these tasks may be completed at a much faster rate thus accelerating our understanding of these fascinating structures that influence bodily functions and diseases. Furthermore, many algorithms are coupled with morphological feature extraction, further streamlining analysis approaches. In addition to this, observers when analysing biological structures may observe different things as our eyes may perceive us but with image analysis algorithms, reproducible results are produced as there is no bias. With image analysis, researchers have the potential to discover new findings that may not be easily seen or detected.

- The exploration of data chapter was intended to be a chapter where multiple image analysis techniques are used to segment cells, actin filaments and fibronectin fibres. The development of pipelines proved to be a useful introduction to image processing techniques and how to use them. The main limitations of the exploratory work was the lack of data, as it was hypothesised that differences may be found between the RBD and WT population but there were only five images for each population, which is not large enough to extract concrete findings from. Additionally, segmenting the actin filaments proved to be difficult as many of the structures of actin filaments extended beyond the image borders therefore the true structure was not segmented. On the other hand, this chapter presented an opportunity through the promising results when comparing the fibronectin populations which set the path for the rest of the work conducted in this thesis.

- Ground Truth data is of high importance for the validation and verification of segmentation techniques. Ground Truth data was used to verify the segmentation of the cell nuclei in the exploration of the data experiment as well as to calculate the distance errors of the tracing algorithms applied on fibre-like structures to compare the algorithms. The hand delineation was conducted by one observer due to time constraints but for best practices it is best done by multiple observers as two people may not agree on an object. Additionally, due to time constraints and having only one observer,

the objective comparison was conducted with only four images, if more images were available then it would have further verified the findings of the study.

- Many algorithms exist for extracellular matrix fibres analysis, yet in most cases seen throughout this thesis is that the images need to be cleaned and processed through a tailored pipeline to prepare the images for tracing through the algorithms, whether to deal with artefacts or separate fibres from cells as seen in the Tissue Micro-arrays experiment, this can be a time consuming task. Additionally, images are not always of good quality some images may contain artefacts such as incorrect staining or acquisition artefacts, which are difficult to deal with therefore limiting datasets for experimentation.

- From the objective comparison, it was seen that Trace Ridges performs best overall in the distance error metrics and computational time yet, it does not always perform the best individually, for certain images U-Net and Graph Based methods performed better, particularly on some of the more crowded images. Another point to add is that although U-Net was relatively fast, deep learning models tend to take some time to process images due to the computational complexities and the depth of their layers. Algorithms such as CT Fire, Twombli, and Graph Based are difficult to run as they require many parameters to be chosen by the user which can lead to additional time being spent to tune the process for an image or a dataset. Additionally, some of the algorithms took a long period of time to process one image due to the complexity and length of their code. One of the main reasons of using image analysis is to process and analyse images faster than if done by a human, therefore one of the aims of any algorithm is to process an image quickly. Another obstacle was the lack of data for U-Net, due to the limited number of images no validation data was used in the training process of U-Net which is generally recommended as this could lead to data leakage as well as overfitting or selection bias for the U-Net. Finally, the qualitative assessment is to be used as an additional layer of analysis but not as the main point of comparison as it is difficult to tell some of the differences between the better performing algorithms visually.

- As previously mentioned, the manual assessment of the ECM is subjective and at times inconsistent that is where image analysis comes and is able to process images quicker as well as extract vital descriptors that aid in biomarker discovery. In literature many examples have used methods such as CT Fire on images of non-small cell lung cancer and correlated collagen fibre descriptors with poor prognosis, similarly researchers have used Twombli to quantify collagen fibres of triple-negative breast cancer tissue and correlated fibre densities with aggressive cancer phenotypes and metastasis risk. Trace Ridges can complement these methods to analyse ECM fibres and aid in our understanding of the microenvironment. Trace Ridges offers a simpler, faster, and easier to use algorithm as well as higher accuracies for certain image types which can further accelerate our understanding of the ECM. Computational tools such as Trace Ridges

can be used to aid in the diagnosis and prognosis of conditions such as cancer based on the ECM remodelling.

- In the modified cell line chapter, Trace Ridges was used to trace fibronectin fibres and extract morphological features, the main issue with this experiment is the dataset size as the comparison was run on five images of the RBD and five images of the WT which might not paint the whole picture or be a solid basis for statistical significance. The results showed an inverse relationship between the number of fibres and the total gap area which has been mentioned before in literature. With a larger dataset, the results could be solidified.

- Trace Ridges was used in the matrix mechano sensing chapter on a large dataset to investigate whether collagen accumulation leads to cytoskeletal remodelling. The findings suggest that longer collagen fibres in a radial arrangement promote higher levels of migration. This work exhibits the effective use of Trace Ridges as an image analysis tool in cancer studies. In the context of breast cancer, it is known that an ECM structure with aligned matrix fibres is associated with poor prognosis, through the tracing of Trace Ridges the alignment of fibres can be determined therefore Trace Ridges can also act as a prognostic tool.

- In the final experiment with the TMA cores, Trace Ridges was used to trace and quantify the fibres as well as compare populations of normal and tumour, which shows how Trace Ridges could be used as a diagnostic tool based on extracted morphological properties. CellPose was used to segment cells of the tissue, one of the limitations of the segmentation is that the network must be tuned in order to improve the segmentation task which can be time consuming. For the classification tasks, the deep learning networks underperformed from expectations due to the lack of data. The clinical features supplied provided another depth to the analysis but had no solid outcomes due to the data size. This experiment highlighted how image analysis can be used to produce reproducible results that can aid in expanding our knowledge in cancer studies, but it also showed how data quality and quantity are essential.

6.5 Limitations:

- If a fibre is dark or has uneven intensity, the watershed algorithm may treat it as a basin and Trace Ridges may not detect it, this was one of the limitations of Trace Ridges. Given more time a work around this problem is to use directional filters to enhance the missed fibres based on their angles.

- The ground truth for both the fibres and the cells were carried out by a single observer due to constraints. Given more time a better approach would be for multiple observers to carry out the hand delineation and the final ground truth would include all of the areas the observers agreed on for both the fibres and the cells. This ensures that there is no bias and further confirms what the observers are seeing.

- The distance error metrics outlined in this thesis had several issues, specifically when an algorithm oversegments inaccurately it appears to perform well in the average and maximum distance errors. Distance error metrics may be used in combination with other metrics to further investigate the performance of the algorithms.

6.6 Future Work:

This section presents a number of suggestions of research objectives that could compliment the work done in this thesis or provide improvements.

- Trace Ridges has been mainly used for tracing of fibrial structures of the extracellular matrix. Trace Ridges could be employed to trace images of retinal scans to investigate eye conditions. Trace Ridges could also be used to investigate vasculature as the blood vessels network is important for the delivery of oxygen, nutrients, and the removal of waste. The algorithm can be more widely used and for purposes other than fibres of the ECM.

- Trace Ridges has been fitted with a Gaussian filter that has performed well in many tasks but as seen in the objective comparison, other filters are better suited for some images. Therefore, it would be better to build in the algorithm an automated filter choosing method based on the type of image. The filter would assess an image and decide on which filter would be best suited. This is to minimise the distance errors and maximise the accuracy of the algorithm.

- Images of dataset 1 were limited in their numbers with just 5 for the mutant and 5 for the wild type. Experimenting with more images of the RBD mutant and WT populations could have led to a better story to tell on the differences and what makes the mutant population special. Similarly, the clinical information provided for images of dataset 4 of tissue micro arrays showed that there were class imbalances when investigating ER and CERB2. A more balanced split between the sub populations would have provided a clearer story on the effect of each receptor status.

- Trace Ridges has mainly been used to find differences between populations of data throughout this thesis. In the future, Trace Ridges can be used as a diagnostic tool that further complements and corroborates biological findings such as that done in the Nature Communications article.

- Another way Trace Ridges could be used is to cross reference extracted morphological features of collagen fibres with spatial transcriptomic data of the cells surrounding fibres which provides genetic expression information with the location of cells. This could answer questions such as do certain collagen alignments correlate with specific cell types or gene signatures and is fibre density higher near invasive, high-expression regions.

- To build a better story Trace Ridges could be used on a dataset to extract morphological features of fibre-like structures, this can be combined with deep learning approaches such as CNNs to extract abstract morphological features.

- In the future, I would also like to apply Graph Neural Networks on the fibre-like images to learn complex spatial relationships as it could be useful to learn and detect subtle

disease associated changes in fibre organisation that traditional approaches might miss. Graph Neural Networks would be able to quantify network connectivity and structural patterns; this can aid in furthering our understanding on the role of the ECM in cancer progression.

- Finally, Trace Ridges analyses two dimensional images so a plan for the future is to modify Trace Ridges to process three dimensional images. Analysing three dimensional images can capture the true fibre branching as well as the fibre orientations in all directions.

Bibliography

[1] H. Gest, "The discovery of microorganisms by Robert Hooke and Antoni Van Leeuwenhoek, fellows of the Royal Society," *Notes and Records of the Royal Society of London*, vol. 58, no. 2, pp. 187–201, May 2004.

[2] A. Bergua, "["The Retina of the Vertebrates" by Santiago Ramn y Cajal–100 years in German translation]," *Klinische Monatsblatter Fur Augenheilkunde*, vol. 205, no. 6, pp. 372–373, Dec. 1994.

[3] M. L. Jones and H. Spiers, "The crowd storms the ivory tower," *Nature Methods*, vol. 15, no. 8, pp. 579–580, Aug. 2018.

[4] B. W. Sperry and W. A. Jaber, "Towards reducing inter- and intra-observer variability: Reasons for optimism?" *Journal of Nuclear Cardiology: Official Publication of the American Society of Nuclear Cardiology*, vol. 29, no. 2, pp. 447–448, Apr. 2022.

[5] S. Sarkar, G. Horn, K. Moulton, A. Oza, S. Byler, S. Kokolus, and M. Longacre, "Cancer Development, Progression, and Therapy: An Epigenetic Overview," *International Journal of Molecular Sciences*, vol. 14, no. 10, pp. 21 087–21 113, Oct. 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3821660/

[6] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, Jan. 2000.

[7] I. F. Tannock and R. P. Hill, Eds., *The basic science of oncology*, 3rd ed., ser. McGraw-Hill international editions. New York: McGraw-Hill, Health Professions Div, 1998.

[8] S. H. Hassanpour and M. Dehghani, "Review of cancer from perspective of molecular," *Journal of Cancer Research and Practice*, vol. 4, no. 4, pp. 127–129, Dec. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2311300617300125

[9] D. Hanahan and R. A. Weinberg, "Hallmarks of Cancer: The Next Generation," *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011, publisher: Elsevier. [Online]. Available: https://www.cell.com/cell/abstract/S0092-8674(11)00127-9

[10] D. Hanahan, "Hallmarks of Cancer: New Dimensions," *Cancer Discovery*, vol. 12, no. 1, pp. 31–46, Jan. 2022. [Online]. Available: https://doi.org/10.1158/2159-8290.CD-21-1059

[11] R. Perona, "Cell signalling: growth factors and tyrosine kinase receptors," *Clinical & Translational Oncology: Official Publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico*, vol. 8, no. 2, pp. 77–82, Feb. 2006.

[12] C. J. Sherr and F. McCormick, "The RB and p53 pathways in cancer," *Cancer Cell*, vol. 2, no. 2, pp. 103–112, Aug. 2002.

[13] J. M. Adams and S. Cory, "The Bcl-2 apoptotic switch in cancer development and therapy," *Oncogene*, vol. 26, no. 9, pp. 1324–1337, Feb. 2007.

[14] M. A. Blasco, "Telomeres and human disease: ageing, cancer and beyond," *Nature Reviews. Genetics*, vol. 6, no. 8, pp. 611–622, Aug. 2005.

[15] F. Arnold, "Tumour angiogenesis," *Annals of the Royal College of Surgeons of England*, vol. 67, no. 5, pp. 295–298, Sep. 1985.

[16] I. J. Fidler, "The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited," *Nature Reviews. Cancer*, vol. 3, no. 6, pp. 453–458, Jun. 2003.

[17] R. Kim, M. Emi, and K. Tanabe, "Cancer immunoediting from immune surveillance to immune escape," *Immunology*, vol. 121, no. 1, pp. 1–14, May 2007.

[18] O. Warburg, "On the origin of cancer cells," *Science (New York, N.Y.)*, vol. 123, no. 3191, pp. 309–314, Feb. 1956.

[19] J. M. Samet, "Tobacco smoking: the leading cause of preventable disease worldwide," *Thoracic Surgery Clinics*, vol. 23, no. 2, pp. 103–112, May 2013.

[20] M. Watson, D. M. Holman, and M. Maguire-Eisen, "Ultraviolet radiation exposure and its impact on skin cancer risk," *Seminars in Oncology Nursing*, vol. 32, no. 3, p. 241254, Aug. 2016. [Online]. Available: http://dx.doi.org/10.1016/j.soncn.2016.05.005

[21] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, May 2024. [Online]. Available: https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21834

[22] E. Alizadeh, J. Castle, A. Quirk, C. D. Taylor, W. Xu, and A. Prasad, "Cellular morphological features are predictive markers of cancer cell state," *Computers in Biology and Medicine*, vol. 126, p. 104044, Nov. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0010482520303759

[23] Z. Mousavikhamene, D. J. Sykora, M. Mrksich, and N. Bagheri, "Morphological features of single cells enable accurate automated classification of cancer from non-cancer cell lines," *Scientific Reports*, vol. 11, no. 1, p. 24375, Dec. 2021. [Online]. Available: https://www.nature.com/articles/s41598-021-03813-8

[24] P.-H. Wu, J. M. Phillip, S. B. Khatau, W.-C. Chen, J. Stirman, S. Rosseel, K. Tschudi, J. Van Patten, M. Wong, S. Gupta, A. S. Baras, J. T. Leek, A. Maitra, and D. Wirtz, "Evolution of cellular morpho-phenotypes in cancer metastasis," *Scientific Reports*, vol. 5, no. 1, p. 18437, Dec. 2015, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/srep18437

[25] M. Hu, M. Li, H. Huang, and C. Lu, "Isolated cancer stem cells from human liver cancer: morphological and functional characteristics in primary culture," *Clinical & Translational Oncology:*, vol. 24, no. 1, pp. 48–56, Jan. 2022.

[26] R. V. Iozzo and M. A. Gubbiotti, "Extracellular Matrix: The driving force of mammalian diseases," *Matrix biology : journal of the International Society for Matrix Biology*, vol. 71-72, pp. 1–9, Oct. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6146050/

[27] N. V. Popova and M. Jücker, "The Functional Role of Extracellular Matrix Proteins in Cancer," *Cancers*, vol. 14, no. 1, p. 238, Jan. 2022.

[28] J. Winkler, A. Abisoye-Ogunniyan, K. J. Metcalf, and Z. Werb, "Concepts of extracellular matrix remodelling in tumour progression and metastasis," *Nature Communications*, vol. 11, no. 1, p. 5120, Oct. 2020, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-020-18794-x

[29] J. Huang, L. Zhang, D. Wan, L. Zhou, S. Zheng, S. Lin, and Y. Qiao, "Extracellular matrix and its therapeutic potential for cancer treatment," *Signal Transduction and Targeted Therapy*, vol. 6, no. 1, pp. 1–24, Apr. 2021, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41392-021-00544-0

[30] C. Walker, E. Mojares, and A. Del Ro Hernndez, "Role of Extracellular Matrix in Development and Cancer Progression," *International Journal of Molecular Sciences*, vol. 19, no. 10, p. 3028, Oct. 2018.

[31] C. Frantz, K. M. Stewart, and V. M. Weaver, "The extracellular matrix at a glance," *Journal of Cell Science*, vol. 123, no. 24, pp. 4195–4200, Dec. 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995612/

[32] C. Bonnans, J. Chou, and Z. Werb, "Remodelling the extracellular matrix in development and disease," *Nature reviews. Molecular cell biology*, vol. 15, no. 12, pp. 786–801, Dec. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4316204/

[33] E. C. Filipe, J. L. Chitty, and T. R. Cox, "Charting the unexplored extracellular matrix in cancer," *International Journal of Experimental Pathology*, vol. 99, no. 2, pp. 58–76, Apr. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6031881/

[34] B. Yue, "Biology of the Extracellular Matrix: An Overview," *Journal of glaucoma*, pp. S20–S23, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4185430/

[35] S. Pompili, G. Latella, E. Gaudio, R. Sferra, and A. Vetuschi, "The Charming World of the Extracellular Matrix: A Dynamic and Protective Network of the Intestinal Wall," *Frontiers in Medicine*, vol. 8, p. 610189, Apr. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8085262/

[36] W. Qin, L. Cao, and I. Y. Massey, "Role of PI3K/Akt signaling pathway in cardiac fibrosis," *Molecular and Cellular Biochemistry*, vol. 476, no. 11, pp. 4045–4059, Nov. 2021.

[37] Y. Arafat, C. CUESTA-APAUSA, E. Castellano, and C. C. Reyes-Aldasoro, "Investigation of the Organisation of the Extracellular Matrix Using Fibre Tracing," *Computer Graphics and Visual Computing (CGVC)*, 2024, artwork Size: 8 pages Edition: 1238 ISBN: 9783038682493 Publisher: The Eurographics Association. [Online]. Available: https://diglib.eg.org/handle/10.2312/cgvc20241238

[38] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, "Cellpose: a generalist algorithm for cellular segmentation," *Nature Methods*, vol. 18, no. 1, pp. 100–106, Jan. 2021, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41592-020-01018-x

[39] I. Novianti and S. Mizukami, "Cellsketch: Simplified Cell Representation for Label-free Cell and Nuclei Segmentation," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2023, pp. 1–4, Jul. 2023.

[40] N. F. Greenwald, G. Miller, E. Moen, A. Kong, A. Kagel, T. Dougherty, C. C. Fullaway, B. J. McIntosh, K. X. Leow, M. S. Schwartz, C. Pavelchek, S. Cui, I. Camplisson, O. Bar-Tal, J. Singh, M. Fong, G. Chaudhry, Z. Abraham, J. Moseley, S. Warshawsky, E. Soon, S. Greenbaum, T. Risom, T. Hollmann, S. C. Bendall, L. Keren, W. Graf, M. Angelo, and D. Van Valen, "Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning," *Nature Biotechnology*, vol. 40, no. 4, pp. 555–565, Apr. 2022.

[41] H. Raymond-Hayling, Y. Lu, K. E. Kadler, and T. Shearer, "A fibre tracking algorithm for volumetric microstructural data - application to tendons," *Acta Biomaterialia*, vol. 154, pp. 335–348, Dec. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1742706122006997

[42] N. S. Rossen, A. Kyrsting, A. J. Giaccia, J. T. Erler, and L. B. Oddershede, "Fiber finding algorithm using stepwise tracing to identify biopolymer fibers in noisy 3D images," *Biophysical Journal*, vol. 120, no. 18, pp. 3860–3868, Sep. 2021.

[43] X. Ma and H. Yu, "Global Burden of Cancer," *The Yale Journal of Biology and Medicine*, vol. 79, no. 3-4, pp. 85–94, Dec. 2006. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1994799/

[44] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[45] T. Lindeberg, "Feature detection with automatic scale selection," *Int J Comp Vision*, vol. 30, no. 2, pp. 79–116, Nov. 1998. [Online]. Available: ://000077990700001

[46] ——, "Edge detection and ridge detection with automatic scale selection," *Int J Comp Vision*, vol. 30, no. 2, pp. 117–154, Nov. 1998. [Online]. Available: ://000077990700002

[47] E. Wershof, D. Park, D. J. Barry, R. P. Jenkins, A. Rullan, A. Wilkins, K. Schlegelmilch, I. Roxanis, K. I. Anderson, P. A. Bates, and E. Sahai, "A FIJI macro for quantifying pattern in extracellular matrix," *Life Science Alliance*, vol. 4, no. 3, Mar. 2021, publisher: Life Science Alliance Section: Methods. [Online]. Available: https://www.life-science-alliance.org/content/4/3/e202000880

[48] J. S. Bredfeldt, Y. Liu, C. A. Pehlke, M. W. Conklin, J. M. Szulczewski, D. R. Inman, P. J. Keely, R. D. Nowak, T. R. Mackie, and K. W. Eliceiri, "Computational segmentation of collagen fibers from second-harmonic generation images of breast cancer," *Journal of Biomedical Optics*, vol. 19, no. 1, p. 016007, Jan. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3886580/

[49] A.-I. Grapa, "Characterization of fibronectin networks using graph-based representations of the fibers from 2D confocal images," PhD thesis, Universite Cote dAzur, Nice, France, Dec 2020, available at https://theses.hal.science/tel-03052167.

[50] A.-I. Grapa, G. Efthymiou, E. V. Obberghen-Schilling, L. Blanc-Féraud, and X. Descombes, "A spatial statistical framework for the parametric study of fiber networks: Application to fibronectin deposition by normal and activated fibroblasts," *Biological Imaging*, vol. 3, p. e25, Jan. 2023.

[51] L. A. Haddad, "Chapter 2 - Cellular structure and molecular cell biology," in *Clinical Molecular Medicine*, D. Kumar, Ed. Academic Press, Jan. 2020, pp. 17–45. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128093566000022

[52] Mediran, "A typical animal cell with labeled organelles." Aug. 2012. [Online]. Available: https://commons.wikimedia.org/wiki/File:Eukaryotic-Cell-(animal).jpg

[53] G. M. Cooper, *The cell: a molecular approach*, 2nd ed. Washington, D.C. Sunderland, Mass: ASM Press, 2000.

[54] I. A. Sawyer and M. Dundr, "10 - Nuclear Bodies," in *Nuclear Architecture and Dynamics*, ser. Translational Epigenetics, C. Lavelle and J.-M. Victor, Eds. Boston: Academic Press, Jan. 2018, vol. 2, pp. 235–256. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128034804000107

[55] M. B. Qumsiyeh, "Structure and function of the nucleus: anatomy and physiology of chromatin," *Cellular and molecular life sciences: CMLS*, vol. 55, no. 8-9, pp. 1129–1140, Jul. 1999.

[56] X. Rambout and L. E. Maquat, "Nuclear mRNA decay: regulatory networks that control gene expression," *Nature Reviews. Genetics*, vol. 25, no. 10, pp. 679–697, Oct. 2024.

[57] Nephron, "Very high magnification micrograph of hairy cell leukemia. H&E stain." [Online]. Available: https://commons.wikimedia.org/wiki/File:Hairy-cell-leukemia-very-high-mag.jpg

[58] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 4th ed. Garland Science, 2002.

[59] T. Hohmann and F. Dehghani, "The Cytoskeleton A Complex Interacting Meshwork," *Cells*, vol. 8, no. 4, p. 362, Apr. 2019. [Online]. Available: https://www.mdpi.com/2073-4409/8/4/362

[60] F. Valderrama, J. M. Durán, T. Babià, H. Barth, J. Renau Piqueras, and G. Egea, "Actin Microfilaments Facilitate the Retrograde Transport from the Golgi Complex to the Endoplasmic Reticulum in Mammalian Cells," *Traffic*, vol. 2, no. 10, pp. 717–726, Oct. 2001. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1034/j.1600-0854.2001.21006.x

[61] P. Collins, "2 - Cellular Mechanisms and Embryonic Tissues," in *Fetal Medicine (Third Edition)*, P. P. Pandya, D. Oepkes, N. J. Sebire, and R. J. Wapner, Eds. London: Elsevier, Jan. 2020, pp. 6–22.e1. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780702069567000026

[62] G. M. Cooper, "Structure of the Plasma Membrane," *The Cell: A Molecular Approach. 2nd edition*, 2000, publisher: Sinauer Associates. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK9898/

[63] W. Crochot, "Schematic three dimensional cross section of a cell membrane." Oct. 2014. [Online]. Available: https://commons.wikimedia.org/wiki/File:Cell-membrane-drawing-en.svg

[64] P. L. Yeagle, "Chapter 15 - Membrane Receptors," in *The Membranes of Cells (Third Edition)*, P. L. Yeagle, Ed. Boston: Academic Press, Jan. 2016, pp. 401–425. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128000472000152

[65] E. J. Miller and S. L. Lappin, "Physiology, Cellular Receptor," in *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2022. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK554403/

[66] W. Pyzynski, "An example of membrane receptors." May 2018. [Online]. Available: https://commons.wikimedia.org/wiki/File:Receptor-(Biochemistry).svg

[67] J. Iqbal, M. Zaidi, and N. G. Avadhani, "Cell signaling," *Annals of the New York Academy of Sciences*, vol. 1211, pp. 3–8, Nov. 2010.

[68] B. N. Kholodenko, "Cell-signalling dynamics in time and space," *Nature Reviews Molecular Cell Biology*, vol. 7, no. 3, pp. 165–176, Mar. 2006. [Online]. Available: http://www.nature.com/articles/nrm1838

[69] B. B. Aldridge, J. M. Burke, D. A. Lauffenburger, and P. K. Sorger, "Physicochemical modelling of cell signalling pathways," *Nature Cell Biology*, vol. 8, no. 11, pp. 1195–1203, Nov. 2006. [Online]. Available: http://www.nature.com/articles/ncb1497

[70] R. a. E. Wikipedia, "Signal transduction pathways. Picture is a new version of old figure self-made in illustrator." Mar. 2007. [Online]. Available: https://commons.wikimedia.org/wiki/File:Signal-transduction-v1.png

[71] S. Chen, M. Zhao, G. Wu, C. Yao, and J. Zhang, "Recent Advances in Morphological Cell Image Analysis," *Computational and Mathematical Methods in Medicine*, vol. 2012, p. 101536, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3261466/

[72] F. Brun, A. Accardo, M. Marchini, F. Ortolani, G. Turco, and S. Paoletti, "Texture analysis of TEM micrographs of alginate gels for cell microencapsulation," *Microscopy Research and Technique*, vol. 74, no. 1, pp. 58–66, Jan. 2011. [Online]. Available: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jemt.20874

[73] S. Amini, D. Veilleux, and I. Villemure, "Tissue and cellular morphological changes in growth plate explants under compression," *Journal of Biomechanics*, vol. 43, no. 13, pp. 2582–2588, Sep. 2010. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0021929010002800

[74] Y. Xiong, C. Kabacoff, J. Franca-Koh, P. N. Devreotes, D. N. Robinson, and P. A. Iglesias, "Automated characterization of cell shape changes during amoeboid motility by skeletonization," *BMC Systems Biology*, vol. 4, no. 33, Mar. 2010. [Online]. Available: ://000277360700001

[75] A. I. Baba and C. Câtoi, *TUMOR CELL MORPHOLOGY*. The Publishing House of the Romanian Academy, 2007, publication Title: Comparative Oncology. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK9553/

[76] C. C. ReyesAldasoro, L. J. Williams, S. Akerman, C. Kanthou, and G. M. Tozer, "An automatic algorithm for the segmentation and morphological analysis of microvessels in immunostained histological tumour sections," *Journal of Microscopy*, vol. 242, no. 3, pp. 262–278, 2011, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2818.2010.03464.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2818.2010.03464.x

[77] J. Cheng, Y. Chou, C. Huang, Y. Chang, C. Tiu, F. Yeh, K. Chen, C. Tsou, and C. Chen, "ACCOMP: Augmented cell competition algorithm for breast lesion demarcation in sonography," *Medical Physics*, vol. 37, no. 12, pp. 6240–6252, Dec. 2010. [Online]. Available: https://aapm.onlinelibrary.wiley.com/doi/10.1118/1.3512799

[78] M. E. Plissiti, C. Nikou, and A. Charchanti, "Automated Detection of Cell Nuclei in Pap Smear Images Using Morphological Reconstruction and Clustering," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 2, pp. 233–241, Mar. 2011. [Online]. Available: https://ieeexplore.ieee.org/document/5601778/

[79] Y. Xiong and P. A. Iglesias, "Tools for analyzing cell shape changes during chemotaxis," *Integrative Biology*, vol. 2, no. 11-12, pp. 561–567, Nov. 2010.

[80] Y.-S. Lin, C.-C. Lin, Y.-S. Tsai, T.-C. Ku, Y.-H. Huang, and C.-N. Hsu, "A spectral graph theoretic approach to quantification and calibration of collective morphological differences in cell images," *Bioinformatics*, vol. 26, no. 12, pp. i29–i37, Jun. 2010. [Online]. Available: https://academic.oup.com/bioinformatics/article/26/12/i29/283584

[81] P. Venkatesan, S. Das, M. M. R. Krishnan, C. Chakraborty, K. Chaudhury, and M. Mandal, "Effect of AEE788 and/or Celecoxib on colon cancer cell morphology using advanced microscopic techniques," *Micron*, vol. 41, no. 3, pp. 247–256, Apr. 2010. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0968432809001681

[82] S. Ohnuki, S. Nogami, and Y. Ohya, "A microfluidic device to acquire high-magnification microphotographs of yeast cells," *Cell Division*, vol. 4, no. 1, p. 5, Dec. 2009. [Online]. Available: https://celldiv.biomedcentral.com/articles/10.1186/1747-1028-4-5

[83] H. Vindin, L. Bischof, P. Gunning, and J. Stehn, "Validation of an Algorithm to Quantify Changes in Actin Cytoskeletal Organization," *SLAS Discovery*, vol. 19, no. 3, pp. 354–368, Mar. 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2472555222073427

[84] H. Rogge, N. Artelt, N. Endlich, and K. Endlich, "Automated segmentation and quantification of actin stress fibres undergoing experimentally induced changes," *Journal of Microscopy*, vol. 268, no. 2, pp. 129–140, Nov. 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/jmi.12593

[85] D. Krentzel, M. I. Gariboldi, M. Juzans, M. Mastrogiovanni, F. Mueller, C. Cuche, V. Di Bartolo, and A. Alcover, "Image processing approaches for microtubule remodeling quantification at the immunological synapse," in *Methods in Cell Biology*. Elsevier, 2025, vol. 193, pp. 39–67. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0091679X24000670

[86] A. Naba, "Mechanisms of assembly and remodelling of the extracellular matrix," *Nature Reviews Molecular Cell Biology*, vol. 25, no. 11, pp. 865–885, Nov. 2024, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41580-024-00767-3

[87] G. Sersa, B. Stabuc, M. Cemazur, B. Jancar, D. Miklavcic, and Z. Rudolf, "Electrochemotherapy with Cisplatin: Potentiation of Local Cisplatin Antitumor Effectiveness by Application of Electric Pulses in Cancer Patients." *Eur. J. Cancer*, vol. 34, pp. 1213–1218, 1998.

[88] T. R. Cox and J. T. Erler, "Remodeling and homeostasis of the extracellular matrix: implications for fibrotic diseases and cancer," *Disease Models & Mechanisms*, vol. 4, no. 2, pp. 165–178, Mar. 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3046088/

[89] I. A. Janson and A. J. Putnam, "Extracellular matrix elasticity and topography: material-based cues that affect cell function via conserved mechanisms," *Journal of biomedical materials research. Part A*, vol. 103, no. 3, pp. 1246–1258, Mar. 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4258536/

[90] M. V. Plikus, X. Wang, S. Sinha, E. Forte, S. M. Thompson, E. L. Herzog, R. R. Driskell, N. Rosenthal, J. Biernaskie, and V. Horsley, "Fibroblasts: Origins, definitions, and functions in health and disease," *Cell*, vol. 184, no. 15, pp. 3852–3872, Jul. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0092867421007923

[91] M. K. Gordon and R. A. Hahn, "Collagens," *Cell and Tissue Research*, vol. 339, no. 1, pp. 247–257, Jan. 2010. [Online]. Available: http://link.springer.com/10.1007/s00441-009-0844-4

[92] J. Myllyharju, "Collagens, modifying enzymes and their mutations in humans, flies and worms," *Trends in Genetics*, vol. 20, no. 1, pp. 33–43, Jan. 2004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0168952503003196

[93] R. Krymchenko, G. Coar Kutluolu, N. Van Hout, D. Manikowski, C. Doberenz, T. H. Van Kuppevelt, and W. F. Daamen, "Elastogenesis in Focus: Navigating Elastic Fibers Synthesis for Advanced Dermal Biomaterial Formulation," *Advanced Healthcare Materials*, vol. 13, no. 27, p. 2400484, Oct. 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/adhm.202400484

[94] J. K. Mouw, G. Ou, and V. M. Weaver, "Extracellular matrix assembly: a multiscale deconstruction," *Nature Reviews Molecular Cell Biology*, vol. 15, no. 12, pp. 771–785, Dec. 2014. [Online]. Available: https://www.nature.com/articles/nrm3902

[95] N. K. Karamanos, A. D. Theocharis, Z. Piperigkou, D. Manou, A. Passi, S. S. Skandalis, D. H. Vynios, V. OrianRousseau, S. RicardBlum, C. E. Schmelzer, L. Duca, M. Durbeej, N. A. Afratis, L. Troeberg, M. Franchi, V. Masola, and M. Onisto, "A guide to the composition and functions of the extracellular matrix," *The FEBS Journal*, vol. 288, no. 24, pp. 6850–6912, Dec. 2021. [Online]. Available: https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.15776

[96] H. Zhou, W. Li, L. Pan, T. Zhu, T. Zhou, E. Xiao, and Q. Wei, "Human extracellular matrix (ECM)-like collagen and its bioactivity," *Regenerative Biomaterials*, vol. 11, p. rbae008, Feb. 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10965421/

[97] A. J. Zollinger and M. L. Smith, "Fibronectin, the extracellular glue," *Matrix Biology: Journal of the International Society for Matrix Biology*, vol. 60-61, pp. 27–37, Jul. 2017.

[98] A. D. Theocharis, D. Manou, and N. K. Karamanos, "The extracellular matrix as a multitasking player in disease," *The FEBS Journal*, vol. 286, no. 15, pp. 2830–2869, 2019, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/febs.14818. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/febs.14818

[99] Joshyoungmozart, "Extracellular matrix," 2019. [Online]. Available: https://commons.wikimedia.org/wiki/File:Extracellular_Matrix.svg

[100] L. B. Mostao-Guidolin, A. C.-T. Ko, F. Wang, B. Xiang, M. Hewko, G. Tian, A. Major, M. Shiomi, and M. G. Sowa, "Collagen morphology and texture analysis: from statistics to classification," *Scientific Reports*, vol. 3, no. 1, p. 2190, Jul. 2013, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/srep02190

[101] S. Z. Despotović and M. Ćosić, "The Morphological Analysis of the Collagen Fiber Straightness in the Healthy Uninvolved Human Colon Mucosa Away From the Cancer," *Frontiers in Physics*, vol. 10, Jul. 2022, publisher: Frontiers.

[102] N. F. Boyd, L. J. Martin, M. J. Yaffe, and S. Minkin, "Mammographic density and breast cancer risk: current understanding and future prospects," *Breast Cancer Research*, vol. 13, no. 6, p. 223, Dec. 2011. [Online]. Available: http://breast-cancer-research.biomedcentral.com/articles/10.1186/bcr2942

[103] J.-W. Shin, S.-H. Kwon, J.-Y. Choi, J.-I. Na, C.-H. Huh, H.-R. Choi, and K.-C. Park, "Molecular Mechanisms of Dermal Aging and Antiaging Approaches," *International Journal of Molecular Sciences*, vol. 20, no. 9, p. 2126, Jan. 2019, number:

9 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1422-0067/20/9/2126

[104] M. M. Atakan, c. N. Koşar, Y. Güzel, H. T. Tin, and X. Yan, "The Role of Exercise, Diet, and Cytokines in Preventing Obesity and Improving Adipose Tissue," *Nutrients*, vol. 13, no. 5, p. 1459, May 2021, number: 5 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2072-6643/13/5/1459

[105] J. K. Burgess and M. C. Harmsen, "Chronic lung diseases: entangled in extracellular matrix," *European Respiratory Review*, vol. 31, no. 163, Mar. 2022, publisher: European Respiratory Society Section: Lung Science Conference Reviews. [Online]. Available: https://publications.ersnet.org/content/errev/31/163/210202

[106] M. W. Conklin, J. C. Eickhoff, K. M. Riching, C. A. Pehlke, K. W. Eliceiri, P. P. Provenzano, A. Friedl, and P. J. Keely, "Aligned Collagen Is a Prognostic Signature for Survival in Human Breast Carcinoma," *The American Journal of Pathology*, vol. 178, no. 3, pp. 1221–1232, Mar. 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0002944010002336

[107] Y.-N. Lee and C.-T. Chien, "Hearing lessons from flies," *eLife*, vol. 5, p. e19285, Aug. 2016. [Online]. Available: https://elifesciences.org/articles/19285

[108] S. Ranjit, A. Dvornikov, M. Stakic, S.-H. Hong, M. Levi, R. M. Evans, and E. Gratton, "Imaging Fibrosis and Separating Collagens using Second Harmonic Generation and Phasor Approach to Fluorescence Lifetime Imaging," *Scientific Reports*, vol. 5, no. 1, p. 13378, Aug. 2015. [Online]. Available: https://www.nature.com/articles/srep13378

[109] Z. Yang, Z. Yue, X. Ma, and Z. Xu, "Calcium Homeostasis: A Potential Vicious Cycle of Bone Metastasis in Breast Cancers," *Frontiers in Oncology*, vol. 10, p. 293, Mar. 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fonc.2020.00293/full

[110] B. Bierie and H. L. Moses, "Tumour microenvironment: TGFbeta: the molecular Jekyll and Hyde of cancer," *Nature Reviews. Cancer*, vol. 6, no. 7, pp. 506–520, Jul. 2006.

[111] D. Hanahan and J. Folkman, "Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis," *Cell*, vol. 86, no. 3, pp. 353–364, Aug. 1996.

[112] O. Warburg, "On respiratory impairment in cancer cells," *Science (New York, N.Y.)*, vol. 124, no. 3215, pp. 269–270, Aug. 1956.

[113] P. A. Jones and S. B. Baylin, "The epigenomics of cancer," *Cell*, vol. 128, no. 4, pp. 683–692, Feb. 2007.

[114] F. Colotta, P. Allavena, A. Sica, C. Garlanda, and A. Mantovani, "Cancer-related inflammation, the seventh hallmark of cancer: links to genetic instability," *Carcinogenesis*, vol. 30, no. 7, pp. 1073–1081, Jul. 2009.

[115] R. A. Weinberg, "How Cancer Arises," *Scientific American*, vol. 275, no. 3, pp. 62–70, Sep. 1996. [Online]. Available: https://www.scientificamerican.com/article/how-cancer-arises

[116] H. Joensuu and S. Dimitrijevic, "Tyrosine kinase inhibitor imatinib (STIS71) as an anticancer agent for solid tumours," *Annals of Medicine*, vol. 33, no. 7, pp. 451–455, Jan. 2001. [Online]. Available: http://www.tandfonline.com/doi/full/10.3109/07853890109002093

[117] M. C. Heinrich, C. D. Blanke, B. J. Druker, and C. L. Corless, "Inhibition of KIT Tyrosine Kinase Activity: A Novel Molecular Approach to the Treatment of KIT-Positive Malignancies," *Journal of Clinical Oncology*, vol. 20, no. 6, pp. 1692–1703, Mar. 2002. [Online]. Available: https://ascopubs.org/doi/10.1200/JCO.2002.20.6.1692

[118] S. W. Chae, J. H. Sohn, D.-H. Kim, Y. J. Choi, Y. L. Park, K. Kim, Y. H. Cho, J.-S. Pyo, and J. H. Kim, "Overexpressions of Cyclin B1, cdc2, p16 and p53 in Human Breast Cancer: The Clinicopathologic Correlations and Prognostic Implications," *Yonsei Medical Journal*, vol. 52, no. 3, p. 445, 2011. [Online]. Available: https://eymj.org/DOIx.php?id=10.3349/ymj.2011.52.3.445

[119] I. K. Bukholm and J. M. Nesland, "Protein expression of p53, p21 (WAF1/CIP1), bcl-2, Bax, cyclin D1 and pRb in human colon carcinomas," *Virchows Archiv*, vol. 436, no. 3, pp. 224–228, Mar. 2000. [Online]. Available: http://link.springer.com/10.1007/s004280050034

[120] G. Matlashewski, P. Lamb, D. Pim, J. Peacock, L. Crawford, and S. Benchimol, "Isolation and characterization of a human p53 cDNA clone: expression of the human p53 gene." *The EMBO Journal*, vol. 3, no. 13, pp. 3257–3262, Dec. 1984. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/j.1460-2075.1984.tb02287.x

[121] S. E. Goelz, B. Vogelstein, S. R. Hamilton, and A. P. Feinberg, "Hypomethylation of DNA from Benign and Malignant Human Colon Neoplasms," *Science*, vol. 228, no. 4696, pp. 187–190, Apr. 1985. [Online]. Available: https://www.science.org/doi/10.1126/science.2579435

[122] A. S. Wilson, B. E. Power, and P. L. Molloy, "DNA hypomethylation and human diseases," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1775, no. 1, pp. 138–162, Jan. 2007. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0304419X06000564

[123] B. W. Futscher, M. M. O'Meara, C. J. Kim, M. A. Rennels, D. Lu, L. M. Gruman, R. E. B. Seftor, M. J. C. Hendrix, and F. E. Domann, "Aberrant Methylation of the

Maspin Promoter Is an Early Event in Human Breast Cancer," *Neoplasia*, vol. 6, no. 4, pp. 380–389, Jul. 2004. [Online]. Available: http://www.ingentaselect.com/rpsv/cgi-bin/cgi?ini=xref&body=linker&reqdoi=10.1593/neo.04115

[124] M. Esteller, "Epigenetic gene silencing in cancer: the DNA hypermethylome," *Human Molecular Genetics*, vol. 16, no. R1, pp. R50–R59, Apr. 2007. [Online]. Available: http://academic.oup.com/hmg/article/16/R1/R50/2355975/Epigenetic-gene-silencing-in-cancer-the-DNA

[125] M. Li, W. D. Chen, N. Papadopoulos, S. N. Goodman, N. C. Bjerregaard, S. Laurberg, B. Levin, H. Juhl, N. Arber, H. Moinova, K. Durkee, K. Schmidt, Y. He, F. Diehl, V. E. Velculescu, S. Zhou, L. A. Diaz, K. W. Kinzler, S. D. Markowitz, and B. Vogelstein, "Sensitive digital quantification of DNA methylation in clinical samples," *Nature Biotechnology*, vol. 27, no. 9, pp. 858–863, Sep. 2009. [Online]. Available: https://www.nature.com/articles/nbt.1559

[126] A. Doi, I.-H. Park, B. Wen, P. Murakami, M. J. Aryee, R. Irizarry, B. Herb, C. Ladd-Acosta, J. Rho, S. Loewer, J. Miller, T. Schlaeger, G. Q. Daley, and A. P. Feinberg, "Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts," *Nature Genetics*, vol. 41, no. 12, pp. 1350–1353, Dec. 2009. [Online]. Available: https://www.nature.com/articles/ng.471

[127] A. Miremadi, M. Z. Oestergaard, P. D. Pharoah, and C. Caldas, "Cancer genetics of epigenetic genes," *Human Molecular Genetics*, vol. 16, no. R1, pp. R28–R49, Apr. 2007. [Online]. Available: http://academic.oup.com/hmg/article/16/R1/R28/2356055/Cancer-genetics-of-epigenetic-genes

[128] P. Marino, M. Mininni, G. Deiana, G. Marino, R. Divella, I. Bochicchio, A. Giuliano, S. Lapadula, A. R. Lettini, and F. Sanseverino, "Healthy Lifestyle and Cancer Risk: Modifiable Risk Factors to Prevent Cancer," *Nutrients*, vol. 16, no. 6, p. 800, Mar. 2024. [Online]. Available: https://www.mdpi.com/2072-6643/16/6/800

[129] H. Arem and E. Loftfield, "Cancer Epidemiology: A Survey of Modifiable Risk Factors for Prevention and Survivorship," *American Journal of Lifestyle Medicine*, vol. 12, no. 3, pp. 200–210, May 2018. [Online]. Available: https://journals.sagepub.com/doi/10.1177/1559827617700600

[130] H. Rumgay, N. Murphy, P. Ferrari, and I. Soerjomataram, "Alcohol and cancer: Epidemiology and biological mechanisms," *Nutrients*, vol. 13, no. 9, p. 3173, Sep 2021.

[131] J. Poorolajal, Y. Mohammadi, M. Fattahi-Darghlou, and F. Almasi-Moghadam, "The association between major gastrointestinal cancers and red and processed meat and fish consumption: A systematic review and meta-analysis of the observational studies," *PLOS ONE*, vol. 19, no. 6, Jun 2024.

[132] H. Seong, R. Izutsu, M. Osaki, and F. Okada, "Cancer prevention: past challenges and future directions," *Genes and Environment*, vol. 47, no. 1, p. 4, Feb. 2025. [Online]. Available: https://genesenvironment.biomedcentral.com/articles/10.1186/s41021-025-00326-y

[133] D. B. D. Silva, M. A. D. Pianovski, and N. P. D. Carvalho Filho, "Environmental pollution and cancer," *Jornal de Pediatria*, vol. 101, pp. S18–S26, Mar. 2025. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0021755724001323

[134] V. Barili, E. Ambrosini, B. Bortesi, R. Minari, E. De Sensi, I. R. Cannizzaro, A. Taiani, M. Michiara, A. Sikokis, D. Boggiani, C. Tommasi, O. Serra, F. Bonatti, A. Adorni, A. Luberto, P. Caggiati, D. Martorana, V. Uliana, A. Percesepe, A. Musolino, and B. Pellegrino, "Genetic Basis of Breast and Ovarian Cancer: Approaches and Lessons Learnt from Three Decades of Inherited Predisposition Testing," *Genes*, vol. 15, no. 2, p. 219, Feb. 2024. [Online]. Available: https://www.mdpi.com/2073-4425/15/2/219

[135] E. Dika, F. Scarf, M. Ferracin, E. Broseghini, E. Marcelli, B. Bortolani, E. Campione, M. Riefolo, C. Ricci, and M. Lambertini, "Basal Cell Carcinoma: A Comprehensive Review," *International Journal of Molecular Sciences*, vol. 21, no. 15, p. 5572, Aug. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7432343/

[136] A. Combalia and C. Carrera, "Squamous Cell Carcinoma: An Update on Diagnosis and Treatment," *Dermatology Practical & Conceptual*, vol. 10, no. 3, p. e2020066, Jun. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7319751/

[137] J. A. Ajani, J. Lee, T. Sano, Y. Y. Janjigian, D. Fan, and S. Song, "Gastric adenocarcinoma," *Nature Reviews Disease Primers*, vol. 3, no. 1, pp. 1–19, Jun. 2017, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nrdp201736

[138] X. Li, W. W. Choi, R. Yan, H. Yu, V. Krasnoperov, S. R. Kumar, A. Schuckman, D. J. Klumpp, C.-X. Pan, D. Quinn, I. S. Gill, P. S. Gill, and R. Liu, "The Differential Expression of EphB2 and EphB4 Receptor Kinases in Normal Bladder and in Transitional Cell Carcinoma of the Bladder," *PLoS ONE*, vol. 9, no. 8, p. e105326, Aug. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4141800/

[139] D. A. Vodanovich and P. F. M Choong, "Soft-tissue Sarcomas," *Indian Journal of Orthopaedics*, vol. 52, no. 1, pp. 35–44, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5791230/

[140] H. Rose-Inman and D. Kuehl, "Acute Leukemia," *Hematology/Oncology Clinics*, vol. 31, no. 6, pp. 1011–1028, Dec. 2017, publisher: Elsevier. [Online]. Available: https://www.hemonc.theclinics.com/article/S0889-8588(17)30130-2/fulltext

[141] L. E. Davis, S. C. Shalin, and A. J. Tackett, "Current state of melanoma diagnosis and treatment," *Cancer Biology & Therapy*, vol. 20, no. 11, pp. 1366–1379, Aug. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6804807/

[142] S. Łukasiewicz, M. Czeczelewski, A. Forma, J. Baj, R. Sitarz, and A. Stanisławek, "Breast CancerEpidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment StrategiesAn Updated Review," *Cancers*, vol. 13, no. 17, p. 4287, Aug. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8428369/

[143] Y.-S. Sun, Z. Zhao, Z.-N. Yang, F. Xu, H.-J. Lu, Z.-Y. Zhu, W. Shi, J. Jiang, P.-P. Yao, and H.-P. Zhu, "Risk Factors and Preventions of Breast Cancer," *International Journal of Biological Sciences*, vol. 13, no. 11, pp. 1387–1397, Nov. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5715522/

[144] E. I. Obeagu and G. U. Obeagu, "Breast cancer: A review of risk factors and diagnosis," *Medicine*, vol. 103, no. 3, p. e36905, Jan. 2024.

