Extracting Meaningful High-Fidelity Knowledge from Convolutional Neural Networks

Kwun Ho Ngan
Data Science Institute
City, University of London
London, EC1 0HB, UK
kwun-ho.ngan@city.ac.uk

Artur d'Avila Garcez
Data Science Institute
City, University of London
London, EC1 0HB, UK
a.garcez@city.ac.uk

Joseph Townsend
Fujitsu Research of Europe Ltd
Slough, SL1 2BE, UK
joseph.townsend@fujitsu.com

***Abstract*—Convolutional Neural Networks (CNNs) have been widely used for complex image recognition tasks. Due to the highly entangled correlations learned by the latent features in the convolutional kernels of CNNs, deriving human-comprehensible knowledge from CNNs has been proven difficult. As such, reasoning from relationships between kernels has been limited, resulting in little knowledge transfer from one task to another related task learned by CNNs. This paper introduces a neural-symbolic approach for providing semantically meaningful explanations to CNNs using logical rules and a shared conceptual representation space to capture the meaning of the knowledge learned. The validity of the proposed approach is demonstrated using benchmark chest x-rays of two respiratory conditions: pleural effusion and COVID-19. Our results show empirically that symbolic rules can be associated with semantically meaningful explanations obtained from different but related CNN models, even in domains requiring specialised knowledge such as medical imaging. This work is expected to aid the analysis of black-box CNNs by associating the predictions obtained from the CNNs with clinical research findings.**

***Index Terms*—Neural-Symbolic Integration, Convolutional Neural Networks, Concept Representation, Knowledge Sharing, Medical Imaging.**

I. Introduction

Convolutional Neural Networks (CNN) are highly effective at image recognition tasks. They have achieved cutting-edge performance in specialised applications such as clinical diagnosis [1], [2]. Interpretability of individual image predictions is typically carried out by visualising activated regions within a trained CNN (e.g. [3], [4]). Such interpretation can be very subjective and it is often insufficient to produce semantically-rich reasoning from the image's features and their relationships. However, semantically-rich reasoning is needed if we are to gain a deeper understanding of the working processes of a trained CNN model. This reasoning process is typically symbolic, requiring relationships between features to be defined explicitly via rules that use symbols to represent abstract concepts. In this paper, our motivation is to extract symbolic knowledge from CNNs in the form of logical rules denoting semantically-rich concepts to help us understand the underlying decision process. This is demonstrated with benchmark face recognition examples and confirmed using medical image data.

Extraction of relevant knowledge from trained CNNs has always been challenging. Approaches for explaining entire networks have encountered computational complexity issues. Deriving very large rule sets from CNNs can result in limited comprehensibility [5]. Similarly, approaches to explaining individual images failed to produce semantically-rich relationships [6]. A novel knowledge extraction approach is introduced in this paper based on the ERIC (Extracting Relations Inferred from Convolutions) method [7]. Symbolic descriptions are extracted from the CNN's feature extraction layer and aggregated to enable reasoning about the CNN's decision process. We demonstrate how the approach can be applied to different related classification tasks, e.g. detection of pleural effusion and classifying COVID-19 from chest X-rays. By analysing related tasks, we are capable of identifying relevant concepts used by the CNN to produce high-accuracy results that, nevertheless, may not be medically justified. We conclude that extracting meaningful knowledge from such tasks can improve our understanding through comparative evaluations of the prediction outcomes produced by CNNs in related domains.

ERIC constructs a compact set of sentence-like explanations for CNNs by extracting logical rules from the feature detectors. It offers a post-hoc decompositional explanation for image classifications. ERIC has been proven to generate logical rules that closely approximate the original CNN model, as measured empirically by high fidelity scores, i.e. the rules' accuracy relative to the CNN [7]. Each logical rule takes the form $L_1 \wedge L_2 \wedge ... \wedge L_n \rightarrow A$ denoting a conjunction of literals $L_i, 1 \leq i \leq n$, each of which is either a propositional atom or its negation, implying atom A. Each atom can be associated with semantic concepts relating to the input image. However, these semantic concepts are manually assigned in ERIC through visual interpretation of the individual images. This paper introduces an approach for aggregating kernel activation outputs to enable a systematic assignment of se-

mantic concepts to the extracted rules. This approach of intuitively interpreting the atoms and their negation will be shown to improve explainability and reasoning about a CNN's classification task, as well as enable a comparative evaluation between CNNs trained for pleural effusion and COVID-19 classification. The results will be the extraction of a highly compact set of interpretable logical rules from CNNs with high fidelity scores trained on both pleural effusion and COVID-19.

In summary, this paper presents four key contributions. Firstly, a novel method is implemented and evaluated to assign appropriate concepts to logical rules derived by the ERIC explainability system. Using the CNN's kernel norms and activation aggregation, one can visualize concepts and seek to interpret the effect of key features and their negation on a target prediction task. With an adequate assignment of concepts to logical rules, it is then possible to identify instances where a CNN can achieve high accuracy for the wrong reasons, prompting the need for changes in model training or data collection. Finally, the extraction of compact, high-fidelity, semantically-relevant explanations from CNNs trained on pleural effusion and COVID-19 data should allow domain experts to comparatively analyze the outcomes of CNNs trained on related radiology tasks.

The paper is organised as follows. The required preliminaries, including a summary of ERIC, and immediate related work are presented in Sections II and III, respectively. Section IV describes the knowledge extraction method and defines the empirical investigations that follow using an illustrative example of benchmark facial image classification. Experimental results on medical image data are presented in Section V. Section VI concludes the paper and discusses directions for future work.

II. Preliminaries

Let **x** denote a set of input images and **t** denote a set of target outputs, each indexed by the subscript i, $1 \leq i \leq n$. A convolutional neural network, M, is trained on examples $\{x_i, t_i\}$ and is made of two components: $g(\cdot)$ mapping x_i to the output of a feature extraction layer, call it $g(x_i)$, and $h(\cdot)$ mapping $g(x_i)$ to the CNN's output, $h(g(x_i))$. Let $A^l_{i,k}$ denote a matrix of activation values $g(x_i)$ at feature extraction layer l, where $1 \leq k \leq k_l$ denotes a *kernel* of the CNN, represented by a square matrix of real numbers that gets vectorized. Let $b^l_{i,k}$ denote a set of truth-values (*true* or *false*) assigned to each kernel (see Eq.1) by a function Q (see Eq.2) mapping the activation matrix to $\{-1, 1\}$, where -1 denotes *false* and 1 denotes *true*. $b^l_{i,k}$ can be expressed symbolically as either a positive literal $L^l_{i,k}$ when $b^l_{i,k} = 1$, or a negative literal $\neg L^l_{i,k}$ when $b^l_{i,k} = -1$. In Eq.2, $a^l_{i,k}$ is the result of calculating the L1-norm of the kernels in $A^l_{i,k}$ (see Eq.3), and θ^l_k is a threshold value calculated for each kernel as the mean L1-norm value for the entire training set (see Eq.4).

