ORIGINAL ARTICLE

IWJ WILEY

Development of an explainable artificial intelligence model for Asian vascular wound images

Zhiwen Joseph Lo[1,2] ⓘ | Malcolm Han Wen Mak[3] ⓘ | Shanying Liang[1] |
Yam Meng Chan[3] | Cheng Cheng Goh[4] | Tina Lai[4] | Audrey Tan[4] |
Patrick Thng[5] | Jorge Rodriguez[5] | Tillman Weyde[5] | Sylvia Smit[5]

[1]Department of Surgery, Woodlands Health, Singapore, Singapore

[2]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore

[3]Department of General Surgery, Tan Tock Seng Hospital, Singapore, Singapore

[4]Wound and Stoma Care, Nursing Speciality, Tan Tock Seng Hospital, Singapore, Singapore

[5]AITIS - Advanced Intelligence and Technology Innovations, London, United Kingdom

Correspondence
Zhiwen Joseph Lo, Department of
Surgery, Woodlands Health,
17 Woodlands Drive 17, Singapore
737628, Singapore.
Email: zhiwen@gmail.com

Abstract

Chronic wounds contribute to significant healthcare and economic burden worldwide. Wound assessment remains challenging given its complex and dynamic nature. The use of artificial intelligence (AI) and machine learning methods in wound analysis is promising. Explainable modelling can help its integration and acceptance in healthcare systems. We aim to develop an explainable AI model for analysing vascular wound images among an Asian population. Two thousand nine hundred and fifty-seven wound images from a vascular wound image registry from a tertiary institution in Singapore were utilized. The dataset was split into training, validation and test sets. Wound images were classified into four types (neuroischaemic ulcer [NIU], surgical site infections [SSI], venous leg ulcers [VLU], pressure ulcer [PU]), measured with automatic estimation of width, length and depth and segmented into 18 wound and peri-wound features. Data pre-processing was performed using oversampling and augmentation techniques. Convolutional and deep learning models were utilized for model development. The model was evaluated with accuracy, F1 score and receiver operating characteristic (ROC) curves. Explainability methods were used to interpret AI decision reasoning. A web browser application was developed to demonstrate results of the wound AI model with explainability. After development, the model was tested on additional 15 476 unlabelled images to evaluate effectiveness. After the development on the training and validation dataset, the model performance on unseen labelled images in the test set achieved an AUROC of 0.99 for wound classification with

mean accuracy of 95.9%. For wound measurements, the model achieved AUROC of 0.97 with mean accuracy of 85.0% for depth classification, and AUROC of 0.92 with mean accuracy of 87.1% for width and length determination. For wound segmentation, an AUROC of 0.95 and mean accuracy of 87.8% was achieved. Testing on unlabelled images, the model confidence score for wound classification was 82.8% with an explainability score of 60.6%. Confidence score was 87.6% for depth classification with 68.0% explainability score, while width and length measurement obtained 93.0% accuracy score with 76.6% explainability. Confidence score for wound segmentation was 83.9%, while explainability was 72.1%. Using explainable AI models, we have developed an algorithm and application for analysis of vascular wound images from an Asian population with accuracy and explainability. With further development, it can be utilized as a clinical decision support system and integrated into existing healthcare electronic systems.

Key Messages

- Vascular wounds assessment remains challenging given its complex and dynamic nature; artificial intelligence and machine learning methods can aid in wounds analysis.
- Utilizing 2957 Asian vascular wound images, machine learning models were developed to analyse wound images. Explainability methods were used to interpret artificial intelligence decision reasoning.
- The wound image analysis model classifies wound images with 95.9% accuracy (AUC 0.99), makes automatic estimated depth classification and wound measurements with 85.0% (AUC 0.97) and 87.1% (AUC 0.92) accuracy, respectively, and performs wound segmentation with 87.8% accuracy (AUC 0.95).
- With further development, it can be utilized as a clinical decision support system and integrated into existing healthcare electronic systems.

1 | INTRODUCTION

Chronic wounds contribute to a significant healthcare and economic burden worldwide.[1,2] This is no exception in multi-ethnic Singapore, with mean cost per patient-year in excess of US $3000 for an episode of diabetic foot ulcer and more than US $30 000 for major amputation.[2] The course of wound healing can have varying trajectories given the complex and dynamic nature of wound care, and there can be significant variability in the assessment and management across different healthcare professionals. Wound imaging remains a key component of wound assessment.[3] Multiple commercial wound imaging systems are available and can be used as adjuncts in the assessment and monitoring of chronic wounds.[4] These include computer applications, mobile applications and specialized imaging devices. Current wound imaging solutions are limited in their ability to provide wound assessment beyond their physical characteristics such as the size, depth and appearance, which can limit their utility beyond a tool for documentation. Some solutions can also be costly to implement with the need for specialized equipment and can be difficult to tailor to specific patient populations depending on how the programme is developed. Artificial intelligence (AI) in healthcare has seen increasing interest and development[5] across multiple fields including machine vision for image-based diagnosis such as in mammography,[6] endoscopy[7] and many

more. The use of AI and machine learning is a promising tool to deliver individualized and data-driven outcomes in wound care.[8,9] By modelling on large datasets, AI and machine learning methods have the potential to provide individualized with increased efficiency,[10] with convoluted neural networks (CNNs) methods showing promise in wound segmentation.[11,12] The use of explainable AI (XAI) modelling can also help with its integration and acceptance in healthcare systems,[13] helping to address the 'black-box' nature of AI algorithms. Currently, there are no trained and validated AI models for the analysis of vascular wound images. Thus, we aim to develop an AI model for analysing Asian vascular wound images with explainability.

2 | METHODS

We utilized 2957 wound images from a vascular wound image registry from a tertiary institution in Singapore to train, validate and test our AI model. The wound image registry is a prospective database of wound images taken during both inpatient and outpatient wound care sessions on mobile phone devices between 2017 and 2020 as part of routine clinical wound photo-documentation of vascular wounds. Of the 2957 images, a total of 2549 images were used for wound classification, comprising 1212 (47.7%) VLU, 1002 (39.5%) NIU, 203 (8.0%) SSI and 122 (4.8%) PU. For depth measurement, 1366 depth annotated images were used, out of which 544 (39.8%) were dermal layer wounds, 520 (38.1%) epidermal layer wounds and 302 (22.1%) deep wounds. For width and length measurements, 1096 images with rulers were used to obtain the number of pixels per centimetre (ppcm) out of the rulers to train the model to predict the scale in ppcm. Finally, for wound segmentation, 973 annotated segmented images were used. Wound location, depth and segmentation annotation were performed by a clinician (Vascular surgeon) or qualified wound nurse using a wound image annotator application (AITIS Image Annotator Version 1.1.1.0, 2021). Figure 1A, B demonstrate the AITIS wound annotator application. Image annotation components are as listed in Table 1, with 18 features used for wound segmentation. Additional 15 476 unlabelled images from the wounds registry were then subsequently utilized for the model effectiveness and confidence. This study was approved by our institution domain-specific review board (DSRB Ref 2020/01062) and AI committee approval.