[145] G. Menon, F. M. Alkabban, and T. Ferguson, "Breast Cancer," in *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK482286/

[146] N. Fakhri, M. A. Chad, M. Lahkim, A. Houari, H. Dehbi, A. Belmouden, and N. El Kadmiri, "Risk factors for breast cancer in women: an update review," *Medical Oncology (Northwood, London, England)*, vol. 39, no. 12, p. 197, Sep. 2022.

[147] T. A. Buchholz and D. E. Wazer, "Molecular biology and genetics of breast cancer development: A clinical perspective," *Seminars in Radiation Oncology*, vol. 12, no. 4, pp. 285–295, Oct. 2002. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053429602800303

[148] C. O. Madu, S. Wang, C. O. Madu, and Y. Lu, "Angiogenesis in Breast Cancer Progression, Diagnosis, and Treatment," *Journal of Cancer*, vol. 11, no. 15, pp. 4474–4494, 2020. [Online]. Available: http://www.jcancer.org/v11p4474.htm

[149] A. A. Velez and M. Howard, "Tumor-suppressor genes, cell cycle regulatory checkpoints, and the skin," *North American Journal of Medical Sciences*, vol. 7, no. 5, p. 176, 2015. [Online]. Available: http://www.najms.org/text.asp?2015/7/5/176/157476

[150] A. B. Williams and B. Schumacher, "p53 in the DNA-Damage-Repair Process," *Cold Spring Harbor Perspectives in Medicine*, vol. 6, no. 5, p. a026070, May 2016. [Online]. Available: http://perspectivesinmedicine.cshlp.org/lookup/doi/10.1101/cshperspect.a026070

[151] S. Shiovitz and L. Korde, "Genetics of breast cancer: a topic in evolution," *Annals of Oncology*, vol. 26, no. 7, pp. 1291–1299, Jul. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0923753419344813

[152] L. E. Rosenberg and D. D. Rosenberg, "The Genetics of Cancer," in *Human Genes and Genomes*. Elsevier, 2012, pp. 259–288. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780123852120000160

[153] L. Ottini, P. Rizzolo, V. Silvestri, and M. Falchetti, "Inherited and acquired alterations in development of breast cancer," *The Application of Clinical Genetics*, p. 145, Nov. 2011. [Online]. Available: http://www.dovepress.com/inherited-and-acquired-alterations-in-development-of-breast-cancer-peer-reviewed-article-TACG

[154] M. Giussani, G. Merlino, V. Cappelletti, E. Tagliabue, and M. G. Daidone, "Tumor-extracellular matrix interactions: Identification of tools associated with breast cancer progression," *Seminars in Cancer Biology*, vol. 35, pp. 3–10, Dec. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1044579X15000930

[155] B. Weng, M. Li, W. Zhu, J. Peng, X. Mao, Y. Zheng, C. Zhang, S. Pan, H. Mao, and J. Zhao, "Distinguished biomimetic dECM system facilitates early detection of metastatic breast cancer cells," *Bioengineering & Translational Medicine*, vol. 9, no. 1, p. e10597, Jan. 2024. [Online]. Available: https://aiche.onlinelibrary.wiley.com/doi/10.1002/btm2.10597

[156] S. Heydari, F. Tajik, S. Safaei, F. Kamani, B. Karami, S. Dorafshan, Z. Madjd, and R. Ghods, "The association between tumor-stromal collagen features and the clinical outcomes of patients with breast cancer: a systematic review," *Breast Cancer Research*, vol. 27, no. 1, p. 69, May 2025. [Online]. Available: https://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-025-02017-6

[157] A. Akinpelu, T. Akinsipe, L. A. Avila, R. D. Arnold, and P. Mistriotis, "The impact of tumor microenvironment: unraveling the role of physical cues in breast cancer progression," *Cancer and Metastasis Reviews*, vol. 43, no. 2, pp. 823–844, Jun. 2024. [Online]. Available: https://link.springer.com/10.1007/s10555-024-10166-x

[158] Q. Tan, L. Xu, J. Zhang, L. Ning, Y. Jiang, T. He, J. Luo, J. Chen, Q. Lv, X. Yang, and H. Xie, "Breast cancer cells interact with tumor-derived extracellular matrix in a molecular subtype-specific manner," *Biomaterials Advances*, vol. 146, p. 213301, Mar. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2772950823000249

[159] A. N. Wilkinson, "Cancer diagnosis in primary care: Six steps to reducing the diagnostic interval," *Canadian Family Physician*, vol. 67, no. 4, p. 265268, Apr. 2021. [Online]. Available: http://dx.doi.org/10.46747/cfp.6704265

[160] M. Hallek, "Chronic lymphocytic leukemia: 2020 update on diagnosis, risk stratification and treatment," *American Journal of Hematology*, vol. 94, no. 11, p. 12661287, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1002/ajh.25595

[161] R. S. Matulewicz, J. O. DeLancey, E. Pavey, E. M. Schaeffer, O. Popescu, and J. J. Meeks, "Dipstick urinalysis as a test for microhematuria and occult bladder cancer," *Bladder Cancer*, vol. 3, no. 1, p. 4549, Jan. 2017. [Online]. Available: http://dx.doi.org/10.3233/BLC-160068

[162] P. R. Galle, F. Foerster, M. Kudo, S. L. Chan, J. M. Llovet, S. Qin, W. R. Schelman, S. Chintharlapalli, P. B. Abada, M. Sherman, and A. X. Zhu, "Biology and significance of alphafetoprotein in hepatocellular carcinoma," *Liver International*, vol. 39, no. 12, p. 22142229, Sep. 2019. [Online]. Available: http://dx.doi.org/10.1111/liv.14223

[163] W. Hamilton, "Cancer diagnosis in primary care," *British Journal of General Practice*, vol. 60, no. 571, p. 121128, Feb. 2010. [Online]. Available: http://dx.doi.org/10.3399/bjgp10X483175

[164] H. Semaan, M. F. Bazerbashi, G. Siesel, P. Aldinger, and T. Obri, "Diagnostic accuracy of non-contrast abdominal ct scans performed as follow-up for patients with an established cancer diagnosis: a retrospective study," *Acta Oncologica*, vol. 57, no. 3, p. 426430, Aug. 2017. [Online]. Available: http://dx.doi.org/10.1080/0284186X.2017.1360512

[165] L. Nicosia, G. Gnocchi, I. Gorini, M. Venturini, F. Fontana, F. Pesapane, I. Abiuso, A. C. Bozzini, M. Pizzamiglio, A. Latronico, F. Abbate, L. Meneghetti, O. Battaglia, G. Pellegrino, and E. Cassano, "History of mammography: Analysis of breast imaging diagnostic achievements over the last century," *Healthcare*, vol. 11, no. 11, p. 1596, May 2023. [Online]. Available: http://dx.doi.org/10.3390/healthcare11111596

[166] J. Frutos Díaz-Alejo, I. González Gómez, and J. Earl, "Ultrasounds in cancer therapy: A summary of their use and unexplored potential," *Oncology Reviews*, vol. 16, no. 1, Feb. 2022. [Online]. Available: http://dx.doi.org/10.4081/oncol.2022.531

[167] M. D. Guimaraes, A. Schuch, B. Hochhegger, J. L. Gross, R. Chojniak, and E. Marchiori, "Functional magnetic resonance imaging in oncology: state of the art," *Radiologia Brasileira*, vol. 47, no. 2, p. 101111, Apr. 2014. [Online]. Available: http://dx.doi.org/10.1590/S0100-39842014000200013

[168] H. Brody, "Cancer diagnosis," *Nature*, vol. 579, no. 7800, pp. S1–S1, Mar. 2020. [Online]. Available: http://www.nature.com/articles/d41586-020-00840-9

[169] L. Fass, "Imaging and cancer: A review," *Molecular Oncology*, vol. 2, no. 2, pp. 115–152, Aug. 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5527766/

[170] L. He, L. R. Long, S. Antani, and G. R. Thoma, "Histology image analysis for carcinoma detection and grading," *Computer Methods and Programs in Biomedicine*, vol. 107, no. 3, pp. 538–56, Sep. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22436890

[171] R. Alyamany, R. El Fakih, A. Alnughmush, A. Albabtain, M. A. Kharfan-Dabaja, and M. Aljurf, "A comprehensive review of the role of bone marrow biopsy and pet-ct in the evaluation of bone marrow involvement in adults newly diagnosed with dlbcl," *Frontiers in Oncology*, vol. 14, Mar. 2024. [Online]. Available: http://dx.doi.org/10.3389/fonc.2024.1301979

[172] J. Dinnes, J. J. Deeks, N. Chuchu, L. Ferrante di Ruffano, R. N. Matin, D. R. Thomson, K. Y. Wong, R. B. Aldridge, R. Abbott, M. Fawzy, S. E. Bayliss, M. J. Grainge, Y. Takwoingi, C. Davenport, K. Godfrey, F. M. Walter, and H. C. Williams, "Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults," *Cochrane Database of Systematic Reviews*, vol. 2018, no. 12, Dec. 2018. [Online]. Available: http://dx.doi.org/10.1002/14651858.CD011902.pub2

[173] D. T. Debela, S. G. Muzazu, K. D. Heraro, M. T. Ndalama, B. W. Mesele, D. C. Haile, S. K. Kitui, and T. Manyazewal, "New approaches and procedures for cancer treatment: Current perspectives," *SAGE Open Medicine*, vol. 9, p. 20503121211034366, Jan. 2021. [Online]. Available: https://journals.sagepub.com/doi/10.1177/20503121211034366

[174] R. B. Mokhtari, T. S. Homayouni, N. Baluch, E. Morgatskaya, S. Kumar, B. Das, and H. Yeger, "Combination therapy in combating cancer," *Oncotarget*, vol. 8, no. 23, pp. 38 022–38 043, Jun. 2017. [Online]. Available: https://www.oncotarget.com/lookup/doi/10.18632/oncotarget.16723

[175] M. Arruebo, N. Vilaboa, B. Sáez-Gutierrez, J. Lambea, A. Tres, M. Valladares, and A. González-Fernández, "Assessment of the Evolution of Cancer Treatment Therapies," *Cancers*, vol. 3, no. 3, pp. 3279–3330, Aug. 2011. [Online]. Available: https://www.mdpi.com/2072-6694/3/3/3279

[176] M. A. Moses, H. Brem, and R. Langer, "Advancing the field of drug delivery," *Cancer Cell*, vol. 4, no. 5, pp. 337–341, Nov. 2003. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1535610803002769

[177] Z. Lei, Q. Tian, Q. Teng, J. N. D. Wurpel, L. Zeng, Y. Pan, and Z. Chen, "Understanding and targeting resistance mechanisms in cancer," *MedComm*, vol. 4, no. 3, p. e265, Jun. 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/mco2.265

[178] M. Z. El-Readi and M. A. Althubiti, "Cancer Nanomedicine: A New Era of Successful Targeted Therapy," *Journal of Nanomaterials*, vol. 2019, pp. 1–13, Oct. 2019. [Online]. Available: https://www.hindawi.com/journals/jnm/2019/4927312/

[179] A. Naji, M. Eitoku, B. Favier, F. Deschaseaux, N. Rouas-Freiss, and N. Suganuma, "Biological functions of mesenchymal stem cells and clinical implications," *Cellular and Molecular Life Sciences*, vol. 76, no. 17, pp. 3323–3348, Sep. 2019. [Online]. Available: http://link.springer.com/10.1007/s00018-019-03125-1

[180] E. Dietrich and K. Antoniades, "Molecularly targeted drugs for the treatment of cancer: oral complications and pathophysiology," *Hippokratia*, vol. 16, no. 3, pp. 196–199, Jul. 2012.

[181] I. Adjei and S. Blanka, "Modulation of the Tumor Microenvironment for Cancer Treatment: A Biomaterials Approach," *Journal of Functional Biomaterials*, vol. 6, no. 1, pp. 81–103, Feb. 2015. [Online]. Available: https://www.mdpi.com/2079-4983/6/1/81

[182] K. Valter, L. Chen, B. Kruspig, P. Maximchik, H. Cui, B. Zhivotovsky, and V. Gogvadze, "Contrasting effects of glutamine deprivation on apoptosis induced by conventionally used anticancer drugs," *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, vol. 1864, no. 3, pp. 498–506, Mar. 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S016748891630341X

[183] K. M. Bailey, J. W. Wojtkowiak, A. I. Hashim, and R. J. Gillies, "Targeting the Metabolic Microenvironment of Tumors," in *Advances in Pharmacology*. Elsevier, 2012, vol. 65, pp. 63–107. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B978012397927800004X

[184] H.-Y. Min and H.-Y. Lee, "Molecular targeted therapy for anticancer treatment," *Experimental & Molecular Medicine*, vol. 54, no. 10, pp. 1670–1694, Oct. 2022. [Online]. Available: https://www.nature.com/articles/s12276-022-00864-3

[185] W. A. Weber, "Positron Emission Tomography As an Imaging Biomarker," *Journal of Clinical Oncology*, vol. 24, no. 20, pp. 3282–3292, Jul. 2006. [Online]. Available: https://ascopubs.org/doi/10.1200/JCO.2006.06.6068

[186] J. Luo, B. D. Manning, and L. C. Cantley, "Targeting the PI3K-Akt pathway in human cancer," *Cancer Cell*, vol. 4, no. 4, pp. 257–262, Oct. 2003. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1535610803002484

[187] S. Yang, Y. Fang, Y. Ma, F. Wang, Y. Wang, J. Jia, Y. Yang, W. Sun, Q. Zhou, and Z. Li, "Angiogenesis and targeted therapy in the tumour microenvironment: From basic to clinical practice," *Clinical and Translational Medicine*, vol. 15, no. 4, p. e70313, Apr. 2025. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/ctm2.70313

[188] H. Hurwitz, L. Fehrenbacher, W. Novotny, T. Cartwright, J. Hainsworth, W. Heim, J. Berlin, A. Baron, S. Griffing, E. Holmgren, N. Ferrara, G. Fyfe, B. Rogers, R. Ross, and F. Kabbinavar, "Bevacizumab plus Irinotecan, Fluorouracil, and Leucovorin for Metastatic Colorectal Cancer," *New England Journal of Medicine*, vol. 350, no. 23, pp. 2335–2342, Jun. 2004. [Online]. Available: http://www.nejm.org/doi/abs/10.1056/NEJMoa032691

[189] C. Pucci, C. Martinelli, and G. Ciofani, "Innovative approaches for cancer treatment: current perspectives and new challenges," *ecancermedicalscience*, vol. 13, Sep. 2019. [Online]. Available: https://ecancer.org/en/journal/article/961-innovative-approaches-for-cancer-treatment-current-perspectives-and-new-challenges

[190] S. Chikara, L. D. Nagaprashantha, J. Singhal, D. Horne, S. Awasthi, and S. S. Singhal, "Oxidative stress and dietary phytochemicals: Role in cancer chemoprevention and treatment," *Cancer Letters*, vol. 413, pp. 122–134, Jan. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0304383517306948

[191] G. B. Celli, E. L. Denchi, and T. De Lange, "Ku70 stimulates fusion of dysfunctional telomeres yet protects chromosome ends from homologous recombination," *Nature Cell Biology*, vol. 8, no. 8, pp. 885–890, Aug. 2006. [Online]. Available: https://www.nature.com/articles/ncb1444

[192] K. R. Levental, H. Yu, L. Kass, J. N. Lakins, M. Egeblad, J. T. Erler, S. F. Fong, K. Csiszar, A. Giaccia, W. Weninger, M. Yamauchi, D. L. Gasser, and V. M. Weaver, "Matrix Crosslinking Forces Tumor Progression by Enhancing Integrin Signaling," *Cell*, vol. 139, no. 5, pp. 891–906, Nov. 2009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0092867409013531

[193] K. A. Aldinger and S. Qiu, "New mouse genetic model duplicates human 15q1113 autistic phenotypes, or does it?" *Disease Models & Mechanisms*, vol. 3, no. 1-2, pp. 3–4, Jan. 2010. [Online]. Available: https://journals.biologists.com/dmm/article/3/1-2/3/2379/New-mouse-genetic-model-duplicates-human-15q11-13

[194] D. Zink, A. H. Fischer, and J. A. Nickerson, "Nuclear structure in cancer cells," *Nature Reviews Cancer*, vol. 4, no. 9, pp. 677–687, Sep. 2004. [Online]. Available: https://www.nature.com/articles/nrc1430

[195] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot, "Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1196–1206, May 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7399414/

[196] Y. Arafat and C. C. Reyes-Aldasoro, *Computational Image Analysis Techniques, Programming Languages and Software Platforms Used in Cancer Research: A Scoping Review*. Springer International Publishing, 2022, p. 833847.

[197] C. C. Reyes-Aldasoro, "The proportion of cancer-related entries in PubMed has increased considerably; is cancer truly "The Emperor of All Maladies"?" *PloS One*, vol. 12, no. 3, p. e0173671, 2017.

[198] Y. Xie, J. Zhang, and Y. Xia, "Semi-supervised adversarial model for benignmalignant lung nodule classification on chest CT," *Medical Image Analysis*, vol. 57, pp. 237–248,

Oct. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841519300611

[199] M. K. Abd-Ellah, A. I. Awad, A. A. M. Khalaf, and H. F. A. Hamed, "A review on brain tumor diagnosis from MRI images: Practical implications, key achievements, and lessons learned," *Magnetic Resonance Imaging*, vol. 61, pp. 300–318, Sep. 2019.

[200] S. Y. Ko, J. H. Lee, J. H. Yoon, H. Na, E. Hong, K. Han, I. Jung, E.-K. Kim, H. J. Moon, V. Y. Park, E. Lee, and J. Y. Kwak, "Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound," *Head & Neck*, vol. 41, no. 4, pp. 885–891, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hed.25415

[201] F. Mahmood, D. Borders, R. J. Chen, G. N. Mckay, K. J. Salimian, A. Baras, and N. J. Durr, "Deep Adversarial Training for Multi-Organ Nuclei Segmentation in Histopathology Images," *IEEE transactions on medical imaging*, vol. 39, no. 11, pp. 3257–3267, Nov. 2020.

[202] N. Tomita, B. Abdollahi, J. Wei, B. Ren, A. Suriawinata, and S. Hassanpour, "Attention-Based Deep Neural Networks for Detection of Cancerous and Precancerous Esophagus Tissue on Histopathological Slides," *JAMA Network Open*, vol. 2, no. 11, p. e1914645, Nov. 2019. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2019.14645

[203] J.-H. Tang, F.-H. Yan, M.-L. Zhou, P.-J. Xu, J. Zhou, and J. Fan, "Evaluation of computer-assisted quantitative volumetric analysis for pre-operative resectability assessment of huge hepatocellular carcinoma," *Asian Pacific journal of cancer prevention: APJCP*, vol. 14, no. 5, p. 30453050, 2013.

[204] A. Yung, J. Kay, P. Beale, K. A. Gibson, and T. Shaw, "Computer-based decision tools for shared therapeutic decision-making in oncology: Systematic review," *JMIR cancer*, vol. 7, no. 4, p. e31616, Oct 2021.

[205] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber, L. Jansen, C. C. Reyes-Aldasoro, I. Zrnig, D. Jger, H. Brenner, J. Chang-Claude, M. Hoffmeister, and N. Halama, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLOS Medicine*, vol. 16, no. 1, p. e1002730, Jan. 2019, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002730

[206] J. Fourier, "Memoire sur la propagation de la chaleur dans les corps solides," *Nouveau Bulletin des sciences par la Societe philomatique de Paris*, vol. I, p. 112116, 1808.

[207] J. Serra, "Introduction to mathematical morphology," *Computer Vision, Graphics, and Image Processing*, vol. 35, no. 3, p. 283305, Sep 1986.

[208] B. B. Mandelbrot, *The fractal geometry of nature*. Freeman, 1983.

[209] E. Oczeretko, M. Juczewska, and I. Kasacka, "Fractal geometric analysis of lung cancer angiogenic patterns," *Folia histochemica et cytobiologica*, vol. 39 Suppl 2, p. 7576, Jan 2001.

[210] A. W. Partin, J. S. Schoeniger, J. L. Mohler, and D. S. Coffey, "Fourier analysis of cell motility: correlation of motility with metastatic potential," *Proceedings of the National Academy of Sciences*, vol. 86, no. 4, p. 12541258, Feb 1989.

[211] C. C. Reyes-Aldasoro, L. J. Williams, S. Akerman, C. Kanthou, and G. M. Tozer, "An automatic algorithm for the segmentation and morphological analysis of microvessels in immunostained histological tumour sections," *Journal of microscopy*, vol. 242, no. 3, pp. 262–278, Jun. 2011.

[212] M. D. J. Peters, C. M. Godfrey, H. Khalil, P. McInerney, D. Parker, and C. B. Soares, "Guidance for conducting systematic scoping reviews," *JBI Evidence Implementation*, vol. 13, no. 3, p. 141146, Sep 2015.

[213] A. C. Tricco, J. Antony, W. Zarin, L. Strifler, M. Ghassemi, J. Ivory, L. Perrier, B. Hutton, D. Moher, and S. E. Straus, "A scoping review of rapid review methods," *BMC Medicine*, vol. 13, no. 1, p. 224, Sep 2015.

[214] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. A. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher, "The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration," *BMJ*, vol. 339, p. b2700, Jul 2009.

[215] D. Moher, L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, L. A. Stewart, and P.-P. Group, "Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement," *Systematic Reviews*, vol. 4, no. 1, p. 1, Jan 2015.

[216] T. Theodosiou, I. S. Vizirianakis, L. Angelis, A. Tsaftaris, and N. Darzentas, "MeSHy: Mining unanticipated PubMed information using frequencies of occurrences and concurrences of MeSH terms," *Journal of Biomedical Informatics*, vol. 44, no. 6, pp. 919–926, Dec. 2011.

[217] N. Jamali, E. T. A. Dobson, K. W. Eliceiri, A. E. Carpenter, and B. A. Cimini, "2020 bioimage analysis survey: Community experiences and needs for the future," *Biological Imaging*, vol. 1, 2022.

[218] P. Bankhead, M. B. Loughrey, J. A. Fernndez, Y. Dombrowski, D. G. McArt, P. D. Dunne, S. McQuaid, R. T. Gray, L. J. Murray, H. G. Coleman, J. A. James, M. Salto-Tellez, and P. W. Hamilton, "Qupath: Open source software for digital pathology image analysis," *Scientific Reports*, vol. 7, no. 1, p. 16878, Dec 2017.

[219] M. R. Lamprecht, D. M. Sabatini, and A. E. Carpenter, "Cellprofiler: free, versatile software for automated biological image analysis," *BioTechniques*, vol. 42, no. 1, p. 7175, Jan 2007.

[220] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J. Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona, "Fiji: an open-source platform for biological-image analysis," *Nature Methods*, vol. 9, no. 7, pp. 676–682, Jul. 2012. [Online]. Available: ://000305942200021

[221] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "Nih image to imagej: 25 years of image analysis," *Nature Methods*, vol. 9, no. 7, p. 671675, Jul 2012.

[222] F. de Chaumont, S. Dallongeville, N. Chenouard, N. Herv, S. Pop, T. Provoost, V. Meas-Yedid, P. Pankajakshan, T. Lecomte, Y. Le Montagner, T. Lagache, A. Dufour, and J.-C. Olivo-Marin, "Icy: an open bioimage informatics platform for extended reproducible research," *Nature Methods*, vol. 9, no. 7, p. 690696, Jun 2012.

[223] J. N. Kather, A. S. Berghoff, D. Ferber, M. Suarez-Carmona, C. C. Reyes-Aldasoro, N. A. Valous, R. Rojas-Moraleda, D. Jger, and N. Halama, "Large-scale database mining reveals hidden trends and future directions for cancer immunotherapy," *Oncoimmunology*, vol. 7, no. 7, p. e1444412, 2018.

[224] C.-W. Lee, Y. J. Ren, M. Marella, M. Wang, J. Hartke, and S. S. Couto, "Multiplex immunofluorescence staining and image analysis assay for diffuse large b cell lymphoma," *Journal of Immunological Methods*, vol. 478, p. 112714, Mar 2020.

[225] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, Sep. 2018. [Online]. Available: https://joss.theoj.org/papers/10.21105/joss.00861

[226] T. S. Gurina and L. Simms, "Histology, Staining," in *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK557663/

[227] H. A. Alturkistani, F. M. Tashkandi, and Z. M. Mohammedsaleh, "Histological Stains: A Literature Review and Case Study," *Global Journal of Health Science*, vol. 8, no. 3, pp. 72–79, Jun. 2015.

[228] A. T. Feldman and D. Wolfe, "Tissue processing and hematoxylin and eosin staining," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 1180, pp. 31–43, 2014.

[229] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: a review," *IEEE reviews in biomedical engineering*, vol. 2, pp. 147–71, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20671804

[230] M. Hggstrm., "Histopathology of well differentiated neuroendocrine tumor (grade 1, from the lung) with salt-and-pepper chromatin, H&E stain," Aug. 2022.

[231] S. Magaki, S. A. Hojat, B. Wei, A. So, and W. H. Yong, "An Introduction to the Performance of Immunohistochemistry," in *Biobanking*, W. H. Yong, Ed. New York, NY: Springer New York, 2019, vol. 1897, pp. 289–298, series Title: Methods in Molecular Biology.

[232] H. M. Hussaini, B. Seo, and A. M. Rich, "Immunohistochemistry and Immunofluorescence," in *Oral Biology*, G. J. Seymour, M. P. Cullinan, N. C. Heng, and P. R. Cooper, Eds. New York, NY: Springer US, 2023, vol. 2588, pp. 439–450, series Title: Methods in Molecular Biology.

[233] W. Grace-Jones, "25 - Tissue Microarray," in *Theory and Practice of Histological Techniques (Sixth Edition)*, J. D. Bancroft and M. Gamble, Eds. Edinburgh: Churchill Livingstone, Jan. 2008, pp. 527–535. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780443102790500324

[234] M. Sjöbeck, M. Haglund, A. Persson, K. Sturesson, and E. Englund, "Brain tissue microarrays in dementia research: white matter microvascular pathology in Alzheimer's disease," *Neuropathology: Official Journal of the Japanese Society of Neuropathology*, vol. 23, no. 4, pp. 290–295, Dec. 2003.

[235] D. Dum, A. Menz, C. Völkel, N. De Wispelaere, A. Hinsch, N. Gorbokon, M. Lennartz, A. M. Luebke, C. Hube-Magg, M. Kluth, C. Fraune, K. Möller, C. Bernreuther, P. Lebok, T. S. Clauditz, F. Jacobsen, G. Sauter, R. Uhlig, W. Wilczak, S. Steurer, S. Minner, A. H. Marx, R. Simon, E. Burandt, and T. Krech, "Cytokeratin 7 and cytokeratin 20 expression in cancer: A tissue microarray study on 15,424 cancers," *Experimental and Molecular Pathology*, vol. 126, p. 104762, Jun. 2022.

[236] D. Van De Vlekkert, E. Machado, and A. d'Azzo, "Analysis of Generalized Fibrosis in Mouse Tissue Sections with Massons Trichrome Staining," *Bio-protocol*, vol. 10, no. 10, p. e3629, May 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7842772/

[237] D. Sridharan, N. Pracha, J. A. Dougherty, A. Akhtar, S. B. Alvi, and M. Khan, "A One-Stop Protocol to Assess Myocardial Fibrosis in Frozen and Paraffin Sections," *Methods and Protocols*, vol. 5, no. 1, p. 13, Jan. 2022.

[238] Q. Guo, D. Sun, A. S. Barrett, S. Jindal, N. D. Pennock, M. W. Conklin, Z. Xia, E. Mitchell, R. Samatham, N. Mirza, S. Jacques, S. Weinmann, V. F. Borges, K. C. Hansen, and P. J. Schedin, "Mammary collagen is under reproductive control with implications for breast cancer," *Matrix Biology: Journal of the International Society for Matrix Biology*, vol. 105, pp. 104–126, Jan. 2022.

[239] C. M. López De Padilla, M. J. Coenen, A. Tovar, R. E. De la Vega, C. H. Evans, and S. A. Müller, "Picrosirius Red Staining: Revisiting Its Application to the Qualitative and Quantitative Assessment of Collagen Type I and Type III in Tendon," *Journal of Histochemistry and Cytochemistry*, vol. 69, no. 10, pp. 633–643, Oct. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8504258/

[240] S. J. Rodig, "Detecting Fluorochrome-Labeled Cells," *Cold Spring Harbor Protocols*, vol. 2021, no. 6, Jun. 2021.

[241] B. I. Tarnowski, F. G. Spinale, and J. H. Nicholson, "DAPI as a Useful Stain for Nuclear Quantitation," *Biotechnic & Histochemistry*, vol. 66, no. 6, pp. 296–302, Jan. 1991, publisher: Taylor & Francis _eprint: https://doi.org/10.3109/10520299109109990. [Online]. Available: https://doi.org/10.3109/10520299109109990

[242] V. DesMarais, R. J. Eddy, V. P. Sharma, O. Stone, and J. S. Condeelis, "Optimizing leading edge F-actin labeling using multiple actin probes, fixation methods and imaging modalities," *BioTechniques*, vol. 66, no. 3, pp. 113–119, Mar. 2019, publisher: Future Science. [Online]. Available: https://www.future-science.com/doi/10.2144/btn-2018-0112

[243] F. Mualla, M. Aubreville, and A. Maier, "Microscopy," in *Medical Imaging Systems: An Introductory Guide*, A. Maier, S. Steidl, V. Christlein, and J. Hornegger, Eds. Cham (CH): Springer, 2018. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK546149/

[244] A. Cougnard-Gregoire, B. M. J. Merle, T. Aslam, J. M. Seddon, I. Aknin, C. C. W. Klaver, G. Garhfer, A. G. Layana, A. M. Minnella, R. Silva, and C. Delcourt, "Blue Light Exposure: Ocular Hazards and PreventionA Narrative Review," *Ophthalmology and Therapy*, vol. 12, no. 2, pp. 755–788, Apr. 2023. [Online]. Available: https://link.springer.com/10.1007/s40123-023-00675-3

[245] P. F. Holt and H. J. Callow, "Rapid Determination of Water in Animals and Plants," *Nature*, vol. 148, no. 3764, pp. 755–756, Dec. 1941. [Online]. Available: https://www.nature.com/articles/148755a0

[246] M. Carter and J. Shieh, "Microscopy," in *Guide to Research Techniques in Neuroscience*. Elsevier, 2015, pp. 117–144. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780128005118000058

[247] Jmgrants, "Electron microscopes." [Online]. Available: https://commons.wikimedia.org/wiki/Category:Electron_microscopes

[248] M. Carter and J. Shieh, "Chapter 5 - Microscopy," in *Guide to Research Techniques in Neuroscience (Second Edition)*, M. Carter and J. Shieh, Eds. San Diego: Academic Press, Jan. 2015, pp. 117–144. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128005118000058

[249] C. A. Noble, A. P. Biesemier, S. F. McClees, A. M. Alhussain, S. E. Helms, and R. T. Brodell, "The history of the microscope reflects advances in science and medicine," *Seminars in Diagnostic Pathology*, vol. 42, no. 2, p. 150831, Mar. 2025. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0740257024000030

[250] A. J. M. Wollman, R. Nudd, E. G. Hedlund, and M. C. Leake, "From *Animaculum* to single molecules: 300 years of the light microscope," *Open Biology*, vol. 5, no. 4, p. 150019, Apr. 2015. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsob.150019

[251] S. Bradbury and B. Bracegirdle, *Introduction to light microscopy*, ser. Microscopy handbooks. Oxford (GB): BIOS Scientific Publ, 1998, no. 42.

[252] B. Caballero, P. M. Finglas, and F. Toldrá, Eds., *Encyclopedia of food and health*, 3rd ed. Burlington: Elsevier Science, 2015.

[253] R. A. Meyers, *Encyclopedia of physical science and technology*, 3rd ed. San Diego, Calif. London: Academic, 2001.

[254] D. B. Murphy and M. W. Davidson, *Fundamentals of Light Microscopy and electronic imaging*. Wiley-Blackwell, 2013.

[255] J. Mertz, *Introduction to optical microscopy*, second edition ed. Cambridge, United Kingdom ; New York, NY: Cambridge University Press, 2019.

[256] Zephyris, "Kohler illumination en.svg." [Online]. Available: https://commons.wikimedia.org/wiki/File:Kohler_Illumination_en.svg

[257] B. R. Masters, *Confocal Microscopy and Multiphoton Excitation Microscopy*. 1000 20th Street, Bellingham, WA 98227-0010 USA: SPIE, Feb. 2006. [Online]. Available: http://link.aip.org/link/doi/10.1117/3.660403

[258] T. J. Evans and P. A. Riley, "Principles of microscopy, culture and serology-based diagnostics," *Medicine*, vol. 49, no. 10, pp. 648–653, Oct. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1357303921002012

[259] J. Sanderson, "Fundamentals of Microscopy," *Current Protocols in Mouse Biology*, vol. 10, no. 2, p. e76, Jun. 2020. [Online]. Available: https://currentprotocols.onlinelibrary.wiley.com/doi/10.1002/cpmo.76

[260] J. D. Bancroft and M. Gamble, Eds., *Theory and practice of histological techniques*, 6th ed., ser. Expertconsult. Philadelphia, PA: Churchill Livingstone/Elsevier, 2008.

[261] B. Herman, *Fluorescence Microscopy*. London: Garland Science, Aug. 2020.

[262] M. J. Sanderson, I. Smith, I. Parker, and M. D. Bootman, "Fluorescence Microscopy," *Cold Spring Harbor protocols*, vol. 2014, no. 10, p. pdb.top071795, Oct. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4711767/

[263] H. Mhlpfordt, "Fluorescence filters," 2008. [Online]. Available: https://commons.wikimedia.org/wiki/File:FluorescenceFilters_2008-09-28.svg

[264] J. W. Lichtman and J.-A. Conchello, "Fluorescence microscopy," *Nature Methods*, vol. 2, no. 12, pp. 910–919, Dec. 2005, number: 12 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nmeth817

[265] J. Lippincott-Schwartz, E. Snapp, and A. Kenworthy, "Studying protein dynamics in living cells," *Nature Reviews Molecular Cell Biology*, vol. 2, no. 6, pp. 444–456, Jun. 2001. [Online]. Available: https://www.nature.com/articles/35073068

[266] A. Ettinger and T. Wittmann, "Fluorescence live cell imaging," *Methods in Cell Biology*, p. 7794, Jan 2015.

[267] X. Wang, Q. Ding, R. R. Groleau, L. Wu, Y. Mao, F. Che, O. Kotova, E. M. Scanlan, S. E. Lewis, P. Li, and et al., "Fluorescent probes for disease diagnosis," *Chemical Reviews*, vol. 124, no. 11, p. 71067164, May 2024.

[268] D. J. Wolff, A. Bagg, L. D. Cooley, G. W. Dewald, B. A. Hirsch, P. B. Jacky, K. W. Rao, and P. N. Rao, "Guidance for Fluorescence in Situ Hybridization Testing in Hematologic Disorders," *The Journal of Molecular Diagnostics*, vol. 9, no. 2, pp. 134–143, Apr. 2007. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S152515781060373X

[269] Pmx, "Fish," 2005. [Online]. Available: https://commons.wikimedia.org/wiki/File:Bcrablmet.jpg

[270] C. M. Kraning-Rush, S. P. Carey, J. P. Califano, and C. A. Reinhart-King, "Chapter 6 - Quantifying Traction Stresses in Adherent Cells," in *Methods in Cell Biology*, ser. Computational Methods in Cell Biology, A. R. Asthagiri and A. P. Arkin, Eds. Academic Press, Jan. 2012, vol. 110, pp. 139–178. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123884039000060

[271] C. J. R. Sheppard, "CHAPTER TWENTY-ONE - Confocal Microscopy Principles, Practice and Options," in *Fluorescent and Luminescent Probes for Biological Activity (Second Edition)*, ser. Biological Techniques Series, W. Mason, Ed. London: Academic Press, Jan. 1999, pp. 303–309. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780124478367500233

[272] N. F. Braghiroli, S. Sugerik, L. A. R. D. Freitas, M. Oliviero, and H. Rabinovitz, "The skin through reflectance confocal microscopy Historical background, technical principles, and its correlation with histopathology," *Anais Brasileiros de Dermatologia*, vol. 97, no. 6, pp. 697–703, Nov. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0365059622001970

[273] R. S. Ramani, I. Tan, L. Bussau, C. M. Angel, M. McCullough, and T. Yap, "Confocal microscopy in oral cancer and oral potentially malignant disorders: A systematic review," *Oral Diseases*, vol. 29, no. 8, pp. 3003–3015, Nov. 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/odi.14291

[274] R. Malik, "Clinical applications of corneal confocal microscopy," *Clinical Ophthalmology*, p. 435, Jun. 2008. [Online]. Available: http://www.dovepress.com/clinical-applications-of-corneal-confocal-microscopy-peer-reviewed-article-OPTH

[275] S. Vicent, L. C. Sayles, D. Vaka, P. Khatri, O. Gevaert, R. Chen, Y. Zheng, A. K. Gillespie, N. Clarke, Y. Xu, J. Shrager, C. D. Hoang, S. Plevritis, A. J. Butte, and E. A. Sweet-Cordero, "Cross-species functional analysis of cancer-associated fibroblasts identifies a critical role for CLCF1 and IL-6 in non-small cell lung cancer in vivo," *Cancer Research*, vol. 72, no. 22, pp. 5744–5756, Nov. 2012.

[276] L. Parisi, A. Toffoli, B. Ghezzi, B. Mozzoni, S. Lumetti, and G. M. Macaluso, "A glance on the role of fibronectin in controlling cell response at biomaterial interface," *Jpn Dent Sci Rev*, vol. 56, no. 1, pp. 50–55, Dec 2020.

[277] S. Gupta, A. R. Ramjaun, P. Haiko, Y. Wang, P. H. Warne, B. Nicke, E. Nye, G. Stamp, K. Alitalo, and J. Downward, "Binding of ras to phosphoinositide 3-kinase p110alpha is required for ras-driven tumorigenesis in mice," *Cell*, vol. 129, no. 5, pp. 957–968, Jun. 2007.

[278] A. M. Stein, D. A. Vader, L. M. Jawerth, D. A. Weitz, and L. M. Sander, "An algorithm for extracting the network geometry of threedimensional collagen gels," *Journal of Microscopy*, vol. 232, no. 3, pp. 463–475, Dec. 2008. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2818.2008.02141.x

[279] Y. Liu and et al., "Fibrillar Collagen Quantification With Curvelet Transform Based Computational Methods," *Frontiers in Bioengineering and Biotechnology*, vol. 8, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00198

[280] B. A. Nerger, P.-T. Brun, and C. M. Nelson, "Marangoni flows drive the alignment of fibrillar cell-laden hydrogels," *Science Advances*, vol. 6, no. 24, p. eaaz7748, Jun. 2020, publisher: American Association for the Advancement of Science. [Online]. Available: https://www.science.org/doi/10.1126/sciadv.aaz7748

[281] S. Han, Y. Y. Li, and B. P. Chan, "Protease inhibitors enhance extracellular collagen fibril deposition in human mesenchymal stem cells," *Stem Cell Research & Therapy*, vol. 6, no. 1, p. 197, Dec. 2015. [Online]. Available: https://stemcellres.biomedcentral.com/articles/10.1186/s13287-015-0191-1

[282] A. Pinhero, A. M l, V. P, C. A. Visaggio, A. N, A. S, and A. S, "Malware detection employed by visualization and deep neural network,"

Computers & Security, vol. 105, p. 102247, Jun. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404821000717

[283] P. E. Pavlova, K. P. Cyrrilov, and I. N. Moumdjiev, "Application of HSV colour system in identification by colour of biological objects on the basis of microscopic images," *Computerized Medical Imaging and Graphics*, vol. 20, no. 5, pp. 357–364, Sep. 1996. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895611196000584

[284] P. S. Ganney, S. Pisharody, and E. Claridge, "Software engineering," in *Clinical Engineering*. Elsevier, 2020, pp. 131–168. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780081026946000097

[285] T. Randen and J. Husoy, "Filtering for texture classification: a comparative study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 4, pp. 291–310, 1999.

[286] R. Bracewell, *The Fourier Transform and its Applications*, 2nd ed. Tokyo: McGraw-Hill Kogakusha, Ltd., 1978.

[287] G. Hemalatha and C. P. Sumathi, "Preprocessing techniques of facial image with median and gabor filters," in *2016 International Conference on Information Communication and Embedded Systems (ICICES)*, 2016, pp. 1–6.

[288] A. E. Ilesanmi and T. O. Ilesanmi, "Methods for image denoising using convolutional neural network: a review," *Complex & Intelligent Systems*, vol. 7, no. 5, pp. 2179–2198, Oct. 2021. [Online]. Available: https://doi.org/10.1007/s40747-021-00428-4

[289] M. Mafi, H. Martin, M. Cabrerizo, J. Andrian, A. Barreto, and M. Adjouadi, "A comprehensive survey on impulse and Gaussian denoising filters for digital images," *Signal Processing*, vol. 157, pp. 236–260, Apr. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168418303979

[290] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Upper Saddle River, NJ: Pearson, Aug. 2007.

[291] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7839189/

[292] J. S. Sevak, A. D. Kapadia, J. B. Chavda, A. Shah, and M. Rahevar, "Survey on semantic image segmentation techniques," in *2017 International Conference on Intelligent Sustainable Systems (ICISS)*. Palladam: IEEE, Dec. 2017, pp. 306–313. [Online]. Available: https://ieeexplore.ieee.org/document/8389420/

[293] T. Wen, B. Tong, Y. Liu, T. Pan, Y. Du, Y. Chen, and S. Zhang, "Review of research on the instance segmentation of cell images," *Computer Methods and Programs in Biomedicine*, vol. 227, p. 107211, Dec. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260722005922

[294] S. Hao, Y. Zhou, and Y. Guo, "A Brief Survey on Semantic Segmentation with Deep Learning," *Neurocomputing*, vol. 406, pp. 302–321, Sep. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220305476

[295] W. Gaihua, L. Jinheng, C. Lei, D. Yingying, and Z. Tianlun, "Instance segmentation convolutional neural network based on multi-scale attention mechanism," *PLOS ONE*, vol. 17, no. 1, p. e0263134, Jan. 2022, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0263134

[296] D. Kaur and Y. Kaur, "Various image segmentation techniques: a review," *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 5, pp. 809–814, 2014.

[297] R. Sarma and Y. K. Gupta, "A comparative study of new and existing segmentation techniques," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, p. 012027, Jan. 2021. [Online]. Available: https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012027

[298] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[299] G. Matheron and J. Serra, "The birth of mathematical morphology," *International Symposium on Mathematical Morphology*, 11 2001.

[300] R. M. Haralick, S. R. Sternberg, and X. Zhuang, "Image Analysis Using Mathematical Morphology," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-9, no. 4, pp. 532–550, Jul. 1987, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. [Online]. Available: https://ieeexplore.ieee.org/document/4767941

[301] R. van den Boomgaard and R. van Balen, "Methods for fast morphological image transforms using bitmapped binary images," *CVGIP: Graphical Models and Image Processing*, vol. 54, no. 3, pp. 252–258, May 1992. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/1049965292900553

[302] E. Trucco, "Computer and Robot Vision," *AI Communications*, vol. 8, no. 1, pp. 50–51, 1995.

[303] L. Vincent, "Morphological Area Openings and Closings for Grey-scale Images," in *Shape in Picture*, ser. NATO ASI Series, Y.-L. O, A. Toet, D. Foster, H. J. A. M. Heijmans, and P. Meer, Eds. Berlin, Heidelberg: Springer, 1994, pp. 197–208.

[304] F.-C. Shih and O. Mitchell, "A mathematical morphology approach to Euclidean distance transformation," *IEEE Transactions on Image Processing*, vol. 1, no. 2, pp. 197–204, Apr. 1992, conference Name: IEEE Transactions on Image Processing.

[305] T. Lee, R. Kashyap, and C. Chu, "Building Skeleton Models via 3-D Medial Surface Axis Thinning Algorithms," *CVGIP: Graphical Models and Image Processing*, vol. 56, no. 6, pp. 462–478, Nov. 1994. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S104996528471042X

[306] M. Kerschnitzki, P. Kollmannsberger, M. Burghammer, G. N. Duda, R. Weinkamer, W. Wagermaier, and P. Fratzl, "Architecture of the osteocyte network correlates with bone material quality," *Journal of Bone and Mineral Research: The Official Journal of the American Society for Bone and Mineral Research*, vol. 28, no. 8, pp. 1837–1845, Aug. 2013.