$$b^l_{i,k} = Q(A^l_{i,k}, \theta^l_k) \tag{1}$$

$$Q(A^l_{i,k}, \theta^l_k) = \begin{cases} 1, \text{ if } a^l_{i,k} > \theta^l_k \\ -1, \text{ otherwise} \end{cases} \tag{2}$$

$$a^l_{i,k} = \|A^l_{i,k}\| \tag{3}$$

$$\theta^l_k = \sum_{i=1}^{n} (a^l_{i,k})/n \tag{4}$$

In ERIC, a set of symbolic rules R is generated as an approximation M^* of M using a decision tree-based rule extraction algorithm similar to the C4.5 algorithm [8] trained on instances $\{b^l_{i,k}, h(g(x_i))\}$. Each rule R_r takes the form of a conjunction of literals $L_1 \wedge L_2 \wedge ... \wedge L_{k_l}$, obtained from the feature extraction layer, which implies a target output t_i denoting a classification of the CNN, that is, $L_1 \wedge L_2 \wedge ... \wedge L_{k_l} \rightarrow t_i$. A rule defines a path in the extracted decision tree from the root node to a leaf node in the tree. Tree pruning is applied to avoid overfitting. Tree node branching is calculated using the Gini index. If a leaf node has multiple outcomes after pruning, a class is chosen based on the majority class. The accuracy of the CNN is measured in the usual way as the percentage of input images that are classified correctly w.r.t. t_i. The accuracy of the extracted rules is determined by the percentage of input images classified correctly by the rules also w.r.t. t_i, i.e. the number of times that $R(M^*, x_i) = t_i$ divided by the number of examples, where R denotes the extracted set of rules. The *fidelity* of the rules to the network is defined as the percentage of rule-based classifications that match the CNN's classification as measured by $R(M^*, x_i) = h(g(x_i))$. Qualitative evaluations of the rules are also performed through up-sampling and inspection of literals in the rules against the input images, as illustrated later.

III. Related Work

Convolutional neural networks learn image features via their convolutional and pooling layers using gradient-based parameter searching on a large volume of image data. While the learned features enable trained CNNs to perform well across a range of predictive tasks, the relationships between the features are hidden within the model's large number of training parameters. Numerous techniques have been proposed for interpreting the internal feature representations of CNNs, either by visualising the highest activation of the hidden units at a layer [9], [10] or by up-sampling to the input image to identify the salient features [3], [4], [11]. These feature visualisations have enhanced understanding of relevant pixels in a CNN classification. However, the conceptual interpretations of such visualisations tend to be subjective and may vary between different but related images.

The work in [12]–[14] examined the interpretation of a CNN's hidden units at a global level. Highly activated regions in a network layer were associated with interpretable concepts using a data set compiled to relate images with different concepts at various levels of abstraction (the *Broden* dataset). This approach showed that a level of feature disentanglement can be achieved w.r.t. objects, object parts, texture and colour. Certain features remained uninterpretable. The evaluation was

confined to the vocabulary of the Broden dataset, with concepts beyond the vocabulary rendered uninterpretable. While this bolsters the case for *semantic meaning* in CNNs, a manual compilation of such concept dataset is laborious. It may also not be possible to specify the relevant concepts a-priori in a specialised field, such as radiology. Furthermore, understanding concepts in isolation may not be sufficient to explain how they relate to one another and the target output. Our purpose in this paper of having logical rules $L_1 \wedge L_2 \wedge ... \wedge L_k \rightarrow t_i$ is to specify that the combination of concepts L_1, L_2,..., and L_k imply a specific outcome, t_i.

Representation learning is a sub-field of machine learning that focuses on techniques for transforming raw data (e.g. image pixels) into the appropriate representation required by a learning model [15]. This data transformation automates the process of encapsulating the input data and other contextual information into a tabular form [16]. This tabular form can be regarded as a knowledge-base that retains semantic information about the given data. In [17], three levels of cognitive representations are introduced that bridge the gap between stored numerical representations for deep learning and symbolic representations for logic. At the neural network level, representations are embedded within the activation patterns of a densely connected network. The symbolic level encodes knowledge in logical rules through symbols and their relationships. A third (spatial) level, referred to as a conceptual space, represents data in geometrical, topological or ordinal dimensions. Concepts are learned at this level by comparing data similarities to neighbouring data points within the conceptual space [17]. This general idea will be applied here in the context of our proposed clustering of kernel activation aggregations.

Efforts have also been made over the years to convert knowledge encoded in a neural network into interpretable rules, which can be summarised in three broad classes of *knowledge extraction* methods [18]: (1) *pedagogical* methods explain the output in terms of the input without evaluating the network's internal mechanisms; (2) *decompositional* methods divide the network to extract knowledge from its internal mechanisms (e.g. groups of neurons and weights); (3) *eclectic* methods are those that combine elements of (1) and (2). In this paper, a decompositional method is applied that is suitable for CNNs to achieve efficiency in the rule extraction process given the natural decomposition of functions within CNNs into $g(\cdot)$ and $h(\cdot)$, as mentioned in our problem setting.

As an example of a closely-related pedagogical approach, [19] sought to mimic the input-output function of a neural network by building a soft decision tree based on the hierarchical probability distribution of a class. The decision tree did not rely on the hierarchical features within the network. While [19] reported high accuracy, their evaluation used soft targets from network predictions as training patterns. As a result, obtaining meaningful fidelity measurements is not possible in the case of [19].

ERIC yielded global explanations for one or more convolutional layers as a decompositional rule extraction method. A quantisation process was used to binarize kernels into logical literals. The literals were then used to generate symbolic rules via a logic program that approximates the behaviour of that convolutional layer w.r.t. the CNN's output. High classification accuracy and high fidelity of the approximation M^* of the original CNN were reported in [7]. The results were also evaluated w.r.t. the sizes of extracted rule sets, with smaller sets considered to be more comprehensible by humans. The work in [20] investigated a global layer-wise extraction of rules from CNNs. Kernel outputs were translated into literals for the extraction of *M-of-N* rules, where a rule is interpreted as being *true* if and only if any combination of M literals is *true* out of a set of N literals. Kernels were represented by the outputs of neurons that yielded the maximum information gain. The rule extraction process was accomplished using a heuristic search that prioritised literals according to the weights associated with the respective neurons leading to the target output. Although theoretically sound, this approach can become inefficient for large networks.

The work in [21] described a post-hoc approach in which representations were part disentangled from the trained CNN and rearranged into a hierarchical AND-OR graph. Interpretability was illustrated qualitatively and quantitatively, but the explanations were not converted into a separate, simpler classifier. As a result, no fidelity evaluation could be conducted. In [22], the work was extended to include extraction of decision trees, with kernels specifically trained using a loss function proposed in [23]. We regard the ability to measure fidelity and to apply to any CNN irrespective of the training protocol as key requirements of any knowledge extraction. For this reason, the work presented in this paper has been chosen to build upon ERIC, taking also into consideration ERIC's efficiency at extracting global rules from CNNs.