2.1 | Data pre-processing

All models were trained using oversampling techniques[14] to ensure classes that were less frequent in the dataset had enough representation compared to the other classes so that the model did not overfit. Other tools tested and used during the development process were CutMix,[15] mixup[16] and swin transformer.[17,18] CutMix allows the model to generalize better and have better object localization capability by randomly replacing a patch from an image with a patch from the training dataset. Mixup generates a weighted combination of random image pairs from training images to improve generalization. An object detection model (YOLO) was used to detect the wound region of interest (ROI), removing unnecessary background.[19]

Images were pre-processed using augmentation techniques such as rotation, vertical and horizontal flip and image sharpening. This ensures that the model accuracy was independent of camera position and rotation (rotations, vertical and horizontal flips), as well as camera resolution (sharpening).

2.2 | Model development

The image dataset was divided into training, validation and test sets of 80%, 10% and 10% respectively. Figure 2 illustrates the process of model development using the training-validation-testing framework. In the training phase, the model is first trained on the training dataset using different parameters and features. In the validation phase, the model's hyperparameters are further finetuned and evaluated on the validation dataset and the best one is chosen based on the validation statistics, accuracy and F1 scores. In the testing phase, the final model is then evaluated on the unseen data of the test dataset. This process generates the final statistics which determine the accuracy of the model.

For model training and validation, convolutional models with pretrained weights from the ImageNet dataset were used.[20] The wound segmentation models were pretrained on a subset of COCO train2017, on the 20 categories present in the Pascal VOC dataset.[21,22] Although the ImageNet and COCO train2017 datasets have very different classes (e.g. horses, cats, dogs, planes, boats, cars, etc.), the weights of these models have been trained to recognize real-life images, so we utilized those weights using transfer learning for our own purposes and classes.[23] Models used during the development process include different versions of DenseNet, MobileNet and ResNet for classification[24–26] (wound type and measurements) and DeepLab, FPN or U-Net for segmentation.[27–29]

2.3 | Model evaluation

Model effectiveness was evaluated using accuracy, F1 score and area under receiver operating characteristic (AUROC)
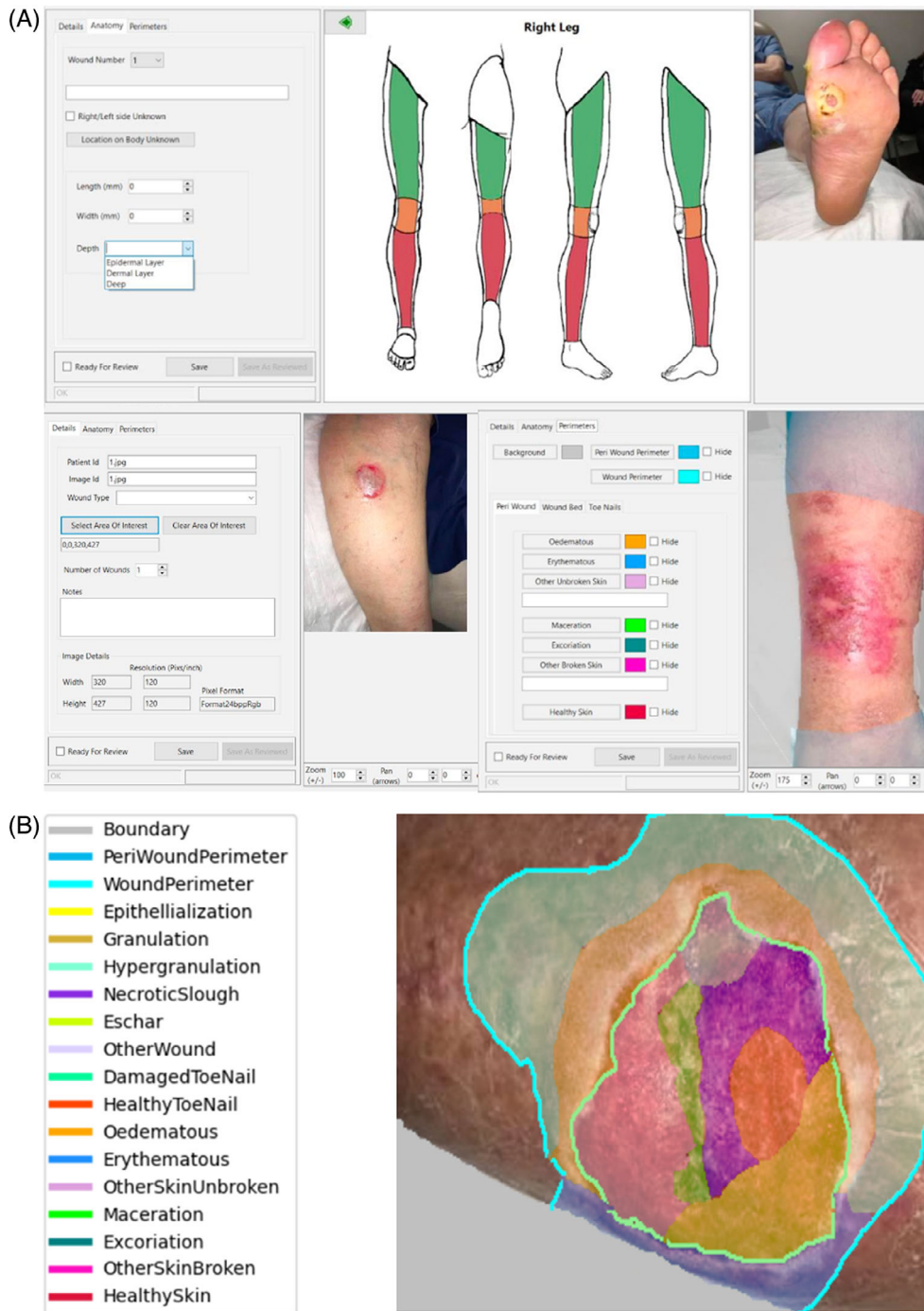
FIGURE 1 (A, B) Wound image annotation using annotator software, specifying wound location, depth and segmentation.

curves. The correct wound image samples were determined by a vascular surgeon and trained wound care specialist nurse, that is, 'ground-truth'. Accuracy is defined as the ratio between correctly classified samples and total number of samples in the evaluation dataset. F1 score is defined as the harmonic mean of the precision and recall.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$F1\, score = 2 \times \frac{precision \times recall}{precision + recall} = \frac{2 \times TP}{2\, X\, TP + FN + FN}$$

2.4 | Model application and explainability

After the model was developed, it was then tested on an additional 15, 476 unlabelled images to evaluate model confidence and explainability. The confidence score is obtained from the predicted class probabilities of the model output, while the explainability score is obtained by determining the attention the model pays to wound area compared to the rest of the image while making its prediction. GradCAM (gradient-weighted class activation mapping),[30] LIME (local interpretable model-agnostic

TABLE 1 Wound image annotation categories and features.