[307] W. Guan and S. Ma, "A list-processing approach to compute Voronoi diagrams and the Euclidean distance transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 7, pp. 757–761, Jul. 1998, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[308] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars, *Computational Geometry: Algorithms and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. [Online]. Available: http://link.springer.com/10.1007/978-3-540-77974-2

[309] A. Pennisi, D. Bloisi, D. Nardi, A. Giampetruzzi, C. Mondino, and A. Facchiano, "Skin Lesion Image Segmentation Using Delaunay Triangulation for Melanoma Detection," *Computerized Medical Imaging and Graphics*, vol. 52, May 2016.

[310] E. Meijering, "Cell Segmentation: 50 Years Down the Road [Life Sciences]," *IEEE Signal Processing Magazine*, vol. 29, no. 5, pp. 140–145, Sep. 2012, conference Name: IEEE Signal Processing Magazine.

[311] G. Danuser, "Computer Vision in Cell Biology," *Cell*, vol. 147, no. 5, pp. 973–978, Nov. 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867411012906

[312] T. Vicar, J. Balvan, J. Jaros, F. Jug, R. Kolar, M. Masarik, and J. Gumulec, "Cell segmentation methods for label-free contrast microscopy: review and comprehensive comparison," *BMC Bioinformatics*, vol. 20, no. 1, p. 360, Jun. 2019. [Online]. Available: https://doi.org/10.1186/s12859-019-2880-8

[313] E. Blum, "The computation of eigenvalues and eigenvector of a completely continuous self-adjoint operator," *Journal of Computer and System Sciences*, vol. 1, no. 4, p. 362370, Dec. 1967. [Online]. Available: http://dx.doi.org/10.1016/S0022-0000(67)80026-6

[314] H. Blum and R. N. Nagel, "Shape description using weighted symmetric axis features," *Pattern Recognition*, vol. 10, no. 3, p. 167180, Jan. 1978. [Online]. Available: http://dx.doi.org/10.1016/0031-3203(78)90025-0

[315] P. K. Saha, G. Borgefors, and G. Sanniti di Baja, "A survey on skeletonization algorithms and their applications," *Pattern Recognition Letters*, vol. 76, p. 312, Jun. 2016. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2015.04.006

[316] S. B. Meyer, F., "The Morphological Approach to Segmentation: The Watershed Transformation," in *Mathematical Morphology in Image Processing*. CRC Press, 1992, pp. 433–481, num Pages: 49.

[317] F. Meyer, "Topographic distance and watershed lines," *Signal Processing*, vol. 38, no. 1, pp. 113–125, Jul. 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0165168494900604

[318] D. S. Fritsch, S. M. Pizer, B. S. Morse, D. H. Eberly, and A. Liu, "The multiscale medial axis and its applications in image registration," *Pattern Recognition Letters*, vol. 15, no. 5, p. 445452, May 1994. [Online]. Available: http://dx.doi.org/10.1016/0167-8655(94)90135-X

[319] F. P. Preparata, "The medial axis of a simple polygon," in *Mathematical Foundations of Computer Science 1977*. Springer Berlin Heidelberg, 1977, p. 443450. [Online]. Available: http://dx.doi.org/10.1007/3-540-08353-7_166

[320] S. Wang, A. Rosenfeld, and A. Y. Wu, "A medial axis transformation for grayscale pictures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-4, no. 4, p. 419421, Jul. 1982.

[321] E. Meijering, "Neuron tracing in perspective," *Cytometry Part A*, vol. 77A, no. 7, p. 693704, Jun. 2010. [Online]. Available: http://dx.doi.org/10.1002/cyto.a.20895

[322] L. Acciai, P. Soda, and G. Iannello, "Automated neuron tracing methods: An updated account," *Neuroinformatics*, vol. 14, no. 4, p. 353367, Jul. 2016. [Online]. Available: http://dx.doi.org/10.1007/s12021-016-9310-0

[323] Y. Liu, G. Wang, G. A. Ascoli, J. Zhou, and L. Liu, "Neuron tracing from light microscopy images: automation, deep learning and bench testing," *Bioinformatics*, vol. 38, no. 24, p. 53295339, Oct. 2022. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btac712

[324] K. K. Delibasis, A. I. Kechriniotis, C. Tsonos, and N. Assimakis, "Automatic model-based tracing algorithm for vessel segmentation and diameter estimation," *Computer Methods and Programs in Biomedicine*, vol. 100, no. 2, p. 108122, Nov. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.cmpb.2010.03.004

[325] V. Mahadevan, H. Narasimha-Iyer, B. Roysam, and H. Tanenbaum, "Robust model-based vasculature detection in noisy biomedical images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 8, no. 3, p. 360376, Sep. 2004. [Online]. Available: http://dx.doi.org/10.1109/titb.2004.834410

[326] F. J. Valiente-Soriano, M. Salinas-Navarro, J. Di Pierdomenico, D. García-Ayuso, F. Lucas-Ruiz, I. Pinilla, N. Cuenca, M. Vidal-Sanz, M. P. Villegas-Pérez, and M. Agudo-Barriuso, "Tracing the retina to analyze the integrity and phagocytic capacity of the retinal pigment epithelium," *Scientific Reports*, vol. 10, no. 1, Apr. 2020. [Online]. Available: http://dx.doi.org/10.1038/s41598-020-64131-z

[327] M. Koziski, A. Mosinska, M. Salzmann, and P. Fua, "Tracing in 2d to reduce the annotation effort for 3d deep delineation of linear structures," *Medical Image Analysis*, vol. 60, p. 101590, Feb. 2020. [Online]. Available: http://dx.doi.org/10.1016/j.media.2019.101590

[328] Y. Cheng, X. Hu, J. Wang, Y. Wang, and S. Tamura, "Accurate vessel segmentation with constrained b-snake," *IEEE Transactions on Image Processing*, vol. 24, no. 8, p. 24402455, Aug. 2015. [Online]. Available: http://dx.doi.org/10.1109/TIP.2015.2417683

[329] W. Wei, Q. Zhang, S. G. Rayner, W. Qin, Y. Cheng, F. Wang, Y. Zheng, and R. K. Wang, "Automated vessel diameter quantification and vessel tracing for ¡scp¿oct¡/scp¿ angiography," *Journal of Biophotonics*, vol. 13, no. 12, Sep. 2020. [Online]. Available: http://dx.doi.org/10.1002/jbio.202000248

[330] M. A. Abdul-Karim, K. Al-Kofahi, E. B. Brown, R. K. Jain, and B. Roysam, "Automated tracing and change analysis of angiogenic vasculature from in vivo multiphoton confocal image time series," *Microvasc Res*, vol. 66, no. 2, pp. 113–25, Sep. 2003.

[331] P. R. Barber, B. Vojnovic, S. M. Ameer Beg, R. J. Hodgkiss, G. M. Tozer, and J. Wilson, "Semi automated software for the threedimensional delineation of complex vascular networks," *Journal of Microscopy*, vol. 211, no. 1, pp. 54–62, Jul. 2003. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1046/j.1365-2818.2003.01205.x

[332] C. C. Reyes-Aldasoro, M. A. Bjrndahl, S. Akerman, J. Ibrahim, M. K. Griffiths, and G. M. Tozer, "Online chromatic and scale-space microvessel-tracing analysis for transmitted light optical images," *Microvascular research*, vol. 84, no. 3, pp. 330–339, Nov. 2012.

[333] J. Wu, S. L. Voytik-Harbin, D. L. Filmer, C. M. Hoffman, B. Yuan, C.-S. Chiang, J. Sturgis, and J. P. Robinson, "Modeling ECM fiber formation: structure information extracted by analysis of 2D and 3D image sets," in *Three-Dimensional and Multidimensional Microscopy: Image Acquisition*

and Processing IX, vol. 4621. SPIE, May 2002, pp. 52–56. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/4621/0000/Modeling-ECM-fiber-formation--structure-information-extracted-by-analysis/10.1117/12.467831.full

[334] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv:1505.04597 [cs]*, May 2015, arXiv: 1505.04597. [Online]. Available: http://arxiv.org/abs/1505.04597

[335] C. Steger, "An unbiased detector of curvilinear structures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 2, pp. 113–125, Feb. 1998, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/659930

[336] D. J. Barry, G. A. Williams, and C. Chan, "Automated analysis of filamentous microbial morphology with AnaMorf," *Biotechnology Progress*, vol. 31, no. 3, pp. 849–852, 2015, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/btpr.2087. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/btpr.2087

[337] E. Cands, L. Demanet, D. Donoho, and L. Ying, "Fast discrete curvelet transforms," *Multiscale Modeling & Simulation*, vol. 5, no. 3, p. 861899, Jan. 2006.

[338] K. Lai, N. Twine, A. OBrien, Y. Guo, and D. Bauer, "Artificial Intelligence and Machine Learning in Bioinformatics," in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schnbach, Eds. Oxford: Academic Press, Jan. 2019, pp. 272–286. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128096338203257

[339] Q. An, S. Rahman, J. Zhou, and J. J. Kang, "A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges," *Sensors (Basel, Switzerland)*, vol. 23, no. 9, p. 4178, Apr. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10180678/

[340] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 2015, arXiv:1409.1556 [cs]. [Online]. Available: http://arxiv.org/abs/1409.1556

[341] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 2021, arXiv:2010.11929. [Online]. Available: http://arxiv.org/abs/2010.11929

[342] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Aug. 2023, arXiv:1706.03762 [cs]. [Online]. Available: http://arxiv.org/abs/1706.03762

[343] H. Blockeel, L. Devos, B. Frénay, G. Nanfack, and S. Nijssen, "Decision trees: from efficient prediction to responsible AI," *Frontiers in Artificial Intelligence*, vol. 6, p. 1124553, Jul. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10411911/

[344] T. Daniya, M. Geetha, and K. Suresh Kumar, "CLASSIFICATION AND REGRESSION TREES WITH GINI INDEX," *Advances in Mathematics: Scientific Journal*, vol. 9, no. 10, pp. 8237–8247, Sep. 2020. [Online]. Available: https://www.research-publication.com/amsj/all-issues/vol-09/iss-10

[345] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[346] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995. [Online]. Available: https://doi.org/10.1007/BF00994018

[347] S. Sperandei, "Understanding logistic regression analysis," *Biochemia Medica*, vol. 24, no. 1, pp. 12–18, Feb. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3936971/

[348] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

[349] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967. [Online]. Available: http://ieeexplore.ieee.org/document/1053964/

[350] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain." *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958. [Online]. Available: https://doi.apa.org/doi/10.1037/h0042519

[351] B. Preim and C. Botha, "Chapter 4 - Image Analysis for Medical Visualization," in *Visual Computing for Medicine (Second Edition)*, B. Preim and C. Botha, Eds. Boston: Morgan Kaufmann, Jan. 2014, pp. 111–175. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780124158733000043

[352] J. Rogowska, "5 - Overview and Fundamentals of Medical Image Segmentation," in *Handbook of Medical Imaging*, ser. Biomedical Engineering, I. N. Bankman, Ed. San Diego: Academic Press, Jan. 2000, pp. 69–85. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780120777907500096

[353] D. Bailey, "An Efficient Euclidean Distance Transform," vol. 3322, Dec. 2004, pp. 394–408.

[354] D. Xu and H. Li, "Euclidean Distance Transform of Digital Images in Arbitrary Dimensions," in *Advances in Multimedia Information Processing - PCM 2006*, ser. Lecture Notes in Computer Science, Y. Zhuang, S.-Q. Yang, Y. Rui, and Q. He, Eds. Berlin, Heidelberg: Springer, 2006, pp. 72–79.

[355] R. Haddad and A. Akansu, "A class of fast Gaussian binomial filters for speech and image processing," *IEEE Transactions on Signal Processing*, vol. 39, no. 3, pp. 723–727, Mar. 1991. [Online]. Available: http://ieeexplore.ieee.org/document/80892/

[356] P. Soille, *Morphological Image Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999. [Online]. Available: http://link.springer.com/10.1007/978-3-662-03939-7

[357] Y. Arafat, C. Cuesta-Apausa, E. Castellano, and C. C. Reyes-Aldasoro, "Fibre tracing in biomedical images: An objective comparison between seven algorithms," *submitted to biorxiv*, vol. submitted to biorxiv, Apr. 2024.

[358] O. Maiques, M. C. Sallan, R. Laddach, P. Pandya, A. Varela, E. Crosas-Molist, J. Barcelo, O. Courbot, Y. Liu, V. Graziani, Y. Arafat, J. Sewell, I. Rodriguez-Hernandez, B. Fanshawe, Y. Jung-Garcia, P. R. Imbert, E. M. Grasset, J. Albrengues, M. Santacana, A. Macià, J. Tarragona, X. Matias-Guiu, R. M. Marti, S. Tsoka, C. Gaggioli, J. L. Orgaz, G. O. Fruhwirth, F. Wallberg, K. Betteridge, C. C. Reyes-Aldasoro, S. Haider, A. Braun, S. N. Karagiannis, A. Elosegui-Artola, and V. Sanz-Moreno, "Matrix mechano-sensing at the invasive front induces a cytoskeletal and transcriptional memory supporting metastasis," *Nature Communications*, vol. 16, no. 1, Feb. 2025. [Online]. Available: http://dx.doi.org/10.1038/s41467-025-56299-7

[359] D. Fukumura, D. G. Duda, L. L. Munn, and R. K. Jain, "Tumor microvasculature and microenvironment: novel insights through intravital imaging in pre-clinical models," *Microcirculation*, vol. 17, no. 3, pp. 206–25, Apr. 2010.

[360] P. Lu, K. Takai, V. M. Weaver, and Z. Werb, "Extracellular matrix degradation and remodeling in development and disease," *Cold Spring Harbor Perspectives in Biology*, vol. 3, no. 12, p. a005058, Dec. 2011.

[361] A. D. Theocharis, S. S. Skandalis, C. Gialeli, and N. K. Karamanos, "Extracellular matrix structure," *Advanced Drug Delivery Reviews*, vol. 97, pp. 4–27, Feb. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169409X15002574

[362] E. Almici, M. Arshakyan, J. L. Carrasco, A. Martínez, J. Ramírez, A. B. Enguita, E. Monsó, J. Montero, J. Samitier, and J. Alcaraz, "Quantitative Image Analysis of Fibrillar Collagens Reveals Novel Diagnostic and Prognostic Biomarkers and Histotype-Dependent Aberrant Mechanobiology in Lung Cancer," *Modern Pathology*, vol. 36, no. 7, p. 100155, Jul. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0893395223000601

[363] Z. Dinkel, A. Baker, A. Akins, K. L. King, M. Harrington, D. Kunkel, Z. Gao, T. Ye, and H. Dunn, "Collagen architecture in triple negative breast cancer," *PLOS One*, vol. 20, no. 5, p. e0324655, May 2025. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0324655

[364] X. Fu, E. Sahai, and A. Wilkins, "Application of digital pathologybased advanced analytics of tumour microenvironment organisation to predict prognosis and therapeutic response," *The Journal of Pathology*, vol. 260, no. 5, pp. 578–591, Aug. 2023. [Online]. Available: https://pathsocjournals.onlinelibrary.wiley.com/doi/10.1002/path.6153

[365] T. Ohshiro, A. Asai, M. Konno, M. Ohkawa, Y. Komoto, K. Ofusa, H. Ishii, and M. Taniguchi, "Direct observation of DNA alterations induced by a DNA disruptor," *Scientific Reports*, vol. 12, no. 1, p. 6945, Apr. 2022, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-022-10725-8

[366] J. Graham, M. Raghunath, and V. Vogel, "Fibrillar fibronectin plays a key role as nucleator of collagen I polymerization during macromolecular crowding-enhanced matrix assembly," *Biomaterials Science*, vol. 7, no. 11, pp. 4519–4535, Oct. 2019, publisher: The Royal Society of Chemistry. [Online]. Available: https://pubs.rsc.org/en/content/articlelanding/2019/bm/c9bm00868c

[367] J. A. Eble and S. Niland, "The extracellular matrix in tumor progression and metastasis," *Clinical & Experimental Metastasis*, vol. 36, no. 3, pp. 171–198, Jun. 2019. [Online]. Available: https://doi.org/10.1007/s10585-019-09966-1

[368] A. Kaur, B. L. Ecker, S. M. Douglass, C. H. Kugel, III, M. R. Webster, F. V. Almeida, R. Somasundaram, J. Hayden, E. Ban, H. Ahmadzadeh, J. Franco-Barraza, N. Shah, I. A. Mellis, F. Keeney, A. Kossenkov, H.-Y. Tang, X. Yin, Q. Liu, X. Xu, M. Fane, P. Brafford, M. Herlyn, D. W. Speicher, J. A. Wargo, M. T. Tetzlaff, L. E. Haydu, A. Raj, V. Shenoy, E. Cukierman, and A. T. Weeraratna, "Remodeling of the Collagen Matrix in Aging Skin Promotes Melanoma Metastasis and Affects Immune Cell Motility," *Cancer Discovery*, vol. 9, no. 1, pp. 64–81, Jan. 2019. [Online]. Available: https://doi.org/10.1158/2159-8290.CD-18-0193

[369] B. L. Ecker, A. Kaur, S. M. Douglass, M. R. Webster, F. V. Almeida, G. E. Marino, A. J. Sinnamon, M. G. Neuwirth, G. M. Alicea, A. Ndoye, M. Fane, X. Xu, M. S. Sim, G. B. Deutsch, M. B. Faries, G. C. Karakousis, and A. T. Weeraratna, "Age-Related Changes in HAPLN1 Increase Lymphatic Permeability and Affect Routes of Melanoma Metastasis," *Cancer Discovery*, vol. 9, no. 1, pp. 82–95, Jan. 2019.

[370] B. Piersma, M.-K. Hayward, and V. M. Weaver, "Fibrosis and cancer: A strained relationship," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1873, no. 2, p. 188356, Apr. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304419X19301702

[371] M. W. Pickup, J. K. Mouw, and V. M. Weaver, "The extracellular matrix modulates the hallmarks of cancer," *EMBO reports*, vol. 15, no. 12, pp. 1243–1253, Dec. 2014, num Pages: 1253 Publisher: John Wiley & Sons, Ltd. [Online]. Available: https://www.embopress.org/doi/full/10.15252/embr.201439246

[372] P. P. Provenzano, K. W. Eliceiri, J. M. Campbell, D. R. Inman, J. G. White, and P. J. Keely, "Collagen reorganization at the tumor-stromal interface facilitates local invasion," *BMC Medicine*, vol. 4, no. 1, p. 38, Dec. 2006. [Online]. Available: https://doi.org/10.1186/1741-7015-4-38

[373] P. P. Provenzano, K. W. Eliceiri, L. Yan, A. Ada-Nguema, M. W. Conklin, D. R. Inman, and P. J. Keely, "Nonlinear Optical Imaging of Cellular Processes in Breast Cancer," *Microscopy and Microanalysis*, vol. 14, no. 6, pp. 532–548, Dec. 2008. [Online]. Available: https://academic.oup.com/mam/article/14/6/532/6920118

[374] P. P. Provenzano, D. R. Inman, K. W. Eliceiri, S. M. Trier, and P. J. Keely, "Contact Guidance Mediated Three-Dimensional Cell Migration is Regulated by Rho/ROCK-Dependent Matrix Reorganization," *Biophysical Journal*, vol. 95, no. 11, pp. 5374–5384, Dec. 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0006349508789616

[375] A. Ray, O. Lee, Z. Win, R. M. Edwards, P. W. Alford, D.-H. Kim, and P. P. Provenzano, "Anisotropic forces from spatially constrained focal adhesions mediate contact guidance directed cell migration," *Nature Communications*, vol. 8, p. 14923, Apr. 2017.

[376] I. Acerbi, L. Cassereau, I. Dean, Q. Shi, A. Au, C. Park, Y. Y. Chen, J. Liphardt, E. S. Hwang, and V. M. Weaver, "Human breast cancer invasion and aggression correlates with ECM stiffening and immune cell infiltration," *Integrative Biology: Quantitative Biosciences from Nano to Macro*, vol. 7, no. 10, pp. 1120–1134, Oct. 2015.

[377] Y.-J. Liu, M. Le Berre, F. Lautenschlaeger, P. Maiuri, A. Callan-Jones, M. Heuzé, T. Takaki, R. Voituriez, and M. Piel, "Confinement and low adhesion induce fast amoeboid migration of slow mesenchymal cells," *Cell*, vol. 160, no. 4, pp. 659–672, Feb. 2015.

[378] Z. Miskolczi, M. P. Smith, E. J. Rowling, J. Ferguson, J. Barriuso, and C. Wellbrock, "Collagen abundance controls melanoma phenotypes through lineage-specific microenvironment sensing," *Oncogene*, vol. 37, no. 23, pp. 3166–3182, Jun. 2018, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41388-018-0209-0

[379] O. M. Pearce, R. M. Delaine-Smith, E. Maniati, S. Nichols, J. Wang, S. Böhm, V. Rajeeve, D. Ullah, P. Chakravarty, R. R. Jones, A. Montfort, T. Dowe, J. Gribben, J. L. Jones, H. M. Kocher, J. S. Serody, B. G. Vincent, J. Connelly, J. D. Brenton, C. Chelala, P. R. Cutillas, M. Lockley, C. Bessant, M. M. Knight, and F. R. Balkwill,

"Deconstruction of a Metastatic Tumor Microenvironment Reveals a Common Matrix Response in Human Cancers," *Cancer Discovery*, vol. 8, no. 3, pp. 304–319, Mar. 2018. [Online]. Available: https://doi.org/10.1158/2159-8290.CD-17-0284

[380] D. Park, E. Wershof, S. Boeing, A. Labernadie, R. P. Jenkins, S. George, X. Trepat, P. A. Bates, and E. Sahai, "Extracellular matrix anisotropy is determined by TFAP2C-dependent regulation of cell collisions," *Nature materials*, vol. 19, no. 2, pp. 227–238, Feb. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6989216/

[381] A. Shellard and R. Mayor, "Collective durotaxis along a self-generated stiffness gradient in vivo," *Nature*, vol. 600, no. 7890, pp. 690–694, Dec. 2021, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41586-021-04210-x

[382] M. S. Samuel, J. I. Lopez, E. J. McGhee, D. R. Croft, D. Strachan, P. Timpson, J. Munro, E. Schrder, J. Zhou, V. G. Brunton, N. Barker, H. Clevers, O. J. Sansom, K. I. Anderson, V. M. Weaver, and M. F. Olson, "Actomyosin-mediated cellular tension drives increased tissue stiffness and catenin activation to induce epidermal hyperplasia and tumor growth," *Cancer Cell*, vol. 19, no. 6, pp. 776–791, Jun. 2011.

[383] S. I. Fraley, P.-h. Wu, L. He, Y. Feng, R. Krisnamurthy, G. D. Longmore, and D. Wirtz, "Three-dimensional matrix fiber alignment modulates cell migration and MT1-MMP utility by spatially and temporally directing protrusions," *Scientific Reports*, vol. 5, no. 1, p. 14580, Oct. 2015, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/srep14580

[384] D. Idossa, M. Borrero, and A. Blaes, "ERBB2-Low (Also Known as HER2-Low) Breast Cancer," *JAMA Oncology*, vol. 9, no. 4, p. 576, Apr. 2023. [Online]. Available: https://jamanetwork.com/journals/jamaoncology/fullarticle/2801797

Appendices

Appendix A

RBD and WT Morphological Metrics:

	Num Edges	Rel Gap Area
Mutant_1	**1503**	**0.5461**
Mutant_2	**1677**	**0.5215**
Mutant_3	**1708**	**0.5542**
Mutant_4	**1536**	**0.5353**
Mutant_5	**1851**	**0.5385**
WT_1	**2084**	**0.5193**
WT_2	**2129**	**0.5095**
WT_3	**2363**	**0.4926**
WT_4	**2068**	**0.5278**
WT_5	**1964**	**0.5188**

Table A.1: Morphological analysis on the five sets of mutant RBD and WT images. Num Fibres is number of edges detected in an image. Rel Gap Area is the relative gap area size.

Appendix B

Significant fibre morphological features:

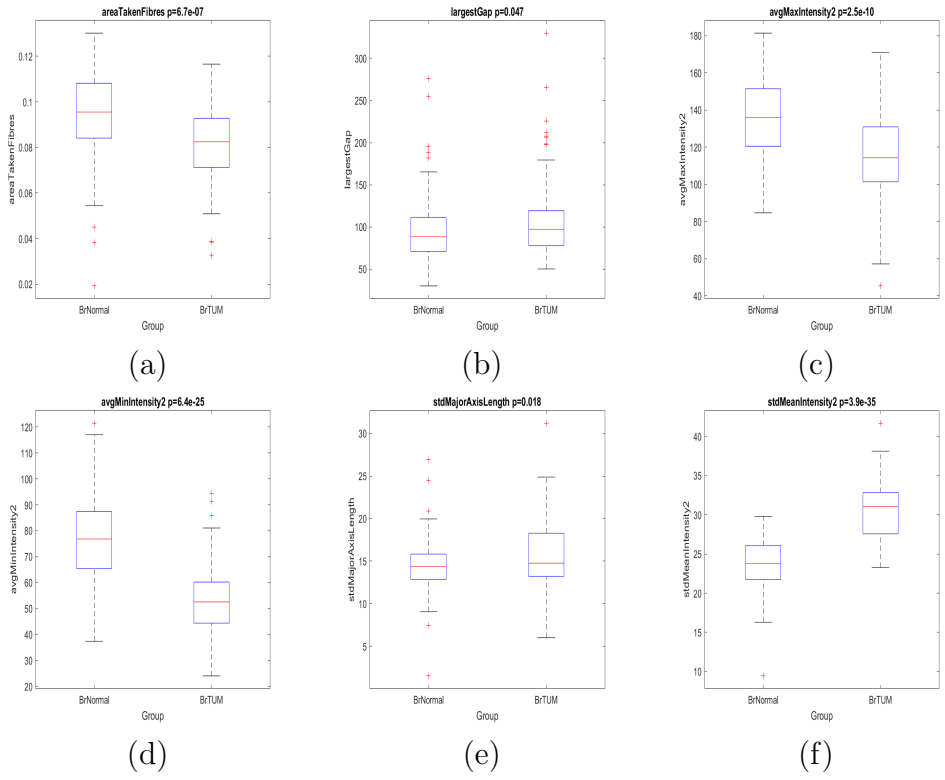


Figure B.0.1: Boxplots of morphological metrics that showed statistical significance when comparing the Normal to Tumour.(a) The area taken up by fibres had a $p-value$ of $6.7e-07$. (b) The largest gap area had a $p-value$ of 0.047. (c) The average maximum intensity had a $p-value$ of $2.5e^{-}10$. (d) The average minimum intensity had a $p-value$ of $2.4e^{-}09$. (e) The std of the major axis length had a $p-value$ of 0.018 . (f) The std of the mean intensity had a $p-value$ of $3.9e^{-}35$.

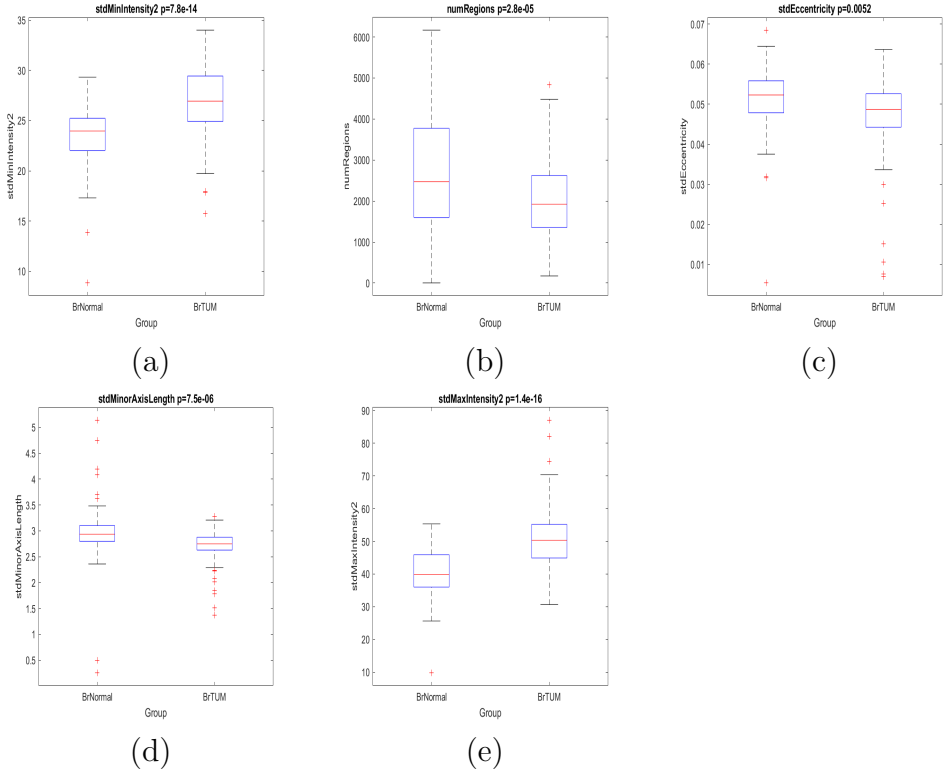


Figure B.0.2: Boxplots of morphological metrics that showed statistical significance when comparing the Normal to Tumour.(a) The std of the min intensity had a $p-value$ of $2.4e^{-}09$. (b) The number of regions had a $p-value$ of $2.8e^{-}05$. (c) The std of the eccentricity had a $p-value$ of 0.0052. (d) The std of the minor axis length has a $p-value$ of $7.5e-06$. (e) The std of the maximum intensity had a $p-value$ of $2.4e^{-}09$.

The area taken up by fibres in the ECM (Fig. B.0.1a) has decreased when looking at the tumour in comparison to the normal, the normal has a wider range than the tumour population as well as a higher median.

The largest gap also (Fig. B.0.1b) increased in the tumour samples. The median of the populations is comparable. The normal population had a larger range than the tumour population. The tumour was positively skewed, therefore, there were more large gaps below the median confined to a small range.

The average maximum intensity (Fig. B.0.1c) is higher in normal than the tumour but the tumour has a wider range than that of the normal. They both have a normal distribution but the median of the normal is higher than that of the tumour.

The average minimum intensity (Fig. B.0.1d) of the tumour fibres has also decreased in comparison to the normal. The range of the normal was bigger than the tumour. The median value of the normal was higher than that of the tumour. The highest value of the tumour is just above the median of the normal. These two metrics and the average mean intensity work together cohesively, where the normal has a larger maximum intensity and a larger minimum intensity and therefore has a larger mean intensity.

The standard deviation of the tumour fibre lengths (Fig. B.0.1e) is also higher than that of the normal fibres. The tumour possesses a larger range than that of the normal, but the median value is comparable.

The mean intensity (Fig. B.0.1f) of fibres of the tumour also deviates from the mean more than the normal fibres. Both populations have a similar range while the median for the tumour is higher than that of the normal.

Similarly, the minimum intensity (Fig. B.0.2a) of fibres of the tumour also deviates from the mean more than the normal fibres. The tumour population has a wider range of deviation from the mean compared to the normal, where the median of the tumour population deviation is also higher than that of the normal.

The number of bright regions (Fig. B.0.2b) is also lower in the tumour populations than that of the normal. Regions are the areas where they exhibit an intensity value 1 degree higher than the standard deviation of the mean intensities. Therefore, they are the bright regions. The normal population exhibited a wider range and a higher median than the tumour. Although the lowest number of regions in a tumour core was higher than the lowest number of regions in a normal core.

The standard deviation of the eccentricity (Fig. B.0.2c) was lower in the tumour. Both populations exhibited similar ranges but the median of the normal was higher than that of the tumour.

The std of the minor axis length (Fig. B.0.2d) is lower in the tumour. The tumour has a small range of values therefore it is easy to predict the deviation value of the width of a tumour fibre sample. The median of both populations was comparable.

The maximum intensity (Fig. B.0.2e) of fibres of the tumour deviates from the mean of the maximum intensity more than the normal, where it also has a wider range of deviations, and the median of those values is higher than the normal. The standard deviation of the normal maximum intensity is confined to a small range.

Appendix C

Significant cell nuclei morphological features:

The deviation of the eccentricity (Fig. C.0.1a) values from the mean in the tumour population was lower than that of the normal population. The range of the tumour deviations was smaller than that of the normal. Both populations follow a normal distribution.

The number of bright regions (Fig. C.0.1b) of the cell nuclei is higher in tumour. The tumour population was positively skewed. The larger number of regions and number of cell nuclei that are larger and wider than those found in the normal population coincides with the smaller gaps present in the tumour as well in comparison to the normal.

Similarly, the average minimum intensity (Fig. C.0.1c) of nuclei in the tumour population was higher. Both populations exhibited a comparable range. The median of tumour was higher than the median for normal. The tumour population was positively skewed while the normal population was negatively skewed.

The average maximum intensity (Fig. C.0.1d) of nuclei in tumour was also higher, the tumour population had a smaller range than the normal population. The median values of normal and tumour are close in value. The tumour population follows a normal distribution while the normal intensity is negatively skewed.

There were more nuclei in the tumour population where their length deviated (Fig. C.0.1e) from the mean in comparison to the normal cell nuclei. The tumour and normal populations had similar deviation ranges. While the median of the tumour population was higher than that of the normal population.

Similarly, the std of the minor axis length (Fig. C.0.1f) was also higher in the cell nuclei of the tumour population. Both populations have a comparable range. The median tumour deviation is higher than normal.

The std of the minimum intensity (Fig. C.0.1g) of the tumour nuclei was higher than the normal population. The range of the normal population is almost double that of the tumour. Both populations are normally distributed with the median tumour population being higher than that of the normal population.

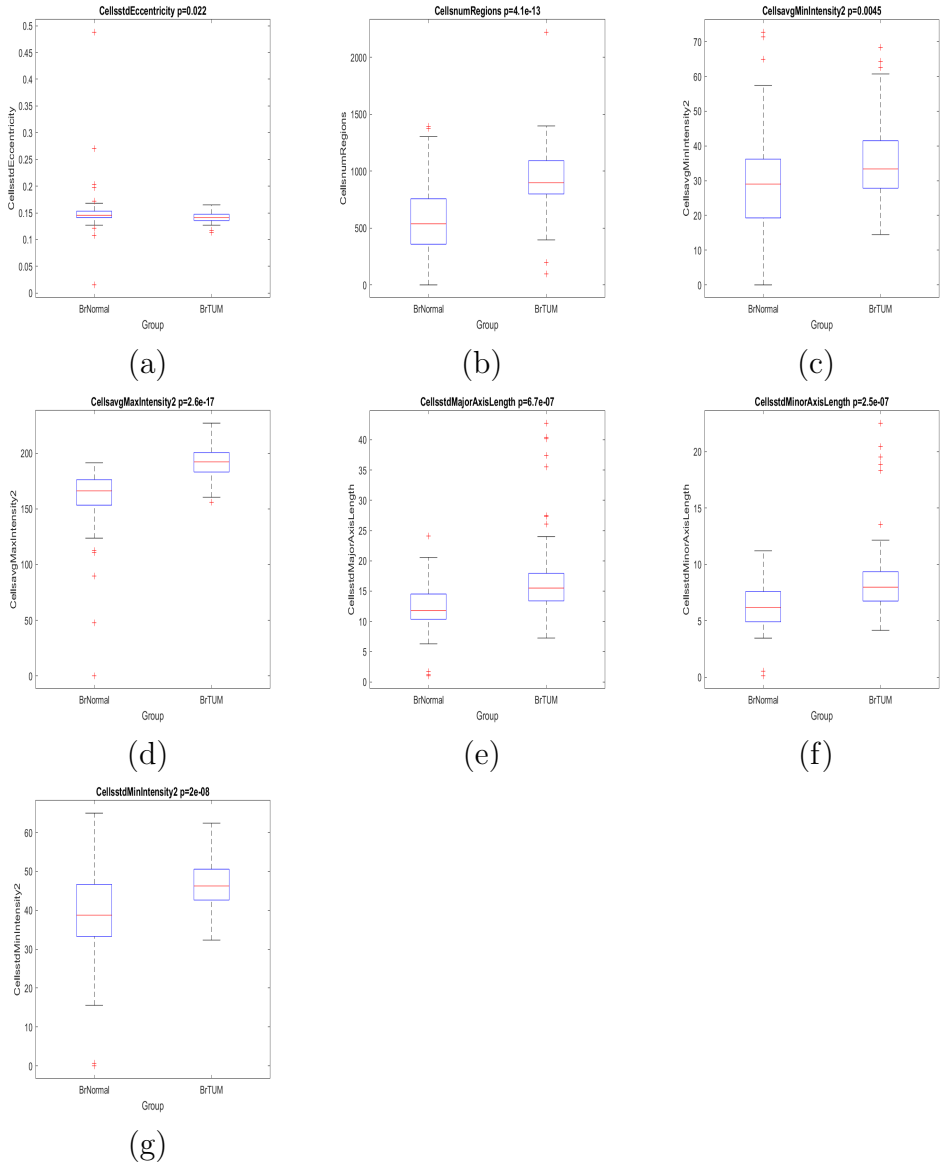


Figure C.0.1: Boxplots of increasing cell nuclei metrics when comparing Normal and Tumour cell nuclei. (a) The std of eccentricity had a $p-value$ of 0.022. (b) The number of regions had a $p-value$ of $4.1e-13$. (c) The average minimum intensity had a $p-value$ of 0.0045. (d) The average maximum intensity had a $p-value$ of $2.6e-17$. (e) The std of the major axis length had a $p-value$ of $6.7e-07$. (f) The std of the minor axis length had a $p-value$ of $2.5e-07$. (g) The std of the minimum intensity had a $p-value$ of $2e-08$.

Appendix D

Previous publications:

PLOS ONE

RESEARCH ARTICLE

Fibre tracing in biomedical images: An objective comparison between seven algorithms

Youssef Arafat[1*], **Cristina Cuesta-Apausa**[2], **Esther Castellano**[2], **Constantino Carlos Reyes-Aldasoro**[1,3]

1 Department of Computer Science, School of Science and Technology, City St George's, University of London, London, United Kingdom, **2** Tumour-Stroma Signalling Lab, Universidad de Salamanca, Salamanca, Spain, **3** Integrated Pathology Unit, Division of Molecular Pathology, The Institute of Cancer Research, Sutton, United Kingdom

* [redacted]

Check for updates

OPEN ACCESS

Citation: Arafat Y, Cuesta-Apausa C, Castellano E, Reyes-Aldasoro C C. (2025) Fibre tracing in biomedical images: An objective comparison between seven algorithms. PLoS ONE 20(4): e0320006. https://doi.org/10.1371/journal.pone.0320006

Editor: Xiaohui Zhang, Bayer Crop Science United States: Bayer CropScience LP, UNITED STATES OF AMERICA

Received: April 25, 2024

Accepted: February 01, 2025

Published: April 10, 2025

Copyright: © 2025 Arafat et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: All data and code are publicly available. - Second Harmonic Generation (SHG) images of tumour bearing mouse mammary glands (Bredfeldt JS, Liu Y, Pehlke CA, Conklin MW, Szulczewski JM, Inman DR, et al. Computational segmentation of collagen fibers from second-harmonic generation images of breast cancer. Journal of Biomedical Optics. 2014;19(1):016007. https://doi.org/10.1117/1.JBO.19.1.016007 and, Liu Y, et al. Fibrillar Collagen Quantification With

Abstract

Obtaining the traces and the characteristics of elongated structures is an important task in computer vision pipelines. In biomedical applications, the analysis of traces of vasculature, nerves or fibres of the extracellular matrix can help characterise processes like angiogenesis or the effect of a certain treatment. This paper presents an objective comparison of six existing methodologies (Edge detection, CT Fire, Scale Space, Twombli, U-Net and Graph Based) and one novel approach called *Trace Ridges* to trace biomedical images with fibre-like structures. Trace Ridges is a fully automatic and fast algorithm that combines a series of image-processing algorithms including filtering, watershed transform and edge detection to obtain an accurate delineation of the fibre-like structures in a rapid time. To compare the algorithms, four biomedical data sets with vastly distinctive characteristics were selected. Ground truth was obtained by manual delineation of the fibre-like structures. Three pre-processing filtering options were used as a first step: no filtering, Gaussian low-pass and DnCnn, a deep-learning filtering. Three distance error metrics (total, average and maximum distance from the obtained traces to the ground truth) and processing time were calculated. It was observed that no single algorithm outperformed the others in all metrics. For the total distance error, which was considered the most significative, Trace Ridges ranked first, followed by Graph Based, U-Net, Twombli, Scale Space, CT Fire and Edge Detection. In terms of speed, Trace Ridges ranked second, only slightly slower than Edge Detection. Code is freely available at *github.com/youssefarafat/Trace_Ridges*.

Introduction

The morphological features of cells and their surrounding environment have been studied with interest in cancer and other conditions [1]. However, it is now recognised that the microenvironment that surrounds these cells has considerable influence [2]. The extracellular matrix (ECM) plays a key role in the development of cancer [3–6] and other diseases [7].

Curvelet Transform Based Computational Methods. Frontiers in Bioengineering and Biotechnology. 2020;8.) - Breast Cancer Biopsy (BCB) images of collagen (Wershof E, Park D, Barry DJ, Jenkins RP, Rullan A, Wilkins A, et al. A FIJI macro for quantifying pattern in extracellular matrix. Life Science Alliance. 2021;4(3). https://doi.org/10.26508/lsa.202000880.) - Disease mimicking ECM (Grapa AI, Efthymiou G, Obberghen-Schilling EV, Blanc-Féraud L, Descombes X. A spatial statistical framework for the parametric study of fiber networks: Application to fibronectin deposition by normal and activated fibroblasts. Biological Imaging. 2023;3:e25. https://doi.org/10.1017/S2633903X23000247.) - Fluorescently labelled Fibronectin images and Trace Ridges algorithm code available at https://github.com/youssefarafat/Trace_Ridges.

Funding: The author(s) received no specific funding for this work.

Competing interests: No competing interests.

The ECM provides essential structural support [8] to the cellular constituents of tissue. Every organ has a different ECM structure that serves its purpose. The ECM constantly changes its shape and remodels by a process known as ECM proteinases [9,10]. Despite this, the extracellular matrix (ECM) has received less attention [11] than other factors of the microenvironment like macrophages, soluble factors, or vasculature. This could be due to the complicated 3D network of the ECM [12,13], which is formed by many elements such as enzymes, glycoproteins, and collagen.

Whilst visual observation of the extracellular matrix structure can reveal interesting characteristics, it is far too difficult to compare two populations by just observing images. Structural characteristics of the ECM such as the length, distribution, shapes and orientation of the fibre-like structure of these glycoproteins could provide valuable information about the microenvironment or treatments if these were to be objectively analysed. Quantification of the fibres and their characteristics is required and for this purpose, a robust tracing algorithm that identifies individual fibre-like structures is necessary.

There have been many manual, semi-automatic and automatic methodologies proposed to delineate and trace elongated structures some of which are based on algorithms that can be considered more general as they can be applied in a context other than tracing, for instance, skeletonisation [14–16], watershed transform [17,18], medial axis transform [19–21], Scale Space [22,23] and even Edge Detection [24]. However, many of these methodologies have been proposed for tracing in specific applications, which may imply that these have been fine-tuned to the characteristics of the datasets being analysed, e.g., neuron tracing [25–27], retinal images [28–30], angiography [31–33], intravital vasculature observed with confocal, fluorescence or light microscopy [34–36], and extracellular matrix [37,38]. In addition, deep learning architectures like the well-known U-Net [39] have been widely used in image analysis and provided excellent results without the need of hand-crafted features or algorithm. Instead, a combination of pairs of data and labels can be used to train an architecture to perform a specific task, which for the purposes of this paper is the segmentation of fibre-like structures.

In this paper an objective comparison of algorithms that perform tracing of fibre-like structures is presented. Four types of datasets with vastly distinctive characteristics were selected to provide a wide variety of conditions of noise and strength of the fibre structures. The images were hand-delineated to create a ground truth for objective comparisons and to provide training data for the deep learning architecture. Five existing algorithms that have been previously used for similar applications were selected: Edge detection [24], CT Fire[40], Scale Space [22,23], Twombli [38], and Graph Based [41,42]. A U-Net [39] architecture was trained with pairs of patches of data and labels. In addition, a simple, yet quick and effective image processing-based tracing algorithm called *Trace Ridges* is proposed and compared with the other algorithms.

The main contributions of this work are the following:

- The objective comparison of seven fibre tracing methods. Each algorithm was tested in four images with vastly different characteristics. To analyse the effect of noise, three filtering approaches were applied as a pre-processing step for each of the algorithms. Three error metrics and computational processing time were extracted from each of the previous combinations.
- A new tracing pipeline called Trace Ridges was proposed. Trace ridges combines filtering, Watershed transforms, Edge detection and mathematical morphology to trace ridges in an image with fibre-like structures.

Materials and methods

Data sets

Four biological data sets with vastly distinctive characteristics were selected to compare the tracing algorithms (Fig 1). Three of the four data sets were selected as they had been used by the previously published algorithms and were freely available. Details of the data sets have been published previously, but a brief description for each will follow.