IV. Method and Experimental Setup

We start by describing the data sets used in this paper in Section IV-A. The training of the CNN models is described in Section IV-B. Symbolic rules are extracted from the trained CNN in Section IV-C. Once symbolic rules are obtained, the processes of kernel activation aggregation (Section IV-D) and concept assignment (Section IV-E) are applied, as described. Depending on the use case, a deeper analysis of the literals used by the rules is required. As an example, an evaluation of the significance of the concept clusters created for the Chest X-ray use case is presented in Section IV-F.

A. Datasets

Three data sets were used in this work to (1) identify gender from facial images, (2) detect pleural effusion, and (3) detect COVID-19. CelebA-HQ [24] was used in (1) to illustrate the proposed knowledge extraction method. The training set included 2,000 images of *male* and *female* faces in equal proportions, as well as 200 images for validation. Only front-facing images were used to simplify image feature interpretation. Fig. A.1 shows sample images from this dataset.

The CheXpert dataset [25] was used in task (2). Frontal X-rays with the classification *pleural effusion* versus *no finding*

were chosen. Images with the class *no finding* were considered to be healthy. Images with artefacts or severely obstructed by supporting devices were removed. Additionally, images with nearly square aspect ratio were chosen to minimise scaling distortion as input to the CNNs. 400 images were used for training and 80 for validation, evenly distributed between the two target classes.

Finally, images of COVID-19 were collected from the IEEE8023 dataset [26] for task (3). A similar image filtering as done in task (2) was performed. 200 images classified as *COVID-19* were combined with healthy images from the CheXpert dataset for training; 40 additional COVID-19 images were used for validation. This established a common baseline of healthy X-rays for our comparative work[1]. All X-ray images were pre-processed with contrast-limited adaptive histogram equalisation (CLAHE) [27] to improve image contrast. Fig.B.1 shows example frontal chest X-rays from each dataset.

B. CNN Model Training

A CNN model (M) based on the VGG-16 architecture was trained using the Adam optimiser and a learning rate of 10^{-6}. The model was trained in batches of 32 images. Elite back-propagation (EBP) was also implemented to improve *class-wise activation sparsity* [28]. This was accomplished when each class was associated with a small number of kernels that activated rarely but strongly for related images, by assigning these as top kernels using a ranking and penalty function for the kernels' activation probabilities during training. EBP was shown to produce a clearer separation of kernel concepts and thus arguably more interpretable representations. When seeking to attach semantic meaning to kernels, the above separation of concepts via EBP can become very useful. As part of our evaluation, CNN models were trained multiple times by shuffling the input images for training and validation using the same proportions but different random seeds. Following the experiments conducted in this paper using VGG-16, the plan is to consider different models and other datasets in future.

C. Symbolic Rule Extraction

Based on ERIC [7], the last convolutional layer, l, of the trained VGG16, M, was quantised and binarized to produce literals and generate rules as a measurable approximation M^* of M. We found that rules with a maximum of 3 literals in the body were sufficient to produce a good approximation, i.e. a high-fidelity score of the rules w.r.t. the original CNN model.

D. Kernel Activation Aggregation

As the first innovation of this paper w.r.t. ERIC, we sought to gain a better understanding of the representation of CNN kernels as logical literals (atoms and negated atoms). To do so, we chose a selection of training images with the highest L1-norm values and we normalised an aggregation of activation values per kernel at the feature extraction layer l. We called this method of aggregating kernel activations *kernel fingerprint*. In this work, the number of images aggregated was set at 10, which appeared to be adequate to provide an initial indication of key locations in the up-sampled images. This value may be adjusted as a hyper-parameter in future. It should be noted that this aggregation was only useful because the images were front-facing, with each anatomical feature (e.g. right eye, left lung, etc.) located in reasonably similar regions within the images. As illustrated in Fig. 1, kernel fingerprints will provide an initial guide to the most relevant regions in a set of images (as opposed to an individual image). When considered in the context of an extracted rule, kernel fingerprints can be seen as a global heat-map of each kernel's contribution to an output class. These kernel fingerprints taken together with the corresponding plots of L1-norm values (e.g. Fig. 2) will provide useful visualisations of atoms, L_i, and their negation, $\neg L_i$, towards offering a contrastive explanation for the meaning of L_i.

For our illustrative CelebA-HQ data set with target classes *male* and *female*, the CNN model achieved an accuracy of 93.0%. In comparison, rules extracted from the CNN obtained an accuracy of 87.3% and fidelity to the CNN of 91.1%. The fingerprints in Fig. 1 represent literals QE, DD, LC and MR, which, without their associated fingerprints, are only meaningless symbol assignments. These fingerprints therefore provide a first indication of the locations of facial features used by the extracted rules in the classification task. The following is an example of an extracted rule relating QE, DD and LC with the *female* class: $\neg$QE $\wedge$ $\neg$DD $\wedge$ LC $\rightarrow$ *female*.

Fig. 1: Kernel fingerprints (showing the intensity of red above 0.5 on a normalised scale) on a sketch processed from a sample image in the CelebA-HQ data set [24]. This provides an initial indication of the location of facial features used by the model explanation for a CNN trained on this dataset. Each fingerprint is associated with a logical literal via a random assignment of symbols: QE, DD, LC and MR from left to right. Our goal is to assign meaning to such symbols.

Taken together with plots of L1-norm values showing the changes in each facial feature (Fig. 2), kernel fingerprints can help domain experts assign meaning to the logic literals. Fig. 2 shows the images associated with the highest and lowest L1-norms in the training set for literal QE. The red threshold line denotes the mean value of the L1-norms for the kernel associated with QE (θ_k^l).[2] To facilitate visualisation, the 1000 examples of images labelled as *female* are plotted on the left, and the 1000 examples labelled as *male* are plotted on the right of the image (blue dots). Examples above the threshold line are

[1]We are aware of the risk of spurious features based on image quality discrepancy when combining datasets. Our work will show that analysis of the generated rules will allow identification of features for further attention.

[2]Following ERIC, the average L1-norm was used in this paper, although it is possible to use (or even learn) different custom threshold values to determine the positive and negative literals.