Category	Feature
Wound type	Neuroischaemic ulcer (NIU)
	Surgical site infections (SSI)
	Venous leg ulcers (VLU)
	Pressure ulcer (PU)
Anatomical location of wound	Right or Left
	Thigh/Knee/Calf/Foot
	Medial/Lateral
	Anterior/Posterior
	Malleolus/Ankle/Heel/Midfoot/Forefoot/Toes
	Dorsum/Plantar
Wound measurements	Width/Length (in centimetres to 1 decimal point)
	Measurements: Width/Length (in centimetres to 1 decimal point)
	Depth: Epidermal, Dermal, Deep
Wound segmentation (18 features)	Area of Interest: Wound Boundary, Periwound Perimeter, Wound Perimeter
	Wound Bed Characteristics: Epithelialization, Granulation, Hypergranulation, Necrotic Slough, Eschar, Others
	Peri-wound Characteristics: Oedematous, Erythematous, Maceration, Excoriation, Other Unbroken Skin, Other Broken Skin, Healthy Skin
	Nail characteristics: Healthy, Damaged
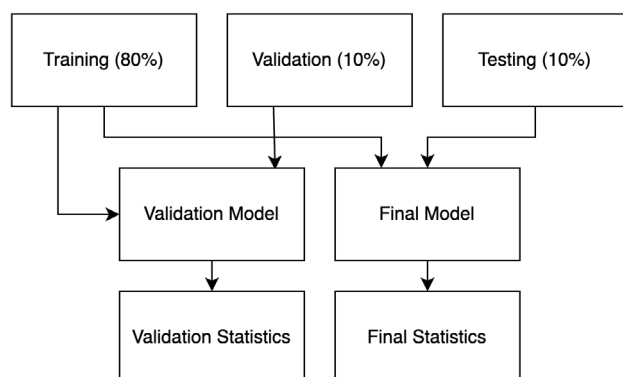
FIGURE 2 Flow diagram of model development.

explanations)[31] and SHAP (Shapley Additive ExPlanations[32] were tested as viable explainability methods, with SHAP used as the preferable explainability method. Figure 3 demonstrates the evaluation of wound image explainability using SHAP. The equation for the derivation of the explainability score is in Appendix A. The higher the explainability score the more certainty that the model is looking into the wound area to make its prediction.

A web browser application was then developed to demonstrate results of the wound AI model with explainability. The application allows the end-user to upload images locally and presents to the user results of wound classification, wound measurements and wound segmentation with accuracy of the model results and confidence of explainability.

3 | RESULTS

During model development in the training and validation phase, different deep learning models were tested, and the best wound classification model obtained an overall 96.3% accuracy with an F1 score of 0.96. The measurements model obtained 92.6% accuracy and F1 score of 0.93 for depth and 87.8% accuracy for width and length estimation. The segmentation model obtained an 88.7% in pixel accuracy. After development on the training and validation dataset, the model was tested for performance on unseen labelled images in the test dataset and achieved an average AUROC of 0.99 for wound classification with a mean accuracy of 95.9%. For wound measurements, the model achieved an average AUROC of 0.97 with a mean accuracy of 85.0% for depth classification, and average AUROC of 0.92 with a mean accuracy of 87.1% for width and length determination. For wound segmentation, an average AUROC of 0.95 and accuracy of 87.8% was obtained.

Figure 4A–D demonstrates the ROC curves for model prediction for wound classifications, wound measurement and wound segmentation respectively. To further evaluate the model after its development, we tested the model on 15 476 unlabelled images and evaluated the model confidence and explainability scores in prediction. The model confidence score for wound classification was 86.8%, 94.9%, 75.2% and 66.6% for NIU, SSI, VLU and PU, respectively, with mean confidence of 82.8%, while explainability score was 54.7%, 76.6%, 39.0% and 73.0% respectively. Measurement of width and length obtained a 93.0% confidence score with 76.6% explainability, with depth confidence score of 75.9%, 93.3% and 61.4% for deep, dermal and epidermal layers, respectively, and explainability score of 90.3%, 67.8% and 23.7% respectively. Segmentation confidence score was 83.9% on average for the 18 classes, while the explainability score was 72.1%. Table 2 demonstrates a summary of model development results. Table 3 demonstrates the model results
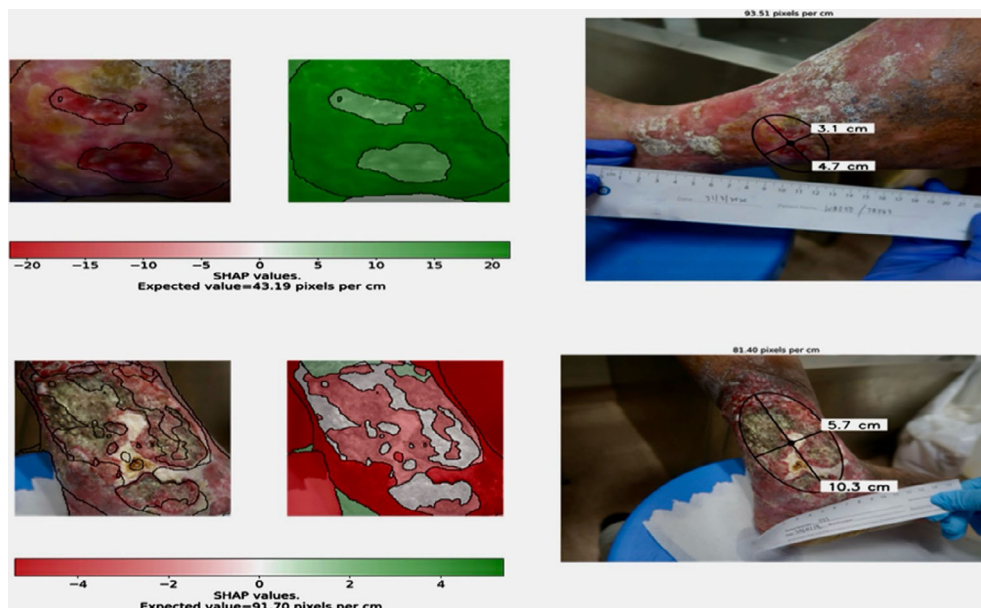
FIGURE 3 Demonstration of model explainability using SHAP. SHAP: Shapley Additive exPlanations.
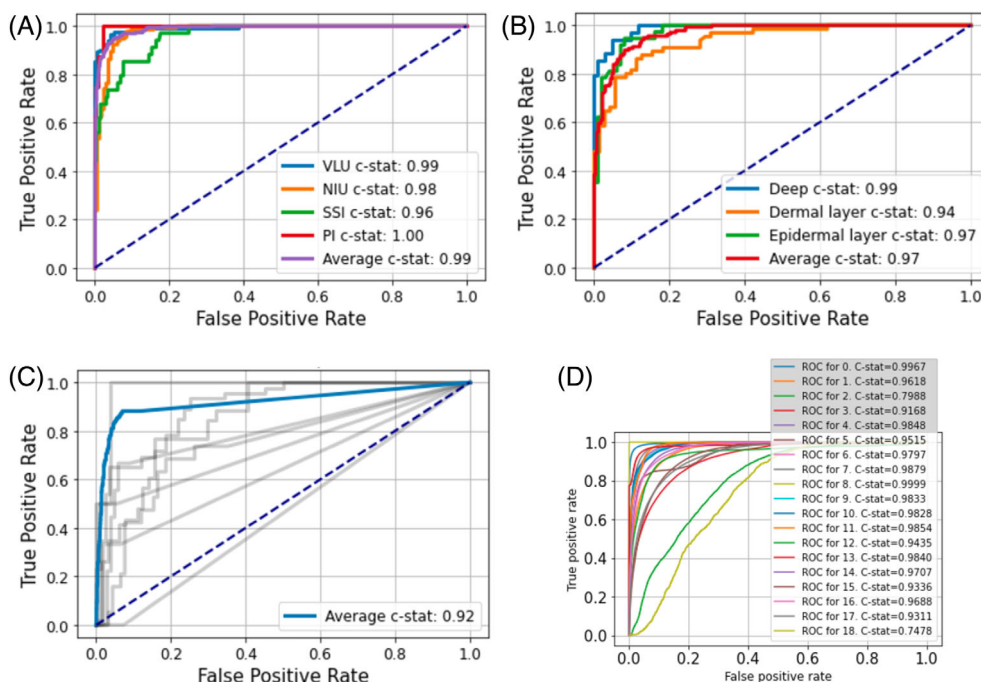