- Second Harmonic Generation (SHG) images of tumour bearing mouse mammary glands were published with the CT Fire algorithm [40,43] (Fig 1a). To obtain the images, *in vivo* acquisition was performed through a glass imaging window that was placed at proximity to palpable tumours inside the mammary glands of live 8 weeks old PyMT mice. The excitation wavelength of the SHG images was 890 nm while the

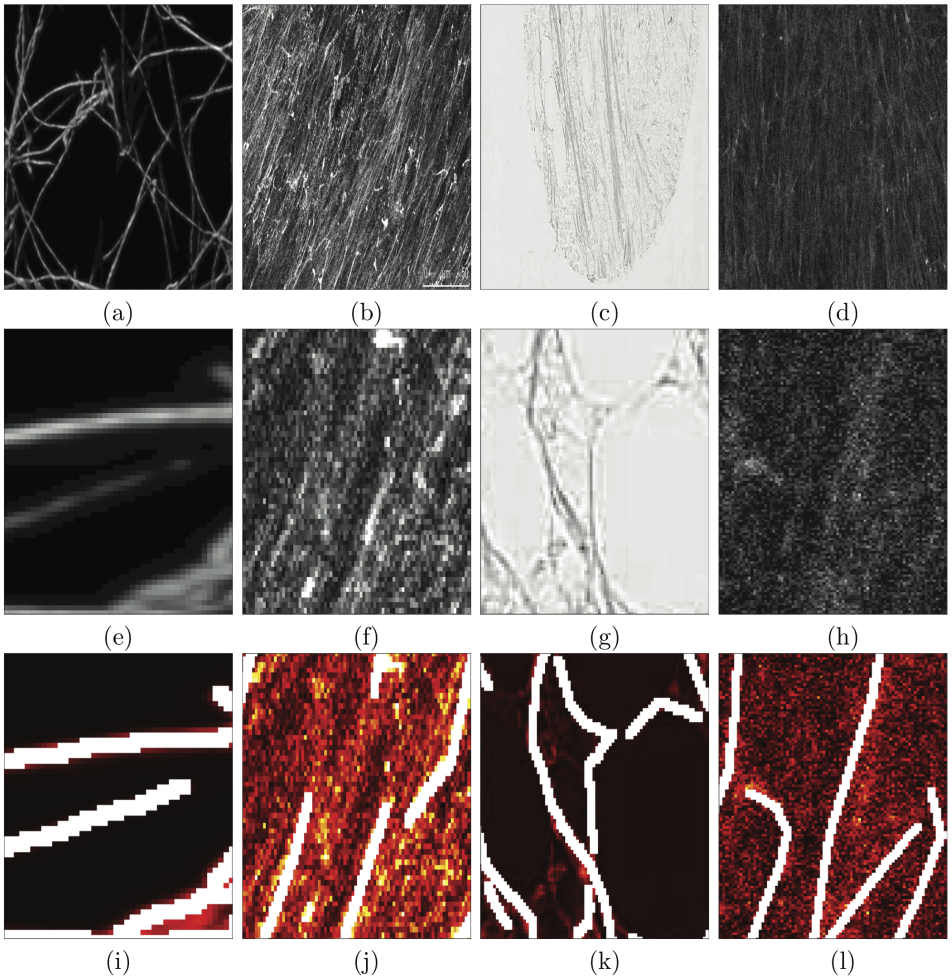


Fig 1. Illustration of the datasets that will be used to compare the tracing algorithms. (a) Second Harmonic Generation Collagen (SHG). (b) Fluorescent Fibronectin (FF). (c) Breast Cancer Biopsy slide (BCB). (d) Disease Mimicking Extra Cellular Matrix (DME). (e-h) Regions of interest (ROI) of the images in (a-d). (i-l) Manually delineated ground truth (GT) for the ROIs of images in (e-h). White lines correspond to the GT and intensities are displayed in a colormap with hot colours (black-red-orange-yellow-white).

https://doi.org/10.1371/journal.pone.0320006.g001

emission filter was 445 nm with a 20nm bandwidth. The pulse length was roughly 100 fs. All the SHG images were verified using cellular autofluorescence from the coenzyme Flavin Adenine Dinucleotide, in slides they were verified by white light images of H&E. Images from this data set are clean and fibres are distinguishable from each other.

- Fluorescently labelled fibronectin images (Fig 1b). Fibronectin fibroblasts were cultivated on a gelatine coated surface bonded with glutaraldehyde, used to create the extracellular matrix. Fluorescent images were captured to observe the fibronectin expression. Fibres appear bright and have high intensity, background is darker with lower intensity. Image is available at *github.com/youssefarafat/Trace_Ridges/blob/main/Fluorescent_Fibronectin*.
- Breast Cancer Biopsy (BCB) images of collagen (Fig 1c) were published with the Twombli algorithm [38]. The samples were stained with Picrosirius red. The slides were deparaffinised and hydrated prior to the application of Picrosirius red mixture for 60 minutes, after which the samples were rinsed twice in acetic acid followed by alcohol dehydration and mounting. Finally, slides are scanned using a Zeiss Axio Scan.Z1 at 10x magnification. This dataset was chosen as the images are stained with Picrosirius red, including variety to the previous monotone images. Fibres have a filamentous structure, which makes it unique.
- Disease mimicking ECM (DME) (Fig 1d) images of fibroblasts were published with the Graph based algorithm [42]. Normal resting fibroblasts were treated with TGF-B1 cytokine to activate the fibroblast in the tumour micro-environment. The Fibronectin rich matrice was created by exposing recombinant cFn to the Fibronectin null mouse embryo fibroblasts. The cultures finally are decellularised after 7 days. The resultant matrices are viewed with a confocal microscope after they are fluorescently stained, at a 10x/0.45 magnification. Images from this data set are noisy. Brightness variation is low.

Ground truth

Fibres on (Fig 1a–1d) were manually delineated to create the ground truth to be later used in a quantitative comparison.

Pre-processing of the data sets

Filtering data sets may reduce the levels of noise and thus improve subsequent steps in a pipeline.

The four data sets present very different noise characteristics, and three different filtering approaches were followed: No pre-processing filtering, low pass filtering with a Gaussian filter and a deep learning de-noising filtering [44]. It should be noted that some methodologies (e.g., Scale Space) include pre-processing steps themselves, but for consistency and ease of comparison all data sets were pre-processed in the same way before applying tracing algorithms.

Gaussian filtering. A Gaussian filter removes noise and detail by combining the values of neighbouring pixels weighted with a Gaussian kernel to replace the original value of a pixel.

A Gaussian filter with a kernel size of 3×3 and a standard deviation value $\sigma = 1$ was applied to all the images before applying the tracing algorithms.

DnCnn. DnCnn [44] is a de-noising deep convolution neural network for noisy images. The model adopts residual learning to separate noise from noisy observations. Batch normalisation and residual learning are adopted to improve the de-noising performance. The DnCnn model can handle blind Gaussian de-noising with unknown noise level. The DnCnn model

was trained using the images in (Fig 1). Each image was split into half, half was used for training and the other half was unseen. The images were converted to 2-D grayscale where needed as the model only accepts 2-D grayscale images. The model was trained with a learning rate of 0.001 and a batch size of 32.

Tracing algorithms

The following seven tracing algorithm were compared: Edge Detection (ED) [24], CT Fire (CTF) [40,43], Scale Space (SS) [22,23], Twombli (TW) [38], U-NET [39], Graph Based (GB) [41,42] and the proposed Trace Ridges (TR). They vary in their complexity and methodology with some leveraging filters as part of the pre-processing (GB with Gabor filters, CTF with Curvelet transform for image enhancement). U-Net utilises deep learning techniques. Trace Ridges employs Watershed [17] and Edge Detection for fibre tracing and uses morphological features for image refinement.

Edge Detection. The Canny Edge Detection [24] extracts edges of objects in image by identifying changes in pixel intensities. It employs Gaussian filtering to reduce noise then applies non-maximum suppression to thin the edges. The thresholding process, also known as hysteresis, is performed to exclude or include edges. In this study, the Canny Edge Detection used a filter size of 2×2 to extract edges as proxies for the fibres. Subsequently, the edges were labelled and their metrics extracted.

CT Fire. CT Fire [40,43] was developed to extract and quantify collagen fibres from SHG images. It utilises a curvelet transform (CT) [45] de-noising filter, with Gaussian filtering to remove noise. The curvelet filter represents the image as stacked elements following ridge lines and wavelets, with a frequency wrapping based de-noising technique. Curvelets commonly follow the rule in (Eq 1). The fibre extraction algorithm processes the binary image, with foreground pixels representing fibres. It computes a distance transform to determine the proximity of the foreground pixels to the nearest background pixel, identifies maximal ridges of the smoothed image, and generates nucleation points. After which fibres extend from each nucleation point based on trajectory, followed by removal of short branches.

$$\text{width} = \text{length}^2 \tag{1}$$

Scale Space. Scale Space [22,23] blurs an image across multiple scales to reveal its characteristics. Significant features emerge at lower scales, while finer details are visible at first scales and are progressively smoothed out. Scale Space treats an image as a series of Gaussian smoothed images, represented by the function f(x,y) and Gaussian g(x,y,t), where t denotes the width of the Gaussian filter. First and second derivatives of x and y dimensions are represented as

$$(L_x, L_{xx}, L_y, L_{yy}),$$

forming a matrix, dictating intensity changes in the image. For Scale Space, the collection of images is represented as L(x,y;t) and Gaussian kernel in (Eq 2). This method was implemented as seen in [36].

$$L(x,y;t) = g(x,y;t) \times f(x,y) = \left(\frac{1}{2\pi t}\right)^{e^{-\frac{x^2+y^2}{2t}}} \times f(x,y) \tag{2}$$

Twombli. Twombli [38] a FIJI macro tool for quantifying fibre-like patterns. It employs existing techniques, including Ridge Detection [46] and AnaMorf [47]. Ridge Detection

assesses first and second directional derivatives using Gaussian kernels to detect line points and extract line width with sub-pixel accuracy. Bias is reduced by inverting the Scale Space behaviour of the asymmetric line model. AnaMorf quantifies the structures identified. Twombli also incorporates OrientationJ to obtain the global alignment metric, determining the dominant direction of the fibrial structure. By analysing changes in intensity signal as a function of nearby pixel distances, Twombli constructs segmentation masks using fibre edges.

U-Net.

Properties:

U-Net [39] is a deep fully convolutional network widely applied in biomedical image segmentation. The architecture compromises of a contracting path for capturing fine details through convolutions, and a symmetric expanding path for accurate localisation. The segmentation map is generated via skip connections between contraction and expanding layers. The network's architecture is symmetrical to the letter "U". U-Net requires minimal training data and can achieve high performance. The U-Net used in this comparative study (Fig 2) had a depth of 3, consisting of 46 layers, including encoder, bridge, decoder, convolution, softmax and segmentation layers, with skip connections linking the contracting and expanding paths.

U-Net Training: Initially, various training parameters were tested, including batch size (16 and 32), and learning rate (0.0001 and 0.001), to optimise training parameters and select the best-performing network. Each image from (Fig 1a–1d) was split into halves for training and testing. BCB Collagen image (Fig 1c) was complemented to highlight fibre features in white, and background in black. To increase the number of images for training, augmentation was employed. Augmentation techniques (Fig 3) included horizontal and vertical flipping, rotation by 90 degrees and addition of Gaussian noise. A total of 8910 images of size 32x32 were used for training. The U-Net was trained with a batch size of 16, learning rate of 0.0001 and the "adam" optimisation algorithm, for 15 epochs. Training duration was 15 minutes.

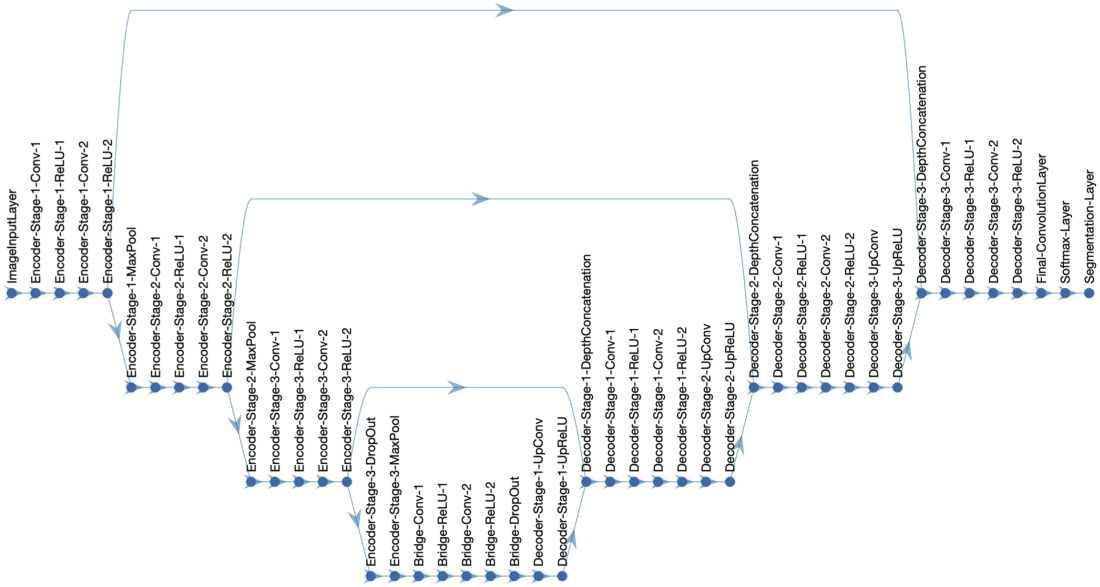


Fig 2. Illustration of the U-Net architecture used to segment the fibres. The architecture consists of three encoder levels and 3 decoder levels.

https://doi.org/10.1371/journal.pone.0320006.g002

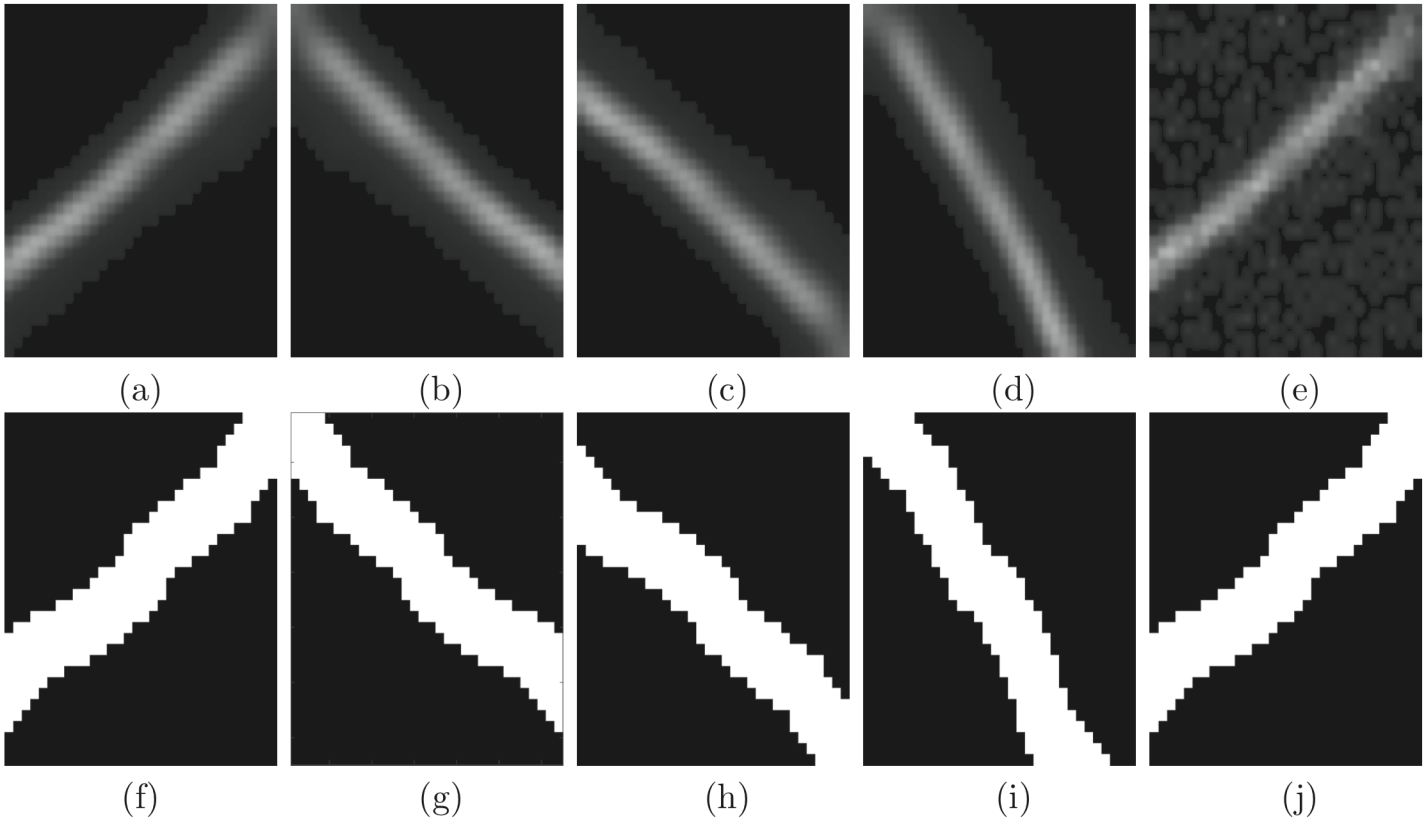


Fig 3. Illustration of the data augmentation approaches used to train the U-Net. (a) Original patch (b) Horizontally flipped (c) Vertically flipped (d) Rotation by 90 degrees (e) Gaussian blurred, with zero mean and a standard deviation of 0.05. (f–j) labels corresponding to augmented patches. It should be noted that (j) is the same as (f) as the Gaussian blur does not affect the label.

https://doi.org/10.1371/journal.pone.0320006.g003

Graph based. The Graph-based algorithm [41,42] combines fast discrete curvelet transform filters to detect curvilinear anisotropic objects, in combination with a fibre extraction algorithm. The algorithm uses Gabor filters to enhance fibres that have been observed with confocal microscopy. The algorithm offers flexibility by detecting various fibre elements at different frequencies and orientations, avoiding translation and rotation errors. The algorithm associates graph networks with fibre morphological skeletons, followed by fibre pruning and post-processing fibre re-connection. Two graph types are used to measure fibre properties: one representing fibre crosslinks or ends, and the other simplifying the first graph into a skeleton. These representations were used to extract geometrical and topological properties from the confocal images. Then, specific properties, e.g., orientation for Gabor and fibre length as graph parameters were used to relate to the physical properties of the fibres.

Trace Ridges. In this paper, a methodology called *Trace Ridges* is proposed. The methodology combines Watershed transform and Edge detection. First, the Watershed is used to delineate ridges. The *main* ridges will correspond to the fibres to be traced, but in addition, *minor* ridges, product of the well-known over segmentation of watersheds [48,49], will appear as coming down from the main ridges. Then, the Edge detection process detects regions where there are sudden changes of intensity, i.e., close to the ridges, ideally one on each side of the centre line of the ridge. The intersection between the edges and watersheds is used to *break* those *minor* ridges that run down from a main ridge through the sides of the basins,

thus separating them from the major ridges. In this way, the well-known problem of over-segmentation of Watershed is counteracted and the methodology prioritises longer ridges along the brighter sections of an image for cleaner results. The combination of these two steps of Trace Ridges is illustrated in (Fig 4a). The region contains three strong and bright ridges with a small noisier region highlighted with a green rectangle. The results of Watershed and Edge Detection are shown separately in (Fig 4b, 4c) and the overlap in (Fig 4d) where the edges are shown in red, and the watershed is shown in yellow. The points where the two overlap are shown as white. These points are discarded thus breaking the small watersheds and maintaining the main ridges. Next, the edges and small segments with a major axis length smaller than a threshold (5 pixels) are removed. Segments that lie within darker regions are also removed as it may be possible that there are watersheds in dark regions that are hardly visible and not of interest.

In some cases, a thicker ridge may produce two parallel watersheds with a very small hole in between. These cases are identified by their Euler characteristic [50]. Only very small holes are filled in, and then the resulting object is thinned, this process leaves a trace that runs through the position of that small hole.

Also, there are cases where Watershed will not detect a ridge, specifically when both sides of the ridge feed to a single basin surrounding it. Those cases are detected by using only Edge detection, filling in the space between the edges, and then thinning it to create a single line where the ridge is located. In some cases, Edge Detection results in only one edge for a ridge, these are discarded and treated as low level ridges or noise.

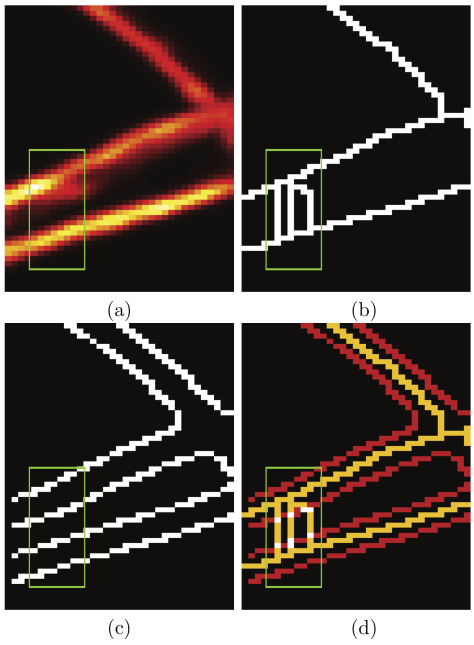


Fig 4. Illustration of the Trace Ridges methodology. (a) Region of interest of one SHG image that shows some strong ridges and a region with some noise highlighted with a green rectangle. (b) Output of a Watershed transform, which traces the strong ridges but also some smaller ones that are not of interest. (c) Edges detected with Canny's algorithm highlight the sides of the strong ridges. (d) The combination of Edge Detection (red lines) with Watershed (yellow lines). White pixels correspond to the overlap of both techniques. The watershed will be broken by removing these overlapping pixels.

https://doi.org/10.1371/journal.pone.0320006.g004

Finally, Trace Ridges discards small and curved fibres and keeps longer and straighter lines. In the case that Trace Ridges detects less than 100 fibres from the missed ridges. Trace Ridges retains from those ridges only those that have a major axis length above 15 pixels and a minor axis length less than 10 pixels. If the number of new fibres detected is higher, a stricter condition is applied whereby the fibres must have a major axis length of at least 25 pixels and the minor axis length must be less than 10 pixels.

Once all the fibres are detected, Trace Ridges combines the ridges detected with the combination of Watershed and Edge Detection as well as the missed ridges detected by Edge Detection, prioritising longer and brighter fibres.

Many morphological metrics can be extracted from the traced lines for instance, orientation, length, width, count, eccentricity, and mean intensities. Furthermore, the gaps between fibres can be calculated and from there, metrics such as gap area and largest gaps can be extracted.

Finally, three dimensional datasets are projected to two dimensions with a maximum intensity projection.

Quantitative comparison with distance metrics

Quantitative comparisons between the GT and results of the algorithms were calculated using distance metrics as has been previously reported [36]. It is important to highlight that traditional pixel-wise metrics such as accuracy, true/false positive rate, sensitivity, and specificity are not suitable for comparing traces against a ground truth. This is illustrated in (Fig 5), which shows three cases of lines that overlap in one single pixel and then extend from there in very different conditions. Whilst the cases are quite different, pixel-wise metrics would identify the same number of true positives, one, in all cases. Similarly, the true negatives (black pixels) could be similar and not useful to distinguish between cases. Pixel-wise metrics are useful with larger structures such as nuclei and cells where the comparison is between areas and not how far from each other the lines can be.

Calculation of the distance metrics has been previously described [36] but will be briefly here for completeness and illustrated with a simple example in (Fig 6) and the Second Harmonic Generation image in (Fig 7). First, a distance map from each of the set of traces to be compared was calculated (Fig 6b, 6e) and (Fig 7b, 7e). Then the map of one set was multiplied by the opposite set. Thus, the traces that are further away from the traces from the opposite set will be in regions of higher distance and close to their opposites in regions of low distance (Fig 6c, 6f) and (Fig 7c, 7f).

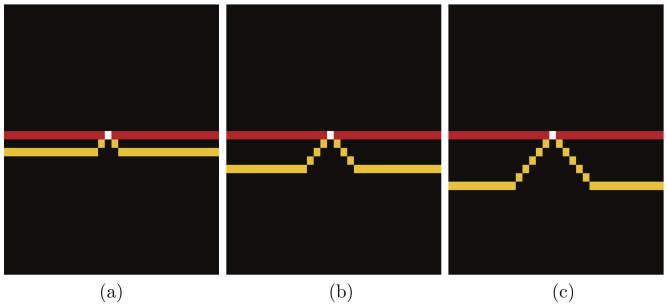


Fig 5. Illustration of three cases of two lines where the pixel-wise metrics of true positives, true negatives, false positives, false negatives are identical, yet the distance between the lines is different.

https://doi.org/10.1371/journal.pone.0320006.g005

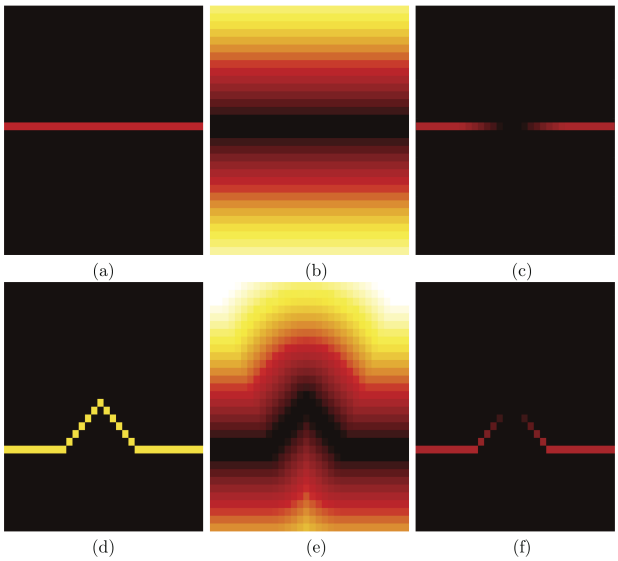


Fig 6. Illustration of the distance error measurement between the two lines of (Fig 5c). (a, d) Two lines that overlap only at a single point. (b) Distance map computed from the red line in (a). Distances follow a hot colour map: black-red-orange-yellow-white. (c) Product of the distance map shown in (e) with the line shown in (a). (e) Distance map computed from the yellow line in (d). (f) Product of the distance map shown in (b) with the yellow line shown in (d).

https://doi.org/10.1371/journal.pone.0320006.g006

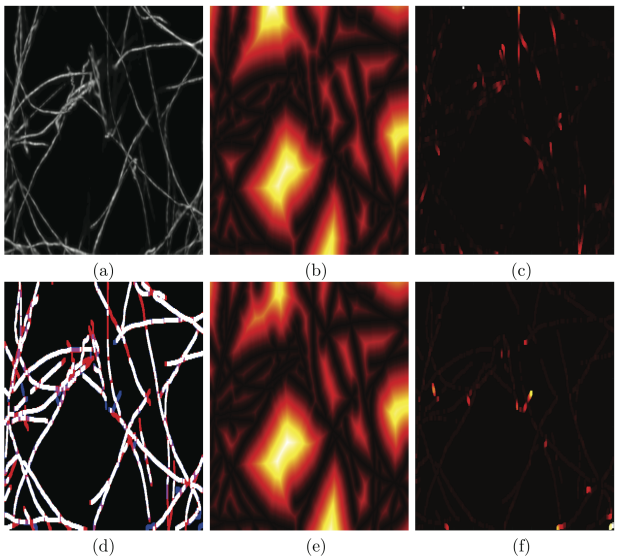


Fig 7. Illustration of the calculation of the error distance maps on the Second Harmonic Generation image shown in (Fig 1a). (a) One representative image with a variety of fibres. (b) Distance map calculated from the traces calculated with Trace Ridges. (c) Product of the ground truth and the distance map of (b). (d) Comparison between the results of Trace Ridges and the ground truth. White lines indicate pixels from GT and trace. Blue lines correspond to Trace Ridges. Red lines are from the GT. (e) Distance map calculated from the ground truth. (f) Product of the traces from Trace Ridges and the distance map of (e). It should be noticed how the cases where the lines overlap (white) in (d) do not appear in any of the distance maps (c, f) whilst the lines that do not overlap (blue and red) appear in the distance maps with increased brightness as they are further away from the lines generated by the opposite method.

https://doi.org/10.1371/journal.pone.0320006.g007

From those two maps many metrics can be calculated, for instance, the average distance error per trace, maximum distance error per trace and total distance error.

Results and discussion

In this work, images of fibre-like structures were traced by seven different image processing algorithms and were filtered by three different methods. To compare and observe which method traces fibres better, four quantitative measurements were calculated: the total distance error, average distance error and maximum distance error, as well as computational time. The distance errors are measured by comparing how close the trace results of each algorithm were compared to the manually delineated ground truth.

Quantitative comparison of the algorithms

Total distance error. The first analysis focuses on total distance errors, the accumulation of the distance errors from GT to algorithm trace generated by the distance maps, as explained. Qualitative results are summarised in Table 1, comparing algorithms with and without filtering. Filtering notably improves Edge Detection, Scale Space, Twombli, Graph Based, and Trace Ridges. DnCnn enhances Edge Detection, Scale Space, and Graph Based, while Gaussian filtering benefits Twombli and Trace Ridges. However, filtering worsens Scale Space with Gaussian and Twombli with DnCnn. U-Net and CT Fire exhibit increased errors post-filtering, likely due to U-Net training on non-filtered images and CT Fire's inherent filtering. Notably, U-Net's performance improved when DnCnn was used on the BCB and SHG images, while U-Net with DnCnn yielded larger errors for the other images, impacting its

Table 1. Quantitative Comparison by total error distance in pixels of the following Tracing methods: Edge Detection (ED), CT Fire (CTF), Scale Space (SS), Twombli (TW), U-Net, Graph based (GB) and Trace Ridges (TR), with three filtering options on the following images: Disease Mimicking ECM (DME), Second Harmonic (SHG), Fluorescent Fibronectin (FF) and Breast Cancer Biopsy (BCB) images. In addition to the errors per image, average per filtering option, average of the methodology and rank are presented. Results are sorted by rank with the worst results at the top (7) and best results at the bottom (1).

Method		Image							
	Filter	DME	SHG	FF	BCB	Average	Rank	avg/Method	Rank
ED	No Filt	947,550	35,905	754,120	224,120	490,424	21	429,173	7
	Gaus Filt	834,660	36,248	714,530	208,090	448,382	20		
	DnCnn	620,970	32,591	612,600	128,690	348,713	19		
CTF	No Filt	311,580	13,594	232,140	83,653	160,242	16	171,529	6
	Gaus Filt	342,290	22,687	269,290	79,694	178,490	18		
	DnCnn	330,190	29,437	266,950	76,848	175,856	17		
SS	No Filt	206,620	15,654	318,770	72,760	153,451	14	146,592	5
	Gaus Filt	209,020	15,493	326,510	75,720	156,686	15		
	DnCnn	215,930	15,334	212,400	74,890	129,639	7		
TW	No Filt	218,870	14,160	286,120	60,969	145,030	11	141,988	4
	Gaus Filt	186,810	14,328	280,020	49,700	132,715	8		
	DnCnn	248,420	14,087	280,280	50,091	148,220	12		
UNET	No Filt	180,610	12,938	129,040	47,940	92,632	1	129,457	3
	Gaus Filt	268,560	12,119	248,620	45,840	143,785	9		
	DnCnn	292,440	12,220	260,070	43,090	151,955	13		
GB	No Filt	247,040	14,953	209,990	103,750	143,933	10	123,648	2
	Gaus Filt	191,540	14,693	203,640	68,019	119,473	6		
	DnCnn	154,756	19,502	188,520	67,372	107,538	2		
TR	No Filt	208,580	9,141	168,170	70,320	114,053	5	112,120	1
	Gaus Filt	198,810	9,537	166,690	63,520	109,639	3		
	DnCnn	195,270	13,727	162,760	78,910	112,667	4		

https://doi.org/10.1371/journal.pone.0320006.t001

overall performance. Considering filtering + algorithm pairs as separate methods, U-Net without filtering performs best, followed by Graph Based with DnCnn, Trace Ridges with Gaussian filtering, Trace Ridges with DnCnn and Trace Ridges with no filtering. However, averaging total distance errors across all images regardless of filtering reveals Trace Ridges as the top-performing method, followed by Graph Based, U-Net, Twombli, Scale Space, CT Fire and Edge Detection.

Average distance error. This quantitative metric calculates total errors divided by the number of traces to provide an average, as shown in Table 2. Filters enhance the performance of various methods including Edge Detection, CT Fire, Scale Space, Twombli, Graph based and Trace Ridges. However, filters do not improve U-Net's values due to its training on non-filtered images. Notably DnCnn improves U-Net's performance for the DME and BCB images, while Gaussian filtering benefits only the BCB image. Filtering yields consistent results for Edge Detection, with DnCnn slightly improving it. Both Gaussian and DnCnn filters significantly enhance CT Fire. Gaussian filtering marginally improves Scale Space, while DnCnn enhances it. Twombli's results are improved by Gaussian filtering but worsened by DnCnn, in agreement with the total distance error metric. Filtering improves the Graph based method, with DnCnn yielding the best performance. Gaussian filtering improves Trace Ridges, but DnCnn slightly increases the value, impacting performance. Filtering + algorithm pairs are ranked, with Edge Detection + DnCnn performing best, followed by Edge Detection + no filter, Trace Ridges + Gaussian, Edge Detection + Gaussian and Twombli + Gaussian. Analysing averages per method, Edge Detection ranks highest, followed closely by Trace Ridges and Twombli. Edge Detection performs well due to the proximity of its traces to the GT, although it detects unwanted edges as fibres. Conversely, Trace Ridges, while achieving the best total

Table 2. Quantitative Comparison by average error distance in pixels of the following Tracing algorithms; Edge Detection (ED), CT Fire (CTF), Scale Space (SS), Twombli (TW), U-Net, Graph based (GB) and Trace Ridges (TR) with three filtering options on Disease Mimicking ECM (DME), Second Harmonic (SHG), Fluorescent Fibronectin (FF) and Breast Cancer Biopsy (BCB) images. In addition to errors per image, average per filtering option, average of the methodology and rank are presented. Results are sorted by rank with the worst results at the top (7) and best results at the bottom (1).

Method		Image							
	Filter	DME	SHG	FF	BCB	Average	Rank	avg/Method	Rank
GB	No Filt	19.3	15.0	17.7	10.5	15.6	21	12.0	7
	Gaus Filt	15.3	4.1	17.3	6.6	10.8	18		
	DnCnn	11.0	7.4	13.0	7.0	9.6	13		
UNET	No Filt	10.7	12.9	10.7	8.2	10.6	16	11.0	6
	Gaus Filt	11.8	13.6	14.4	7.5	11.8	20		
	DnCnn	12.7	9.7	14.3	6.1	10.7	17		
CTF	No Filt	8.6	15.7	14.3	8.4	11.7	19	9.6	5
	Gaus Filt	11.4	3.8	11.4	7.9	8.6	11		
	DnCnn	9.5	6.4	10.1	7.5	8.4	9		
SS	No Filt	10.7	5.2	16.1	8.6	10.1	15	9.5	4
	Gaus Filt	10.4	4.7	15.8	8.2	9.8	14		
	DnCnn	10.7	5.0	10.4	8.2	8.6	10		
TW	No Filt	10.0	5.0	10.6	6.9	8.1	8	8.2	3
	Gaus Filt	9.7	4.7	10.3	5.7	7.6	5		
	DnCnn	14.5	5.5	10.3	5.4	8.9	12		
TR	No Filt	11.5	3.1	9.8	7.6	8.0	6	7.8	2
	Gaus Filt	10.2	3.4	8.7	6.7	7.3	3		
	DnCnn	10.5	4.3	9.3	8.3	8.1	7		
ED	No Filt	8.4	4.8	9.4	6.4	7.3	2	7.2	1
	Gaus Filt	8.5	4.9	9.1	6.6	7.3	4		
	DnCnn	8.3	5.1	8.8	5.7	7.0	1		

https://doi.org/10.1371/journal.pone.0320006.t002

Table 3. Quantitative Comparison by maximum error distance in pixels of the following Tracing algorithms; Edge Detection (ED), CT Fire (CTF), Scale Space (SS), Twombli (TW), U-Net, Graph based (GB) and Trace Ridges (TR) with three filtering options on Disease Mimicking ECM (DME), Second Harmonic (SHG), Fluorescent Fibronectin (FF) and Breast Cancer Biopsy (BCB) images. In addition to errors per image, average per filtering option, average of the methodology and rank are presented. Results are sorted by rank with the worst results at the top (7) and the best results at the bottom (1).

Method		Image							
	Filter	DME	SHG	FF	BCB	Average	Rank	avg/Method	Rank
SS	No Filt	65.2	50.4	69.6	202.4	96.9	21	94.9	7
	Gaus Filt	63.6	55.6	66.4	191.8	94.4	20		
	DnCnn	61.4	57.4	59.2	195.5	93.4	19		
CTF	No Filt	68.9	89.0	81.0	73.6	78.1	16	77.5	6
	Gaus Filt	57.5	98.5	75.5	82.7	78.5	17		
	DnCnn	58.4	96.5	75.7	73.1	75.9	15		
TR	No Filt	68.5	55.0	95.2	102.1	80.2	18	76.1	5
	Gaus Filt	84.3	50.4	91.9	70.4	74.3	14		
	DnCnn	76.3	48.8	88.2	81.7	73.7	13		
TW	No Filt	62.4	65.8	57.8	68.7	63.7	8	66.3	4
	Gaus Filt	69.8	66.9	60.5	60.5	64.4	9		
	DnCnn	79.2	89.0	62.7	51.8	70.7	12		
ED	No Filt	45.4	59.4	52.3	119.7	69.2	10	66.3	3
	Gaus Filt	46.1	59.4	52.3	119.7	69.4	11		
	DnCnn	50.3	86.0	58.3	46.1	60.2	4		
GB	No Filt	58.9	72.1	65.8	56.2	63.2	7	60.7	2
	Gaus Filt	59.6	53.8	62.9	47.8	56.0	1		
	DnCnn	71.6	55.8	70.6	53.2	62.8	6		
UNET	No Filt	58.8	63.4	66.4	50.9	59.9	3	59.8	1
	Gaus Filt	56.1	63.4	57.6	50.4	56.9	2		
	DnCnn	57.2	89.0	56.0	47.8	62.5	5		

https://doi.org/10.1371/journal.pone.0320006.t003

distance error and the second-best average, is considered a more robust and accurate tracing method.

Maximum distance error. This metric examines the maximum singular error in each image, as shown in Table 3. Filtering enhances the performance of Edge Detection, CT Fire, Scale Space, U-Net, Graph Based and Trace Ridges. Filtering hinders Twombli. Gaussian filtering produces comparable results to no filtering for Edge Detection, while DnCnn enhances results. CT Fire shows slight improvement with Gaussian and more significant improvement with DnCnn, both Gaussian and DnCnn improve Scale Space, with DnCnn yielding the lowest maximum error. Interestingly, Gaussian filtering improves U-Net but worsens with DnCnn filtering, DnCnn hinders the performance of U-Net on the SHG image, heavily impacting the average. For Graph based, DnCnn's impact resembles that of no filtering, whereas Gaussian improves results, both Gaussian and DnCnn improve Trace Ridges, with Gaussian having a similar effect to DnCnn but slightly lesser. Among filtering + algorithm pairs, Graph based + Gaussian performs the best, followed by U-Net + Gaussian, U-Net + no filtering, Edge Detection + DnCnn and U-Net + DnCnn. Observing averages per method, U-Net achieves the best results, followed by Graph Based and Edge Detection. However, Trace Ridges does not perform as well in this metric due to its preference for brighter and longer fibres, resulting in higher singular maximum errors. While this metric provides insights, total and average distance errors are better indicators of algorithm robustness and accuracy.

Time. The processing time for each trace of every algorithm with various filter combinations on each image was recorded and is presented in Table 4. Surprisingly, filtering improved the speed of every algorithm, with DnCnn consistently achieving the fastest results across all algorithms and filters. However, Edge Detection exhibited consistent speed with Gaussian

Table 4. Quantitative Comparison by time taken in seconds of the following Tracing algorithms; Edge Detection (ED), CT Fire (CTF), Scale Space (SS), Twombli (TW), U-Net, Graph based (GB) and Trace Ridges (TR) with three filtering options on Disease Mimicking ECM (DME), Second Harmonic (SHG), Fluorescent Fibronectin (FF) and Breast Cancer Biopsy (BCB) images. In addition to time taken per image, average per filtering option, average of the methodology and rank are presented. Results are sorted by rank with the worst results at the top (7) and the best results at the bottom (1).

Method		Image							
	Filter	DME	SHG	FF	BCB	Average	Rank	avg/Method	Rank
GB	No Filt	1387.0	88.0	550.0	146.0	542.75	21	433.08	7
	Gaus Filt	836.0	55.0	664.0	153.0	427.00	20		
	DnCnn	527.0	84.0	573.0	134.0	329.50	16		
TW	No Filt	383.0	211.0	777.0	306.0	419.25	19	385.92	6
	Gaus Filt	681.0	340.0	438.0	152.0	402.75	18		
	DnCnn	520.0	303.0	380.0	140.0	335.75	17		
CTF	No Filt	180.0	60.0	145.0	77.0	115.50	15	102.42	5
	Gaus Filt	167.0	57.0	110.0	71.0	101.25	14		
	DnCnn	135.0	49.0	113.0	65.0	90.50	13		
SS	No Filt	98.6	4.8	72.7	45.4	55.40	11	55.78	4
	Gaus Filt	86.0	5.2	71.8	82.5	61.36	12		
	DnCnn	87.0	5.1	68.0	42.2	50.57	10		
UNET	No Filt	23.0	5.2	21.1	20.1	17.33	9	12.23	3
	Gaus Filt	20.1	5.0	18.9	23.7	16.92	8		
	DnCnn	2.5	2.2	2.4	2.6	2.44	5		
TR	No Filt	8.0	0.5	4.0	2.7	3.81	7	3.00	2
	Gaus Filt	4.5	0.8	3.1	3.0	2.87	6		
	DnCnn	4.0	0.6	2.9	1.8	2.32	4		
ED	No Filt	1.2	0.1	0.5	0.5	0.58	2	0.53	1
	Gaus Filt	0.7	0.5	0.5	0.6	0.58	3		
	DnCnn	0.8	0.1	0.4	0.4	0.43	1		

https://doi.org/10.1371/journal.pone.0320006.t004

and no filtering, and Gaussian filtering slowed down Scale Space. Each subsequent algorithm and filter pair in Table 4 generally performed better and faster, except for Graph Based with DnCnn, which outperformed Twombli. U-Net with DnCnn was faster than Trace Ridges with Gaussian or no filtering. Edge Detection was the fastest algorithm overall, followed closely by Trace Ridges, which is known for its efficiency. U-Net was ×3 slower than Trace Ridges, followed by Scale Space and CT Fire. Speed variations were significantly and heavily influenced by image noise levels, with noisier images (FF and DME) taking longer to process than less noisy images.

Visual assessment of the algorithms

A comparison between the results and the GT is presented in (Fig 8), (Fig 9), (Fig 10), (Fig 11), (Fig 12), (Fig 13), (Fig 14) and (Fig 15) with the GT in red, the results in blue. White lines correspond to pixels where the result of the algorithm and GT overlap, i.e., correct traces. It should be noted that in some cases a purple colour may be perceived, but this is an artefact when a blue line is remarkably close to a red line.

Second harmonic generation. The SHG is a clear image, detected and missed fibres are easily identifiable. Edge Detection (Fig 8) consistently produces duplicate lines for each fibre. From (Fig 12), DnCnn application improves the tracing, increasing white regions compared to the original. CT Fire algorithm has demonstrated notable success on the original image (Fig 8) but struggles with filtered images, evidenced by decreased number of white lines regions (Fig 12), suggesting hindrance by filtering. Scale Space remains unaffected by filtering, showing consistent results (Fig 12). Scale Space does not perform well. Twombli exhibits

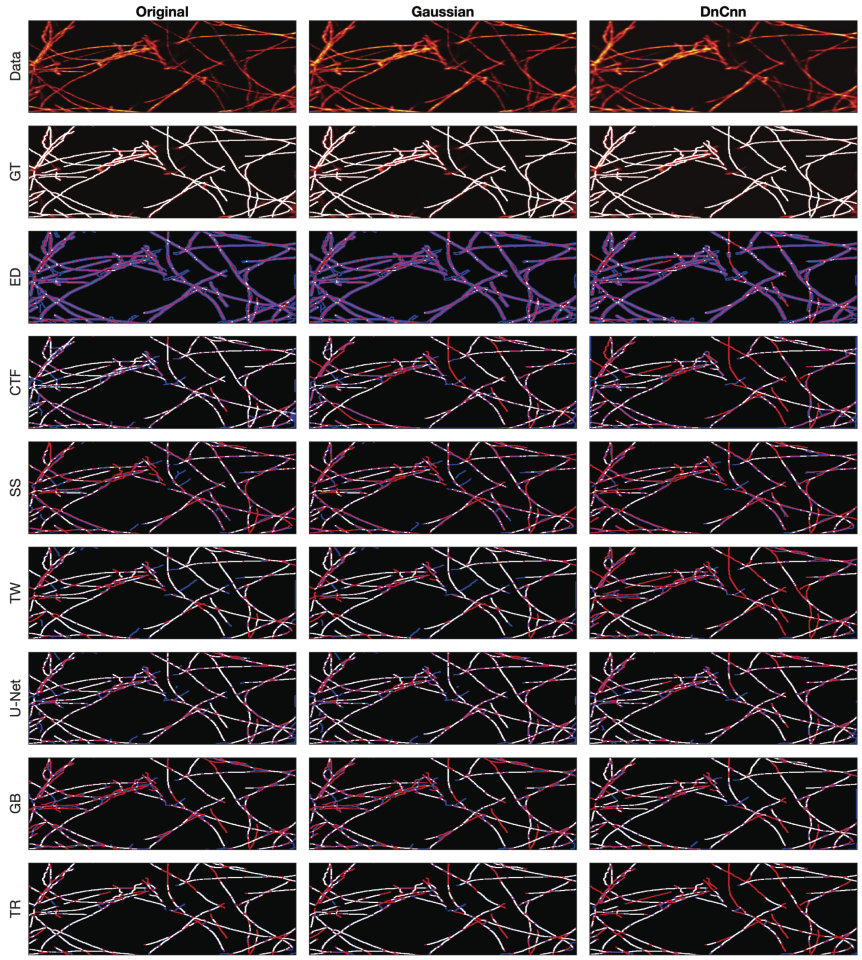


Fig 8. Comparison of tracing algorithms on SHG image (full view) [40]. Image was applied to algorithms with three different filters: Original/no filtering, Gaussian and DnCnn. Data corresponds to the tracing images; GT is the manually delineated ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is depicted in blue and areas where they overlap is white.

https://doi.org/10.1371/journal.pone.0320006.g008

sporadic traces and misses major fibres (Fig 8). Application of DnCnn appears to worsen Twombli's performance (Fig 12). U-Net performs well, with significant white line regions (Fig 8). Filtering does not visibly affect U-Net's segmentation. The Graph based method misses several fibres shown in red. It also detects spurious objects with DnCnn application (Fig 8). Trace Ridges shows the best visual performance, with the highest number of correct fibre regions (Fig 8). However, visually, Trace Ridges with DnCnn performs the worst amongst filter methods (Fig 12).