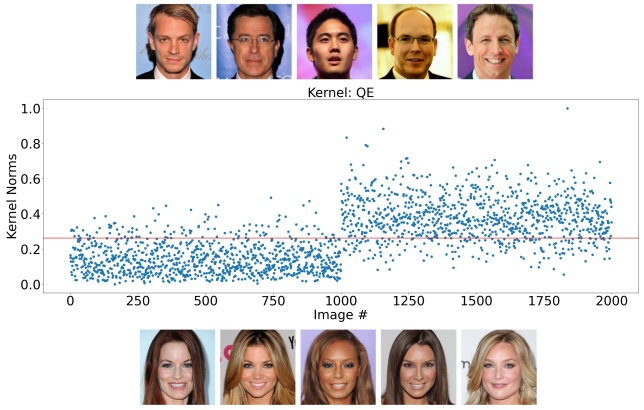


Fig. 2: An example plot of the L1-norm values obtained from the kernel of a CNN associated with literal QE. The first 1000 data points are labelled as *female* and the next 1000 as *male*. A threshold value (red line) separates positive literals QE (above the line) and negative literals ¬QE. The images shown are examples with the highest and lowest L1-norms in the training set. Taken alongside the kernel fingerprint for QE (Fig. 1) and inspecting the most extreme (contrastive) images shown here, this plot reveals the angularity of the lower jawbone changing from angular (QE) to round (¬QE) towards the lower end.

considered to be QE, while those on or below the threshold line are considered to be ¬QE. It can be seen that the majority of *female* images are associated with ¬QE, whereas the majority of *male* images are associated with QE. Upon inspection of the images with the highest and lowest L1-norms, with a focus provided by the QE kernel fingerprint, it is observed that the positive literal (QE) tends to be associated with faces having angular jawbones, as opposed to faces that are oval-shaped with less defined jawbones (¬QE). Visualisations such as shown in Figs.1 and 2 provide a rough definition for literal QE (changes at the jawbone); similarly for DD (changes around the cheekbones just below the eyes), LC (the right brow and eye lid), and MR (the shape of the jaw and chin). Additional contrastive examples of kernel plots are provided in Appendix A.

The relationships between the concepts (QE, DD, etc.) implying a given CNN's output class are defined by the set of rules extracted from the decision tree produced by ERIC. Fig. 3 shows one such tree explaining the CelebA-HQ dataset. The figure shows an image being classified as female given that QE = *false*, DD = *false*, and LC = *true*. The features identified appear to relate closely with some relevant findings from cosmetic surgery research on gender identification (see Appendix A for details). Interestingly, the features identified do not relate closely with available metadata purported to explain the CNN's classification w.r.t. the wearing of makeup or accessories (e.g. lipsticks or earrings).

E. Concept Assignment

The second innovation of this paper is the use of unsupervised learning to translate kernel fingerprints into clinically-relevant concepts. While facial features can be interpreted in principle

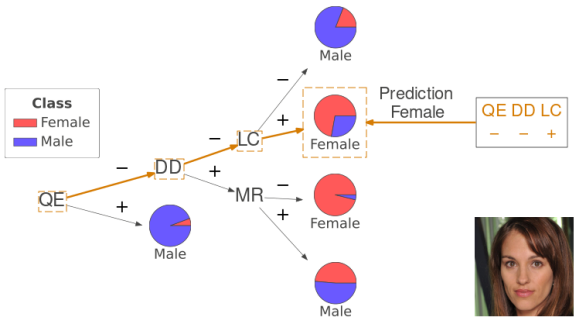


Fig. 3: An example set of symbolic rules extracted from a CNN in the form of a decision tree (max. depth of 3) relating literals QE, DD, LC and MR from Fig. 1 with a *male* or *female* classification. The pie charts shown near the tree leaves indicate the gender composition of the training data given the rule (*male* in violet and *female* in red), defining the majority class. The gender of the inset image is predicted correctly as female with the associated semantic explanation: round jawbone (¬QE), shallow cheeks (¬DD) and prominent brow ridge (LC).

without specialist knowledge, inspecting the many fingerprints in the medical X-ray use cases discussed later in this paper can be challenging without prior medical knowledge. A conceptual representation space was therefore needed to cluster these kernel fingerprints.

First, the fingerprints were resized to a common resolution of 12x12.[3] This resolution was determined empirically to be appropriate for capturing the pixel intensity and spatial information in the fingerprints for the data sets used: a lower resolution was insufficient to capture the necessary information to generate the interpretable clusters; increasing the resolution made no noticeable changes to clustering results. These re-scaled fingerprints were then converted to a three-dimensional space via Principal Component Analysis (PCA) and clustered using K-Means[4]. Fig. 5 shows a clustering of kernel fingerprints.

F. Significance of Concept Clusters

The third innovation of this paper is the analysis of concept clusters in medical diagnosis to (1) enable the identification of wrong reasons for high-classification accuracy and (2) promote a better understanding of the reasons for a CNN's classification of COVID-19 images based on a CNN's classification of images of pleural effusion. Three sets of investigations were carried out to determine the significance of the concept clusters. The contributions of selected kernels to the CNN's output were evaluated by *muting*, replacing those kernel's outputs in the CNN with zero values.

1) First, an increasing group of kernels was selected randomly at 10% intervals to be muted. This aimed to

[3]Future work using more complex CNN architectures with fingerprints will investigate various resolutions being adapted to this common resolution.

[4]Other more complex approaches were tested but did not produce the desired clustering of interpretable anatomical concepts.

determine a minimum set of kernels required to achieve the accuracy of the trained CNN model.

2) Then, kernels belonging to a specific anatomical cluster or a specific combination of clusters were muted. This evaluated the dependence of other kernels on the selected cluster(s) to maintaining the trained CNN's accuracy.
3) Finally, all kernels were muted except for a cluster or combination of clusters being evaluated. This aimed to establish if the cluster(s) being evaluated were sufficient to achieve the accuracy of the original CNN.

The experiments that follow, evaluating CNNs trained for pleural effusion and COVID-19 classifications, will show that the illustrated knowledge extraction of meaningful concepts and rules is capable of identifying concepts responsible for high-accuracy results that are not medically justified. This should prompt a revised learning of such CNN models. Additionally, this method is also shown to be capable of identifying meaningful clusters from the analysis of pleural effusion to aid in the analysis of COVID-19, as discussed next.

V. Experimental Results

The previous illustrative example, using CelebA-HQ, has shown that the proposed approach can assign relevant meaning to logical model explanations derived from a CNN. However, applying the same approach to images in a specialised domain, such as radiology, will pose a challenge as the interpretation of activations alone becomes difficult without domain-specific (medical) knowledge. In this section, we report the results on two chest X-ray data sets to demonstrate how kernel fingerprint clustering can improve the explanatory power of the rules for these complex applications (while clustering was previously not necessary for the case of CelebA-HQ).

A. Model Explanation for Pleural Effusion

Representative kernel fingerprints from multiple VGG16 models are shown in Fig. 4 (see Section IV-A for details of the pleural effusion dataset and task). While the fingerprint generated by a particular kernel might differ for each VGG16 model, the types of anatomical regions identified by each model were roughly the same. A large number of fingerprints, with minor variations, were generated and clustered with the objective of defining an appropriate conceptual space.

These fingerprints were mapped onto a three-dimensional space using PCA and K-means clustering (Fig. 5(a)). The generated clusters were identified to associate with the anatomical regions of Fig. 4 plus a cluster of unspecified fingerprints (labelled as 'redundant' in Fig. 5), where the fingerprints did not have a clear associated region. This cluster might contain fingerprints highlighting irrelevant regions of X-rays (e.g. image borders) or non-interpretable, incoherent group of regions. This cluster accounted for 66% of all fingerprints. We shall analyse the contribution of this cluster to the model's accuracy by performing a kind of ablation study, the muting of the kernels associated with the fingerprints in the cluster.