FIGURE 4 (A) Receiver operating characteristic curve for model performance on wound classification. (B) Receiver operating characteristic curve for model performance on wound depth classification. C-stat: C-statistics. (C) Receiver operating characteristic curve for model performance on wound width and length measurement. (D) Receiver operating characteristic curve for model performance on wound segmentation. C-stat, C-statistics; NIU, neuroischaemic ulcer; PU, pressure ulcer; SSI, surgical site infection; VLU, venous leg ulcer.

after testing on unlabelled images. Figures 5, 6 and 7 demonstrate the interface from the web browser application developed for clinical application.

4 | DISCUSSION

Wound care remains an important global healthcare problem. Wound care starts with accurate wound assessment and documentation,[33] and wound characteristics have been shown to predict eventual wound outcomes.[34] Wound classification is a critical first step in assessing severity, healing potential and determining the correct treatment across all wound types.[33] Wound evaluation has traditionally relied on visual assessment by a trained clinician or wound care specialist. Imaging technologies can help to provide measurements of the optical properties of wound components, which can be analysed and interpreted to assess wound severity, healing potential and progress in a rapid, objective and non-invasive

TABLE 2   Summary of model results on the validation (development) and test (final model) datasets. AUROC: area under receiver operating characteristic curve.

		Accuracy (%)	F1 Score	AUROC
Validation (model development) statistics	**Classification**	96.3	0.96	
	Measurement (depth)	92.6	0.93	
	Measurement (width and length)	87.8		
	Wound segmentation	88.7		
Test (final model) statistics	**Classification**	95.9		0.99
	Measurement (depth)	85.0		0.97
	Measurement (width and length)	87.1		0.92
	Wound segmentation	87.8		0.95

TABLE 3   Summary of model results on unlabelled images.

		Confidence score (%)	Explainability score (%)
Wound classification	**Overall, Mean**	82.8	60.6
	NIU	86.8	54.7
	SSI	94.9	76.6
	VLU	75.2	39.0
	PU	66.6	73.0
Measurement (depth)	**Overall, Mean**	87.6	68.0
	Deep	75.9	90.3
	Dermal	93.3	67.8
	Epidermal	61.4	23.7
Measurement (width and length)		93.0	76.6
Wound segmentation		83.9	72.1

Abbreviations: NIU, neuroischaemic ulcer; PU, pressure ulcer; SSI, surgical site infection; VLU, venous leg ulcer.