Fluorescent fibronectin. For the Fluorescent Fibronectin image (Fig 9), several notable findings emerge. Despite the increased noise on the image as compared to the previous example (SHG), similar trends persist, that is, both CT Fire and Edge Detection generate a large number of traces, more than there are actual fibres. Filtering enhances Edge Detection (Fig 13), but degrades the results CT Fire. Scale Space and Trace Ridges yield fewer traces, with Scale Space detecting darker fibres and Trace Ridges favouring the bright ones. Trace Ridges exhibits fewer spurious lines (blue) and more white regions, indicative of its

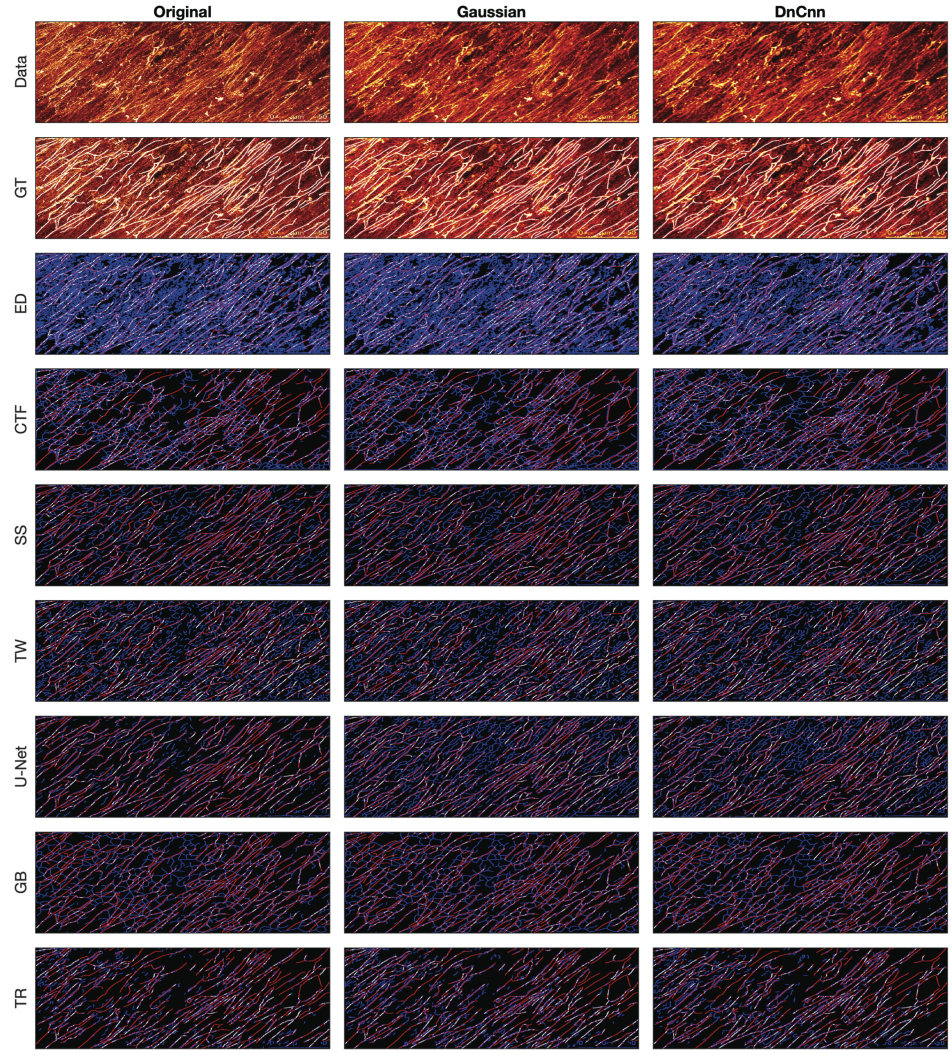


Fig 9. Comparison of tracing algorithms on FF image (full view). Image was applied to algorithms with three different filters: Original/no filtering, Gaussian and DnCnn. Data corresponds to the images traced; GT is the manually delineated ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is depicted in blue and areas where they overlap is white.

https://doi.org/10.1371/journal.pone.0320006.g009

strict nature (Fig 13), showing improvement with image filtering. Graph based method and Twombli exhibit over-segmentation, visible as blue regions (Fig 9) and (Fig 13). Pre-filtering notably improves their performance. Conversely, U-Net's results with filtered images are inaccurate, as it was trained on non-filtered images, yielding optimal results with the original image as is visible in (Figs 9 and 13).

Breast cancer biopsy. For the BCB image, similar trends persist, with Edge Detection over-segmenting and producing inaccuracies, as seen in (Fig 10) and (Fig 14). However, pre-filtering results in cleaner images and reduces inaccuracies. CT Fire misses major fibres while detecting noise as spurious objects, evident in (Fig 10) and (Fig 14). Filtering enhances the algorithm's performance, as observed in (Fig 14). Scale Space shows inaccuracies, detecting noise and spurious objects, contrary to its usual strict behaviour, as shown in (Fig 14).

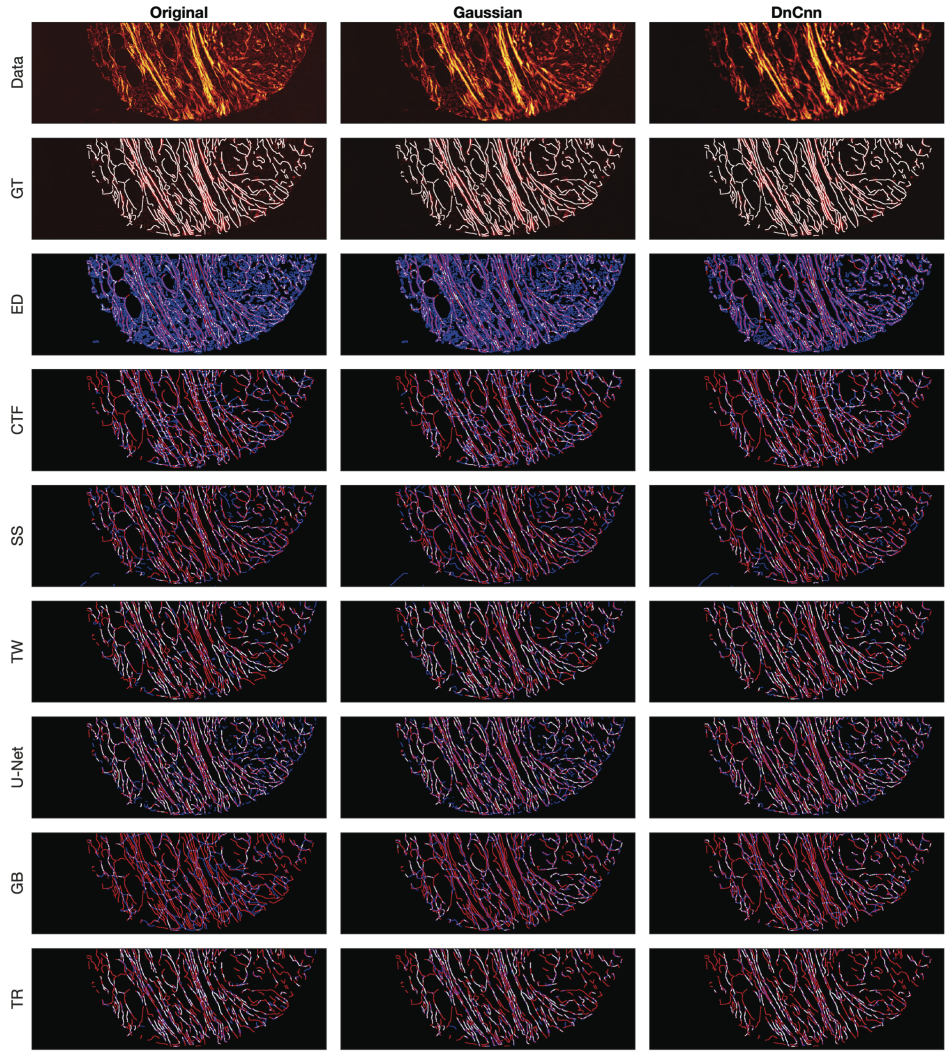


Fig 10. Comparison of tracing algorithms on Picrosirius red Breast Cancer Biopsy image (full view) [38]. Image was applied to algorithms with three different filters: Original/no filtering, Gaussian and DnCnn. Data corresponds to the images traced; GT is the manually delineated ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is depicted in blue and areas where they overlap is white.

https://doi.org/10.1371/journal.pone.0320006.g010

Filter application does not alter the results. Twombli shows notable performance, especially in detecting fibres at the top right corner in (Fig 14). Gaussian filtering improves Twombli, while DnCnn hinders its performance, causing loss of white lines. U-Net's results are intriguing, with many white regions indicating accurate delineation (Fig 10). Filtering improves U-Net's performance, reducing spurious objects, contrary to its behaviour on other images. Trace Ridges maintains strictness, yielding cleaner delineations (Fig 14), albeit missing some key fibres. Graph based algorithm performs similarly with Trace Ridges, with filtering reducing spurious elements (Fig 14).

Disease mimicking ECM. A visual examination of the disease mimicking ECM (DME) image (Fig 11) reveals notable insights. By looking at the data image, both Gaussian and DnCnn filtering visibly enhance image clarity. Edge Detection does not perform well.

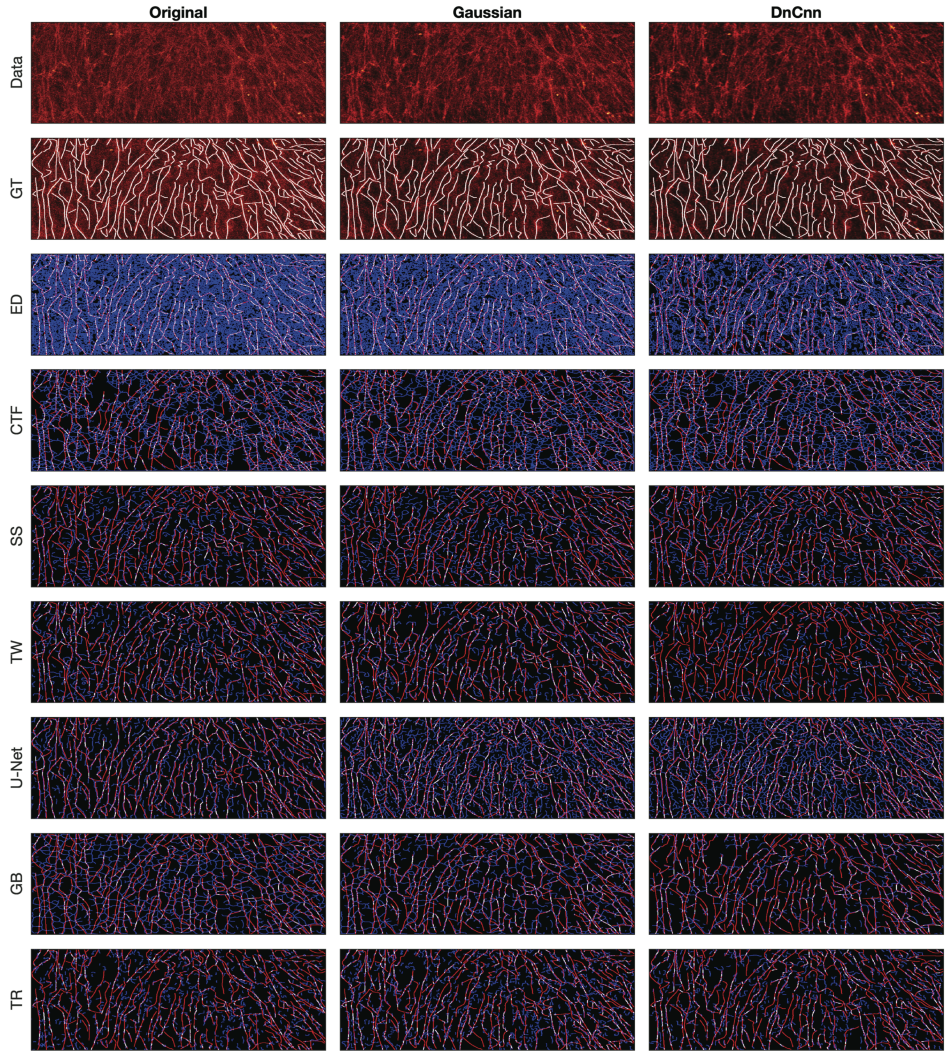


Fig 11. Comparison of tracing algorithms on DME image (full view) [42]. Image was applied to algorithms with three different filters: Original/no filtering, Gaussian and DnCnn. Data corresponds to images traced; GT is the manually delineated ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is depicted in blue and areas where they overlap is white.

https://doi.org/10.1371/journal.pone.0320006.g011

The main drawback of Edge Detection is that there are two lines for each fibre. The results show many blue lines, where many fibres were detected that are not there. Filtering notably enhances Edge Detection's performance, particularly evident in the ROI (Fig 15), where DnCnn offers the most significant improvement. CT Fire performed better than Edge Detection, it also presented spurious objects. Gaussian and DnCnn filtering resulted in more non-present traces; this is easily viewed in the ROI (Fig 15). Scale Space displays fewer traces, benefiting from its multi resolution capabilities, but may merge separate fibres occasionally. Gaussian and DnCnn filtering have increased the number of fibres, but the latter has resulted in the long fibres separating. Twombli yields mediocre results, segmenting non-existent fibres, yet outperforming certain algorithms. Filtering enhances Twombli's delineation, with Gaussian filtering notably improving the algorithms performance, this is evident in the number of

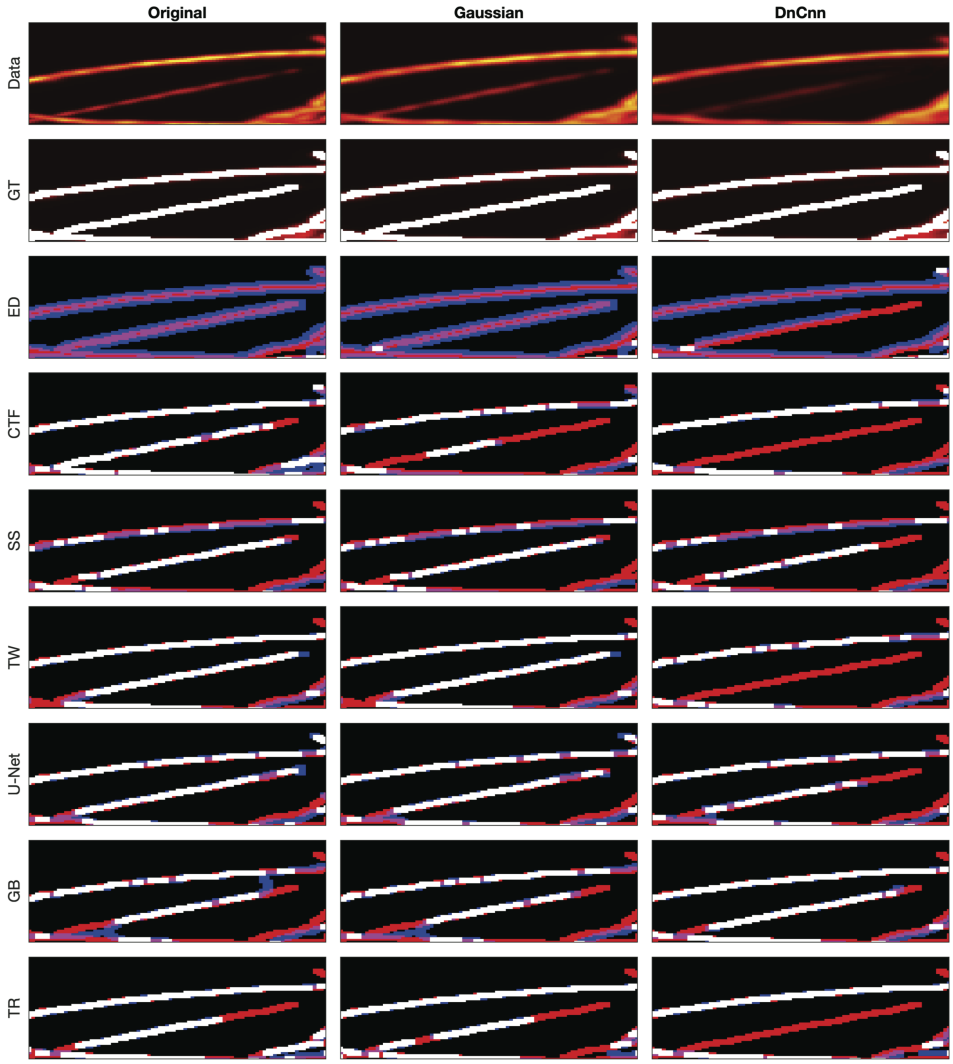


Fig 12. Comparison of tracing algorithms on SHG image (region of interest) [40]. Image was applied to algorithms with three different filters: Original/no filtering, Gaussian and DnCnn. Data corresponds to the images traced; GT is the manually delineated ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is depicted in blue and areas where they overlap is white.

https://doi.org/10.1371/journal.pone.0320006.g012

white lines in (Fig 15). However, DnCnn leads to spurious lines. U-Net performs well, especially on the original image, but shows degraded performance on filtered images due to training on non-filtered data. The Graph based algorithm demonstrates robust performance, further improved with filtering, as seen in (Fig 15), where DnCnn enhances accuracy and cleanliness. Trace Ridges, akin to Scale Space, exhibits fewer traces, deliberately separating intersecting fibres for improved accuracy. Filtering enhances Trace Ridges' performance, particularly evident in (Fig 15).

Conclusions

In this paper we introduced *Trace Ridges*, a novel method for tracing fibre-like structures, compared against six existing methods (Edge Detection, CT Fire, Scale Space, Twombli,

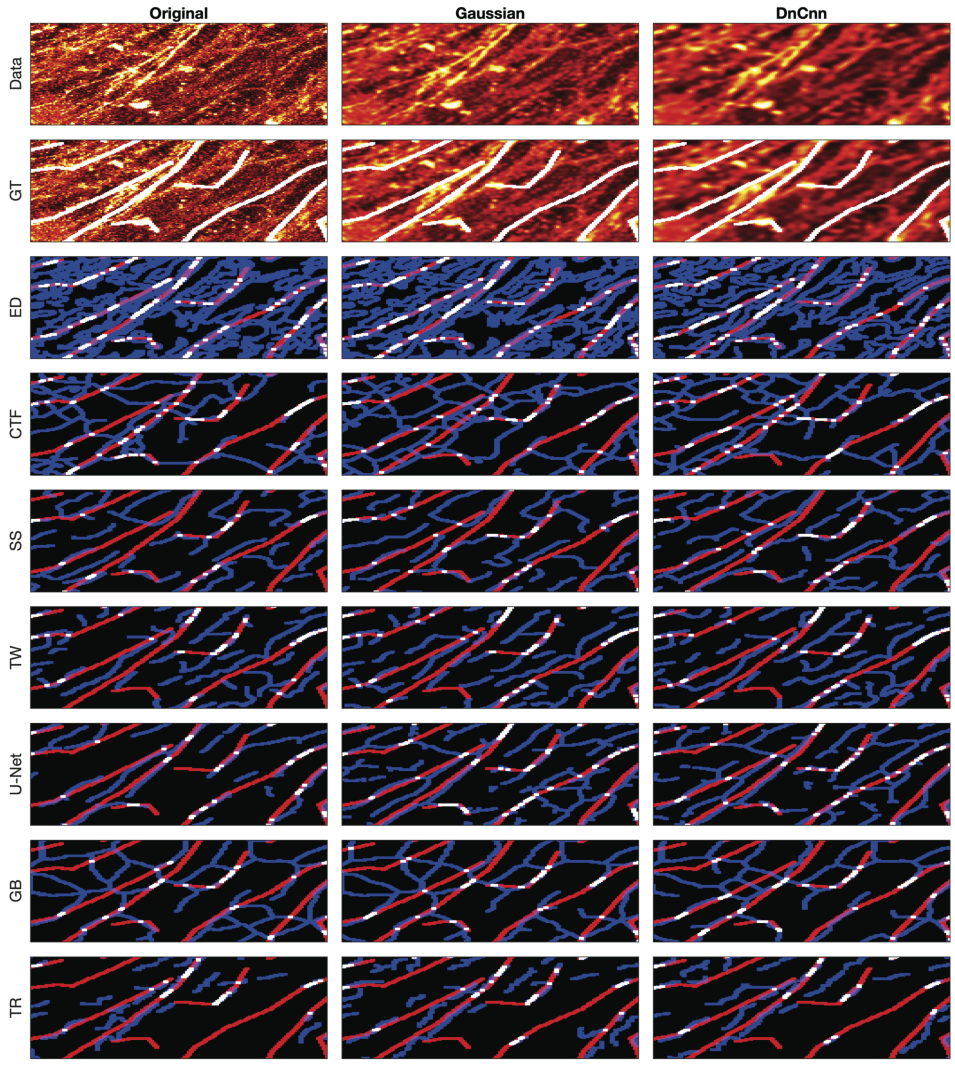


Fig 13. Comparison of tracing algorithms on FF image (region of interest). Image was applied to algorithms with three different filters: Original/no filtering, Gaussian and DnCnn. Data corresponds to images traced; GT is the manually delineated ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is depicted in blue and areas where they overlap is white.

https://doi.org/10.1371/journal.pone.0320006.g013

U-Net and Graph based) with three filtering options (no filtering, Gaussian filtering and DnCnn). Gaussian filtering is a traditional filtering technique that uses Gaussian kernels while DnCnn is a deep neural network. Images representing SHG of tumor bearing mouse mammary glands images, fluorescently labelled fibronectin images, Picrosirius red breast cancer collagen fibres images and disease mimicking ECM images. These images were used for evaluation, compromising both noisy and clean samples.

Manual ground truth delineations were created for each image and distance maps were used to quantify distance errors between the ground truth and algorithmic traces.

Trace Ridges exhibited the lowest total distance error, followed by Graph based and U-Net. Edge Detection achieved the lowest average distance error, attributed to its tendency to oversegment, thus its traces are always close to a fibre from the ground truth. Edge Detection was

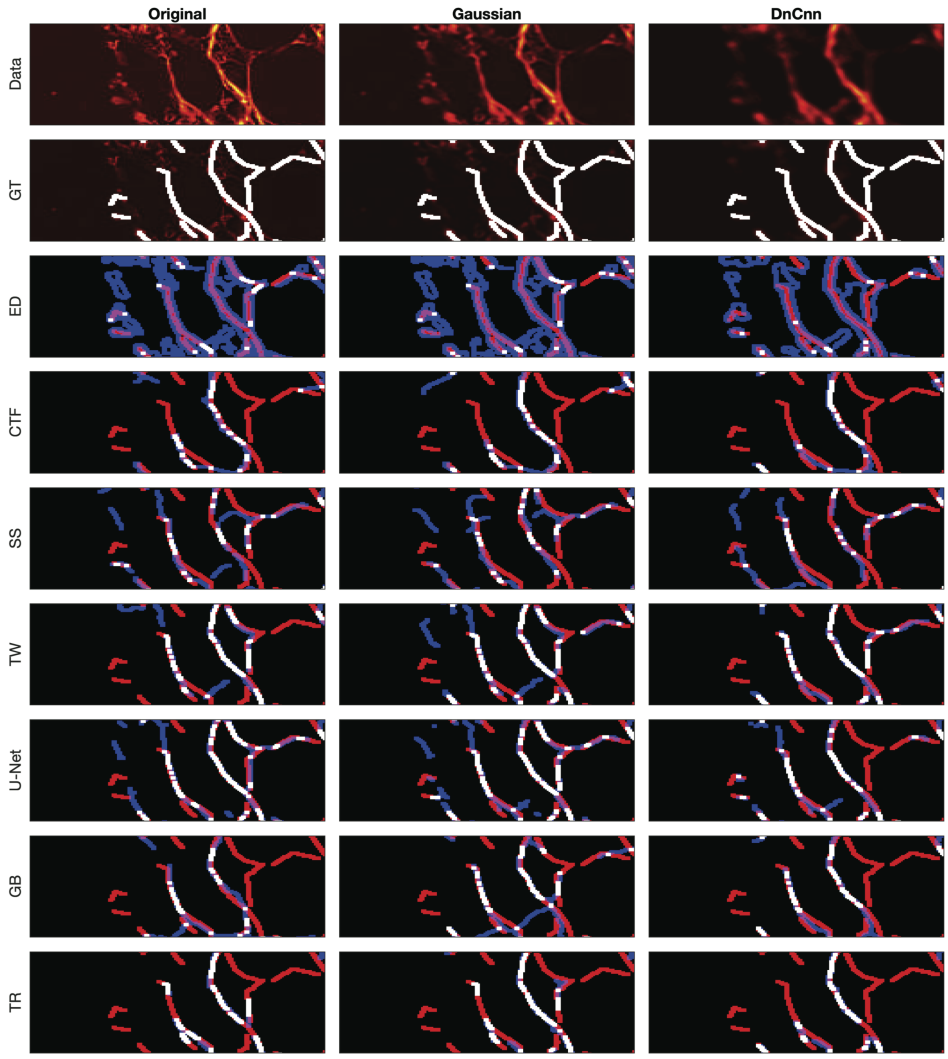


Fig 14. Comparison of tracing algorithms on Picrosirius red Breast Cancer Biopsy image (region of interest) [38]. Image was applied to algorithms with three different filters: Original/no filtering, Gaussian and DnCnn. Data corresponds to images traced; GT is the manually delineated ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is depicted in blue and areas where they overlap is white.

https://doi.org/10.1371/journal.pone.0320006.g014

followed by Trace Ridges and Twombli. U-Net attained the lowest maximum distance error followed by Graph based and Edge Detection. Trace Ridges did not perform well in this metric as it favours brighter and longer fibres, so it is prone to have a higher singular maximum error.

Interestingly, filtering improved results for all algorithms except U-Net, which was trained on non-filtered patches. However, filtering enhanced U-Net's performance specifically with the Breast Cancer Biopsy image.

Edge Detection was the fastest on average, followed by Trace Ridges and U-Net. Filtering generally expedited algorithms except for U-Net, although it improved processing time with the Breast Cancer Biopsy image.

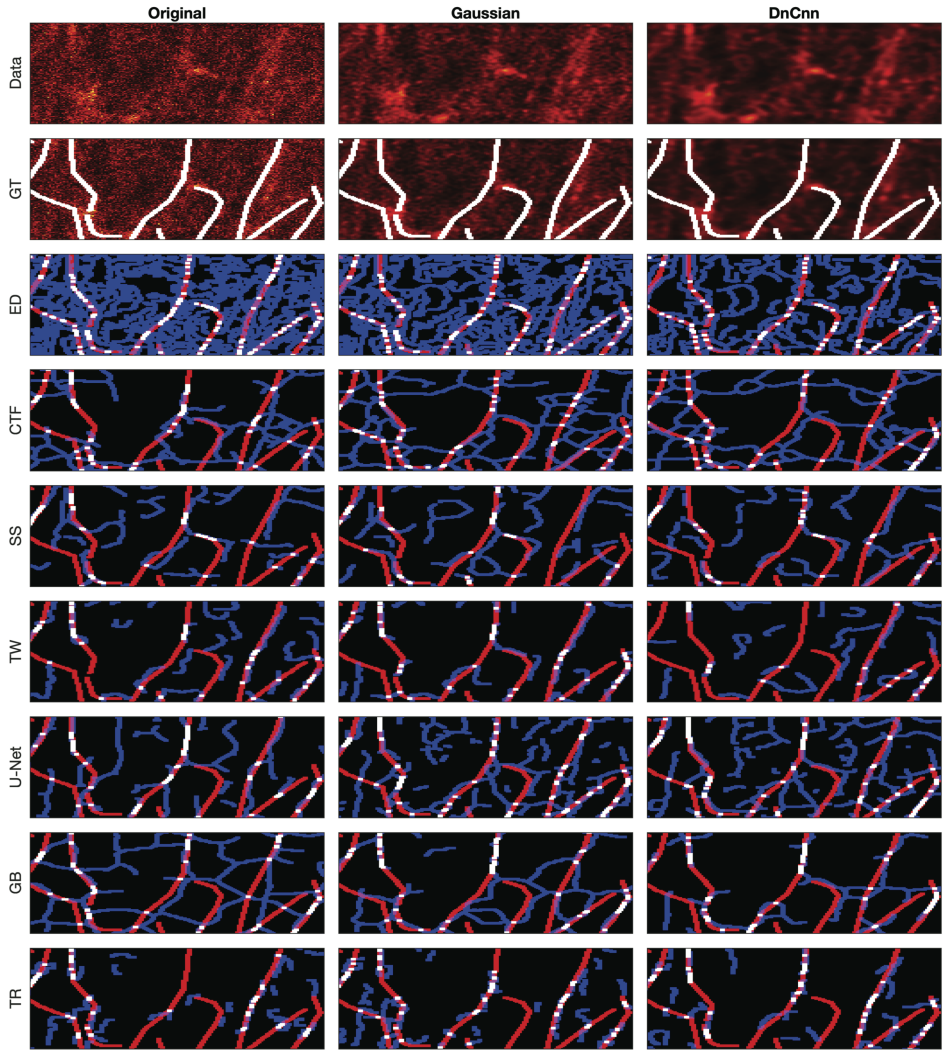


Fig 15. Comparison of tracing algorithms on DME image (region of interest) [42]. Image was applied to algorithms with three different filters: Original/no filtering, Gaussian and DnCnn. Data corresponds to images traced; GT is the manually delineated ground truth. For the evaluation of Tracing Algorithms, the GT fibres are represented in red, algorithm tracing is depicted in blue and areas where they overlap is white.

https://doi.org/10.1371/journal.pone.0320006.g015

Overall, Trace Ridges demonstrated superior performance, achieving the lowest total distance error, second-lowest average distance error, and second-fastest processing time. It offers rapid, automated tracing and extracts essential topological and morphological information, including fibre length, fibre count, orientation, circularity, and gap characteristics. Thus, Trace Ridges emerges as the preferred algorithm for tracing various fibre-like structures, including extracellular matrix and collagen fibres.

Author contributions

Conceptualization: Youssef Arafat, Constantino Carlos Reyes-Aldasoro.

Data curation: Cristina Cuesta-Apausa, Esther Castellano.

Formal analysis: Youssef Arafat, Cristina Cuesta-Apausa, Esther Castellano, Constantino Carlos Reyes-Aldasoro.

Software: Youssef Arafat.

Supervision: Constantino Carlos Reyes-Aldasoro.

Writing – original draft: Youssef Arafat, Constantino Carlos Reyes-Aldasoro.

Writing – review & editing: Cristina Cuesta-Apausa, Esther Castellano.

References

1. Hu M, Li M, Huang H, Lu C. Isolated cancer stem cells from human liver cancer: morphological and functional characteristics in primary culture. Clin Transl Oncol. 2022;24(1):48–56. https://doi.org/10.1007/s12094-021-02667-w PMID: 34169442
2. Fukumura D, Duda DG, Munn LL, Jain RK. Tumor microvasculature and microenvironment: novel insights through intravital imaging in pre-clinical models. Microcirculation. 2010;17(3):206–25. https://doi.org/10.1111/j.1549-8719.2010.00029.x PMID: 20374484
3. Popova NV, Jücker M. The functional role of extracellular matrix proteins in cancer. Cancers (Basel). 2022;14(1):238. https://doi.org/10.3390/cancers14010238 PMID: 35008401
4. Winkler J, Abisoye-Ogunniyan A, Metcalf KJ, Werb Z. Concepts of extracellular matrix remodelling in tumour progression and metastasis. Nat Commun. 2020;11(1):5120. https://doi.org/10.1038/s41467-020-18794-x PMID: 33037194
5. Huang J, Zhang L, Wan D, Zhou L, Zheng S, Lin S, et al. Extracellular matrix and its therapeutic potential for cancer treatment. Signal Transduct Target Ther. 2021;6(1):153. https://doi.org/10.1038/s41392-021-00544-0 PMID: 33888679
6. Walker C, Mojares E, Del Río Hernández A. Role of extracellular matrix in development and cancer progression. Int J Mol Sci. 2018;19(10):3028. https://doi.org/10.3390/ijms19103028 PMID: 30287763
7. Iozzo RV, Gubbiotti MA. Extracellular matrix: the driving force of mammalian diseases. Matrix Biol. 2018;71–72:1–9. https://doi.org/10.1016/j.matbio.2018.03.023 PMID: 29625183
8. Frantz C, Stewart KM, Weaver VM. The extracellular matrix at a glance. J Cell Sci. 2010;123(Pt 24):4195–200. https://doi.org/10.1242/jcs.023820 PMID: 21123617
9. Lu P, Takai K, Weaver VM, Werb Z. Extracellular matrix degradation and remodeling in development and disease. Cold Spring Harb Perspect Biol. 2011;3(12):a005058. https://doi.org/10.1101/cshperspect.a005058 PMID: 21917992
10. Theocharis AD, Skandalis SS, Gialeli C, Karamanos NK. Extracellular matrix structure. Adv Drug Deliv Rev. 2016;97:4–27. https://doi.org/10.1016/j.addr.2015.11.001 PMID: 26562801
11. Filipe EC, Chitty JL, Cox TR. Charting the unexplored extracellular matrix in cancer. Int J Exp Pathol. 2018;99(2):58–76. https://doi.org/10.1111/iep.12269 PMID: 29671911
12. Yue B. Biology of the extracellular matrix: an overview. J Glaucoma. 2014;23(8 Suppl 1):S20-3. https://doi.org/10.1097/IJG.0000000000000108 PMID: 25275899
13. Pompili S, Latella G, Gaudio E, Sferra R, Vetuschi A. The charming world of the extracellular matrix: a dynamic and protective network of the intestinal wall. Front Med (Lausanne). 2021;8:610189. https://doi.org/10.3389/fmed.2021.610189 PMID: 33937276
14. Blum EK. The computation of eigenvalues and eigenvector of a completely continuous self-adjoint operator. J Comput Syst Sci. 1967;1(4):362–70. https://doi.org/10.1016/s0022-0000(67)80026-6
15. Blum H, Nagel RN. Shape description using weighted symmetric axis features. Pattern Recognition. 1978;10(3):167–80. https://doi.org/10.1016/0031-3203(78)90025-0
16. Saha PK, Borgefors G, Sanniti di Baja G. A survey on skeletonization algorithms and their applications. Pattern Recognit Lett. 2016;76:3–12. https://doi.org/10.1016/j.patrec.2015.04.006
17. Meyer SBF. The morphological approach to segmentation: the watershed transformation. In: Mathematical morphology in image processing. CRC Press; 1992. p. 433–81.
18. Meyer F. Topographic distance and watershed lines. Signal Processing. 1994;38(1):113–25. https://doi.org/10.1016/0165-1684(94)90060-4
19. Fritsch DS, Pizer SM, Morse BS, Eberly DH, Liu A. The multiscale medial axis and its applications in image registration. Pattern Recognit Lett. 1994;15(5):445–52. https://doi.org/10.1016/0167-8655(94)90135-x

20. Preparata FP. The medial axis of a simple polygon. In: Mathematical foundations of computer science 1977. Berlin, Heidelberg: Springer; 1977. p. 443–50.
21. Wang S, Rosenfeld A, Wu AY. A medial axis transformation for grayscale pictures. IEEE Trans Pattern Anal Mach Intell. 1982;4(4):419–21. https://doi.org/10.1109/tpami.1982.4767274 PMID: 21869057
22. Lindeberg T. Feature detection with automatic scale selection. Int J Comp Vision. 1998;30(2):79–116.
23. Lindeberg T. Edge detection and ridge detection with automatic scale selection. Int J Comp Vision. 1998;30(2):117–54.
24. Canny J. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell. 1986;8(6):679–98. PMID: 21869365
25. Meijering E. Neuron tracing in perspective. Cytometry A. 2010;77(7):693–704. https://doi.org/10.1002/cyto.a.20895 PMID: 20583273
26. Acciai L, Soda P, Iannello G. Automated neuron tracing methods: An updated account. Neuroinformatics. 2016;14(4):353–67.
27. Liu Y, Wang G, Ascoli GA, Zhou J, Liu L. Neuron tracing from light microscopy images: automation, deep learning and bench testing. Bioinformatics. 2022;38(24):5329–39. https://doi.org/10.1093/bioinformatics/btac712 PMID: 36303315
28. Delibasis KK, Kechriniotis AI, Tsonos C, Assimakis N. Automatic model-based tracing algorithm for vessel segmentation and diameter estimation. Comput Methods Programs Biomed. 2010;100(2):108–22. https://doi.org/10.1016/j.cmpb.2010.03.004 PMID: 20363522
29. Mahadevan V, Narasimha-Iyer H, Roysam B, Tanenbaum HL. Robust model-based vasculature detection in noisy biomedical images. IEEE Trans Inf Technol Biomed. 2004;8(3):360–76. https://doi.org/10.1109/titb.2004.834410 PMID: 15484442
30. Valiente-Soriano FJ, Salinas-Navarro M, Di Pierdomenico J, García-Ayuso D, Lucas-Ruiz F, Pinilla I, et al. Tracing the retina to analyze the integrity and phagocytic capacity of the retinal pigment epithelium. Sci Rep. 2020;10(1):7273. https://doi.org/10.1038/s41598-020-64131-z PMID: 32350384
31. Koziński M, Mosinska A, Salzmann M, Fua P. Tracing in 2D to reduce the annotation effort for 3D deep delineation of linear structures. Med Image Anal. 2020;60:101590. https://doi.org/10.1016/j.media.2019.101590 PMID: 31841949
32. Yuanzhi Cheng, Xin Hu, Ji Wang, Yadong Wang, Tamura S. Accurate vessel segmentation with constrained B-snake. IEEE Trans Image Process. 2015;24(8):2440–55. https://doi.org/10.1109/TIP.2015.2417683 PMID: 25861085
33. Wei W, Zhang Q, Rayner SG, Qin W, Cheng Y, Wang F, et al. Automated vessel diameter quantification and vessel tracing for OCT angiography. J Biophotonics. 2020;13(12):e202000248. https://doi.org/10.1002/jbio.202000248 PMID: 32857462
34. Abdul-Karim M-A, Al-Kofahi K, Brown EB, Jain RK, Roysam B. Automated tracing and change analysis of angiogenic vasculature from in vivo multiphoton confocal image time series. Microvasc Res. 2003;66(2):113–25. https://doi.org/10.1016/s0026-2862(03)00039-6 PMID: 12935769
35. Barber PR, Vojnovic B, Ameer-Beg SM, Hodgkiss RJ, Tozer GM, Wilson J. Semi-automated software for the three-dimensional delineation of complex vascular networks. J Microsc. 2003;211(Pt 1):54–62. https://doi.org/10.1046/j.1365-2818.2003.01205.x PMID: 12839551
36. Reyes-Aldasoro CC, Björndahl MA, Akerman S, Ibrahim J, Griffiths MK, Tozer GM. Online chromatic and scale-space microvessel-tracing analysis for transmitted light optical images. Microvasc Res. 2012;84(3):330–9. https://doi.org/10.1016/j.mvr.2012.09.002 PMID: 22982542
37. Wu J, Voytik-Harbin SL, Filmer DL, Hoffman CM, Yuan B, Chiang C-S, et al. Modeling ECM fiber formation: structure information extracted by analysis of 2D and 3D image sets. SPIE Proceedings. 2002;4621:52. https://doi.org/10.1117/12.467831
38. Wershof E, Park D, Barry DJ, Jenkins RP, Rullan A, Wilkins A, et al. A FIJI macro for quantifying pattern in extracellular matrix. Life Sci Alliance. 2021;4(3):e202000880. https://doi.org/10.26508/lsa.202000880 PMID: 33504622
39. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical image computing and computer-assisted intervention – MICCAI 2015. Cham: Springer; 2015. p. 234–41.
40. Bredfeldt JS, Liu Y, Pehlke CA, Conklin MW, Szulczewski JM, Inman DR, et al. Computational segmentation of collagen fibers from second-harmonic generation images of breast cancer. J Biomed Opt. 2014;19(1):16007. https://doi.org/10.1117/1.JBO.19.1.016007 PMID: 24407500
41. Grapa AI. Characterization of fibronectin networks using graph-based representations of the fibers from 2D confocal images [PhD thesis]. Universite Cote d'Azur. Nice, France; 2020.

42. Grapa A-I, Efthymiou G, Van Obberghen-Schilling E, Blanc-Féraud L, Descombes X. A spatial statistical framework for the parametric study of fiber networks: application to fibronectin deposition by normal and activated fibroblasts. Biol Imaging. 2023;3:e25. https://doi.org/10.1017/S2633903X23000247 PMID: 38510171

43. Liu Y, et al. Fibrillar collagen quantification with curvelet transform based computational methods. Front Bioeng Biotechnol. 2020;8.

44. Zhang K, Zuo W, Chen Y, Meng D, Zhang L. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans Image Process. 2017;26(7):3142–55. https://doi.org/10.1109/TIP.2017.2662206 PMID: 28166495

45. Candès E, Demanet L, Donoho D, Ying L. Fast discrete curvelet transforms. Multiscale Model Simul. 2006;5(3):861–99. https://doi.org/10.1137/05064182x

46. Steger C. An unbiased detector of curvilinear structures. IEEE Trans Pattern Anal Machine Intell. 1998;20(2):113–25. https://doi.org/10.1109/34.659930

47. Barry DJ, Williams GA, Chan C. Automated analysis of filamentous microbial morphology with AnaMorf. Biotechnol Prog. 2015;31(3):849–52. https://doi.org/10.1002/btpr.2087 PMID: 25864556

48. Frucci M. Oversegmentation reduction by flooding regions and digging watershed lines. Int J Patt Recogn Artif Intell. 2006;20(01):15–38. https://doi.org/10.1142/s0218001406004533

49. Frucci M, Baja GS di. Oversegmentation reduction in watershed-based grey-level image segmentation. IJSISE. 2008;1(1):4. https://doi.org/10.1504/ijsise.2008.017767

50. Pratt W. Digital image processing. Wiley; 1978. Available from: https://www.amazon.co.uk/Digital-Image-Processing-William-Pratt/dp/0471018880

nature communications

Article

https://doi.org/10.1038/s41467-025-56299-7

Matrix mechano-sensing at the invasive front induces a cytoskeletal and transcriptional memory supporting metastasis

Received: 30 October 2024

Accepted: 13 January 2025

Published online: 14 February 2025

Check for updates

Oscar Maiques[1,2,3,22], Marta C. Sallan[1,4,23], Roman Laddach[5,6,23], Pahini Pandya[3,23], Adrian Varela[1,2], Eva Crosas-Molist[1,2,3], Jaume Barcelo[1,2], Olivia Courbot[7], Yanbo Liu[1], Vittoria Graziani[1,2], Youssef Arafat[8], Joanne Sewell[2,3], Irene Rodriguez-Hernandez[2,3], Bruce Fanshawe[3,9], Yaiza Jung-Garcia[1,2], Paul RC Imbert[10], Eloise M. Grasset[11], Jean Albrengues[11], Maria Santacana[12], Anna Macià[13], Jordi Tarragona[12], Xavier Matias-Guiu[12,13,14], Rosa M. Marti[15], Sophia Tsoka[6], Cedric Gaggioli[11], Jose L. Orgaz[2,3,16], Gilbert O. Fruhwirth[9], Fredrik Wallberg[17,18], Kai Betteridge[18], Constantino Carlos Reyes-Aldasoro[8,19], Syed Haider[20], Andrejs Braun[4], Sophia N. Karagiannis[5,21], Alberto Elosegui-Artola[8] & Victoria Sanz-Moreno[1,2,3]

The extracellular matrix (ECM) controls tumour dissemination. We characterise ECM organization in human and mouse tumours, identifying three regions: tumour body, proximal invasive front and distal invasive front. Invasive areas show increased matrix density, fibre thickness, length, and alignment, with unique radial fibre orientation at the distal invasive front correlating with amoeboid invasive features. Using patient samples and murine models, we find that metastases recapitulate ECM features of the primary tumour. Ex vivo culture of murine cancer cells isolated from the different tumour regions reveals a spatial cytoskeletal and transcriptional memory. Several in vitro models recapitulate the in vivo ECM organisation showing that increased matrix induces 3D confinement supporting Rho-ROCK-Myosin II activity, while radial orientation enhances directional invasion. Spatial transcriptomics identifies a mechano-inflammatory program associated with worse prognosis across multiple tumour types. These findings provide mechanistic insights into how ECM organization shapes local invasion and distant metastasis.

Solid tumours are characterized by heterogeneous populations of cancer cells and their tumour microenvironment (TME) composed of fibroblasts, immune cells, vessels, and the extracellular matrix (ECM)[1]. ECM is composed of fibrillar collagens, fibronectin, elastin, and laminins[2]. Aberrant ECM deposition promotes tumour cell proliferation, invasion and metastasis[3]. Homeostatic control of the ECM is compromised during cancer progression, and tumour cells become exposed to a collagen-rich stroma. Besides collagen accumulation, matrix is differentially organized in pathological conditions such as cancer, aging and fibrosis[4–7]. Furthermore, aligned ECM fibres generate tracks for the invasion of tumour cells[8,9] and collagen fibre organization correlates with disease onset and progression[10]. Tumour-

A full list of affiliations appears at the end of the paper. e-mail: [redacted]

Associated Collagen Signatures (TACS) have revealed that collagen structures differ according to location in the tumour, presenting higher levels of alignment at the tumour edge, enabling for contact guidance directed cell migration[8,11,12]. Nevertheless, how disseminating cancer cells sense ECM features at the invasive front and translate them into pro-metastatic signals is unclear.

Fibre alignment correlates with increased tissue stiffness and tumour cell invasion[8,13,14]. While sensing localised changes in substrate stiffness surrounding the cell is enabled by cell adhesions and is crucial for translating cytoskeletal forces into motion[15], it is the actomyosin machinery that is responsible for the determination of overall stiffness and adoption of different migratory modes in response to varying ECM stiffness[15]. Moderate Myosin II activity is required for actin-rich protrusions and retraction of the cell rear, while higher Rho-ROCK-Myosin II engagement drives rounded-amoeboid migration requiring lower levels of adhesion[16]. Many chemical stimuli such as cytokines and growth factors support amoeboid behaviour[17–21]. However, physical confinement coupled to lower levels of adhesion induced fast amoeboid bleb-based migration[16,22,23]. Such in vitro studies led to the hypothesis that confinement imposed by surrounding tissue on a growing tumour could trigger amoeboid migration, but the mechanism by which physical confinement imposed by the matrix at the edge of tumours induces amoeboid invasion is not fully understood.

In this study, we describe three distinct ECM areas within Melanoma and Breast cancer primary tumours, each characterized by different properties that influence cancer cell cytoskeleton and behaviour. Crucially, ECM distinct organisation at these areas supports a transcriptional program in cancer cells that is important for local invasion in primary lesions and for distant metastases formation. Our findings reveal a complex interplay between ECM structure and cancer cell transcriptional responses.