The proportions of anatomically-relevant fingerprints, excluding those in the redundant cluster, compared between

Fig. 4: Kernel fingerprints (showing the intensity of red above 0.5 on a normalised scale) providing an initial indication of regions used in the explaination of pleural effusion classification. These fingerprints corresponded to the following anatomical regions (from left): Mediastinum (M), Cardiophrenic (A), Central & Peripheral (P), and Diaphragm & Subphrenic (D).

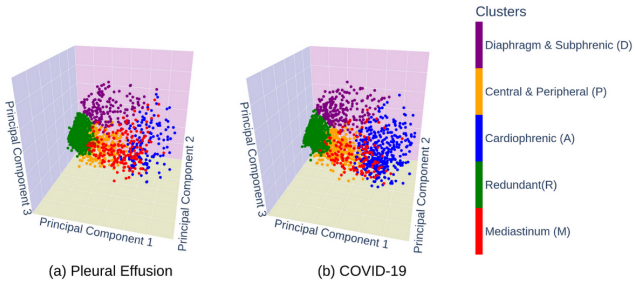


Fig. 5: (a) K-means clustering of kernel fingerprints generated from VGG16 models trained on pleural effusion data. Cluster labels were assigned by the proximity of fingerprints in the clusters to the anatomical regions shown in Fig. 4. (b) Cluster assignment of kernel fingerprints generated from VGG16 models trained on COVID-19 data. Cluster labels were inferred to be the same anatomical regions used for pleural effusion.

pleural effusion and COVID-19 models are shown in Fig. 6. Repeated runs revealed a consistent distribution of fingerprints associated with specific anatomical regions. A significant number of fingerprints were assigned to the central & peripheral (P) regions in the case of pleural effusion (in blue). These regions extended in the periphery from the shoulders (top of image) to the costophrenic angles (bottom left and right of image), as well as the inner lung space on both sides of the spine. The mediastinum (M) (including the heart) formed the second most frequent cluster. These regions are recognised as critical anatomical landmarks for chest X-ray interpretation [29]. Additionally, it is noted that pleural effusion is frequently associated with the blunting of the costophrenic angles [30], [31] and heart failure [30], [32], [33].

The average accuracy and fidelity of the extracted rules for pleural effusion were 93.6% and 97.9%, respectively, compared to the average accuracy of 95.4% from the original CNN classification. Fig. 7 shows one extracted set of symbolic rules. Interpreted alongside the kernel norm plots[5] and kernels (e.g. DH, IB and N) for the anatomical regions (e.g. M, A, P and D) via the trained K-Mean clusters, the input image shown in Fig.7 is classified as *healthy* because it shows a clear peripheral region with no evidence of costophrenic

[5]Please refer to the kernel norm plots in Appendix B-C for a contrastive explanation for kernels DH, IB and N.

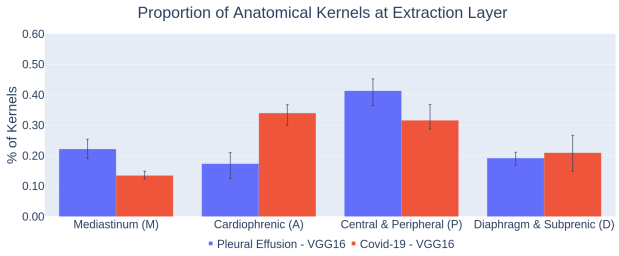


Fig. 6: Proportions of anatomically relevant kernel fingerprints identified in five trained VGG16 models for pleural effusion (in blue) and COVID-19 classification (in orange). Comparing to fingerprint proportions for pleural effusion, the proportion in the mediastinum region (M) decreases significantly, while that of the cardiophrenic region (A) increases for COVID-19.

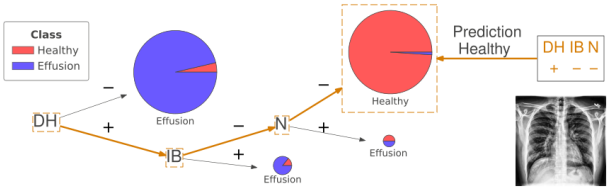


Fig. 7: A decision path for classifying the inset chest X-ray as normal using the symbolic rule DH $\wedge \neg$IB $\wedge \neg$N $\rightarrow$ *healthy*, through the decision tree extracted from a VGG16 pleural effusion classification model. Literal DH is associated with the central & peripheral anatomical region. Literals IB and N are associated with the mediastinum. A contrastive analysis of the kernel norm plots allows associating DH with the concept of costophrenic angle blunting, IB with a clear pear-shaped heart, and N with fluid accumulation around the diaphragm.

angle blunting (DH), a clear pear-shaped heart (IB), and no opacity/haziness (i.e. fluid accumulation) around the heart and diaphragm (N).

B. Model Explanation Extended to COVID-19

Given the relevance of anatomical knowledge in chest X-ray interpretation, as illustrated earlier, it is reasonable to assume that COVID-19 X-rays classification will require similar knowledge about anatomical regions and the use of these features. Fig. 5(b) shows the clustering of the fingerprints obtained from five CNN models trained on the COVID-19 data set (see Section IV-A for details about this dataset and task) inferred from the same anatomical regions for pleural effusion. A significant increase in fingerprint counts belonging to cluster A (cardiophrenic region) can be seen compared with Fig. 5(a). This is also shown in Fig. 6 where the proportion of fingerprints (in orange) for cluster A is similar to those in cluster P for COVID-19, with cluster M becoming less prominent.

Five runs on COVID-19 data generated rule sets with an average accuracy of 98.6% and fidelity of 99.7%, compared to the 98.6% accuracy of the original CNN model. Fig. 8 illustrates one of the model's explanations for a correctly classified image for COVID-19. The classification made use of kernels NY and ET in the central & peripheral regions (P).

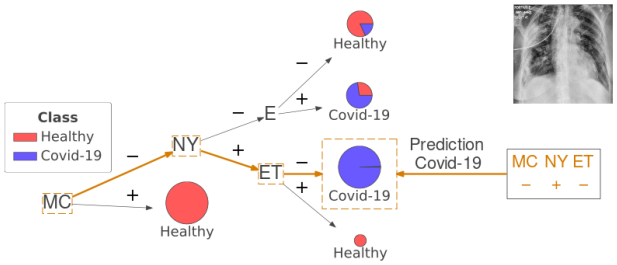


Fig. 8: A decision path for classifying the inset chest x-ray as COVID-19 using symbolic rule $\neg$MC $\wedge$ NY $\wedge \neg$ET $\rightarrow$ *COVID-19*, through the decision tree extracted from a VGG16 COVID-19 classification model. Literals NY and ET are associated with the central & peripheral region, but literal MC is associated with the image greyness according to the kernel norm plot shown in Fig. B.9. The identification of MC as a spurious concept should prompt an intervention to eliminate the reliance on MC in the classification task.