FIGURE 5   Demonstration of Wound App showing model results, accuracy and explainability across domains of wound classification, measurement and segmentation.
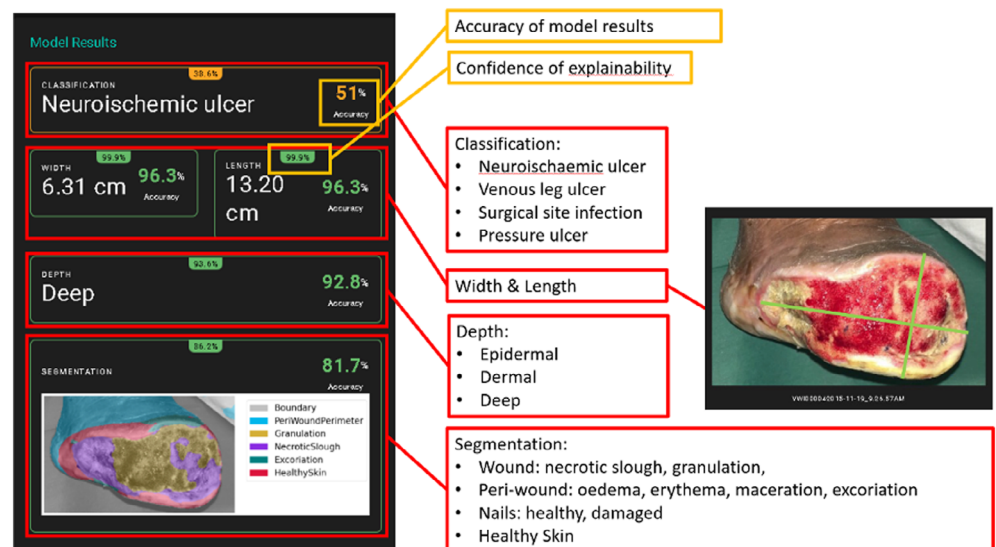
manner. Previous studies have reported the use of AI in wound assessment[11,12,35,36] but are limited by data size and applicability outside their population set. By training our model with a large number of wound images from a multi-ethnic population, this enhances its applicability in wound care in the tropics.
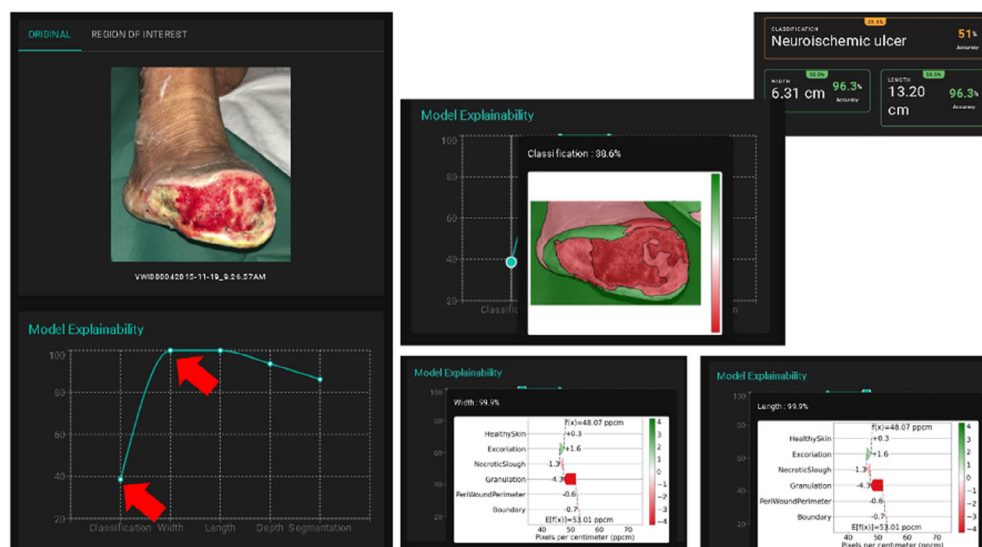
FIGURE 6 Demonstration of Wound App showing model explainability in domains of classification, wound measurement and segmentation.
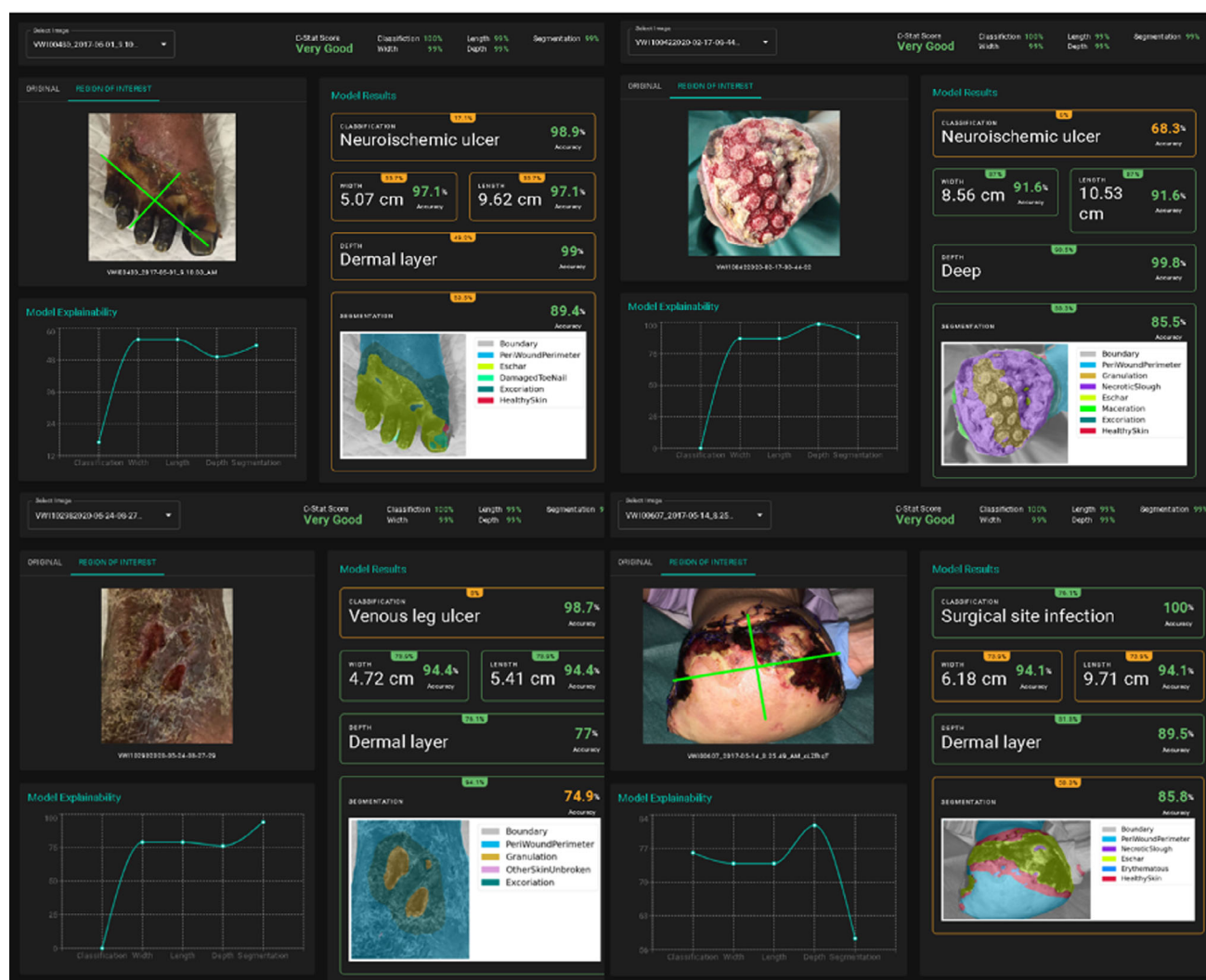

FIGURE 7 Further demonstration of Wound App showing model explainability in domains of classification, wound measurement and segmentation.

Our model performed well in the task of wound classification with a mean accuracy of 95.9%. This is comparable with other wound image classification deep learning models, with mean accuracy values which range between 73.0% and 97.1%.[37,38] Width and length measurements were also accurate with the mean accuracy of 87.1%, while depth accuracy was 85.0%. Wound measurements achieved a mean accuracy of 87.8%. These results likely reflect the difference in dataset numbers, with the largest number of images utilized for wound classifications and are likely to be improved further with a larger dataset. AUROC results of 0.99 for wound classification, 0.97 for wound depth measurements, 0.92 for width and length measurements and 0.95 for wound segmentation were achieved, reflecting the model's discriminative ability as a classifier. Similar results have been obtained in other machine learning models with large datasets, with AUROC of up to 0.96 indicating strong model performance.[39]

Evaluating the model on unlabelled images returns the predicted class probabilities in each category, instead of traditional accuracy metrics which would require a labelled 'ground-truth' to determine the true positive and negative rates. This evaluation allows an estimate of the model performance on a separate, unseen image set. This is further augmented by the explainability score, which provides a measure of determining the attention the model pays to a wound area compared with the rest of the image while making its prediction. The overall confidence score for wound classification was 82.8%, with the highest confidence for SSI at 94.9% and lowest for PU at 66.6%. Although the images are unclassified, this difference is likely due to the high numbers of post-surgical wounds and low numbers of PU in the dataset, which comprises vascular wound images. This similar finding is found with epidermal wounds with the lowest confidence score at 61.4%, again likely due to the lower numbers in a vascular wound image dataset which mostly consists of dermal and deep wounds. Establishing the model confidence scores can also help define a threshold score to accept or reject classifications, which need further clinical studies to determine the level of confidence a model is required to provide to either make a correct clinical decision or avoid clinical error. Future improvements include utilizing this additional dataset to increase the number of annotated images in our original model development image set, which would require further time and resources. Given that the wound classification task had attained 95.9% accuracy, this could be used to further improve wound measurement and wound segmentation performance.