Results

Regional characterisation of the extracellular matrix in solid tumours

We combined digital pathology and image analysis to study matrix organisation using Trichrome staining[24–26] and second harmonic generation (SHG)[27] microscopy in tissue samples. Several tissue-microarrays of tumour sections from melanoma and breast cancer (BRCA) were constructed and divided in different areas: the centre or tumour body (TB) and the invasive front (IF)-area of the tumour in contact with the matrix (Supplementary Data 1-3). Within this invasive region we could further sub-classify the matrix into two separate areas: proximal invasive front (PIF) comprising between 50-100 µm of boundary with the extracellular matrix; and the distal invasive front (DIF) 100-200 µm further away from the PIF (Supplementary Table 1).

QuPath[24] was used to assess global changes in ECM organization in the different areas (Supplementary Table 2, See methods section). In melanoma, the TB showed the lowest percentage of matrix, and this matrix was less intense and less anisotropic measured by Haralick Entropy[28], while both invasive areas had much higher levels of all parameters (Fig. 1b, c). Matrix topography was assessed using curvelet transform algorithm[29] (Supplementary Table 2). We observed that fibres in PIF and DIF were significantly longer, thicker, straighter, showed reduced mean distance between fibres, and were more aligned compared to TB. Additionally, the fibre density of TB was significantly lower (Fig. 1d, e). Importantly, fibres were oriented parallel to the tumour boundary in the PIF while they were oriented perpendicularly or radially to the tumour border in the DIF (Fig. 1f). We next analysed matrix organization in BRCA samples and found these three separate zones with the same pattern, matrix abundance and topography we had found in melanoma (Fig. 1g–l). We validated these observations using SHG imaging and curvelet transform analysis in the same human tissue samples (Supplementary Fig. 1a-d). Collagen-I staining showed correlation in the areas of high matrix density and high fibre alignment such as PIF and DIF, while collagen IV was only observed in the basement membranes of vessels and not in invasive regions (Supplementary Fig. 1e, f).

Overall, we observe similar regional ECM organisation in melanoma and breast cancer tissues: TB with low matrix abundance/low fibre alignment; PIF with high matrix abundance/high fibre alignment and parallel angles, and DIF with intermediate matrix abundance/high fibre alignment but perpendicular angles. Furthermore, this data shows that the key difference in the invasive areas is the orientation of the fibres at the DIF which could be steering invasive processes in cancer cells leaving the tumour mass.

Regional organisation of ECM is associated with cytoskeletal remodelling, local invasion and metastatic spread in human tumours

We next investigated if ECM regional distribution was associated with specific invasion patterns, cytoskeletal characteristics, and metastatic spread (Supplementary Table 1). A gradual increase in circularity and pMLC2 (MLC2 activity) was observed from TB to PIF and then even further to DIF (Fig. 2a, b). Interestingly, we did not observe changes in total MLC2 levels (Fig. 2b). Both groups of cells and individual cancer cells invading were observed at PIF, but amoeboid single cell invasion was highest at the DIF- which also correlated with highest invasion levels (Fig. 2c–e, Supplementary Fig. 2a).

Very similar observations were made regarding cytoskeletal changes and invasive patterns using BRCA tissues (Fig. 2f–i, Supplementary Fig. 2b), while we observed that groups of cells were larger than in melanoma. Moreover, increased amoeboid invasion correlated with higher disease stage in melanoma and lymph node engagement in breast cancer (Fig. 2k).

These results suggest that the three different ECM regions are associated with specific invasive patterns and cytoskeletal rearrangements in melanomas and breast cancers. Importantly the PIF to DIF area selects for amoeboid cytoskeletal characteristics that we find in association with specific matrix topographies, invasive patterns (Fig. 2l), and lymph node metastasis.

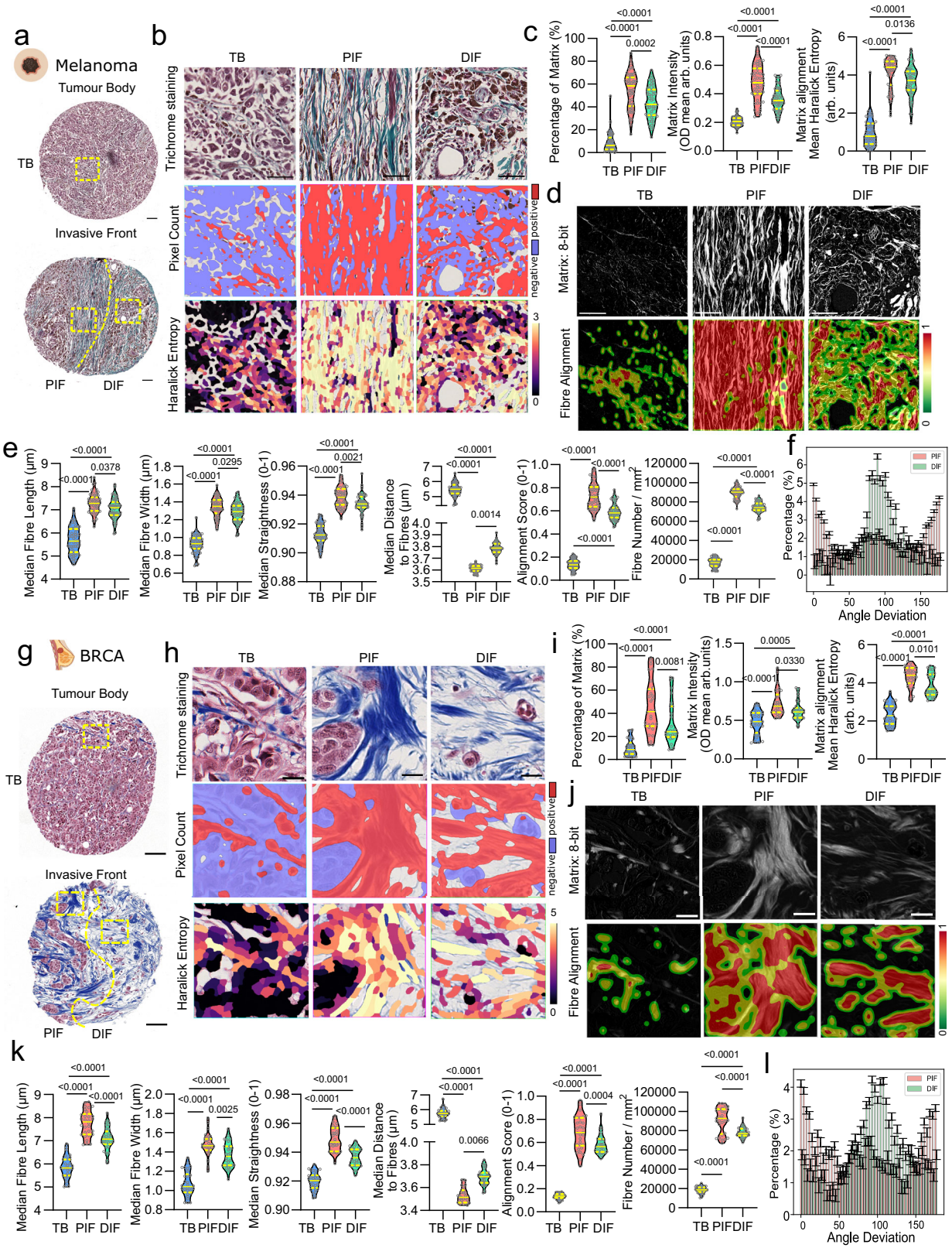
ECM in vivo organisation is associated with a cytoskeletal memory

To study the spatial temporal relationship between matrix organisation and cytoskeletal dynamics, we used 4599 $BRAF^{V600E}$ mouse melanoma cells[30] injected ortho-topically into mouse dermis that develop lung metastasis after 30 days (both micro- and macro-metastasis) (Fig. 3a).

ECM organisation at the primary murine tumour showed a higher percentage, intensity, and alignment of matrix at the PIF and DIF compared to the TB (Fig. 3b, c). Matrix characteristics in this mouse model (Supplementary Fig. 2c) recapitulated those observed in primary human cancer tissues (Fig. 1). Importantly, similar angle orientations at DIF and PIF were observed when comparing fibres in murine tumours (Fig. 3d) to human tumours (Fig. 1f and 1l). Increased overall invasion and individual cell invasion was observed at the DIF of murine tumours (Fig. 3e) as observed in human tumours (Fig. 2).

In the metastatic tumours, we defined micro-metastasis as lesions smaller than ~0.02 mm^2 and found that they occupied 5–30% of metastatic lesions (Supplementary Fig. 2d). Metastasis larger than this size developed a TB and an IF and were classified as macro-metastasis. We could not detect a DIF in the macro-metastasis, likely requiring longer periods of time to invade further away. Lung micro- and macro-metastasis were characterised by different matrix characteristics (Fig. 3b, c). In macro-metastasis: there was a higher percentage of matrix at the IF, it was more intense and more aligned (Fig. 3b, c; Supplementary Fig. 2e) compared to the matrix at the TB, very much resembling the primary tumour. Importantly, micro-metastasis and the macro-metastasis IF presented similar features. We next analysed

https://doi.org/10.1038/s41467-025-56299-7



tissues from human melanoma metastasis patients and observed similar matrix organisation patterns when compared to the primary lesions. Indeed, different regional organization was observed at the IF of metastasis compared to the TB (Supplementary Fig. 3a–d). We confirmed that fibres at the IF of metastatic lesions are longer, thicker, more compact and aligned compared to their TB (Supplementary Fig. 3a–d). These data shows that both human and mouse tumours show matrix topographical features in the metastatic lesions that resemble those found in the primary tumour.

We then quantified the 4599 mCherry[31] morphology and pMLC2 and MLC2 levels in tissues derived from both the primary lesions and the lung metastasis using multiplex IHC[32] (Fig. 3f). We observed a gradual increase in cell rounding and pMLC2 when comparing the TB to the PIF, and then to the DIF of 4599 primary tumours (Fig. 3f, g;

Fig. 1 | Regional organisation of the extracellular matrix in solid tumours. **a** TMA cores from Tumour Body (TB) and Invasive Front (IF) of primary melanoma ($n = 53$), stained with Gomori trichrome. Insets show regions of interest (ROIs). Yellow lines separate proximal (PIF) and distal invasive fronts (DIF). Scale: 200 μm. **b** ROIs from TB, PIF, DIF: trichrome staining, QuPath Pixel Count maps, Haralick Entropy alignment maps. Scale: 50 μm. **c** Quantification of Matrix Percentage (%), Matrix Intensity (OD Mean arb. units), and Matrix Alignment (Haralick Entropy arb. units) in melanoma ($n = 53$ patients). **d** ROIs from TB, PIF, DIF: 8-bit matrix images and CurveAlign alignment maps. Scale: 50 μm. **e** ROI quantification: Median Fibre Length, Width, Straightness, Distance to Fibres, Alignment Score, Fibre number/mm² ($n = 53$ patients). **f** Angle deviation histogram comparing PIF and DIF ($n = 53$ patients). **g** TMA cores from TB and IF of primary breast cancer (BRCA) ($n = 30$ patients), Masson trichrome stained. Insets show ROIs. Yellow lines separate PIF from DIF. Scale: 100 μm. **h** BRCA ROIs: trichrome staining, QuPath Pixel Count maps, Haralick Entropy alignment maps. Scale: 20 μm. **i** Quantification of Matrix Percentage (%), Matrix Intensity (OD Mean arb. units), and Matrix Alignment (Haralick Entropy arb. units) in BRCA ($n = 30$ patients). **j** BRCA ROIs: 8-bit matrix images and CurveAlign alignment maps. Scale: 20 μm. **k** BRCA ROI quantification: Median Fibre Length, Width, Straightness, Distance to Fibres, Alignment Score, Fibre number/mm² ($n = 30$ patients). **l** Angle deviation histogram comparing PIF and DIF in BRCA ($n = 30$ patients). One-way ANOVA with Tukey post-hoc (c, e, i, k), displayed as truncated violin plots showing medians and quartiles. Histograms (f, l) show mean and standard error of the mean (SEM). Cartoons in Fig. 1(a, g) created in BioRender. Maiques, O. (2025) https://BioRender.com/f02u038 https://BioRender.com/n61t687.

Supplementary Fig. 3f) as seen in human tumours. Circularity and pMLC2 levels were very high in micro-metastasis and at the IF of the macro-metastasis. Interestingly, TB of macro-metastasis were less rounded and with lower levels of pMLC2 and no changes in total MLC2 (Fig. 3f, g. Supplementary Fig. 3e). This data shows that this murine model can be used to study spatial-temporal dynamics of matrix and cytoskeleton both at the primary and at the metastatic site.

Next, we investigated the possibility that cells grown ex vivo could retain a cytoskeletal memory associated with their previous location within the tumour's heterogeneous topography. We isolated 4599 melanoma cells from the TB, PIF, and DIF in the primary tumour and from the spontaneous metastasis (MET) in the lungs. When grown on collagen ex vivo, we observed that melanoma cells retained the cytoskeletal features we observed in the tumours- that is increased cell rounding and pMLC2 levels in cells from the invasive areas and from the lung metastasis (Fig. 3h, i). We then grew tumour-spheroids mixing equal numbers of cells (labelled with different CellTracker dyes) extracted from the different tumour areas (cells from the TB mixed with cells from the DIF or cells from the TB mixed with cells from the MET). Both cells derived from DIF and MET outcompeted cells derived from the TB in 3D invasion assays (Fig. 3j). We next argued that cells derived from the DIF- that retain cytoskeletal amoeboid features and invasive properties- are those likely to reach the vasculature and arrive to the lungs. To test if they would also survive better in the lung micro-environment, we co-injected equal numbers of TB and DIF derived cells using tail vein injection. Importantly, cells derived from DIF outperformed cells derived from the TB and survived better in the lung parenchyma after 48 h, while no difference was observed at 30 min post injection (Fig. 3k, Supplementary Fig. 3h, i). These data indicate that cancer cells extracted from the DIF retain amoeboid cytoskeletal features and pro-survival ability when they reach the lung parenchyma.

Overall, these data suggest strong associations between matrix organisation, cytoskeletal remodelling, and invasive and survival properties during metastasis.

Collagen I accumulation induces cytoskeletal remodelling

We next hypothesised that blocking ECM degradation in vitro would lead to increased local matrix density as observed in the invasive front of tumours in vivo. Addition of protease inhibitors (PIs) to cells grown in collagen I matrices resulted in increased length, width, straightness, and alignment of fibres, while reducing the distance between them (Fig. 4a, b) as seen in vivo. Both collagen I fragments and MMP activity decreased with PI treatment, indicating matrix accumulation around cells (Supplementary Fig. 4a–c). These changes in organisation led to increased matrix confinement, as shown by gap analysis (Fig. 4c). Using melanoma (4599, A375P, SKMEL28), fibrosarcoma (HT1080) and breast cancer (MDA-MB-231) cell lines, we observed an increase in amoeboid cytoskeletal features when cells grown on collagen I were treated with PI (Fig. 4d, e, Supplementary Fig. 4d–g). Moreover, 5-day organotypic assays were used to generate an invasive front and measure invasion patterns. This approach revealed an increase in amoeboid single cell invasion after PI treatment (Supplementary Fig. 4h).

To further model in vitro the increase in matrix density we observe in vivo, we grew cancer cells in matrices containing increasing amounts of collagen I, that have been shown to display increased stiffness[33,34]. We found that increased collagen concentrations resulted in increased number of fibres, higher thickness of the fibres and lower distance between fibres (Fig. 4f, g), as observed in patients (Fig. 1e, k and Supplementary Fig. 1b, d). This resulted in increased matrix confinement as show by Gap analysis (Fig. 4h). On the other hand, the fibres were shorter and less aligned (Fig. 4g), which has been previously reported for high collagen containing matrices[35].

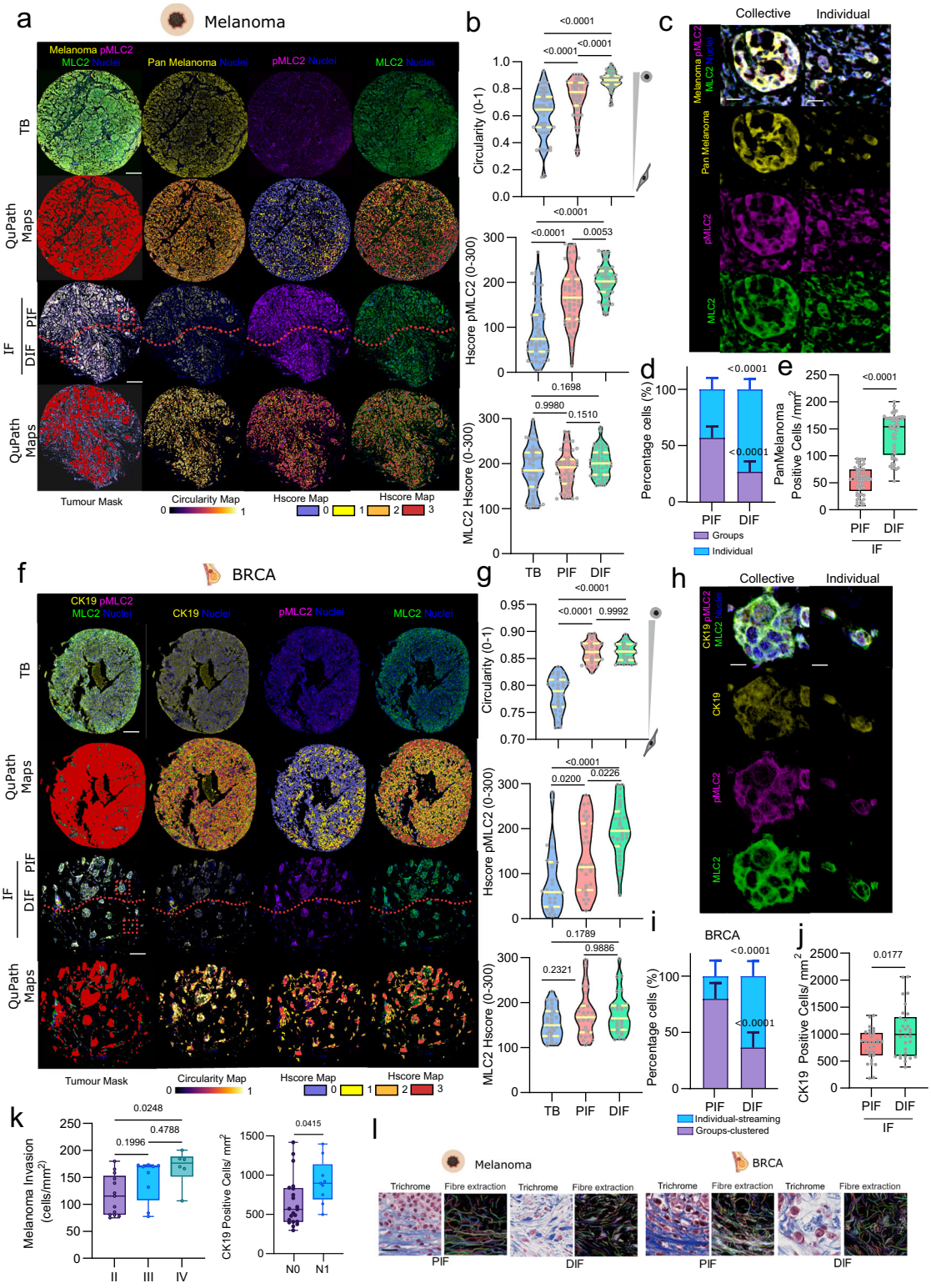
When melanoma and fibrosarcoma cancer cells were grown in increasing collagen I-containing matrices, we observed a gradual increase in cell rounding and pMLC2 levels in higher concentrations (Fig. 4i, Supplementary Fig. 4i) most likely due to the observed increased 3D confinement. Interestingly, the highest collagen concentrations resulted in decreased 3D invasion levels, likely because of shorter and less aligned fibres (Supplementary Fig. 4j).

This data supports the notion that even if matrix induced confinement increases actomyosin levels, an optimal level of collagen I supports invasion in which the fibres should be longer and more aligned.

Collagen I accumulation and radial alignment support Myosin II driven invasive and metastatic behaviour

Matrix alignment coupled to perpendicular fibres is a poor prognostic factor in breast cancer[25]. To test the hypothesis that matrix accumulation has an influence on cytoskeletal features, while the radial alignment supports maximal invasion, we used collagen gels that polymerise following Marangoni flows (Fig. 5a) driven by saturated Magnesium Chloride ($MgCL_2$)[36]. This strategy resulted in increased fibre length, thickness, straightness, density while decreasing the distance between the fibres (Fig. 5b, c, Supplementary Fig. 5a). Importantly fibres were oriented radially towards the centre of the gel (Fig. 5d), therefore recapitulating the topographical features we defined at the distal invasive front in human and mouse tissues (Figs. 1 and 3). These changes in topography also resulted in higher stiffness, as measured by atomic force microscopy (Fig. 5e). Moreover, Gap analysis revealed increased matrix confinement under these conditions (Fig. 5f). Importantly, an increase in roundness and pMLC2 levels was observed when cells were grown in the Marangoni flows-driven matrix both in 4599 and HT1080 cells, (Fig. 5h, i Supplementary Fig. 5b, c). Using tumour spheroids embedded in a collagen matrix, we observed that Marangoni flows gels favoured higher radial invasion in melanoma, fibrosarcoma and breast cancer cells (Fig. 5j, k and Supplementary Fig. 5c). This data shows that Marangoni flows-driven matrices support high actomyosin activity driven by confinement and maximal individual invasion driven by long radially oriented fibres (Supplementary Fig. 5e).

https://doi.org/10.1038/s41467-025-56299-7



We then primed or educated cancer cells by growing them on collagen I polymerised under Marangoni flows or control matrices. After 7 days of priming on collagen I matrices, cancer cells were collected and grown in tissue culture plastic for 6 days (to test if a matrix organisation memory is retained in the absence of the matrix input). Cells derived from Marangoni-priming schedule were mixed with cells primed to control collagen, labelled with different colours and challenged to 3D invasion assays (Fig. 5n). We found that 4599 cells derived from Marangoni priming were more invasive and retained higher pMLC2 levels (Fig. 5n–p). Importantly, when co-injecting a mix of Marangoni-primed with control-primed melanoma cells into the tail vein of mice, we found that that Marangoni priming matrix induced cytoskeletal changes that conferred a survival advantage during lung metastatic colonisation (Fig. 5q–s, Supplementary Fig. 5f).

Overall, our data indicates that matrix organisation alone can induce cytoskeletal changes that are retained over time, generating a cytoskeletal pro-invasive and pro-survival memory.

Fig. 2 | Regional organisation of ECM is associated with cytoskeletal remodelling, local invasion and metastatic spread in human tumours. a Representative multiplex images of Tumour Body (TB), Proximal Invasive Front (PIF), and Distal Invasive Front (DIF) cores in melanoma patients. Images show merged pan melanoma, pMLC2, and MLC2 markers, with individual stainings. Nuclei are stained blue. QuPath maps display tumor mask, circularity, pMLC2 H-score, and MLC2 H-score. Scale bar: 200 μm. **b** Quantification of tumour circularity (0–1), Myosin II activity (pMLC2 H-score, 0–300), and MLC2 H-score (0–300). $n = 53$ patients. **c** Representative images of melanoma invading cells across PIF and DIF. Scale bar: 30 μm. **d** Percentage of individual and grouped cells invading PIF and DIF in melanoma. $n = 53$ patients. **e** Number of invading melanoma cells/mm² in PIF and DIF. $n = 53$ patients per group. **f** Representative images of TB, PIF, and DIF in breast cancer (BRCA). Images show merged CK19, pMLC2, and MLC2 markers, with individual staining. QuPath maps display tumour circularity. Scale bar: 100 μm. **g** Quantification of tumour circularity, Myosin II activity, and MLC2 H-score in BRCA. $n = 30$ patients. **h** Images of BRCA invading cells across PIF and DIF. Scale bar: 20 μm. **i** Percentage of individual and grouped cells invading PIF and DIF in BRCA. $n = 30$ patients. **j** Number of invading BRCA cells/mm² in PIF and DIF. $n = 30$ patients. **k** Melanoma invasion by stage (II, III, IV; $n = 24$ patients) and BRCA invasion by lymph node status (N0, N1; $n = 30$ patients). **l** Trichrome staining and fibre extraction in melanoma and BRCA at PIF and DIF. Scale bar: 30 μm. One-way ANOVA with Tukey post-hoc (b, g, k) displayed as truncated violin plots. Two-way ANOVA with Tukey post-hoc (d, i), and Mann-Whitney U test (e, j). Bar graphs (d, i) show mean percentages with standard deviation error bars. Boxplots (e, j) show medians, interquartile ranges, and individual data points. Cartoons in Fig. 2(a, f, l) created in BioRender. Maiques, O. (2025) https://BioRender.com/f02u038 https://BioRender.com/n61t687.

Matrix organisation changes result in Rho-ROCK engagement and uncoupling of Myosin II from vinculin

Vinculin is required for force transmission between the matrix and the cytoskeleton[37–39] and is necessary for Myosin II-dependent adhesion strength and traction force[40]. Using super-resolution spinning disk microscopy, we analysed endogenous vinculin changes in 4599 or HT1080 cells when challenged to increasing collagen concentration or PIs. We observed that overall signal of vinculin structures decreased progressively when comparing 1.7 to 3.5 and then to 6.6 mg/ml collagen-I (Fig. 6a–c; Supplementary Fig. 6a, b) as well as overall protein levels (Fig. 6c). A drop in vinculin total levels as measured by immunofluorescence and western blot was also apparent while cancer cells where rounding-up after protease inhibition (Fig. 6d, e; Supplementary Fig. 6a, b). Furthermore, we analysed myosin light chain 2 and vinculin dynamics at actin rich protrusions. We could observe pMLC2/Vinculin signal at actin rich protrusions either in collagen 1.7 mg/ml or untreated 4599 and HT1080 cells (Fig. 6a, b, Supplementary Fig. 6a, b). However, in higher collagen concentrations (3.5, 6.6 mg/ml) or in the presence of PIs (4 h, 24 h) while protrusions had retracted and even disappeared, we observed an increase in pMLC signal vs vinculin signal in the actin cortex (Fig. 6d, e, Supplementary Fig. 6a, b). This data suggests that matrix accumulation results in uncoupling of vinculin from cell matrix adhesions/protrusions, while active MLC2 becomes cortical. Importantly, initial cell body retraction was associated with cells establishing fewer adhesions, indicated by decreased points of attachment to the matrix (Supplementary Fig. 6c). Overall, this data suggests that due to collagen I accumulation, cells respond by contracting their actomyosin cortex and lowering their matrix adhesions, uncoupling vinculin from protrusions while blebbing starts to be apparent.

RhoA is activated in response to mechanical stress[41,42] and, accordingly, we measured increased RhoA GTP but also RhoC GTP levels after protease inhibition (Fig. 6g). GEFH1 and LARG (ARHGEF12) are Rho GEFs that selectively activate RhoA[43,44] in response to applied mechanical stress[41]. When GEFH1 or LARG1 expression was reduced using several RNAi, cells could not change their morphology or increase Myosin II levels after protease inhibition (Supplementary Fig. 6d-f). The increase in Myosin II levels was ROCK dependent, as shown using ROCKi, H1152 (Fig. 6h, i).

These results suggest that upon inhibition of peri-cellular proteolysis, changes in matrix organisation led to RhoGEF/RhoA/C/ROCK/MLC2 activation. After Myosin II is uncoupled from vinculin-based adhesions, high RhoA/C-ROCK activity results in Myosin II strongly localising to the cortex, cells lowering adhesion levels with a clear induction of amoeboid blebbing behaviour.

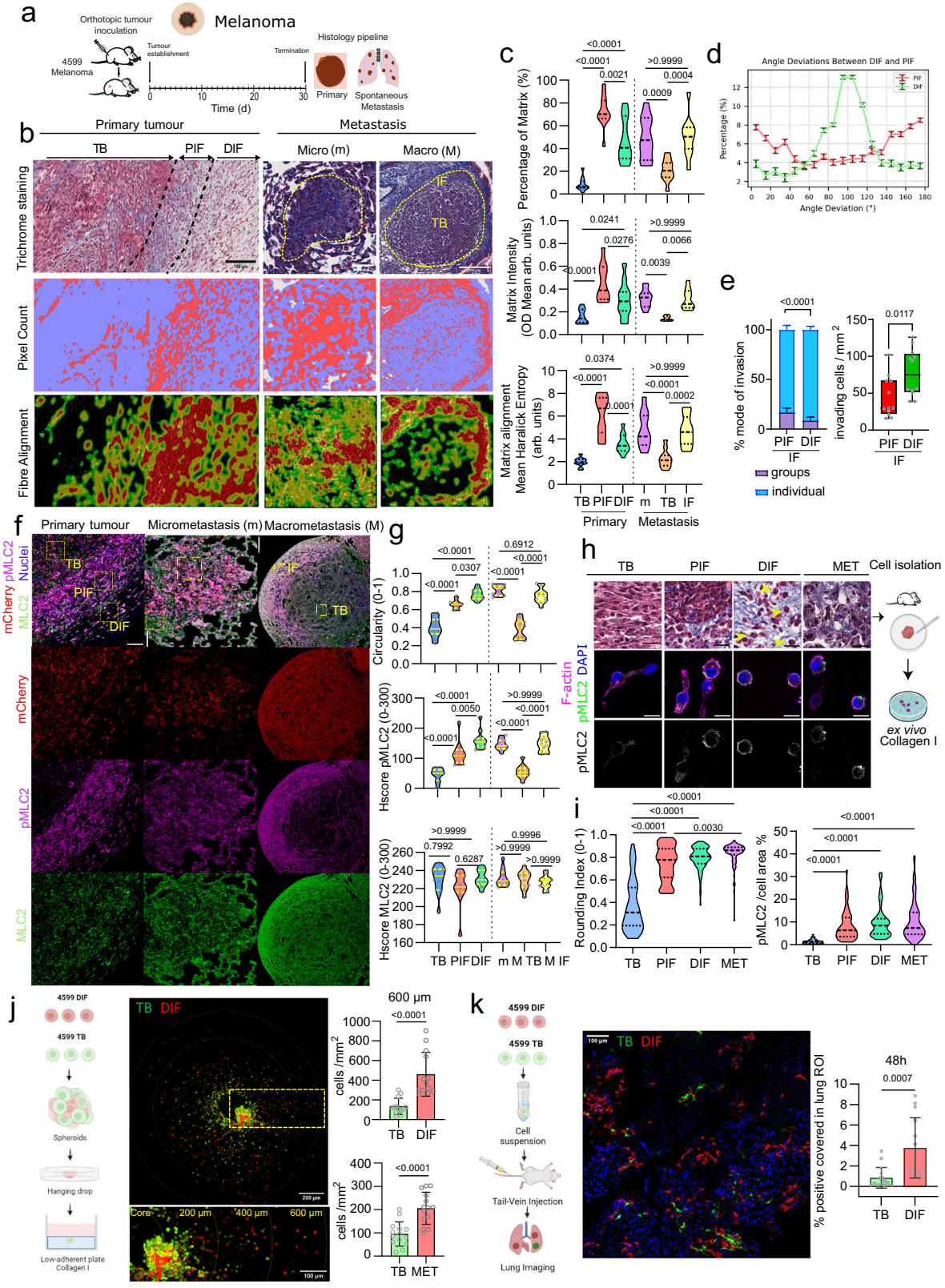
Matrix organisation drives a pro-metastatic transcriptional program

To understand if cytoskeletal changes linked to matrix organisation were coupled to a specific transcriptional program, we performed RNA sequencing on 4599 melanoma cells, which were extracted from the three distinct regions of the primary tumour (TB, PIF, and DIF) and the lung metastasis (MET) (Fig. 7a). Unsupervised PCA plot showed four separate clusters corresponding to each region (Fig. 7b). Gene overrepresentation analysis showed 4 major gene categories that were increased and shared by cells derived from PIF and DIF compared to the TB. We observed enrichment in genes related to ECM remodelling in cancer cells derived from either PIF, DIF or MET compared to cells derived from the TB (Fig. 7c). Interestingly, cancer cells derived from PIF, were enriched in genes encoding for matrix metalloproteases (Supplementary Fig. 7a), while cells extracted from DIF areas expressed higher levels of collagen encoding genes and collagen crosslinking enzymes (Supplementary Fig. 7a). Importantly, both collagen and LOX genes were sustained in cancer cells extracted from metastatic lesions (Supplementary Fig. 7a).

RNAseq analysis showed that PIF, DIF and MET derived cells shared gene expression programmes related to mechano-sensing, cell migration, GTPase activity, amoeboid type migration, chemotaxis, and angiogenesis (Fig. 7c). Such processes were enriched progressively in cells derived from PIF to cells derived from DIF and then to cells derived from MET (Fig. 7c). Transcription factor (TF) analysis based on differential expression genes showed that TEAD4 -a major regulator of YAP mechano-sensing – was enriched in PIF derived cells compared to TB derived cells (Fig. 7d, Supplementary Fig. 7b). STAT3 and inflammation/survival related genes and EMT master regulators (Snai1, Twist1) were enriched in DIF derived cells (Fig. 7d). Further GSEA analysis showed gene upregulation of inflammatory response (Hallmark) in PIF, DIF and MET derived cells when compared to TB derived cells (Fig. 7e). Given that NF-κB, STAT, TGF-β and Interferon signatures were significatively upregulated (Fig. 7e), increased YAP levels, and NF-κB and STAT3 activity were confirmed at protein level in cancer cells extracted from the different areas and grown in collagen ex vivo (Fig. 7f).

Transcriptomic analysis of HT1080 cells treated with protease inhibitors showed similar gene expression changes that were also related to the HIPPO/YAP and NF-κB pathways after 4 h, and an enrichment in survival, inflammation, and metabolism related genes, including JAK-STAT3, PI3K, P38-MAPK, after 24 hours (Supplementary Fig. 7c, d). We confirmed that using all the 3 models: increasing collagen concentrations, Marangoni flows and protease inhibitors induced a similar mechano-inflammatory response in vitro (Fig. 7g, h, Supplementary Fig. 7e–g), suggesting that this is a response to the increased physical confinement induced by the 3 models. Furthermore, priming melanoma cells to Marangoni flow-induced matrices for 7 days and removing the priming for 6 days resulted in a persistent activation of STAT3 activity (pSTAT3) (Fig. 7i). Moreover, we confirmed that knockdown of either YAP or NF-kB in the presence of PIs impaired the activation of STAT3 (Fig. 7j) while RNAi knockdown of YAP, NF-kB, and STAT3 impaired cytoskeletal remodelling after PI

https://doi.org/10.1038/s41467-025-56299-7



treatment, indicating their cooperation in matrix organisation-induced responses (Fig. 7k).

Finally, in tissue sections from mouse primary tumours and lung metastases, progressive increases in nuclear YAP, p65, and pSTAT3 were observed from TB to PIF and to DIF (Fig. 7l, m, Supplementary Fig. 7h). Micro-metastases and macro-metastases showed distinct expression patterns. High levels of the 3 markers were found in micro-metastases and higher levels were found at the border of macro-metastases compared to the tumour bodies of the same lesions (Fig. 7l, m, Supplementary Fig. 7h).

Overall, these results indicate that different transcriptional programs are spatially activated in tumours in response to ECM changes.

Fig. 3 | ECM in vivo organisation is associated with a cytoskeletal memory. **a** Diagram showing 4599 mouse melanoma orthotopic model. Intradermal cancer cell injection leads to development of primary tumours and spontaneous lung metastasis. **b** Histological images of primary tumour (TB, PIF, DIF) and metastasis (micro/macro) showing trichrome staining, QuPath intensity map, and CurveAlign fibre alignment. Scale bars: 100 μm (primary), 50 μm (micro), 200 μm (macro). **c** Quantification of Matrix Percentage (%), Matrix Intensity (OD Mean arb. units), and Matrix Alignment (Haralick Entropy arb. units). $n = 11$ mice. **d** Angle deviation line plot comparing PIF (red) and DIF (green) fibre distributions. $n = 11$ mice. **e** Quantification of individual versus grouped cell invasion at PIF/DIF, and invading cells/mm² count. $n = 11$ mice. **f** Multiplex images of primary tumour and metastases showing merged and individual channels (mCherry, pMLC2, MLC2). Nuclei in blue. Scale bars: 100 μm (primary), 20 μm (micro), 300 μm (macro). **g** Quantification of tumour circularity (0-1), pMLC2 H-score (0-300), and MLC2 H-score (0-300). $n = 11$ mice. **h** Ex vivo analysis of cells isolated from TB, PIF, DIF and metastasis in Collagen I, showing F-actin and pMLC2 staining. Nuclei in blue. **i** Quantification of cell rounding index and pMLC2/cell area from ex vivo isolates. $n = 100$, 76, 90 and 90 cells per group, 3 independent experiments. **j** Spheroid invasion assay comparing TB (green) and DIF (red) cells and TB (green) and Met (red), with quantification at 600 μm from core. Scale bars: 200 μm, inset 100 μm. $n = 13$ and 14 spheroids per group. **k** Lung colonisation assay showing TB (green) versus DIF (red) cells in lung colonization. Nuclei in blue. $n = 16$ mice. One-way ANOVA with Šídák (c, g) and Tukey post-hoc (i) displayed as truncated violin plots showing medians and quartiles. Unpaired *t*-test (e right, j, k) shown as bar graphs with mean ± SD (e left, j) and boxplots showing median, interquartile range and individual data points (e right). Two-way ANOVA with Sidak post-hoc (e left). Angle deviation data (d) presented as line plot with SEM error bars. Cartoons in Fig. 3 (a, h, j, k) created in BioRender. Maiques, O. (2025) https://BioRender.com/z35l297 https://BioRender.com/f02u038.

These programs, initially triggered by matrix organisation, are perpetuated through pro-inflammatory and pro-survival signalling pathways during the metastatic cascade.

Matrix organisation induced transcriptional program in human tumours correlates with worse outcome

We sought to validate our results in the human clinical setting. We combined the signatures derived from collagen in vitro models and the signature derived from cells purified from DIF areas in vivo and we obtained a list of 325 genes that we defined this as "Mechano-Inflammatory Signature" (MIS) in which STAT3, TGF-β and cytoskeletal genes were present (Fig. 8a, Supplementary Fig. 8a, Supplementary Data 4). Using spatial transcriptomics in human melanoma metastasis lesions, we annotated IF and TB in situ, while we observed that MIS was upregulated at the IF (Fig. 8a, b, Supplementary Fig. 8b). Importantly, differential gene expression from IF compared to TB showed an enrichment for ECM components (Collagens, Fibronectin, MMPs) together with STAT and NF-κB genes (Fig. 8c). Furthermore, we validated that Gene-Ontology signatures related to MIS were activated at the IF compared to TB (Fig. 8d).

Finally, we confirmed in TMAs of human primary melanomas that nuclear YAP1, p65 and pSTAT3 levels were enriched in cancer cells at the PIF compared to the TB, while cells at the DIF of the primary lesions had the highest levels (Fig. 8e, f). Moreover, cells at the IF of human melanoma metastasis were also enriched for YAP-p65-pSTAT3 compared to the TB (Fig. 8e, f; Supplementary Fig. 8c). Metastasis confers worse prognosis to patients[45], therefore we hypothesized that the IF of primary tumours could already harbour prognostic information if cancer cells retain the matrix organisation-induced transcriptional program later in metastasis. Indeed, DIFs with higher presence of triple positive (YAP1-p65-pSTAT3) cancer cells (> 150 cells/mm²) were associated with shorter melanoma patient survival (Fig. 8g). Moreover, higher fibre alignment at the DIF was also associated with shorter melanoma patient survival (Fig. 8g). Importantly, using transcriptomics data from TCGA (UVM, PDAC, HNSC, LIHC, SKCM, KIRP, CESC, LUAD, LGG, MESO, KIRP, LUSC, BLCA, GBM) and METABRIC (BRCA) especially in TNBC, we found that MIS is associated with shorter survival in several solid tumour types (Supplementary Fig. 9a, b).

This data shows that matrix organisation at the edge of primary tumours supports cytoskeletal remodelling and a transcriptional program that is conserved in metastatic lesions (Fig. 9) and is associated with worse prognosis.

Discussion

High collagen expression and deposition have been correlated with poor patient survival[46,47]. Increased collagen poses a challenge that can limit cancer cell invasion[1] while aligned matrix fibres facilitate invasion[1,13]. We use digital pathology, second harmonic generation, and curvelet transform methods to analyse ECM organisation, cancer cell cytoskeletal remodelling, and transcriptional reprogramming. We find that human tumours are characterized by 3 different regions in terms of 1) distance to the tumour centre, and 2) matrix deposition, alignment, and fibre orientation: TB, PIF, and DIF. We observe that increased matrix deposition (linked to increased physical confinement) can promote cell body retraction resulting in amoeboid cell features as observed at the PIF and DIF. Importantly, aligned radial/perpendicular orientation of the fibres away from the TB is conductive of maximal amoeboid invasion.

Our work builds upon seminal observations by Provenzano and colleagues that found how high fibre alignment, coupled with perpendicular fibre orientation was associated with poor patient prognosis in BRCA[8,10]. In the current work, we go further to define spatially matrix organisation: PIF areas which share features described as TACS2, while DIF areas share features with TACS3. Moreover, we expand the characterisation of the regional matrix architecture in primary and secondary tumours in two clinical models of invasive and metastatic tumour types (melanoma and breast cancer), while we explain how cytoskeletal remodelling coupled to transcriptional reprogramming is crucial for local invasion and metastatic colonisation.

To model collagen accumulation in vitro, we built gels containing increasing collagen concentration and/or gels that accumulated uncleaved collagen by blocking matrix metallo-protease activity. To specifically recapitulate the DIF, we used Marangoni flows as a method to polymerise collagen fibres radially[36]. We observe that directionality of the radial fibres is also coupled to increased thickness, length, and alignment of fibres. Importantly, using tumour-spheroid invasive assays, we measure increased rounded amoeboid directed invasion when the fibres are oriented in a radial fashion.