Based on the analysis of the kernel norm plots in Appendix B-D, NY indicates haziness around the lower spine in both lung lobes, while $\neg$ET indicates haziness primarily at the outer periphery of the lung air space. Additionally, kernel MC was extracted as a literal from the non-interpretable 'redundant' cluster, which was found to detect image greyness. Image greyness is not a medically justifiable concept but was incorrectly used by the trained CNN, perhaps as a shortcut to achieving high accuracy from the given data. Spurious features are common in trained CNNs [34]. Whether to prompt model re-training or to intervene directly on the rule sets, a meaningful knowledge extraction allows one to identify and quantify the effect of such errors in the system.

C. Cluster Significance on Model Accuracy

In order to measure the importance of a cluster of kernels on a model's outcome, we have systematically evaluated the results of replacing the outputs of kernels with zero values. We started by muting random kernel selections in 20 repeated runs on trained models for both pleural effusion and COVID-19. Figs. 9 and B.16 show the average results of muting an increasing fraction of kernels in the case of pleural effusion and COVID-19, respectively. The results indicate that model accuracy can be maintained with up to 60-70% of kernels muted.

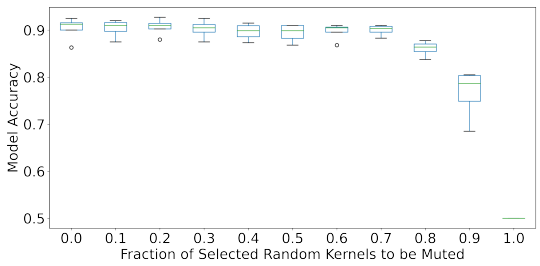
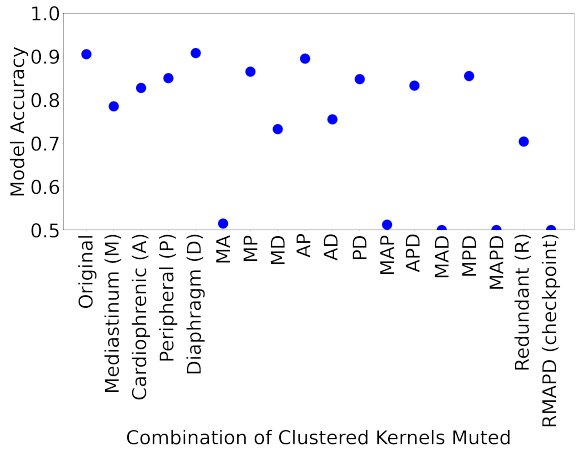
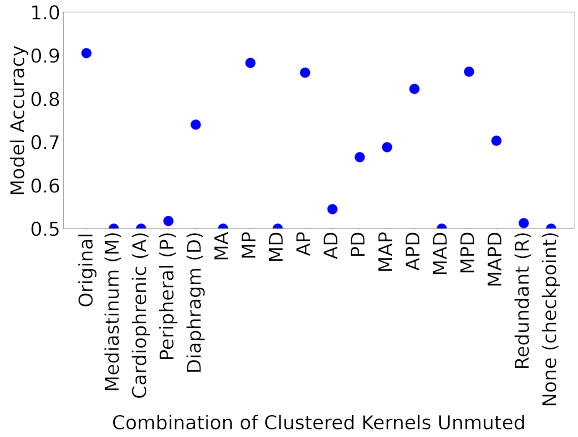


Fig. 9: Evaluation of model accuracy degradation by muting a random selection of kernels at 10% intervals when applied to trained CNN models on pleural effusion classification.

Recall that approximately 60% of kernels were found to be semantically non-interpretable (redundant). Significant change in model accuracy resulted from muting/unmuting kernels of selected anatomical cluster or combinations thereof would indicate these kernels as essential features to the model prediction. Fig. 10(a) shows the drop in accuracy when muting all the kernels in a cluster (e.g. M, A, etc.) and in a combination of clusters (e.g. M and A together (MA)) for pleural effusion. Fig.10(b) shows the model accuracy when muting every kernel other than the kernels in a specified cluster. A similar evaluation for COVID-19 is shown in Fig.B.17.



(a)



(b)

Fig. 10: Evaluation of average model accuracy by (a) muting and (b) unmuting kernels in anatomically relevant clusters (e.g. M, A) and its combintations (e.g. MA, MP) over five trained CNN models for pleural effusion.

Fig.10(a) shows that only the muting of combinations of clusters (MA, MAP, MAD and MAPD) can reduce model accuracy to the level of random guessing. These clusters represent regions that encompass the entire lung periphery. It is conjectured that the combination of clusters M and A contain necessary information about heart failure and fluid accumulation (which other kernels cannot compensate), but insufficient information to provide on their own an accurate prediction of pleural effusion. Conversely, the combination on its own, e.g. of clusters M and P (MP), produced very high accuracy, indicating that the original CNN model can be approximated well by a drastically pruned version of the model. As expected, the contribution from anatomically non-interpretable kernels of the redundant cluster alone was found to be insignificant, despite their removal still resulting in a noticeable reduction in accuracy. In the case of the models trained on the COVID-19 data (Fig.B.17(a)), MA and MAP were the cluster combinations producing the largest loss of accuracy when muted, while APD was found to be a cluster combination capable of maintaining high model accuracy on its own (Fig.B.17(b)). Differently from the pleural effusion case, muting the redundant cluster (R) for COVID-19 had a negligible effect on model accuracy. Muting of cluster A alone caused a considerable drop in accuracy in comparison with cluster M. It is conjectured that the COVID-19 X-rays consisted of more cases with significant fluid accumulation at the lower lobes resulting in substantial opacity at the cardiophrenic region.

VI. Discussion & Conclusion

We proposed a novel method for analyzing symbolic rules extracted from CNNs trained on medical images. By using feature aggregation, clustering and kernel norm plots, we demonstrated that meaningful concepts can be assigned to symbolic rules extracted from black-box CNNs with high fidelity. This method was capable of identifying (medically unjustified) spurious concepts as part of the learned features in CNNs when evaluating with X-rays for pleural effusion and COVID-19 classifications. It was also found that the identified anatomical concepts via clustering could provide meaningful explanations to different but related classification tasks. This method is expected to benefit the use of CNNs in radiology by analysing new diseases in comparison with better understood diseases. While this work is not intended for clinical diagnosis, the findings highlight the value of providing domain-specific, meaningful, logical sentence-like and contrastive explanations that clinicians can act on. In practice, these explanations are more measurable and useful than subjective interpretations of saliency maps or other popular representations.

In our experiments, the conceptual space provides a common ground for knowledge sharing between CNN kernels. This facilitates the comparison of different CNNs using a common set of concepts. In addition, this work will also aid in continual improvement for explaining domain-specific related tasks by comparing kernel norm values with clinical metrics. These findings suggested the following potential research directions: (i) collaboration with clinicians to analyze intra-cluster variance among kernels; (ii) enhancement of cluster labelling to improve concept assignment in the conceptual space; (iii) extending this approach to explain other network models and data sets, with experiments on transfer learning and out-of-distribution learning aided by symbolic knowledge.