Interpreting XAI remains a challenge in healthcare applications,[40] and its adoption into clinical practice presently presents technological, ethical and medicolegal challenges.[41] Deep learning neural networks are connected via many nonlinear intertwined relations and often considered a 'black box' which can be difficult to interpret and trust.[42] XAI methods are useful to explain AI-based decision making process, especially in medical decision making which is inherently high risk and to increase trust in both the clinician and patient.[43] XAI methods can be developed with either intrinsic or post hoc methods,[44] with LIME and SHAP being popular explainable AI techniques, particularly in the field of computer vision and deep learning.[45,46] We utilized SHAP in our study, as it provided a more accurate explanation of the model's prediction by measuring the contribution of each individual pixel to a feature in the image. Utilizing XAI in model evaluation can also help in model troubleshooting and improvement.[47] Insufficient data remain a significant limitation in model performance. If a low explainability score is encountered, it could mean that the model is not able to identify specific features to derive its prediction. This could either be due to a lack of samples in the dataset, or there could be other features in the image not yet defined and the model is looking at other areas to make its prediction. Categories with poor explainability score in our testing include that for predicting NIU (54.7%), VLU (39.0%) and depth measurement for epidermal wounds (23.7%). While the poor explainability score can be accounted for by lack of images in the category, the same cannot be said for those in NIU and VLU which form the majority of wound images in a vascular wound image dataset, but more likely due to specific features the model looks at to derive its predictions.

Commercial systems are available with good inter-rated reliability, but often require the use of proprietary equipment or software. Examples include Tissue Analytics (Net Health Systems Inc, Florida, USA) for venous leg ulcer measurement, which requires the use of green dot and video capture to analyse the wound,[48] and WoundAide (Konica Minolta Inc, Tokyo, Japan) which requires use of dedicated infrared device.[49] The CARES4-WOUNDS (Tetsuyu, Singapore) system for monitoring diabetic foot ulcers requires a dedicated sensor or a specific smartphone model.[50] The images from our dataset consisted of wound images taken by medical personnel for photo documentation during clinical practice, with significant variation in the device used for image capture, lighting conditions, photo angles, image size and quality. This mimics real-world scenarios better where there may be multiple members involved in wound care in different locations, making standardization of the wound image capture device and conditions difficult. Using AI methods in our study, this eliminates the need for device

standardization and the need for a ruler to make wound measurements. This increases its versatility in use in a variety of clinical scenarios, such as in the community or home-based care.

The clinical utility of an AI-enhanced wound image application is manifold. Wound care and diabetic foot ulcer care often involve multidisciplinary teams and multi-disciplinary care shown to improve limb-related outcomes and cost savings.[51] By having a standardized imaging tool trained on images from the same patient population, it enables the end-user to make wound assessments more confidently and consistently and improve communication between healthcare professionals. This can also be used when patients transit from tertiary care to primary care and vice versa. Patient-centred remote wound care monitoring is also a possible application.[52] The model can also aid in wounds education and quality improvement, by serving as a platform to enable patients and caregivers to have the confidence to monitor their wound progress and for training medical staff who may have less experience in wound care management. The use of XAI also helps clinicians understand how the model derives its conclusion and improves trust in the model. It is not designed as a replacement for clinician decision making but as a decision support system and tool for wound care education. Future works to improve the utility of the model include: First, as the current application is web-browser based, future steps would include integration into the hospital electronic medical record for ease of use. Next, the model can be tested on a separate external cohort for external validation. Third, the imaging system can be utilized as part of a multi-modal analysis including patient clinical characteristics and clinical text notes to provide a holistic predictive model for wound healing assessment.

5 | CONCLUSION

Using explainable AI models, we have developed an algorithm and application for analysis of wound images from an Asian cohort with accuracy and explainability. It can be applied for wound classification, automatic estimation of width, length and depth, as well as wound segmentation. With further data and development, it can be utilized as a clinical decision support system and integrated into existing healthcare electronic systems. Other applications include use for wound care education and as a patient empowerment tool to improve confidence in wound assessment.

CONFLICT OF INTEREST STATEMENT
Co-authors Patrick Thng, Jorge Rodriguez, Tillman Weyde and Sylvia Smit are affiliated and/or employees of AITI Solutions.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID
Zhiwen Joseph Lo https://orcid.org/0000-0003-2289-5266
Malcolm Han Wen Mak https://orcid.org/0000-0002-3711-0946

REFERENCES

1. Olsson M, Järbrink K, Divakar U, et al. The humanistic and economic burden of chronic wounds: a systematic review. *Wound Repair Regen Off Publ Wound Heal Soc Eur Tissue Rep Soc.* 2019;27(1):114-125. doi:10.1111/wrr.12683
2. Lo ZJ, Surendra NK, Saxena A, Car J. Clinical and economic burden of diabetic foot ulcers: a 5-year longitudinal multi-ethnic cohort study from the tropics. *Int Wound J.* 2021;18(3): 375-386. doi:10.1111/iwj.13540
3. Queen D, Harding KG. Importance of imaging to wound care practice. *Int Wound J.* 2023;20(2):235-237. doi:10.1111/iwj.14082
4. Chan KS, Lo ZJ. Wound assessment, imaging and monitoring systems in diabetic foot ulcers: a systematic review. *Int Wound J.* 2020;17(6):1909-1923. doi:10.1111/iwj.13481
5. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. *Artif Intell Healthc.* 2020;25-60. doi:10.1016/B978-0-12-818438-7.00002-2
6. Schaffter T, Buist DSM, Lee CI, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Netw Open.* 2020;3(3):e200265. doi:10.1001/jamanetworkopen.2020.0265
7. Okagawa Y, Abe S, Yamada M, Oda I, Saito Y. Artificial intelligence in endoscopy. *Dig Dis Sci.* 2022;67(5):1553-1572. doi:10.1007/s10620-021-07086-z
8. Anisuzzaman DM, Wang C, Rostami B, Gopalakrishnan S, Niezgoda J, Yu Z. Image-based artificial intelligence in wound

assessment: a systematic review. *Adv Wound Care*. 2022;11(12): 687-709. doi:10.1089/wound.2021.0091