Numerous studies have explored how ECM biomechanical properties like stiffness, confinement, and viscoelasticity affect tumour progression[48–52]. In the current study, we find that collagen accumulation after blocking proteases or increasing its concentration (and its stiffness) results in cell individualisation and cell contraction/rounding due to increased 3D confinement. Protease inhibition leads to a switch in the mode of migration to amoeboid strategies, while very high collagen concentration will lead to overall less invasion (but remaining invasion will be individual amoeboid). Importantly, using Marangoni flows that recapitulate all the conditions seen in vivo at the DIF: more fibres, thicker and longer fibres that are closer to each other and are both straighter and more aligned. We then observe: 1.an increase in actomyosin contractility and 2. optimal directed migration. In summary, increasing collagen I around the cells via increasing its concentration, blocking proteases or using Marangoni flows will result in cell body retraction, cell rounding and increased actomyosin contractility due to the increased confinement. However, it is the radial alignment of the longer fibres that supports higher migratory and invasive behaviour (Fig. 5j, Supplementary Fig. 5d). Importantly, we

https://doi.org/10.1038/s41467-025-56299-7

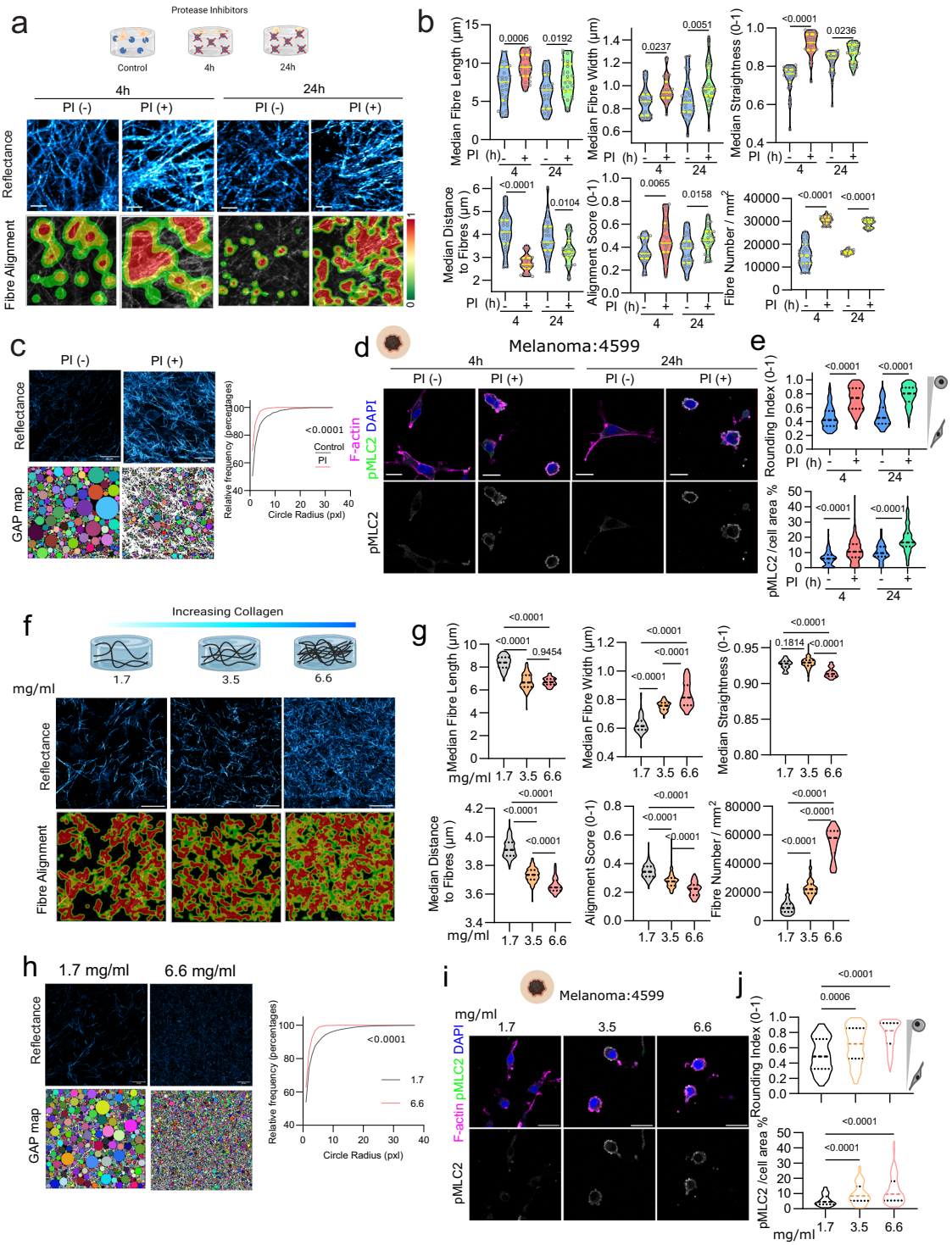
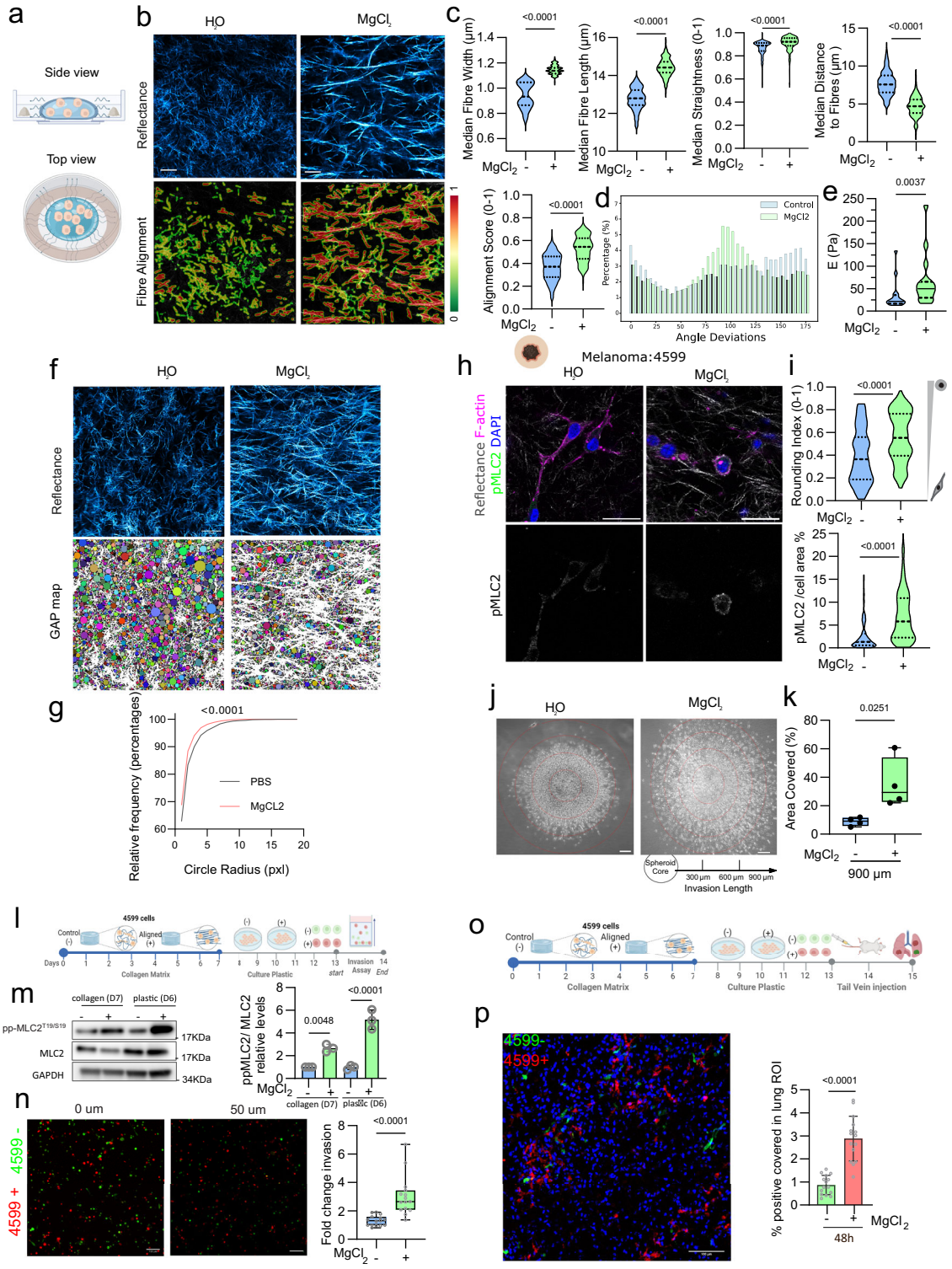


Fig. 4 | Collagen I accumulation induces cytoskeletal remodelling. a Experimental setup showing protease inhibitor (PI) treatment of cells in a collagen I matrix. Images show collagen reflectance and CurveAlign matrix alignment maps at 4 h and 24 h. Scale bar: 5 μm. **b** Matrix analysis quantifying fibre length, width, straightness, distance, alignment score, and fibre number/mm² with/without PI at 4 h and 24 h. $n = 27$, 25, 26 and 27 samples analysed from three independent experiments. Sidak one way. **c** Collagen reflectance images with gap analysis map and quantification comparing control versus PI treatment. $n = 1066$ and 2052 gaps analysed, three independent experiments. Scale bar: 20 μm. **d** Immunofluorescence of F-actin (magenta) and pMLC2 (green) in 4599 cells ±PI. Nuclei in blue. Scale bar: 10 μm. **e** Quantification of cell rounding index and pMLC2/cell area. $n = 134$, 155, 122 and 155 cells from three independent experiments. **f** Collagen images at increasing concentrations (1.7, 3.5, 6.6 mg/mL) by confocal reflectance microscopy. Scale bar: 20 μm. **g** Matrix analysis quantifying fibre length, width, straightness, distance, alignment score, and fibre number/mm² at increasing concentrations (1.7, 3.5, 6.6 mg/mL). $n = 90$, 78 and 55 images, three independent experiments. **h** Collagen reflectance images with gap analysis comparing 1.7 and 6.6 mg/mL collagen. $n = 3584$ and 10493 gaps analysed, three independent experiments. Scale bar: 20 μm. **i** Immunofluorescence showing F-actin and pMLC2 in 4599 cells at different collagen concentrations. Nuclei in blue. **j** Quantification of rounding index and pMLC2/cell area across collagen concentrations. $n = 110$, 95, 151 cells, three independent experiments. One-way ANOVA with Šídák (b, e) and Tukey post-hoc (g, j) displayed as violin plots showing medians and quartiles. Mann-Whitney U test between (c, h) shown as cumulative plots. Cartoons in Fig. 4(a, d, f, i) created with BioRender Maiques, O. (2025) https://BioRender.com/z57v040 https://BioRender.com/f02u038.



link matrix organisation to cellular decisions in different time scales: cytoskeletal dynamics (short term) and transcriptional reprogramming (long term).

Even if we have focussed our efforts on collagen I matrix organization, the ECM is a complex network where other fibrillar proteins, such as fibronectin, laminins, or other collagens, could play a role. Data obtained from RNAseq approaches in cells extracted from different mouse tumour areas and the spatial transcriptomics performed in human tissues reveals that cancer cells produce other collagens (see Supplementary Data 4). Importantly, our RNA-seq analysis showed that different pathways associated with ECM remodelling are activated or downregulated depending on the location within the tumour (Supplementary Fig. 7a). We provide evidence that cells in the PIF region upregulate MMP/TIMP genes, while cells in the DIF region upregulate collagen genes and the lysyl oxidase (LOX) genes. Future work should focus on understanding the role of other fibrillar proteins together with collagen I in controlling the cytoskeletal and transcriptional memory we describe in our studies.

Furthermore, our work highlights the complexity of cell-ECM interactions and the importance of considering the dimensional

Fig. 5 | Collagen I accumulation and radial alignment support Myosin II driven invasive and metastatic behaviour. a Schematic showing how $MgCl_2$ affects collagen matrix alignment, illustrated in side and top views. **b** Collagen reflectance images comparing H_2O and $MgCl_2$ treatment, showing reflectance and CurveAlign matrix alignment maps. Scale bar: 20 μm. **c** Matrix analysis quantifying fibre parameters (length, width, straightness, distance, alignment score, fibre number/mm²). $n = 150$ and 132 images, three independent experiments. **d** Angle deviation histogram comparing control versus $MgCl_2$-treated collagen fibres ($n = 30$ samples per group). **e** AFM elasticity (Pa) measurements of control and $MgCl_2$-treated collagen. $n = 15$ per group, two independent experiments. **f** Reflectance images with gap analysis maps. Scale bar: 20 μm. **g** Gap analysis quantification comparing control and $MgCl_2$ collagen fibres. $n = 6941$ and 6125 gaps analysed, three independent experiments. **h** Immunofluorescence of F-actin (magenta) and pMLC2 (green) in 4599 cells ±$MgCl_2$. Nuclei in blue. Scale bar: 20 μm. **i** Cell rounding index and pMLC2/cell area quantification in 4599 cells ±$MgCl_2$. $n = 140$ and 160 cells, three independent experiments. **j** Spheroid assay showing 4599 cell invasion length in non-aligned (H_2O) /aligned ($MgCl_2$) matrices. Scale bar: 200 μm. **k** Area coverage quantification from 4599 tumour spheroids at 900 μm from spheroid core comparing non-aligned (H_2O) /aligned ($MgCl_2$) matrices. $n = 4$ independent experiments. **l** Diagram of 3D invasion assay schematic using educated 4599 cells in non-aligned (H_2O) /aligned ($MgCl_2$) matrices. **m** Western blot of ppMLC2 and MLC2 in primed cells. $n = 3$ independent experiments. The samples derive from the same experiment but different gels for pp-MLC2$^{T19/S19}$ and GAPDH, and another for MLC2 were processed in parallel. **n** 3D invasion images at 0 μm and 50 μm with quantification. Scale bar: 100 μm. $n = 3$ independent experiments. **o** Diagram of in vivo lung invasion assay schematic using pre-educated 4599 cells. **p** Lung parenchyma representative image at 48 h post-injection. Scale bar: 100 μm. ROI quantification at 30 min and 48 h. $n = 14$ mice, 2 independent experiments. Unpaired t-test (c, i, k, n) and Mann Whitney test (e, f) between -/+ conditions. One-way ANOVA with Šídák post-hoc test (m). Data is displayed as violin plots showing individual data points and medians (c, e, i), boxplots showing median, interquartile range and individual data points (k, n), Bar plot with Mean and SD (m, p), and histogram in (d). Cartoons in Fig. 5(a, h, l, o) created in BioRender. Maiques, O. (2025) https://BioRender.com/z35l297.

context when investigating cell motility and adhesion mechanisms. In agreement with previous work, we show that vinculin dynamics and protein levels can change when cells are under confinement, leading to the loss of strong cell-ECM adhesions[53,54]. Furthermore, we find that the initial changes in matrix organization result in the activation of the RhoA/C-ROCK-Myosin II pathway, leading to cytoskeletal contraction[11,55,56].

On the other hand, melanoma cells isolated from the TB, PIF, DIF, and MET retained actomyosin contractility levels ex vivo and this correlated with their invasive and pro-survival properties. Transcriptomic analysis showed that ECM remodelling, mechano-sensing, pro-survival/pro-inflammatory and cytoskeleton organization- related genes, were all enriched in the cells derived from the invasive areas and the metastatic lesions. We find that matrix organisation-induced responses are linked to YAP-NF-κB activation, while STAT3 signalling sustains cytoskeletal features of amoeboid migration in the long term. STAT3[17,57] and NF-κB[58] regulate Myosin II activity via cytokines such as IL-6/LIF or IL-1α. However, in the current work, we propose that matrix organisation on its own leads to the activation of these signalling pathways, independently of such biochemical signals. Mechanistically, we show that collagen polymerised using Marangoni flows alone can prime cells for invasion and lung colonisation via Myosin II (pro-invasive) and pSTAT3 (pro-survival) activation. These findings suggest that biophysical adaptations induced by collagen mechanosensing are retained throughout the metastatic process via acquisition of a mechanical memory. From our bioinformatic analysis, we can speculate that such memory could be supported by epigenetic factors such as npBAF[59] complex or ARID1A[60], as we find expression levels are increased in the invasive areas (Fig. 7d, Supplementary Fig. 7b). Other types of biophysical adaptations supporting metastasis have been reported, which together with the current work highlight the importance of the physical properties of the micro-environment during cancer progression and metastasis[61].

In conclusion, this work clarifies how physical coupling/uncoupling of the cytoskeleton from the matrix leads to long-term molecular events that sustain efficient metastatic dissemination. Our observations suggest that matrix organisation, cytoskeletal features and transcriptional activation are strong predictors of aggressive behaviour of cancer cells. Patients with these tumour architectures could benefit from therapies targeting these pro-invasive and pro-survival programs linked to collagen dynamics. For example, pan-LOX inhibitors are under active preclinical investigation for highly desmoplastic tumour types, thus PXS-5505 is a strong candidate in future clinical trials[62–64]. Furthermore, YAP1 (NCT04590664) or JAK1[65] inhibitors are currently being tested in preclinical and clinical studies to treat cancer. We have used cancer as a model, but our work will shed light on other physiological conditions in which matrix organisation could influence cell migration and cell survival, such as during development, fibrosis, aging, wound healing, and immune responses.

Methods

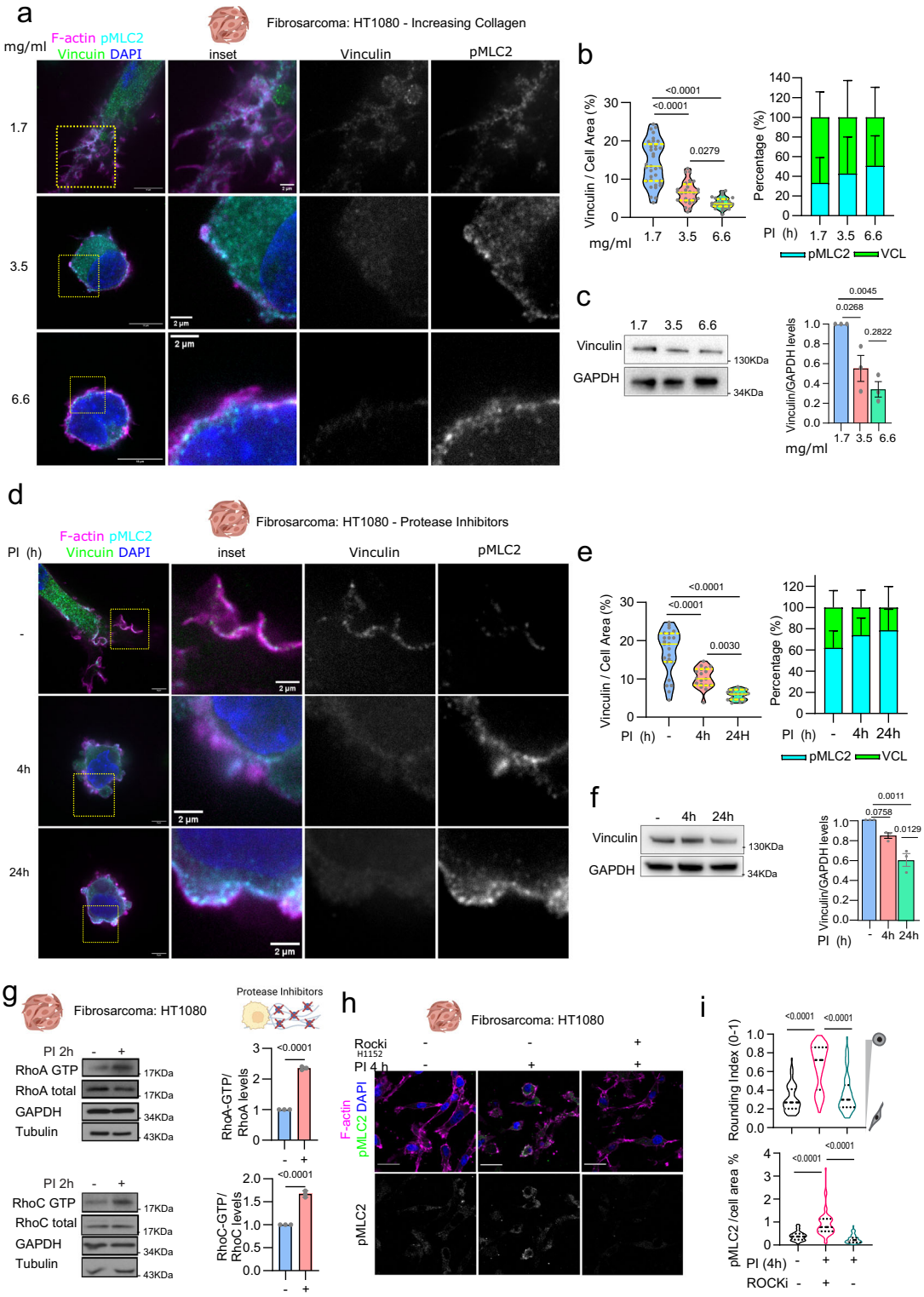
Compliance

This research complies with all relevant ethical regulations; all animals were maintained under specific-pathogen-free conditions and handled in accordance with the Institutional Committees on Animal Welfare of the UK Home Office (The Home Office Animals Scientific Procedures Act, 1986). All animal experiments were approved by the Ethical Review Process Committees at Barts Cancer Institute or King's College London and carried out under license from the Home Office, UK. For human tumour material, tumour samples were processed by IRB Lleida (PT17/0015/0027) and HUB-ICO-IDIBELL (PT17/0015/0024) Biobanks integrated in the Spanish National Biobank Network and Xarxa de Bancs de Tumors de Catalunya following standard operating procedures with approval of their ethics and scientific committee and were collected with specific informed consent, in accordance with the Helsinki Declaration. Consent to report patient data has been obtained and it is included in the study's ethical compliance. Compensation was not provided. This study did not specifically account for sex or gender in the experimental design because the samples were obtained as part of routine clinical practice or were retrospectively collected from existing repositories. Thus, participant recruitment was not influenced by sex or gender.

Antibodies and inhibitors

Antibodies and dilutions used: pSer19-MLC2 (#3671, immunofluorescence, 1:200; immunohistochemistry, 1:100), pThr18/Ser19-MLC2 (#3674, immunoblotting, 1:750), MLC2 (#3672, immunoblotting, 1:750), MLC2[D18E2] (#8505, Immunohistochemistry, 1:300), pY705-STAT3 (#9145, immunoblotting, 1:750; immunohistochemistry, 1:100), NF-κB1 (#4764; immunoblotting, 1:500), RhoA [67B9] (#2117, immunoblotting, 1;750), RhoC[D40E4] (#3430, immunoblotting, 1:750), STAT3[124H6] (#9139, immunoblotting, 1:750), p65[C22B4] (#4764, immunoblotting, 1:750; immunohistochemistry, 1:200), pSer536-p65 (#3036, immunoblotting, 1:750) from Cell Signalling Technology; LARG (sc-25638; immunoblotting, 1:500), YAP1 (sc-101199, immunohistochemistry, 1:100, immunoblotting, 1:750) from Santa Cruz Biotechnology; GAPDH (MAB374; 1:10,000) from Millipore; Melanoma antibody [HMB45 + M2-7C10 + M2-9E3 + T311] (ab733; immunohistochemistry, 1:300), Cytokeratin-19 [EP1580Y] (ab52625, immunohistochemistry, 1:500); Vinculin (V9131; immunofluorescence, 1:200, immunoblotting, 1:5000) from Sigma; GEF-H1 (ab90783; immunoblotting, 1:500) from Abcam; affinity-purified antibody COL1-¾C (no. 0217-050, immunofluorescence; 1:200) directed against the C-terminal cleavage neo-epitope of collagen types I and II, Collagen I

https://doi.org/10.1038/s41467-025-56299-7



(ab34710; immunohistochemistry, 1:1000), Collagen IV (M0785; Immunohistochemistry; 1:25) from Agilent.

Protease inhibitor mix: 100 µM GM6001, 50 µM Pepstatin A, 250 µM E64, 2 µM Aprotinin, 2 µM Leupeptin. GM6001 and H1152 were from Merck Millipore (Hertfordshire, UK); Pepstatin, E-64, Aprotinin and Leupeptin were from Sigma-Aldrich (Dorset, UK).

Cell culture

HT1080 cells were obtained from Prof Chris Marshall (Institute of Cancer Research), A375P cells from Prof Richard Hynes (HHMI, MIT, USA), SKMEL28 cells from Prof Richard Marais (Cancer Research UK Manchester Institute, Manchester, UK) and MDA-MB-231 cells were from Prof Clare Isacke (Institute of Cancer Research). 4599 (from Richard Marais) were established from the C57BL/6 mouse model Braf+/LSL-V600E;Tyr::CreERT2+/o[30].

For the spontaneous model of melanoma metastasis using 4599 were virally transduced to express NIS-mCherry as described in Maiques O in ref. 31. We used them for Fig. 3a–g.

Isolation of murine melanoma cells and ex vivo cultures were generated after intradermal injection using naïve 4599 (no

Fig. 6 | Matrix organisation changes result in Rho-ROCK engagement and uncoupling of Myosin II from vinculin. a Confocal SoRa images of HT1080 Fibrosarcoma cells showing nuclei (blue), F-actin (magenta), pMLC2 (cyan), and Vinculin (green) at increasing collagen concentrations (1.7-6.6 mg/ml). Scale bars: 10 μm, insets 2 μm. **b** Quantification of Vinculin/cell area (%) and percentage of pMLC2/Vinculin levels at the indicated collagen I concentrations. Left, $n = 35$, 28 and 28 cells and right, $n = 23$, 25 and 22 cells, three independent experiments. **c** Western blot analysis of Vinculin levels in HT1080 cells grown in different collagen I concentration gels. $n = 3$ independent experiments. Vinculin and GAPDH were run in the same gel. **d** Confocal SoRa images of HT1080 Fibrosarcoma cells showing nuclei (blue), F-actin (magenta), pMLC2 (cyan), and Vinculin (green) ± protease inhibitors (PI) at 4 h and 24 h. Scale bars: 10 μm, insets 2 μm. **e** Quantification of Vinculin/cell area (%) and percentage of pMLC2/Vinculin levels ± PI. Left, $n = 23$, 18 and 21 cells and right, $n = 30$, 27 and 25 cells, three independent experiments. **f** Western blot analysis of Vinculin levels ±PI. $n = 3$ independent experiments. Vinculin and GAPDH were run in the same gel. **g** RhoA and RhoC activation by pull-down assays in HT1080 cells ±PI after 2 h treatment, with quantification of GTP-bound and total protein levels. n = 3 independent experiments. **h** Immunofluorescence images showing F-actin and pMLC2 in HT1080 cells ±PI and ROCK inhibitor (H1152) at 4 h. Scale bar: 20 μm. **i** Quantification of cell rounding index and pMLC2/cell area from conditions in (h). $n = 61$, 66 and 59 cells, three independent experiments. Unpaired *t*-test for pull-down assays (g) and One-way ANOVA with Tukey post-hoc test for all other comparisons (b left, c, e left, f, i). Two-way ANOVA with Tukey post-hoc (b right and e right). Data displayed as violin plots showing individual data points and medians (b left, e left, i), bar plot with mean and SD (b, right, c, e right, f, g). Cartoons in Fig. 6 (a, d, g, h) created in BioRender. Maiques, O. (2025) https://BioRender.com/c83e894.

plasmid) (Fig. 3h–k). (see animal methods for orthotopic injection). After 28 days from injection animals were sacrificed and primary tumours and lungs were removed transferred in falcon tubes with PBS −/− $CaCl_2$/ $MgCl_2$ and kept in ice before subjecting to tumour cell isolation. Before cell isolation, primary tumours were processed as shown in diagram in Supplementary Table 1. Primary tumours were divided into two parts using a scalpel; one part was kept in 4% formaldehyde solution for paraffinization, while the second part was used for tumour cell isolation. For the latter, scissors and scalpel were used to collect the centre of the tumour or TB. For the IF, the peritumoral semi-transparent area with more fat tissue was collected, which contained the DIF area. The PIF was collected removing TB and DIF areas. To enrich populations, a pool of 6 primary tumours per area was used. For the lung where metastases were observed after postmortem examination, 3 lobules from each mouse were collected (total of 6 mice), the other remaining 2 lobules per mice were used for FFPE. Collected material for cell isolation was transferred to solutions with a mixture of Liberases (TH and TM, 75 μg/ml each, Roche Diagnostics) and 1 μg/ml DNase I (Sigma) in HBSS, digested briefly with gentle MACS Dissociator (Miltenyi Biotec) and transferred for 30 min at 37 °C shaking, and then passed through 100 μM strainers. After digestion, cells were passed through the sorter ARIA Fusion (Becton Dickinson) using naïve 4599 as control. Cells were maintained in DMEM containing 10% foetal calf serum (FCS).

4599 TB, PIF, DIF and MET were isolated from naïve 4599 injected orthotopically into the skin of female 6–8 weeks old NXG (NOD-Prkdcscid-Il2rgtm1/Rj) mice (Janvier-Labs) at the experimental end point (28 days).

Transfection and RNAi

For HT1080 cells, 200,000 cells were plated per 35 mm dish and transfected the next day with 20–40 nM SmartPool (SP) or individual OTs (On Targetplus) siRNA oligonucleotides, using Optimem-I and Lipofectamine 2000 (Invitrogen). Forty-eight hours after transfection cells were seeded on collagen in 10% FBS, allowed to attach overnight and treatments were added were indicated.

siRNA sequences

All siRNA were from Dharmacon (Lafayette, USA). siRNA stocks were centrifuged, diluted in 1X siRNA resuspension buffer (Dharmacon) for a stock concentration of 100 μM and allowed to mix via rotation at room temperature for 30 min. (Supplementary Table 3).

Cell culture on thick layers of collagen I

Standard collagen unless differently stated, was fibrillar bovine dermal atelo Type-I Collagen solution (3 mg/ml) (#5005; PureCol, Advanced BioMatrix) and prepared at 1.7 mg/ml (Supplementary Table 4). For increasing collagen concentrations atelo Type-I Collagen solution (10 mg/ml) (#5133; FibriCol, Advanced BioMatrix) was prepared at 1.7, 3.5 and 6.6 mg/ml (Supplementary Table 4). For radial alignment experiment using Marangoni flows fibrillar bovine dermal telo Type-I Collagen solution (3 mg/ml) (#5026, TeloCol-3, Advanced BioMatrix) and prepared at 1.7 mg/ml (Supplementary Table 4).

PureCol and Fibricol experiments were performed using the following volumes, 100 μl per well in 96-well plates (immunofluorescence); 300 μl per well in 24-well plates (cell morphology), 700 μl per well in 12-well plates (immunoblotting). After collagen gel polymerization for 4 h, cells were seeded on top of collagen in medium containing 10% FCS, allowed to adhere for 24 h and then treatments were added in 1% FBS.

Self-assembly of collagen fibres using Marangoni flows

To generate in vitro aligned collagen matrices, we used the protocol described in Nerger et al.[36], with some minor modifications. In brief, T24 (Corning™ Costar™ 24-well Clear TC-treated) or T6 well plates (Corning™ Costar™ 6-well Clear TC-treated) were prepared drawing a circle with hydrophobic pen (ImmEdge® Hydrophobic Barrier PAP Pen (H-4000)) in the centre of the well, leaving space to add the saturated solution of $MgCl_2$ around the well perimeter.

Simultaneously, type I bovine telo-collagen solution (Telocol) at 1.7 mg/ml was prepared to embed the cells as shown in Supplementary Table 4. Collagen concentration was adjusted using PBS, water and the pH of the resulting mixture was adjusted using the manufacturers' neutralizing solution (Advanced Biomatrix). Cell-laden hydrogels: cells were mixed with collagen final solution and plated according to the following instructions: for morphological assessment and immunofluorescence imaging either T12 or T24 well plates were used at a concentration of 100,000 cells/300 μl well, and 12,000 cells/100 μl well. For biochemistry and memory experiments, T6 were used a concentration of 300,000 cells/ 100 μl well. For spheroids assays, the spheroid was collected and embedded in the centre before collagen had polymerized, using T12 plates.

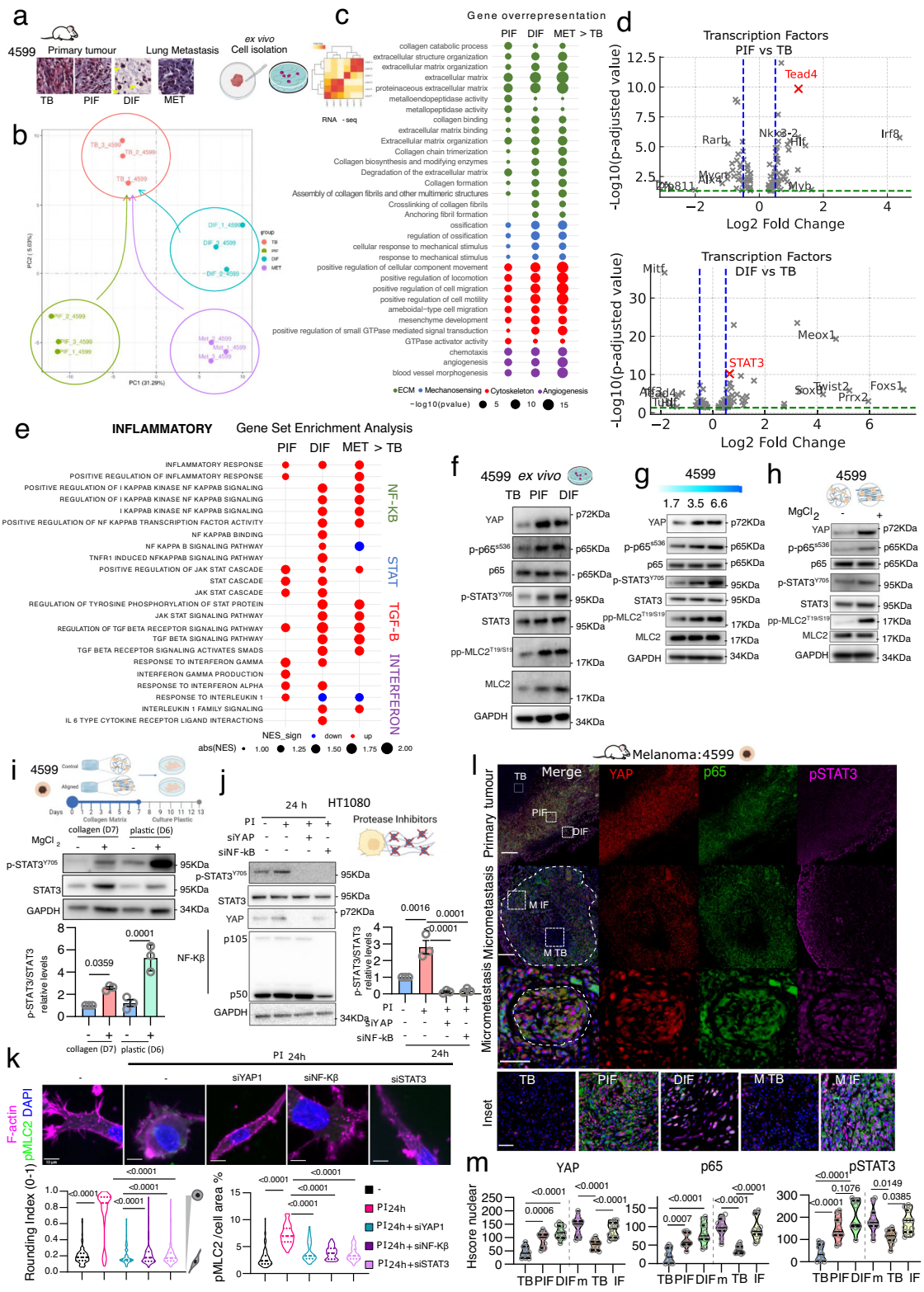
The collagen solution was placed inside the circle created with the PAP-Pen. After collagen polymerisation (40 min -1 h), salt ($MgCl_2$) was carefully removed using PBS 1X avoiding salt touching collagen to finally fill the well with DMEM + 10% FBS.

For immunofluorescence imaging experiments, cells were left for 24 h in complete media and deprived to 1% FBS for 24 h before fixation and staining. For biochemistry experiments and establishment of mechanical memory, cells remained for 6 days in complete media (10% FBS) changed every 48 h and deprived to 1% FBS the 7th Day. Next, hydrogels were digested (see Collagen cell extraction section), and cells were pelleted down. For immunoblotting, cells were lysed with Laemli buffer followed by protein extraction. For mechanical memory experiments, cells were plated in T6 (plastic adherent) plates for expansion.

Organotypics

Five hundred thousand HT1080 cells were plated on the top of 1 ml of matrix gel, which polymerized 1 h at 37 °C before adding the cells. The

https://doi.org/10.1038/s41467-025-56299-7



next day, the gel was mounted on a metal bridge and fed from underneath with changes of 10% FCS media with treatment (vehicle or PImix) every day. After 5 days, cultures were fixed (4% paraformaldehyde + 0.25% glutaraldehyde in PBS) embedded in paraffin blocks, sectioned, and stained with H&E. The type of invasion was calculated from ImageJ measurements.

Cell extraction from collagen gels

Collagen matrices were collected to a falcon tube with HBSS and Collagenase (Gibco). Gels were left at 37 °C during 30 mins on a shaker. Once the collagen was dissolved, the total volume was centrifuged in a falcon tube at 300 g for 5 mins. Cell pellets were resuspended in PBS 1X and cells counted for further downstream analysis.

Fig. 7 | Matrix organisation drives a pro-metastatic transcriptional program. **a** Mouse melanoma model showing intradermal injection of 4599 cells leading to lung metastasis. Cells isolated from different tumour areas (TB, PIF, DIF, Metastasis) for ex vivo experiments and RNA-seq. **b** Principal component analysis of 4599 cells' transcriptome from different tumour regions. $n = 3$ independent experiments. **c** Gene overrepresentation analysis comparing PIF, DIF, and MET versus TB, grouped by ECM, Mechano-sensing, Cytoskeleton, and Angiogenesis signatures. **d** Volcano plots of transcription factors in PIF versus TB and DIF versus TB. Threshold: -0.5<Log2FC < 0.5, -Log10 p-adjusted value. **e** GSEA analysis of inflammatory pathways, comparing PIF, DIF, and Met to TB. Dotplot shows normalized enrichment scores (FDR < 0.1). **f**–**h** Immunoblots showing YAP, p-p65, p65, p-STAT3, STAT3, pp-MLC2, and MLC2 protein levels in 4599 cells from different tumour areas ($n = 5$ independent experiments), collagen concentrations ($n = 4$), and aligned/non-aligned matrix ($n = 3$). GAPDH is used as loading control. Quantification in Supplementary Fig. 7g. The samples for each panel derive from the same experiment but different gels for STAT3, MLC2 and GAPDH, another for p-STAT3^{Y705}, pp-MLC2$^{T19/S19}$, another for p65, another for p-p65^{s536}, and another YAP were processed in parallel. **i** Top, diagram of experimental setting showing priming in ± $MgCl_2$ (7 days), and further culture in plastic (6 days). Bottom, corresponding p-STAT3/STAT3 protein levels in 4599 cells and quantification. GAPDH is used as loading control. $n = 3$ independent experiments. The samples derive from the same experiment but different gels for p-STAT3^{Y705} and GAPDH, and another for STAT3 were processed in parallel. **j** Left, protein levels of p-STAT3/STAT3 in HT1080 cells after YAP or NF-κB knockdown with PI treatment (24 h). GAPDH is used as loading control. Corresponding quantification of p-STAT3/STAT3. $n = 3$ independent experiments. The samples derive from the same experiment but different gels for p-STAT3^{Y705} and GAPDH, another for STAT3, another for YAP1 and another for NF-Kβ1 were processed in parallel. **k** Top, confocal images showing F-actin, pMLC2, and nuclei after PI treatment with siYAP1, siNF-κB1, or siSTAT3. Bottom, corresponding cell rounding index and pMLC2/cell area quantification. Scale bar: 10 μm. $n = 91$, 97, 91, 91 and 92 cells, three independent experiments. **l** Multiplex immunohistochemistry of YAP (red), p65 (green), p-STAT3 (magenta) in 4599 primary tumours and lung metastases. Scale bars: 200 μm (primary, macro), 50 μm (micro, insets). **m** Nuclear H-score quantification (0–300) of transcription factors across tumour regions from (l). $n = 11$/group. One-way ANOVA with Šídák (I, j, k) and Tukey (m) post-hoc test. Data displayed as violin plots showing individual data points and medians (k, m), bar plot with mean and SD (I, j). Cartoons in Fig. 7 (a, f, h, i, j, l) created with BioRender.com. created with BioRender Maiques, O. (2025) https://BioRender.com/z57v040 https://BioRender.com/f02u038.

Immunoblotting

Cells were lysed in Laemmli sample buffer and sonicated for 15 s before centrifugation. Cell lysates were separated using precast 4–12% gradient SDS–polyacrylamide gel electrophoresis (SDS–PAGE) gels (Life Technologies) in non-reducing conditions and transferred subsequently to PVDF filters. ECL Plus or Prime ECL detection Systems (GE Healthcare) with HRP-conjugated secondary antibodies (GE Healthcare) were used for detection. Bands were quantified using Image J.

RhoA/C-GTP pulldown assay

2×10^5 cells/well were seeded on 6-well plates, treated with protease inhibitors for 2 h as indicated. Cells were lysed in buffer containing 50 mM Tris pH7.4, 10% glycerol, 1% NP40, 5 mM MgCl2, 100 mM NaCl, 1 mM DTT and EDTA free protease inhibitor and spun down. 10% of protein lysate was separated for determination of total RhoA and RhoC levels. The remaining protein lysate was incubated with Glutathione S-transferase (GST)-conjugated with Rhotekin RBD beads for 1 h. Beads were collected by centrifugation, washed, and resuspended in LDS loading buffer. All samples were boiled for 5 min and resolved by SDS-polyacrylamide gels electrophoresis. RhoA, RhoC, Tubulin and GAPDH proteins were detected via immunoblot.

Gelatin zymography

HT1080 conditioned media was loaded on 10% SDS-polyacrylamide gels containing 0.3% gelatin (Sigma) in non-reduced protein sample. Following electrophoresis, proteins were renatured by incubating gels in 2.5% Triton X-100 for 2 h at 37 °C. Gels were then washed three times in distilled water and incubated in substrate buffer (50 mM Tris, pH 7.4, and 5 mM $CaCl_2$) at 37 °C for 18 h with gentle shaking. Gels were stained with 0.1% Coomassie Blue R-250 (Sigma) and destained in 7% acetic acid. Enzymatic activities appear as cleared bands in a dark background.

Immunofluorescence

Cells were seeded on collagen I in glass-bottomed dishes (MatTek, Ashland, MA, USA) and immunostained as described[57]. Cells were fixed (4% formaldehyde), permeabilized (0.2% Triton X-100; except for COL1-¾C immunostaining, in which cells were not permeabilized), blocked in 5% bovine serum albumin (BSA) and immunostained with specific antibodies diluted in 5% BSA, which were detected with secondary antibodies conjugated to either AlexaFluor488, -546, or -647 (Life Technologies). F-actin was stained using AlexaFluor546-phalloidin (Life Technologies) or AlexaFluor647-phalloidin (Life Technologies).

Confocal images were analysed using ImageJ

-For cell morphology, the shape descriptor 'roundness' in ImageJ was used after manually drawing around the cell shape using F-actin-staining images.

-Phospho-MLC2 fluorescence signal was quantified calculating the pixel intensity in single cells relative to the cell area.

-Percentage of nuclear YAP was calculated by measuring YAP fluorescence signal.

-Reflectance/matrix signal within 10 μm from the edge of single cells was quantified in single cells and calculated relative to the cell perimeter.

-For measurement of collagen degradation, COL1-¾C fluorescence signal within 15 μm from the edge of single cells was quantified and calculated relative to the cell perimeter[57].

Phase-contrast microscopy

Static phase contrast images were obtained on a 01QIClick FM12 microscope camera attached to a Nikon TX-100 inverted phase-contrast microscope using a Nikon LW 10x objective. Images were acquired using QCapture Pro 7 Software.

Confocal microscopy

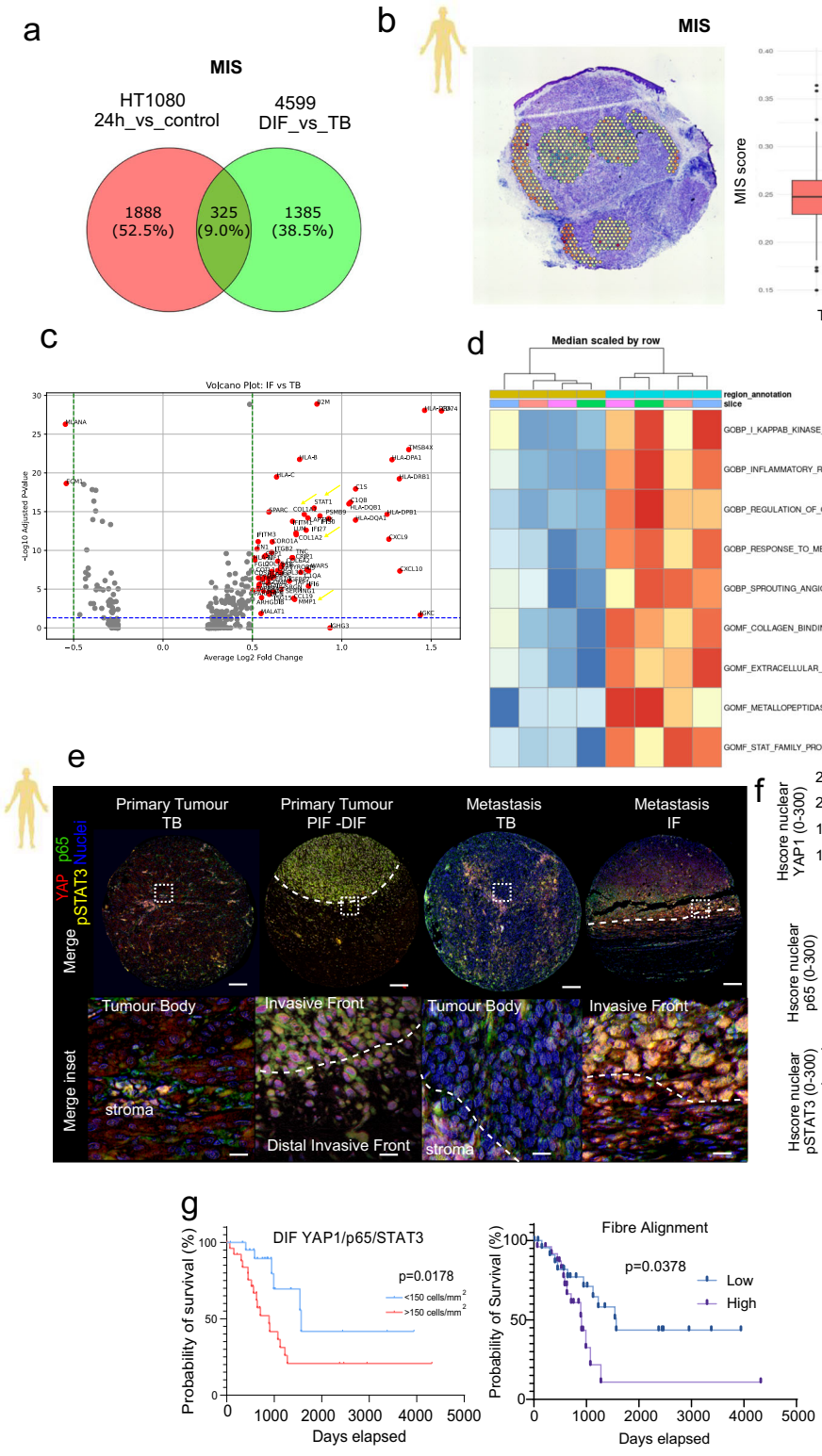
Confocal imaging was performed with Zeiss LSM 510, 710 and 880 with Airyscan Fast confocal microscope (Carl Zeiss) with C-Apochromat x 10/1.2, 40/1.2 (oil) and 63/1.2 (oil) objectives lenses and Zen software (Carl Zeiss). Reflectance imaging of the matrix was performed by collecting the backscattered light. Since reflectance is a surface property, extreme caution was taken to ensure the acquisition of images at the same height across all conditions.

Fixed samples were removed using forceps, turned over a MatTek dish so that the top surface of the collagen with the cells was in contact with the coverslip and then imaged using an inverted setup. Single confocal z-slices (1024×1024 pixels, 8-bit; 0.7x zoom) were taken across the mid-plane of the cells. For z-stacks used in analyses of collagen attachments, 30 μm thick stacks were taken with a fixed step size of 0.5 μm (512 × 512 pixels, 8-bit; 0.7x zoom).

Multiphoton microscopy and second harmonic generation (SHG)

Imaging was performed using a SP8DIVE (Leica) upright multiphoton microscope and using 25x/1.0 W IR Apo with motCorr objective. The SHG was generated with an InSight® DS + ™ a tunable 680–1300 nm IR

https://doi.org/10.1038/s41467-025-56299-7



pulsed solid-state laser plus a 1045 nm fixed line (Spectra-Physics), using a laser pulsation at 900 nm with a band-pass filter spanning 10 to 20 nm at 450 nm, or using laser pulsation beam at 1300 nm with a relatively narrow band-pass filter at 650 nm. Samples were counter stained with Draq5 (1:2000, eBioscience) for nuclear staining, CK19 for BRCA tumours and Melanoma marker for Melanoma (see antibodies and inhibitors section).

Super-resolution spinning disk microscopy

Vinculin and pMLC2 imaging was performed using Super Resolution Spinning Disk (TIRF/STORM/SoRa) with C-Apochromat X 63/1.2 oil objective lens, CSU-W1 camera and i3 Marianas software. z-stacks were used to analyse collagen attachment, 30 µm -thick stacks were taken with a fixed step size of 0.5 µm (51024×1024 pixels, 16-bit; 4x zoom).

Fig. 8 | Matrix organization drives a pro-metastatic transcriptional program. **a** Venn diagram showing Mechano-Inflammatory signature (MIS) derived from 4599 cells extracted from TB and DIF and from HT1080 treated with PI for 24 h. **b** Overlapping genes from (a) - conserved signature- were scored on the Visium array spots in a representative section of human melanoma lymph node metastasis. Annotations correspond to TB and IF areas. $n = 4$ slices were analysed. Box plot shows the comparison between TB and IF annotations. $n = 12$ TB annotations and $n = 12$ IF annotations were analysed. **c** Volcano plot showing DEGs upon IF and TB comparison from spatial transcriptomics defined annotations. Highlighted in red DEGs that present average log2 fold change >0.5 and -$\log_{10}$ adjusted p-value. **d** Heatmap shows median scaled by row, displaying Gene Ontology Biological Pathways (GOBP) and Molecular Functions (GOMF) terms separated by regions (TB and IF) and slices from Visium array analysed in melanoma metastasis samples. Z-score scale. **e** Representative merged multiplex images of primary and metastatic lesions of melanoma tumours. (bottom) Inset of merged multiplex. Nuclei were visualized with Haematoxylin stain (blue). Scale bar: 100 μm. Scale bar insets: 20 μm. **f** H-score quantification (0-300) of YAP1, p65 and pSTAT3 in the primary and metastatic lesions. $n = 53$ primary and 45 metastatic lesions. **g** Kaplan-Meier curve showing overall survival with the quantification of amoeboid cells YAP1, p65 and pSTAT3 positive in the DIF area (left). Kaplan-Meier curve displays the overall survival comparing high and low fibre alignment (right). One-way ANOVA with Tukey post-hoc test (f). Data displayed as violin plots showing individual data points and medians (f). Cartoons in Fig. 8 (b, e) Created in BioRender. Maiques, O. (2024) https://BioRender.com/h15l089.