References

[1] "Nanox AI," https://www.nanox.vision/ai.

[2] "Lunit INC," https://www.lunit.io/en/products/insight-cxr.

[3] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[5] T. D. Nguyen, K. E. Kasmarik, and H. A. Abbass, "Towards interpretable deep neural networks: An exact transformation to multi-class multivariate decision trees," *arXiv preprint*, 2020.

[6] I. van der Linden, H. Haned, and E. Kanoulas, "Global aggregations of local explanations for black box models," *arXiv preprint*, 2019.

[7] J. Townsend, T. Kasioumis, and H. Inakoshi, "ERIC: Extracting relations inferred from convolutions," in *Computer Vision – ACCV 2020*, ser. Lecture notes in computer science, 2021, pp. 206–222.

[8] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[9] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *arXiv preprint arXiv:1412.6856*, 2014.

[11] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[12] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Comparing the interpretability of deep networks via network dissection," in *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer, 2019, pp. 243–252.

[13] ——, "Interpreting deep visual representations via network dissection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41 (9), pp. 2131–2145, 2018.

[14] D. Bau *et al.*, "Understanding the role of individual units in a deep neural network," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 117 (48), pp. 30 071–30 078, 2020.

[15] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35 (8), pp. 1798–1828, 2013.

[16] N. Lavrač, V. Podpečan, and M. Robnik-Šikonja, "Introduction to representation learning," in *Representation Learning: Propositionalization and Embeddings*, 2021, pp. 1–16.

[17] P. Gärdenfors, *Conceptual spaces: The geometry of thought*. The MIT Press, 2000.

[18] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge-Based Systems*, vol. 8 (6), pp. 373–389, 1995.

[19] N. Frosst and G. Hinton, "Distilling a neural network into a soft decision tree," *arXiv preprint*, 2017.

[20] S. Odense and A. d. Garcez, "Layerwise knowledge extraction from deep convolutional networks," *arXiv preprint*, 2020.

[21] Q. Zhang, R. Cao, F. Shi, Y. N. Wu, and S. Zhu, "Interpreting CNN knowledge via an explanatory graph," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[22] Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu, "Interpreting cnns via decision trees," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[23] Q. Zhang, Y. Nian Wu, and S. Zhu, "Interpretable convolutional neural networks," in *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, 2018, pp. 8827–8836.

[24] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint*, 2017.

[25] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 590–597.

[26] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," *arXiv preprint arXiv:2003.11597*, 2020.

[27] K. Zuiderveld, "Viii.5. - contrast limited adaptive histogram equalization," in *Graphics Gems*. Academic Press, 1994, pp. 474–485.

[28] T. Kasioumis, J. Townsend, and H. Inakoshi, "Elite BackProp: Training sparse interpretable neurons," in *CEUR Workshop Proceedings*, vol. 2986, 2021, pp. 83–93.

[29] E. Puddy and C. Hill, "Interpretation of the chest radiograph," *Continuing Education in Anaesthesia, Critical Care and Pain*, vol. 7 (3), pp. 71–75, 2007.

[30] M. J. Na, "Diagnostic tools of pleural effusion," *Tuberc. Respir. Dis.*, vol. 76 (5), pp. 199–210, 2014.

[31] S. Singla, B. Pollack, S. Wallace, and K. Batmanghelich, "Explaining the black-box smoothly-a counterfactual approach," *arXiv preprint arXiv:2101.04230*, 2021.

[32] J. M. Porcel, "Pleural effusions from congestive heart failure," *Semin. Respir. Crit. Care Med.*, vol. 31 (6), pp. 689–697, 2010.

[33] V. S. Karkhanis and J. M. Joshi, "Pleural effusion: diagnosis, treatment, and management," *Open Access Emerg. Med.*, vol. 4, pp. 31–52, 2012.

[34] S. Singla, B. Nushi, S. Shah, E. Kamar, and E. Horvitz, "Understanding failures of deep networks via robust feature extraction," in *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, 2021, pp. 12 853–12 862.

[35] E. Brown and D. I. Perrett, "What gives a face its gender?" *Perception*, vol. 22 (7), pp. 829–840, 1993.

[36] H. Miller, "Ich werde ein perfekter künstler bd. 1." Mannheim, 2006.

[37] L. Capitán and D. Simon, "Facial feminization surgery and facial gender confirmation surgery," *Comprehensive Care of the Transgender Patient E-Book*, p. 54, 2019.

[38] M. Ascha *et al.*, "Nonsurgical management of facial masculinization and feminization," *Aesthet. Surg. J.*, vol. 39(5), pp. NP123–NP137, 2019.

[39] European Society of Radiology, "Good practice for radiological reporting. guidelines from the european society of radiology (ESR)," *Insights Imaging*, vol. 2 (2), pp. 93–96, 2011.

APPENDIX A
SUPPLEMENTARY MATERIALS FOR GENDER IDENTIFICATION

A. Representative Images from CelebA-HQ dataset

The CelebA-HQ dataset [24], a high-resolution facial image dataset, was used to determine the gender of celebrities through frontal facial images. Images were manually screened for mislabeling. These images were used for both model training and evaluation of the performance of the explainable approximation.

(a)

(b)

Fig. A.1: Representative examples from the CelebA-HQ dataset of (a) female and (b) male celebrities. It consists of frontal images of people of various origins, ethnic groups, and facial features.

B. Relevant work on Gender Identification from Facial Images

Human faces carry valuable information needed for social interaction. When a face is encountered, a quick determination of the gender can be established based on visual features. Numerous studies across various disciplines have identified gender-relevant features [35]–[37]. These features have now been used as reference landmarks from figurine drawing [36] to the definition of facial feature changes in gender confirmation clinical management [37].

Fig. A.2 presents the research work [35] on gender perception in which a selected panel of individuals were asked to rate their perception of gender after one (or more) facial parts from a prototype face of either gender were substituted. The study concluded that grafting the jaw, brows with eyes, chin, and brows (in descending order) onto a prototype face of the opposite gender would result in a significant shift in the perceived gender.