9. Woods JS, Saxena M, Nagamine T, et al. The future of data-driven wound care. *AORN J*. 2018;107(4):455-463. doi:10.1002/aorn.12102

10. Cross K, Harding K. Risk profiling in the prevention and treatment of chronic wounds using artificial intelligence. *Int Wound J*. 2022;19(6):1283-1285. doi:10.1111/iwj.13952

11. Ohura N, Mitsuno R, Sakisaka M, et al. Convolutional neural networks for wound detection: the role of artificial intelligence in wound care. *J Wound Care*. 2019;28(Sup10):S13-S24. doi:10.12968/jowc.2019.28.Sup10.S13

12. Goyal M, Reeves ND, Rajbhandari S, Ahmad N, Wang C, Yap MH. Recognition of ischaemia and infection in diabetic foot ulcers: dataset and techniques. *Comput Biol Med*. 2020; 117:103616. doi:10.1016/j.compbiomed.2020.103616

13. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*. 2021;113:103655. doi:10.1016/j.jbi.2020.103655

14. Ming. Imbalanced Dataset Sampler. 2023 Accessed June 6, 2023. https://github.com/ufoym/imbalanced-dataset-sampler

15. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y. CutMix: regularization strategy to train strong classifiers with localizable features. 2019 Accessed June 5, 2023. http://arxiv.org/abs/1905.04899

16. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. Mixup: beyond empirical risk minimization. 2018 Accessed June 5, 2023. http://arxiv.org/abs/1710.09412

17. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2021. doi:10.48550/arXiv.2010.11929

18. Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. 2021. doi:10.48550/arXiv.2103.14030

19. Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. 2022. doi:10.48550/arXiv.2207.02696

20. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009:248-255. doi:10.1109/CVPR.2009.5206848

21. Lin TY, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. 2015. doi:10.48550/arXiv.1405.0312

22. Everingham M, van Gool L, Williams CKI, Winn J, Zisserman A. The Pascal visual object classes (VOC) challenge. *Int J Comput Vis*. 2010;88(2):303-338. doi:10.1007/s11263-009-0275-4

23. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. 2010;22(10):1345-1359. doi:10.1109/TKDE.2009.191

24. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. 2018. doi:10.48550/arXiv.1608.06993

25. Howard A, Sandler M, Chu G, et al. Searching for MobileNetV3. 2019. doi:10.48550/arXiv.1905.02244

26. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2015 Accessed August 12, 2023. http://arxiv.org/abs/1512.03385

27. Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. 2017. doi:10.48550/arXiv.1706.05587

28. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. 2017. doi:10.48550/arXiv.1612.03144

29. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015. doi:10.48550/arXiv.1505.04597

30. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis*. 2020;128(2): 336-359. doi:10.1007/s11263-019-01228-7

31. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the predictions of any classifier. 2016. doi:10.48550/arXiv.1602.04938

32. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, Vol. 30*. Curran Associates, Inc.; 2017 Accessed June 5, 2023. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

33. Russell L. The importance of wound documentation and classification. *Br J Nurs*. 1999;8(20):1342-1343. doi:10.12968/bjon.1999.8.20.1342

34. Bender C, Cichosz SL, Pape-Haugaard L, et al. Assessment of simple bedside wound characteristics for a prediction model for diabetic foot ulcer outcomes. *J Diabetes Sci Technol*. 2020; 15(5):1161-1167. doi:10.1177/1932296820942307

35. Wang L, Pedersen PC, Agu E, Strong D, Tulu B. Boundary determination of foot ulcer images by applying the associative hierarchical random field framework. *J Med Imaging Bellingham Wash* 2019;6(2):024002. doi:10.1117/1.JMI.6.2.024002.

36. Wang L, Pedersen PC, Agu E, Strong DM, Tulu B. Area determination of diabetic foot ulcer images using a cascaded two-stage SVM-based classification. *IEEE Trans Biomed Eng*. 2017; 64(9):2098-2109. doi:10.1109/TBME.2016.2632522

37. Rostami B, Anisuzzaman DM, Wang C, Gopalakrishnan S, Niezgoda J, Yu Z. Multiclass wound image classification using an ensemble deep CNN-based classifier. *Comput Biol Med*. 2021;134:104536. doi:10.1016/j.compbiomed.2021.104536

38. Anisuzzaman DM, Patel Y, Rostami B, Niezgoda J, Gopalakrishnan S, Yu Z. Multi-modal wound classification using wound image and location by deep neural network. *Sci Rep*. 2022;12(1):20057. doi:10.1038/s41598-022-21813-0

39. Echle A, Ghaffari Laleh N, Quirke P, et al. Artificial intelligence for detection of microsatellite instability in colorectal cancer-a multicentric analysis of a pre-screening tool for clinical application. *ESMO Open*. 2022;7(2):100400. doi:10.1016/j.esmoop.2022.100400

40. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy Basel Switz*. 2020;23(1):18. doi:10.3390/e23010018

41. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20:310. doi:10.1186/s12911-020-01332-6

42. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based

medical image analysis. *Med Image Anal*. 2022;79:102470. doi: 10.1016/j.media.2022.102470

43. Chaddad A, Peng J, Xu J, Bouridane A. Survey of explainable AI techniques in healthcare. *Sensors*. 2023;23(2):634. doi:10.3390/s23020634

44. Zhang Y, Weng Y, Lund J. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagn Basel Switz*. 2022; 12(2):237. doi:10.3390/diagnostics12020237

45. Sarp S, Kuzlu M, Wilson E, Cali U, Guler O. The enlightening role of explainable artificial intelligence in chronic wound classification. *Electronics*. 2021;10(12):1406. doi:10.3390/electronics10121406

46. Aldughayfiq B, Ashfaq F, Jhanjhi NZ, Humayun M. Explainable AI for retinoblastoma diagnosis: interpreting deep learning models with LIME and SHAP. *Diagnostics*. 2023;13(11): 1932. doi:10.3390/diagnostics13111932

47. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745-e750. doi:10.1016/S2589-7500(21)00208-9

48. Fong KY, Lai TP, Chan KS, et al. Clinical validation of a smartphone application for automated wound measurement in patients with venous leg ulcers. *Int Wound J*. 2023;20(3):751-760. doi:10.1111/iwj.13918

49. Chan KS, Liang S, Cho YT, et al. Clinical validation of a machine-learning-based handheld 3-dimensional infrared wound imaging device in venous leg ulcers. *Int Wound J*. 2022; 19(2):436-446. doi:10.1111/iwj.13644

50. Chan KS, Chan YM, Tan AHM, et al. Clinical validation of an artificial intelligence-enabled wound imaging mobile application in diabetic foot ulcers. *Int Wound J*. 2022;19(1):114-124. doi:10.1111/iwj.13603

51. Lo ZJ, Chandrasekar S, Yong E, et al. Clinical and economic outcomes of a multidisciplinary team approach in a lower extremity amputation prevention programme for diabetic foot ulcer care in an Asian population: a case-control study. *Int Wound J*. 2022;19(4):765-773. doi:10.1111/iwj.13672

52. Keegan AC, Jeddah D, Lev-Ari E, et al. Feasibility of remote wound care: a pilot study implementing a patient-centered remote wound monitoring system using a smartphone application. *J Vasc Surg*. 2023;77(6):e195-e196. doi:10.1016/j.jvs.2023.03.265

> **How to cite this article:** Lo ZJ, Mak MHW, Liang S, et al. Development of an explainable artificial intelligence model for Asian vascular wound images. *Int Wound J*. 2023;1-14. doi:10.1111/iwj.14565

APPENDIX 1: Derivation of explainability score.