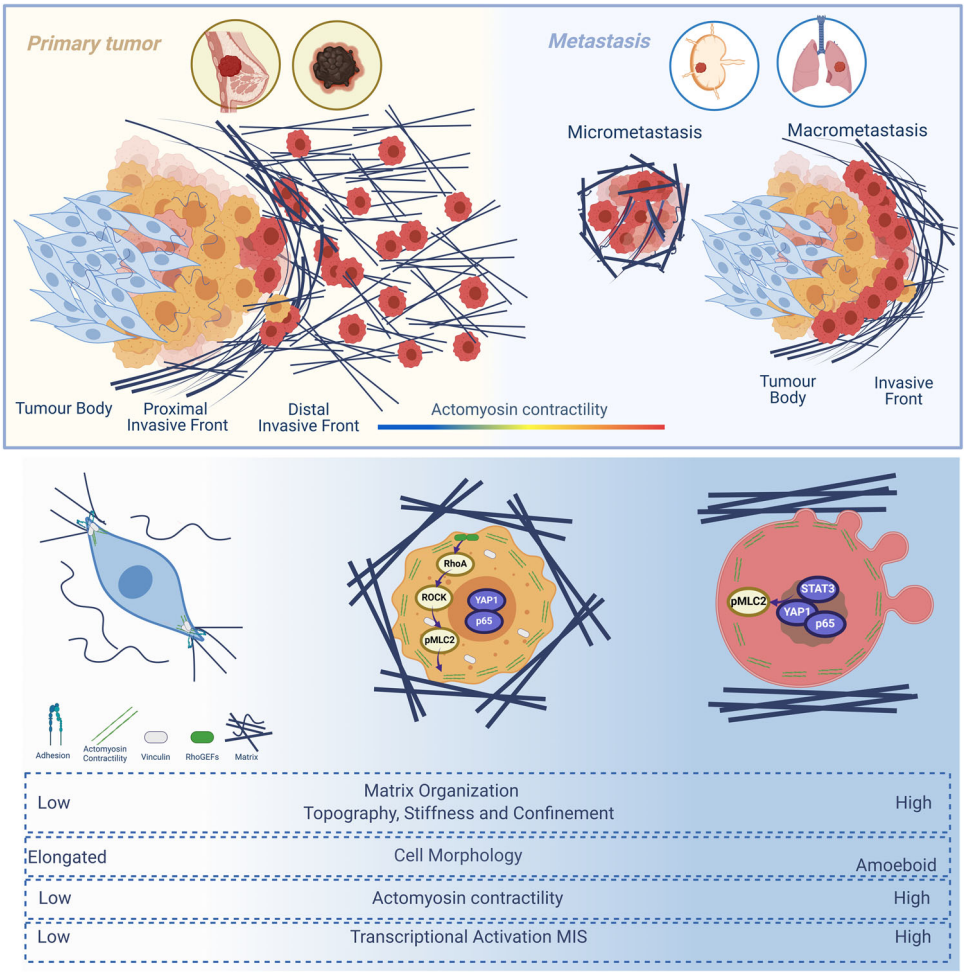


Fig. 9 | Model. Matrix organisation in primary tumours and metastasis: matrix heterogenous distribution in human solid tumours coupled to changes in the cytoskeleton and invasive patterns. The observations in primary tumour are mirrored in metastatic lesions. Mechanism at the cellular level (bottom): cancer cells under confinement, imposed by increased matrix, retract their protrusions and contract their cytoskeleton reducing cell-matrix adhesion. This retraction is accompanied by vinculin uncoupling from adhesions and activation of RhoA-ROCK-Myosin II axis leading to long-term transcriptional reprogramming orchestrated by YAP1-NF-κB and STAT3. Figure created in BioRender. Maiques, O. (2025) https://BioRender.com/v18d918.

Atomic force microscopy

The mechanical properties of collagen I gels were measured in gels exposed to H2O (control, $n = 5$) or MgCl2 ($n = 5$) using a JPK NanoWizard AFM (Bruker). For each gel, 100 points from three different areas were measured using a MLCT-O10-E cantilever (Bruker) manually conjugated to a polystyrene bead of 7.3 μm in radius (F21010, Invitrogen) using UV curing adhesive (267423, Rapid Electronics Ltd). The spring constant was 0.054 N/m. Young's elastic modulus of each gel was estimated using the Hertz model, with a 0.5 Poisson's ratio, using the JPK Data Processing software (v8.0.92, Bruker).

Sphere invasion assay

For sphere preparation, 4599 TB and 4599 DIF or 4599 MET labelled cells using cellTracker dyes (see vital dye cell staining) were mixed to a 1:1 ratio to a final cell number of 80,000 cells in 20% methylcellulose and 80% DMEM + 10% FBS in total volume of 1 ml. 25 μl drops were deposited in the lid of a round petri dish, carefully inverting the lid to natural position and adding PBS in the base. For experiments using HT1080 cells, spheres were made at 40,000 total number of cells per 1 ml of total volume. The spheres were formed in hanging drops for 24 h. After that, spheres were embedded in PureCol 1.7 mg/ml (Advanced BioMatrix). For alignment

experiments, cells were embedded in TeloCol 1.7 mg/ml. The spheroids were fixed in 4% paraformaldehyde (PFA) after 24 h or 72 h.

Brightfield images of spheroids embedded in a collagen matrix were taken at end point. An annotation of the core of the spheroid was created and was then expanded every 300 µm. A pixel classifier was created to differentiate spheroid from background/matrix. Random trees classifier was used for training while resolution of the mask was set as "High resolution", and Features were selected scale 1.0, 2.0 and Gaussian, Hessian determinant, Hessian max eigenvalue, Hessian middle eigenvalue and Hessian min eigenvalue. Mask detecting spheroid was created as annotation.

3D invasion

For 3D invasion assay, cells were extracted from non-aligned (H2O) and aligned ($MgCl_2$) matrices and labelled with different colour dyes (see Colour coding vital dye cell staining). 15 K cells were resuspended in Bovine dermal fibrillar collagen (PureCol, Cat. 5005 Advanced BioMatrix) plated in a 96-well plate and spun down before matrix polymerisation. After the matrix was polymerised, 10% FBS-containing media was added on top of the matrix as chemoattractant. After 24 h, cells were fixed and imaged using a Zeiss LSM 710 confocal microscope. Invasion was calculated as number of invading cells at 50 µm divided by the number of cells at the bottom.

Vital dye cell staining

For vital dye staining, cells were suspended in 10 µM CMFDA-Green (C7025, Life Technologies) or 10 µM CMRA-Orange (C34551, Life Technologies) in 1 ml OPTIMEM (31985070, Gibco). Cells were stained for 15 min at 37 °C and washed twice with PBS 1X. Cells were suspended in 1 ml PBS 1X and counted to use them at 1:1 ratio for further experiments.

Animal experiments

Intradermal tumours. All animals were maintained under specific pathogen-free conditions and handled in accordance with the Institutional Committees on Animal Welfare of the UK Home Office (The Home Office Animals Scientific Procedures Act, 1986). All animal experiments were approved by the Ethical Review Process Committees at Barts Cancer Institute and King's College London, and carried out under licences from the Home Office, UK. 6–12 weeks old male (Figs. 3a–f, 7l, m) and female (Figs. 3h–k, 7a–f) mice were used. Animals were housed in groups of 4–5 mice per cage with access to food and water ad libitum. Mice were maintained on an alternating 12 h light-dark cycle, with controlled room temperature (21 ± 1 °C) and relative humidity (40–60 %). The maximal tumour size permitted in the animal license is 1.5 cm3 for a single tumour. 200,000 4599 mouse melanoma cells in 30 µL PBS were injected intradermally into NOD. Cg-Prkdc^{scid} Il2rg^{tm1Wjl}/SzJ mice (NSG; purchased from Charles River UK). Tumours were allowed to grow during 30 days after intradermal injection[31] and primary tumours and Lungs with apparent presence of macrometastasis were collected. Tissues were fixed in 4% formaldehyde for 24 h at 4 °C, following this, tissues were transferred to histology cassettes and dehydrated for 1 h in 70% ethanol, 2 h in 96% ethanol, 2 h in 100% ethanol and 1.5 h in xylene. Following dehydration, samples were transferred to a paraffin bath set to 60 °C for 1.5 h. Samples were then embedded using an embedding machine and allowed to solidify on a cold plate. Embedded samples were kept at 4 °C prior to sectioning.

Lung colonisation assay. 4599, 4599 TB and 4599 DIF murine melanoma cells were used. 4599 TB were mixed with 4599 DIF with different dyes (see Colour coding vital dye cell staining). 4599 naïve cells were primed for 7 days with TeloCol (1.7 mg/ml) non-aligned (H_2O) and aligned ($MgCl_2$) collagen matrices.

Cells were labelled with different vital dyes following above protocol. Priming experiments required extracting cells from collagen using collagenase (see collagenase section) and expanded on plastic for 6 days. Equal numbers of CMFDA-Green and CMRA-Orange labelled cells were mixed (500,000 cells of each type), resuspended in 0.2 ml PBS 1x and injected into tail vein of 9–10-week-old female NOD/SCID/ IL2Rγ-/- mice (NSG, Charles River). Experiments were repeated inverting the colours to ensure reproducibility regardless of the vital dye used.

Mice were sacrificed after 30 min (to confirm equal cell numbers had been injected and reached the lung) and 48 h after tail vein injection to assess lung colonisation. Lungs were extracted, washed with PBS 1X (with calcium/magnesium) twice and fixed with 4% formaldehyde for 16 h at 4 °C and counterstained with DAPI (1:2000 dilution). Lungs were examined under a Zeiss LSM 880 confocal microscope (Carl Zeiss) with a 20X objective and imaging 2 × 2 tiles. Data are presented as percentage of field of area covered by fluorescence, and 15 fields per mouse were analysed. $n = 7$ mice/condition for each experiment (2 mice sacrificed after 30 min and 5 mice 48 h after tail vein injection), $n = 2$ independent experiments.

QuPath was used for image analysis of the lung tissues. An average thresholder from three channels (Blue, Green and Orange) was created and then used as a general mask and as a single annotation. Two independent thresholders were created for green and Orange using simple thresholder tool. Thresholder masks for Green and Orange were created as detection measurements inside the higher-level annotation.

Histopathology

Human case selection. We used formalin-fixed paraffin-embedded (FFPE) of human primary melanomas (53 patients), and metastasis (45 patients); breast cancer (BRCA) samples (30 patients) obtained from Hospital Arnau de Vilanova, Lleida and Hospital de Bellvitge, Barcelona (Spain) (See supplementary Data 1–3). Tumours were classified following the most recent World Health Organization criteria. The study was approved by the Ethical Committees of both Hospitals, and a specific informed consent was used. A tissue arrayer device (Beecher Instruments, Silver Spring, MD, U.S.A.) was used to construct the tissue microarrays (TMA). Two selected cylinders (1 mm in diameter for Melanoma; 0.4 mm for BRCA) from two different areas targeting TB and IF which contain PIF and DIF were included in each case.

Tissue staining. For Masson trichrome staining, samples were rehydrated, stained with Weigert's iron hematoxylin working solution for 10 min, stained in Biebrich scarlet-acid fuchsin solution for 10–15 min, differentiated in phosphomolybdic-phospho tungstic acid solution for 10–15 min, transferred to directly (without rinse) aniline blue solution and stained for 5–10 min. Finally, samples were rinsed briefly in distilled water and treated with 1% acetic acid solution for 2–5 min. Gomori staining was performed at Hospital Arnau de Vilanova, Lleida according to the manufacturers' instructions.

For immunohistochemistry, TMA blocks were sectioned (3 µm-thick) and dried for 1 h at 65 °C before deparaffinization and rehydration. Antigen retrieval was performed in Antigen Unmasking Solution pH 6 (H-3300, Vector Labs) using a pressure cooker system (110 C for 10 min). Samples were washed in Dako Wash Buffer (S3006). Primary antibody incubation was performed overnight at 4 °C with the antibodies listed below and concentration as explained in the antibodies and inhibitors section. Primary antibodies were diluted in Antibody Diluent Reagent Solution (003218, Invitrogen/ThermoFisher Scientific). Samples were washed and incubated with ImmPRESS® polymer secondary goat anti-rabbit antibody (goat anti-rabbit, RTU, Vector Labs, MP-7451) for 45 min. The reaction was developed using VIP peroxidase substrate solution (Vector Labs, SK-4600) for 10 min. All incubations were carried out at room temperature. Slides were counterstained with haematoxylin and mounted using DPX mounting medium (06522-500 ML, sigma). Slides were imaged using the NanoZoomer S210 slide scanner (Hamamatsu, Japan). Multiplex IHC was

used to assess amoeboid features, while staining and re-probing cycles were performed for Melanoma marker or CK-19, pMLC2 and total MLC2. Staining and re-probing cycles were also performed for YAP1, p65 and pSTAT3 staining (adapted from Maiques O[32]).

Digital pathology

Matrix analysis. Masson's trichrome staining: FFPE tissue slides were processed following the pipeline in Supplementary Table 2. A ROI per zone in either TB, PIF or DIF was selected and created in QuPath[24]. Annotated regions were exported using a custom-made script. ROIs for BRCA were 482×482 pixels and for Melanoma 954 × 954 pixels (4.5251pxl/μm, both).

To analyse global matrix changes, we created a specific project in QuPath. Regions of Interest (ROIs) were loaded, and images were duplicated and categorized under Brightfield (other). Color deconvolution was performed using user-defined values. For Masson's Trichrome:

Colour_1: [r1]=0.43 [g1]=0.745 [b1]=0.51; Colour_2: [r2]=0.758 [g2]=0.601 [b2]=0.255; Colour_3: [r3]=0.501 [g3]=0.571 [b3]=0.65. For Gomori's Trichrome *Colour_1: [r1]=0.517 [g1]=0.682 [b1]=0.518; Colour_2: [r2]=0.688 [g2]=0.577 [b2]=0.439; Colour_3: [r3]=0.443 [g3]=0.583 [b3]=0.681*.

The first batch was used to calculate the percentage of matrix and matrix intensity. Fully annotated images were created, and the 'positive pixel count' plugin was executed as follows:

runPlugin('qupath.imagej.detect.tissue.PositivePixelCounterIJ', '{"downsampleFactor":1,"gaussianSigmaMicrons":2.0,"thresholdStain1":0.2,"thresholdStain2":0.1,"addSummaryMeasurements":true,"clearParentMeasurements":true,"appendDetectionParaMETers":false,"legacyMeasurements0.1.2":false}').

Annotations were exported from all cases and Positive % of total ROI area was calculated. Detections were exported from all cases and DAB OD mean value was used as a proxy of stain 2 intensity or matrix values after colour deconvolution was applied.

The second batch of duplicated ROIs used the same colour deconvolution parameters. Subsequently, superpixel segmentation was performed using the Difference of Gaussian (DoG) method as follows:

runPlugin('qupath.imagej.superpixels.DoGSuperpixelsPlugin', *'{"downsampleFactor":1.0,"sigmaPixels":2.0,"minThreshold":10.0,"maxThreshold":200.0,"noiseThreshold":1.0}')*.

Afterwards, intensity values and Haralick features were added to the segmented masks as follows:

runPlugin('qupath.lib.algorithms.IntensityFeaturesPlugin', *'{"downsample":1.0,"region":"ROI","tileSizePixels":200.0,"colorOD":false,"colorStain1":false,"colorStain2":true,"colorStain3":false,"colorRed":false,"colorGreen":false,"colorBlue":false,"colorHue":false,"colorSaturation":false,"colorBrightness":false,"doMean":true,"doStdDev":false,"doMinMax":false,"doMedian":false,"doHaralick":true,"haralickDistance":1,"haralickBins":32}')*.

Haralick entropy, calculated from the gray-level co-occurrence matrix (GLCM), was used to assess the complexity of collagen matrix texture, with higher entropy values indicating greater alignment and organization of collagen fibres.

Detection measurements from each ROIs (TB, PIF, and DIF) were exported. A custom-made script in Python calculated the average of "ROI: 1.00 px per pixel: DAB: Haralick Entropy (F8)" per region. The script was developed using Python v3.8.2 and the Pandas library.

Matrix topography

For trichrome staining, Channel 2 was extracted using color deconvolution parameters as outlined in the previous section, applying either Masson's or Gomori's trichrome values, as appropriate. Colour_2 (matrix channel) was selected and converted to 8-bit, Enhanced Contrast and saturated 0.35 and set min:68 and max:255 for all ROIs using Fiji-ImageJ in a custom script for all ROIs. For confocal images, reflectance channels were used without any pre-processing step.

Matrix topographical features such as Length, Width, Straightness and Angle orientation were calculated in CT-FIRE V5 stand-alone plugin Matlab following the author's instructions[26,29]. Minimum fibre length was amended to 10pxl, fibre line width to 0.2. A custom-made script using python 3.8.2 and pandas' library calculated the median of fibre length, width, straightness for each ROI analysed using the files' created and named as HistLEN', 'HistWID', 'HistSTR'. Pixel values were corrected to micron scale, obtained from the original image. For angle orientation: histograms from 0–180 angles were generated comparing PIF and DIF showing the percentage and SEM per bin. For matrix alignment, distance between fibres was calculated using CurveAlign. Alignment values were normalized to number of fibres per ROI and an Alignment score was calculated. For distance between fibres, the median per ROI was calculated using a custom-made script using python 3.8.2 and pandas' library. Median distance between fibres was normalized to microns using the scale factor from the original images. For fibre density, pixels^2 per ROI and scale factor were considered and represented in fibres per mm^2. Overlay maps of fibre alignment were created using standard parameters but changing the "Heatmap Gaussian disc filter sigma (in pixels)" from 4 to 12.

Gap analysis. Confocal reflectance images were processed following MATLAB script provided https://www.actonlab.co.uk/gap-analysis-matlab/ and used previously[66].

Multiplex IHC scoring. TMAs were stained, stripped and reprobed consecutively within the same slide for pMLC2, total MLC2 and tumour markers. In case of melanoma, a pan-Melanoma marker was used, for BRCA TMA CK19 was used as a tumour marker. Using Multiview tool in QuPath, the TMAs of the three consecutive sections were displayed and ROIs were created. Using a custom-made script following the adapted protocol of staining, quantification was performed using QuPath 0.4.4[24] using Virtual multiplex pipeline adapted from Maiques O et al.[32].

In summary, ROIs with overlapping regions from the same tissue with the three different markers (panMelanoma or CK19, pMLC2 and MLC2) were transferred to FIJI-ImageJ. Colour Deconvolution using H AEC channels was performed for each channel. VIP colour channels for three markers and nuclear markers were merged using a Stack. Using rigid registration, the images were aligned, and next pseudo-colour was applied and LUTs were inverted. The stack image was saved in TIFF and loaded in QuPath. Simple threshold for the cancer marker (panMelanoma or CK19) was created. Pixels were created as annotation masks and StarDist segmentation was applied[67]. Next, Intensity thresholds were selected for pMLC2 and MLC2. Detection measurements were exported and means per core were calculated for circularity, and H-score for pMLC2 and MLC2. Same pipeline was applied for mCherry, pMLC2 and MLC2. YAP, p65 and pSTAT3 analysis was done similarly measuring nuclear intensity.

Microarray analysis

HT1080 cells treated with protease inhibitors for 4 and 24 h seeded on top of collagen I and RNA was extracted using Trizol (Invitrogen, Carlsbad, CA) followed by Qiagen RNeasy Midi Kit. cRNA was prepared according to the GeneChip Expression Analysis Technical Manual (Affymetrix), hybridized onto HU133A chips (Affymetrix), and scanned by a GeneArray@ 2500 Scanner (Affymetrix). Each sample was run in triplicate. Microarray data analysis was performed using R (version 2.9.2) (http://www.R-project.org) and Bioconductor version (2.4)[68]. Signal data were normalised using the robust multichip average (RMA) algorithm from the "Affy"[69] package with default parameters.

Network and enrichment analysis of the gene lists was performed using Ingenuity pathway analysis (IPA, Qiagen) and Cytoscape using

ClueGo module[70]. All IPA analysis was carried out with a threshold of p-value < 0.05. The comparisons were carried out between HT1080 control vs 4 h of protease inhibitor, HT1080 control vs 24 h of protease inhibitor. Hierarchical clustering was performed using Gennepattern[71].

RNAseq

4599 murine melanoma cells isolated from different tumour parts (TB, PIF, DIF and MET) were seeded on Fibrillar bovine dermal collagen 1.7 mg/ml (PureCol, Cat. 5005 Advanced BioMatrix) for 24 h DMEM + 10% FBS. Media was changed to DMEM + 1% FBS during 24 h before the RNA extraction. RNA was extracted using TRIzol (Invitrogen, Carlsbad, CA) followed by QIAGEN RNeasy Midi Kit. RNA integrity and quality was evaluated using RNA QubitTM assay. Sequencing was performed using 150 bp paired end sequencing strategy using Illumina sequencing platform and PE150 Technology. The data output was 6 Gb.

Data quality control and filtering was performed using Fastq Software. Reads were mapped to *mus musculus* (GRCm38 – mm10) reference genome using Hisat2 v2.0.5 as mapping tool. Feature Counts v1.5.0-p3 function from HTseq software was used to count the reads numbers mapped to each gene. Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced (FPKM) of each gene were calculated based on the length of the gene and reads count mapped to this gene.

DESeq2 R package (1.20.0) was used obtain differential expression analysis. The resulting p-values were adjusted using the Benjamini and Hochberg's approach for controlling the false discovery rate. Genes with an adjusted p-value <=0.05 found by DESeq2 were assigned as differentially expressed. The DEGs were further used for GO enrichment analysis implemented by the ClusterProfiler R package, in which gene length bias was corrected. GO terms with corrected p-value less than 0.05 were considered significantly enriched by DEGs. Gene overrepresentation, KEGG and Reactome pathway investigation was also performed using ClusterProfiler software and GSEA R package to provide the gene set enrichment analysis. The lists of gene signatures used are related to ECM, angiogenesis, mechano-sensing, cytoskeleton and inflammation.

The list of DEGs are from Novogene. List of DEGs was filtered to only include genes which are FDR < = 0.05. The mouse genes were converted to human orthologous genes using mouse/human ortholog with phenotype mapping (https://www.informatics.jax.org/downloads/reports/index.html). Orthologous pairs from the HMD_HumanPhenotype.rpt (accessed 16-Jun-2023) were used. R package ggvenn (0.1.10) was used to create the Venn diagram. The common gene signature list (n = 325) is available in Supplementary Data 4.

Spatial transcriptomics

Optimal Cutting Temperature compound (OCT) frozen tumour samples (n = 4) were obtained from Cresciolli et al.[72]. Patient information included in ref. 72. Corresponding H&E from spatial transcriptomics analysis was used for Digital pathology analysis. Tissue scans were loaded in Qupath and cell detection plus machine-learning classification for Tumour, Stroma and Immune cells was performed. Visium 10x spatial transcriptomics of a melanoma tumour was used, showing the H&E staining, cell annotation and region of interest selection using QuPath. QuPath information was subsequently projected onto Visium spots.

Processed Visium samples from non-consecutive tumour slices were obtained upon request from Cresciolli et al.[72] as Seurat objects. FASTQ files were aligned to human reference genome (GRCh38-2020-A) using SpaceRanger v1.3.1 with default parameters. Seurat package version 4.3.0.1 was used with default parameters to create Seurat object[73]. Regions of interest (tumour body, invasive front) were manually annotated in QuPath and merged with Seurat objects.

The list of common genes (n = 325) as obtained in section (xyz) was used to calculate the enrichment score per spot using the AddModuleScore function of Seurat with default parameters. This was subsequently projected onto the tissue and summarized as boxplots. Additional enrichment scores were calculated for selected gene ontology pathways, using gene overrepresentation pathway definition version c5.all.v2023.1.Hs.symbols.gmt obtained from www.gsea-msigdb.org.

MIS signature in patients

Clinical and transcriptomic data in relative z-scores were downloaded from cBioportal (https://www.cbioportal.org/) from 32 different TCGA projects included in the TCGA PanCancer Atlas Studies collection[74] and breast cancer patients from the METABRIC collection[75]. Patients were given a unique value for the gene signature calculated using the mean value of the individual z-scores of the genes included in the signature and separated between "high" and "low" signature using the median value of the signature expression. Kaplan-Meier survival curves and log-rank statistical values were obtained in R by using the package "Survival" and calculated for each TCGA project independently.

Statistics & reproducibility

Statistical analysis was carried out on 3 or more independent experiments, using Graphpad Prism 9 and 10 (GraphPad Software, San Diego California USA, www.graphpad.com). For western blots and data sets exhibiting normal distribution, two-tailed Student's t-test was performed to compare two sets of data, while ANOVA (Analysis of Variance) with Tukey's post-hoc test was carried out to compare 3 or more independent sets of data. In case of larger data sets exhibiting a non-normal distribution, non-parametric tests were used. A Mann-Whitney test was performed to compare two sets of data, while a non-parametric ANOVA with Kruskal–Wallis multiple comparisons test was carried out to compare 3 or more independent sets of data. The degree of statistical dependence of two variables was determined by Spearman's correlation test and Pearson's p-value calculated using GraphPad Prism. The threshold for statistical significance was set to a P value of less than 0.05. No statistical methods were used to calculate sample size. The investigators were not blinded to allocation during experiments and outcome assessment. The results of the experiments provided quantitative data that was analysed with the appropriate statistical tests to evaluate differences and statistical significance. Immunohistochemical analyses were blinded to clinical patient information.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

RNAseq data and Microarray information have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus. HT1080 microarray was deposited in Gene Expression Omnibus repository with reference number GSE197310. 4599 ex vivo TB, PIF, DIF and MET RNAseq was deposited in Gene Expression Omnibus repository with reference number GSE263097. All study data are included in the article and supporting information. Source data are provided as a Source Data file. Codes for image processing, data curation and plotting have been deposited in github https://github.com/mAIqPath/ECM_Transcription. Source data are provided with this paper.

References

1. Winkler, J., Abisoye-Ogunniyan, A., Metcalf, K. J. & Werb, Z. Concepts of extracellular matrix remodelling in tumour progression and metastasis. *Nat. Commun.* **11**, 5120 (2020).
2. Frantz, C., Stewart, K. M. & Weaver, V. M. *The Extracellular Matrix at a Glance*. *J. Cell Sci.* **123**, 4195–4200 (Company of Biologists, 2010).

3. Eble, J. A. & Niland, S. The extracellular matrix in tumor progression and metastasis. *Clin. Exp. Metastasis* **36**, 171–198 (2019).
4. Kaur, A. et al. Remodeling of the Collagen Matrix in Aging Skin Promotes Melanoma Metastasis and Affects Immune Cell Motility. *Cancer Discov.* **9**, 64–81 (2019).
5. Ecker, B. L. et al. Age-Related Changes in HAPLN1 Increase Lymphatic Permeability and Affect Routes of Melanoma Metastasis. *Cancer Discov.* **9**, 82–95 (2019).
6. Piersma, B., Hayward, M. K. & Weaver, V. M. Fibrosis and cancer: A strained relationship. *Biochim. Biophys. Acta Rev. Cancer* **1873**, 188356 (2020).
7. Pickup, M. W., Mouw, J. K. & Weaver, V. M. The extracellular matrix modulates the hallmarks of cancer. *EMBO Rep.* **15**, 1243–1253 (2014).
8. Provenzano, P. P. et al. Collagen reorganization at the tumor-stromal interface facilitates local invasion. *BMC Med.* **4**, 38 (2006).
9. Provenzano, P. P. et al. Nonlinear optical imaging of cellular processes in breast cancer. *Microsc Microanal.* **14**, 532–548 (2008).
10. Conklin, M. W. et al. Aligned collagen is a prognostic signature for survival in human breast carcinoma. *Am. J. Pathol.* **178**, 1221–1232 (2011).
11. Provenzano, P. P., Inman, D. R., Eliceiri, K. W., Trier, S. M. & Keely, P. J. Contact guidance mediated three-dimensional cell migration is regulated by Rho/ROCK-dependent matrix reorganization. *Biophys. J.* **95**, 5374–5384 (2008).
12. Ray, A. et al. Anisotropic forces from spatially constrained focal adhesions mediate contact guidance directed cell migration. *Nat. Commun.* **8**, 1–17 (2017).
13. Park, D. et al. Extracellular matrix anisotropy is determined by TFAP2C-dependent regulation of cell collisions. *Nat. Mater.* **19**, 227–238 (2020).
14. Acerbi, I. et al. Human breast cancer invasion and aggression correlates with ECM stiffening and immune cell infiltration. *Integrative. Biol. (U. Kingd.)* **7**, 1120–1134 (2015).
15. Charras, G. & Sahai, E. Physical influences of the extracellular environment on cell migration. *Nat. Rev. Mol. Cell Biol.* **15**, 813–824 (2014).
16. Liu, Y. J. et al. Confinement and low adhesion induce fast amoeboid migration of slow mesenchymal cells. *Cell* **160**, 659–672 (2015).
17. Sanz-Moreno, V. et al. ROCK and JAK1 Signaling Cooperate to Control Actomyosin Contractility in Tumor Cells and Stroma. *Cancer Cell* **20**, 229–245 (2011).
18. Cantelli, G. et al. TGF-beta-Induced Transcription Sustains Amoeboid Melanoma Migration and Dissemination. *Curr. Biol.* **25**, 2899–2914 (2015).
19. Rodriguez-Hernandez, I. et al. WNT11-FZD7-DAAM1 signalling supports tumour initiating abilities and melanoma amoeboid invasion. *Nat. Commun.* **11**, 5315 (2020).
20. Laser-Azogui, A., Diamant-Levi, T., Israeli, S., Roytman, Y. & Tsarfaty, I. Met-induced membrane blebbing leads to amoeboid cell motility and invasion. *Oncogene* **33**, 1788–1798 (2014).
21. Odenthal, J., Takes, R. & Friedl, P. Plasticity of tumor cell invasion: governance by growth factors and cytokines. *Carcinogenesis* **37**, 1117–1128 (2016).
22. Ruprecht, V. et al. Cortical contractility triggers a stochastic switch to fast amoeboid cell motility. *Cell* **160**, 673–685 (2015).
23. Logue, J. S. et al. Erk regulation of actin capping and bundling by Eps8 promotes cortex tension and leader bleb-based migration. *Elife* **4**, e08314 (2015).
24. Bankhead, P. et al. QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 1–7 (2017).
25. Bredfeldt, J. S. et al. Automated quantification of aligned collagen for human breast carcinoma prognosis. *J. Pathol. Inf.* **5**, 28 (2014).
26. Bredfeldt, J. S. et al. Computational segmentation of collagen fibers from second-harmonic generation images of breast cancer. *J. Biomed. Opt.* **19**, 16007 (2014).
27. Chen, X., Nadiarynkh, O., Plotnikov, S. & Campagnola, P. J. Second harmonic generation microscopy for quantitative analysis of collagen fibrillar structure. *Nat. Protoc.* **7**, 654–669 (2012).
28. Löfstedt, T., Brynolfsson, P., Asklund, T., Nyholm, T. & Garpebring, A. Gray-level invariant Haralick texture features. *PLoS One* **14**, e0212110 (2019).
29. Liu, Y. et al. Fibrillar Collagen Quantification With Curvelet Transform Based Computational Methods. *Front Bioeng. Biotechnol.* **8**, 198 (2020).
30. Dhomen, N. et al. Oncogenic Braf Induces Melanocyte Senescence and Melanoma in Mice. *Cancer Cell* **15**, 294–303 (2009).
31. Maiques, O. et al. A preclinical pipeline to evaluate migrastatics as therapeutic agents in metastatic melanoma. *Br. J. Cancer* **125**, 699–713 (2021).
32. Maiques, O. & Sanz-Moreno, V. Multiplex chromogenic immunohistochemistry to stain and analyze paraffin tissue sections from the mouse or human. *STAR Protoc* **3**, 101879 (2022).
33. Shellard, A. & Mayor, R. Collective durotaxis along a self-generated stiffness gradient in vivo. *Nature* **600**, 690–694 (2021).
34. Samuel, M. S. et al. Actomyosin-mediated cellular tension drives increased tissue stiffness and β-catenin activation to induce epidermal hyperplasia and tumor growth. *Cancer Cell* **19**, 776–791 (2011).
35. Fraley, S. I. et al. Three-dimensional matrix fiber alignment modulates cell migration and MT1-MMP utility by spatially and temporally directing protrusions. *Sci. Rep.* **5**, 14580 (2015).
36. Nerger, B. A., Brun, P. T. & Nelson, C. M. Marangoni flows drive the alignment of fibrillar cell-laden hydrogels. *Sci. Adv.* **6**, eaaz7748 (2020).
37. Doyle, A. D., Carvajal, N., Jin, A., Matsumoto, K. & Yamada, K. M. Local 3D matrix microenvironment regulates cell migration through spatiotemporal dynamics of contractility-dependent adhesions. *Nat. Commun.* **6**, 8720 (2015).
38. Grashoff, C. et al. Measuring mechanical tension across vinculin reveals regulation of focal adhesion dynamics. *Nature* **466**, 263–266 (2010).
39. Steinwachs, J. et al. Three-dimensional force microscopy of cells in biopolymer networks. *Nat. Methods* **13**, 171–176 (2016).
40. Dumbauld, D. W. et al. How vinculin regulates force transmission. *Proc. Natl. Acad. Sci. USA* **110**, 9788–9793 (2013).
41. Guilluy, C. et al. The Rho GEFs LARG and GEF-H1 regulate the mechanical response to force on integrins. *Nat. Cell Biol.* **13**, 722–727 (2011).
42. Collins, C. et al. Localized tensional forces on PECAM-1 elicit a global mechanotransduction response via the integrin-RhoA pathway. *Curr. Biol.* **22**, 2087–2094 (2012).
43. Reuther, G. W. et al. Leukemia-associated Rho guanine nucleotide exchange factor, a Dbl family protein found mutated in leukemia, causes transformation by activation of RhoA. *J. Biol. Chem.* **276**, 27145–27151 (2001).
44. Lessey-Morillon, E. C. et al. The RhoA guanine nucleotide exchange factor, LARG, mediates ICAM-1-dependent mechanotransduction in endothelial cells to stimulate transendothelial migration. *J. Immunol.* **192**, 3390–3398 (2014).
45. Fares, J., Fares, M. Y., Khachfe, H. H., Salhab, H. A. & Fares, Y. Molecular principles of metastasis: a hallmark of cancer revisited. *Signal Transduct. Target. Ther.* **5**, 1–17 (2020).
46. Miskolczi, Z. et al. Collagen abundance controls melanoma phenotypes through lineage-specific microenvironment sensing. *Oncogene* https://doi.org/10.1038/s41388-018-0209-0 (2018).
47. Pearce, O. M. T. et al. Deconstruction of a Metastatic Tumor Microenvironment Reveals a Common Matrix Response in Human Cancers. *Cancer Discov.* **8**, 304–319 (2018).
48. Barbazan, J. et al. Cancer-associated fibroblasts actively compress cancer cells and modulate mechanotransduction. *Nat. Commun.* **14**, 1–17 (2023).

49. Elosegui-Artola, A. et al. Matrix viscoelasticity controls spatiotemporal tissue organization. *Nat. Mater.* **22**, 117–127 (2022).
50. Stashko, C. et al. A convolutional neural network STIFMap reveals associations between stromal stiffness and EMT in breast cancer. *Nat. Commun.* **14**, 1–16 (2023).
51. Bera, K. et al. Extracellular fluid viscosity enhances cell migration and cancer dissemination. *Nature* **611**, 365–373 (2022).
52. Bansaccal, N. et al. The extracellular matrix dictates regional competence for tumour initiation. *Nature* **623**, 828–835 (2023).
53. Fraley, S. I. et al. A distinctive role for focal adhesion proteins in three-dimensional cell motility. *Nat. Cell Biol.* **12**, 598–604 (2010).
54. Velez, D. O. et al. 3D collagen architecture regulates cell adhesion through degradability, thereby controlling metabolic and oxidative stress. *Integr. Biol. (Camb.)* **11**, 221–234 (2019).
55. Mistriotis, P. et al. Confinement hinders motility by inducing RhoA-mediated nuclear influx, volume expansion, and blebbing. *J. Cell Biol.* **218**, 4093–4111 (2019).
56. Law, R. A. et al. Cytokinesis machinery promotes cell dissociation from collectively migrating strands in confinement. *Sci. Adv.* **9**, eabq6480 (2023).
57. Orgaz, J. L. et al. Diverse matrix metalloproteinase functions regulate cancer amoeboid migration. *Nat. Commun.* **5**, 4255 (2014).
58. Georgouli, M. et al. Regional Activation of Myosin II in Cancer Cells Drives Tumor Progression via a Secretory Cross-Talk with the Immune Microenvironment. *Cell* **176**, 757–774.e23 (2019).
59. Alfert, A., Moreno, N. & Kerl, K. The BAF complex in development and disease. *Epigenetics Chromatin* **12**, 1–15 (2019).
60. Wu, J. N. & Roberts, C. W. M. ARID1A mutations in cancer: another epigenetic tumor suppressor? *Cancer Discov.* **3**, 35–43 (2013).
61. Cambria, E. et al. Linking cell mechanical memory and cancer metastasis. *Nat. Rev. Cancer* **24**, 216–228 (2024).
62. Saatci, O. et al. Targeting lysyl oxidase (LOX) overcomes chemotherapy resistance in triple negative breast cancer. *Nat. Commun.* **11**, 1–17 (2020).
63. Chitty, J. L. et al. A first-in-class pan-lysyl oxidase inhibitor impairs stromal remodeling and enhances gemcitabine response and survival in pancreatic cancer. *Nat. Cancer* **4**, 1326–1344 (2023).
64. Yao, Y. et al. Pan-Lysyl Oxidase Inhibitor PXS-5505 Ameliorates Multiple-Organ Fibrosis by Inhibiting Collagen Crosslinks in Rodent Models of Systemic Sclerosis. *Int. J. Mol. Sci.* **23**, 5533 (2022).
65. Shawky, A. M., Almalki, F. A., Abdalla, A. N., Abdelazeem, A. H. & Gouda, A. M. A Comprehensive Overview of Globally Approved JAK Inhibitors. *Pharmaceutics* **14**, 1001 (2022).
66. Acton, S. E. et al. Dendritic cells control fibroblastic reticular network tension and lymph node expansion. *Nature* **514**, 498–502 (2014).
67. Weigert, M. & Schmidt, U. NUCLEI INSTANCE SEGMENTATION AND CLASSIFICATION IN HISTOPATHOLOGY IMAGES WITH STARDIST. https://doi.org/10.1109/ISBIC56247.2022.9854534 (2022).
68. Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
69. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (2004).
70. G, B. et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).
71. M, R. et al. GenePattern 2.0. *Nat. Genet* **38**, 500–501 (2006).
72. Crescioli, S. et al. B cell profiles, antibody repertoire and reactivity reveal dysregulated responses with autoimmune features in melanoma. *Nat. Commun.* **14**, 1–21 (2023).
73. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
74. Weinstein, J. N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
75. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

Acknowledgements

VSM lab was supported by: the Institute of Cancer Research; Cancer Research UK (CRUK) C33043/A24478; Breast Cancer Now as part of Programme Funding to the Breast Cancer Now Toby Robins Research Centre; Barts Charity; Worldwide Cancer Research 22-0329 and UKRI grant reference EP/X033392/, Fundacion Alfonso Martin Escudero and by Marie Sklodowska-Curie Action (H2020-MSCA-IF-2014-EF-ST) and King's Overseas Scholarship. GOF's lab was supported by Cancer Research UK [C48390/A21153]. BF was supported by a King's Health Partners' studentship to VSM and GOF. Instituto de Salud Carlos III (ISCIII) through the project "PI18/00573 and PI21/00294" (R.M.M. and A.M.) and co-funded by the European Union; CIBERONC CB16/12/0023 (R.M.M. and X.M.G.). Comunidad de Madrid Atracción de Talento Investigador 2019-T1/BMD-13642 (supported J.L.O.). Blood Cancer UK Project 19001 and Cancer Research UK Centre City of London Centre grant C7893/A26233 (M.C.S. and A.B.). This work was [funded/supported] by Barts Charity (MGU0504) (O.M.). Tumour samples were processed by IRBLleida ((B.0000682) PT17/0015/0027 and PT20/00021) and HUB-ICO-IDIBELL (PT17/0015/0024) Biobanks integrated in "Xarxa de Bancs de Tumors de Catalunya sponsored by Pla Director Oncologia de Catalunya (XBTC)", and Biobank and Biomodels Platform ISCIII. Views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health. We acknowledge, BCI-CMR Advanced Bio-Imaging Facility of QMUL and ICR light-microscopy Facility, Chelsea, for the use, help and advice with microscopes, Andrew Clear and Nadia Rahman from BCI-Pathology facility, for the support in tissue processing and advice in tissue staining, BCI-BSU Animal Facility, for the support with the animal studies, Dr. Mathieu-Benoit Voisin from William Harvey Research Institute, QMUL, for support with the SHG setup, and Brian Nerger currently at John A. Paulson School of Engineering and Applied Sciences (Harvard University, Cambridge, US) for the technical advice to perform alignment experiments. We thank I. Malanchi (The Francis Crick Institute) and Anne Claire Dupuis (independent researcher) for scientific discussions. The graphic representation in the figures were created using BioRender.com.

Author contributions

V.S.-M. was principal investigator, designed the study, supervised experiments, performed some experiments and wrote the manuscript; O.M. designed the study, performed most experiments, and wrote the manuscript. M.C.S. performed in vitro and tissue analysis supervised by A.B. P.P. designed parts of the study and performed experiments; R.L. performed spatial transcriptomics and RNAseq analysis supervised by S.T. and S.K. A.V., E.C-M., Y.L, V.G., Y.J-G, J.L.O. performed in vitro experiments. I.R-H., J.B and S.H performed bioinformatics analysis. Y.A. and C.C.R-A. performed image analysis experiments. J.S. performed in vivo experiments. G.O.F. supervised in vivo experiments performed by B.F. O.C performed AFM measurements supervised by A.E-A. E.M.G. and J.A. performed zymography experiments supervised by C.G. Human tissue materials, trichomic, ethics approvals and relevant anonymized clinical information was provided by M.S., A.M., J.T., X.M-G, R.M.M., P.RC.I., F.W. and K.B. provided Imaging support.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-56299-7.

Correspondence and requests for materials should be addressed to Victoria Sanz-Moreno.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025

[1]Cytoskeleton and metastasis Team, The Breast Cancer Now Toby Robins Research Centre Division of Breast Cancer Research, The Institute of Cancer Research, Chester Beatty Laboratories, London SW3 6JB, UK. [2]Centre for Tumour Microenvironment, Barts Cancer Institute, Queen Mary University of London, John Vane Science Building, Charterhouse Square, London EC1M 6BQ, UK. [3]Randall Division of Cell and Molecular Biophysics, King's College London, New Hunt's House, Guy's Campus, London SE1 1UL, UK. [4]Centre for Haemato-Oncology, Barts Cancer Institute, Queen Mary University of London, John Vane Science Building, Charterhouse Square, London EC1M 6BQ, UK. [5]St. John's Institute of Dermatology, School of Basic and Medical Biosciences, King's College London, SE1 9RT London, UK. [6]Department of Informatics, Faculty of Natural and Mathematical Sciences, King's College London, Bush House, London WC2B 4BG, UK. [7]Cell and Tissue Mechanobiology Laboratory, The Francis Crick Institute, London NW1 1AT, UK. [8]Department of Computer Science, City St George's, University of London, London, UK. [9]Comprehensive Cancer Centre, School of Cancer and Pharmaceutical Sciences, King's College London, London SE1 1UL, UK. [10]CMR Advanced Bio-imaging Facility, Centre for Microvascular Research, Queen Mary University of London, John Vane Science Building, Charterhouse Square, London EC1M 6BQ, UK. [11]University Cote d'Azur, CNRS UMR7284, INSERM U1081, Institute for Research on Cancer and Aging, Nice (IRCAN), Nice, France. [12]Department of Pathology and Molecular Genetics, Hospital Universitari Arnau de Vilanova, University of Lleida, IRBLleida, CIBERONC, Lleida 25198, Spain. [13]Oncologic Pathology Group, IRBLleida, Departments of Experimental Medicine and Basic Medical Sciences, University of Lleida, Lleida 25198, Spain. [14]Department of Pathology, Hospital Universitari de Bellvitge University of Barcelona, IDIBELL, CIBERONC, L'Hospitalet-, Barcelona 08907, Spain. [15]Department of Dermatology, Hospital Universitari Arnau de Vilanova, CIBERONC, University of Lleida, CIBERONC, IRB Lleida, Lleida 25198, Spain. [16]Instituto de Investigaciones Biomédicas Sols-Morreale (IIBM), CSIC-UAM, 28029 Madrid, Spain. [17]Quell Therapeutics, Translation & Innovation Hub, 84 Wood Ln, London W12 0BZ, UK. [18]Light Microscopy Facility, The Institute of Cancer Research, Chester Beatty Laboratories, London SW3 6JB, UK. [19]Integrated Pathology Unit, Division of Molecular Pathology, The Institute of Cancer Research, Sutton, UK. [20]Breast Cancer Research Bioinformatics Group, Chester Beatty Laboratories, London SW3 6JB, UK. [21]Breast Cancer Now Research Unit, School of Cancer & Pharmaceutical Sciences, King's College London, Guy's Cancer Centre, London SE1 9RT, UK. [22]Present address: Cancer Biomarkers & Biotherapeutics, Barts Cancer Institute, Queen Mary University of London, London, UK. [23]These authors contributed equally: Marta C. Sallan, Roman Laddach, Pahini Pandya. ✉e-mail: [redacted]