The explainability score was obtained using two segmentation models: segmentation level 1 (wound, periwound perimeter, wound perimeter and background) and level 2 (using 18 different wound segmentation classes).

Segmentation level 1 provides the basic regions of the wound image. Each one is assigned a weight based on its closeness to the wounded tissues as shown in the image below where the brighter the pixel on the image the closest it is to a wound. It goes from 0.3 (background) to 1 (wounded tissue).

Segmentation level 2 provides the different regions of the image to analyse, and each one is given a SHAP value according to its contribution to the total outcome of the model. If a region contributes positively to predict the final class, then the SHAP value is positive (green). If a region contributes negatively to predict the final class, then the SHAP value is negative (red). The greater (or smaller) the SHAP value the more (or less) contribution to the final class prediction. The figure below demonstrates SHAP values associated with each segmentation region of a wound:
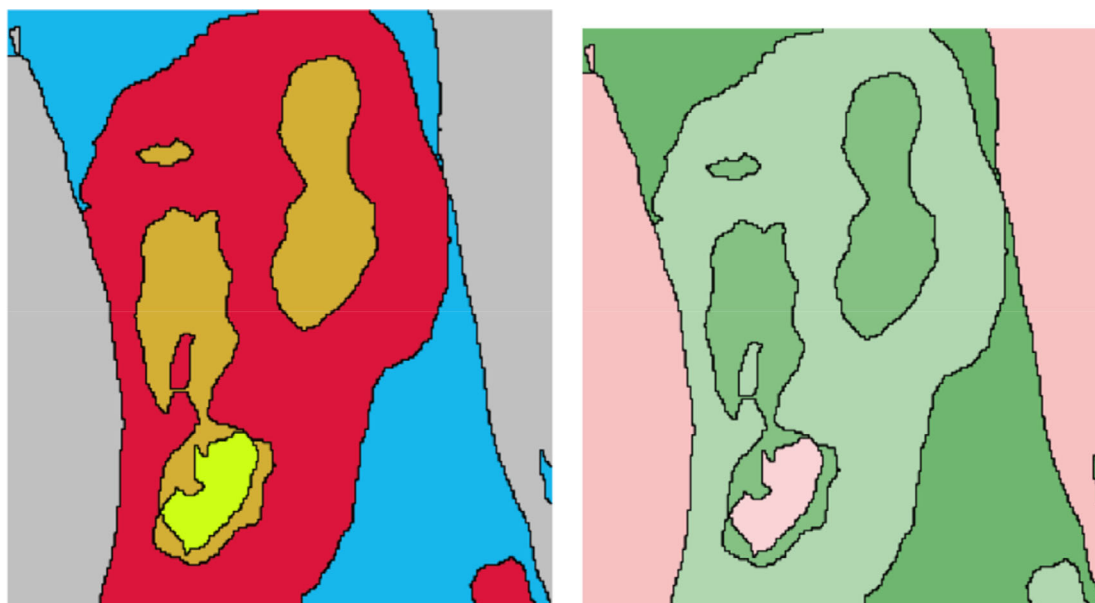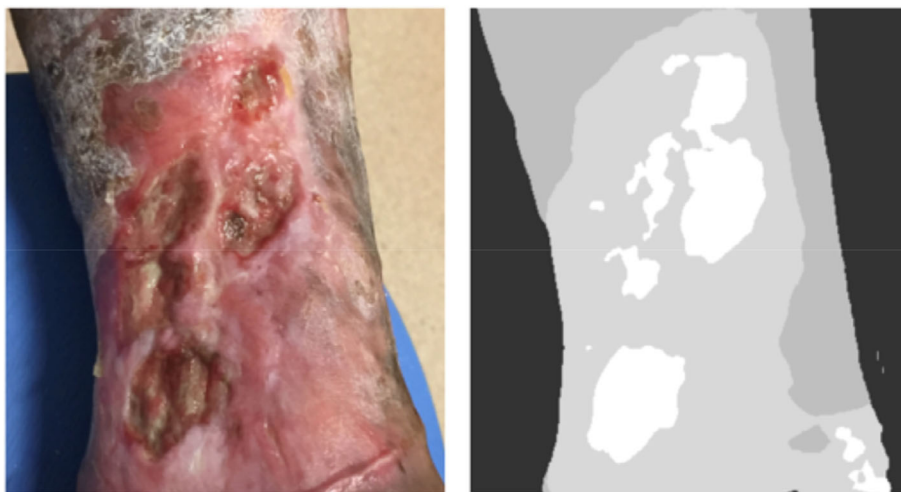
Each pixel is then associated with a SHAP value and a value representing its closeness to the wound. We define the following variables:

$$T = \sum_{i=0}^{P} p_{shap}(i)^4$$

$$f_{neg} = \left(\sum_{\substack{i=0 \\ if\ p_{shap}(i)<0}}^{P} \left(p_{shap}(i) \cdot p_{closeness}(i) \right)^4 \right) / T$$

$$f_{pos} = \left(\sum_{\substack{i=0 \\ if\ p_{shap}(i)>0}}^{P} \left(p_{shap}(i) \cdot p_{closeness}(i) \right)^4 \right) / T$$

Each of these variables represent:

T is the sum of all SHAP values raised to the power of four.

f_{neg} is the fraction of pixels with negative SHAP values weighted by the closeness of each pixel to the wound.

f_{pos} is the fraction of pixels with positive SHAP values weighted by the closeness of each pixel to the wound.

In the end, the explainability score is determined by:

$$exp = 100 \cdot \frac{f_{pos} - f_{neg}}{1 - f_{neg}}$$

This score penalizes negative SHAP values (regions not contribution to the output) and increases if the SHAP values are positive, taking into consideration the closeness of those SHAP values to the wound. Positive SHAP values closer to the wound are more favourable than other SHAP values not closer to the wound. This ensures the model is looking at the right places and that the attention is on the wound and not in other non-related regions.