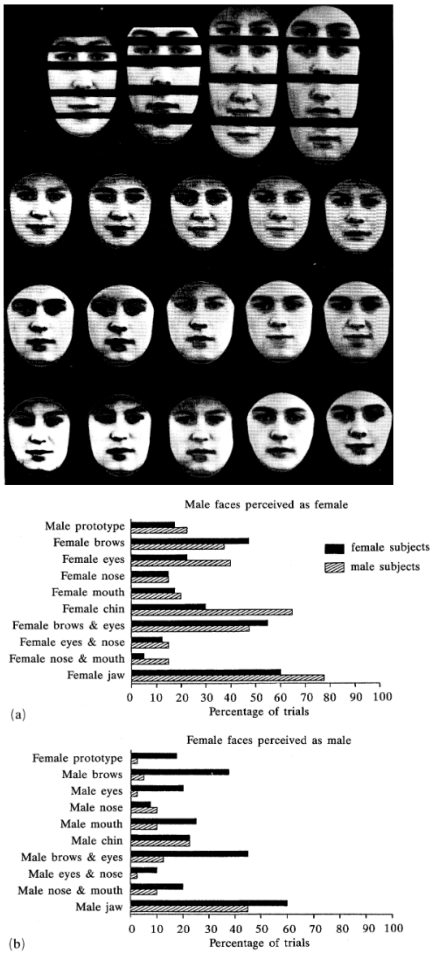


Fig. A.2: Research on gender perception using a mix-n-match method on different facial parts. It was discovered grafting the jaw, brows with eyes, chin and brows (in descending order) onto a prototype face of the opposite gender would significantly alter the perceived gender. (Source: [35])

Fig. A.3 presents facial traits in figurine drawing for a masculine face with numerous angular shapes caused by prominent bones and muscles. It is also characterised by straighter and thicker brows, as well as square jaw and chin. Feminine faces are typically more oval-shaped with curved features. The chin is more pointed, and the jaw less angular. The brows are also more arched. Fig. A.4 further illustrates the most critical facial features (highlighted) in the perception of gender that contribute to the success of facial gender confirmation surgery. [38] also summarised the characteristics of ideal masculine and feminine faces in their study, which corroborate the other referenced findings indicating the significance of the brow, jaw and chin in gender identification.

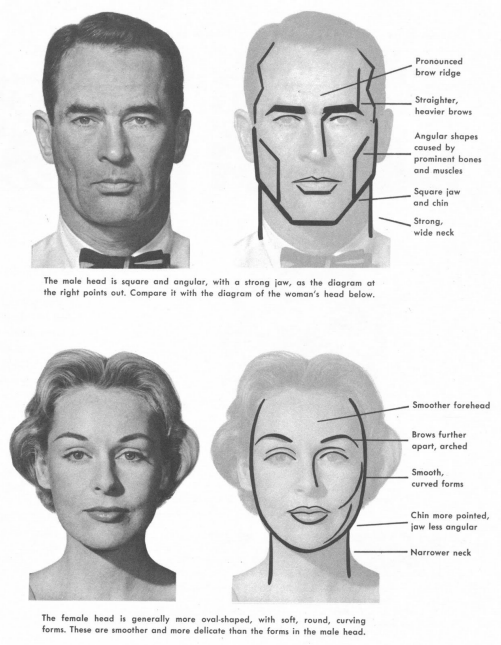


Fig. A.3: Facial traits associated with gender in figurine drawing . (Source: [36]).

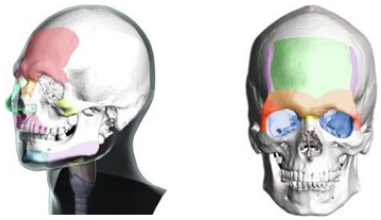


Fig. A.4: Significant facial features (highlighted) that would alter the perception of gender following a facial gender confirmation surgery. (Source: [37])

C. Additional Experimental Results on Gender Identification

To benchmark with the CNN model illustrated in the main text, Fig. A.5 shows a decision tree generated entirely using metadata from the CelebA-HQ dataset [24]. It classified gender based on the choice of makeup and accessories as features and achieved a prediction accuracy of 94.1% compared to 93.0% by the CNN model. The symbolic rules extracted from CNN in this work achieved an accuracy of 87.3% and a fidelity of 91.1%. Despite the slightly lower accuracy, it is argued that the symbolic rules used more anatomically relevant features for the classifications. The following kernel norm plots and additional examples of explainable rules will supplement our novel knowledge extraction approach using symbolic rules presented in the main text.

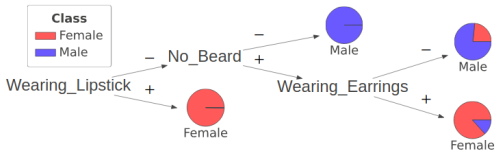


Fig. A.5: An example of decision tree generated using metadata from the CelebA-HQ dataset. It used the choice of makeup and accessories as part of the feature set for gender identification and achieved a prediction accuracy of 94.1%.

Additional kernel norm plots associated with the set of explainable rules generated in the illustrative examples of gender identification are shown in Figs. A.6, A.7 and A.8. These plots provided a quantified representation of the image variations identified in the frequently activated regions of the corresponding kernel fingerprints. Kernel 'DD' recognised the appearance of sunken eyes (e.g. eye bags) and/or the prominence of the cheeks. Higher values indicated fewer traits of sunken eyes but fuller cheeks. Kernel 'LC' determined the archness of the brow and the prominence of the brow ridge, with higher values showing a more arched brow and protruding brow ridge. Finally, kernel 'MR' defined the squareness of the face with higher values indicating a squarer face and wider chin. These findings in appearance for gender perception matched those from domain research in Appendix A-B.

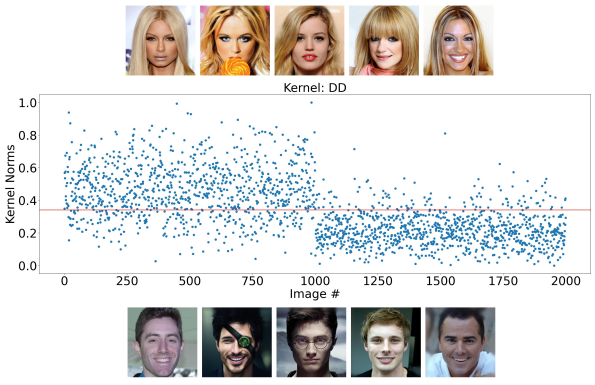


Fig. A.6: The corresponding kernel norm plot (L1-norm values) for kernel 'DD'. This plot presented the appearance variations of sunken eyes and/or shallow cheeks. Higher values indicated fewer traits of sunken eyes but fuller cheeks, whereas lower values revealed signs of sunken eyes and/or shallower cheeks. Images for the top 5 and bottom 5 kernel norm values are shown above and below the kernel norm plot respectively.

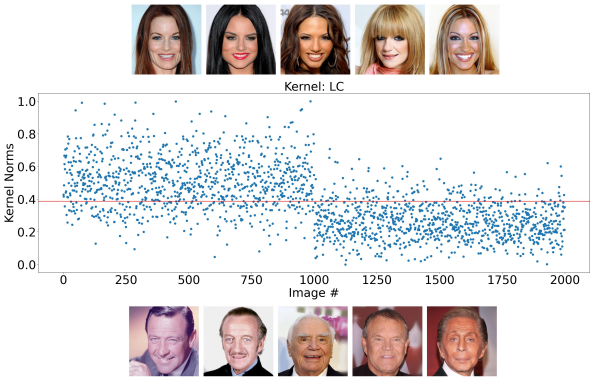


Fig. A.7: The corresponding kernel norm plot (L1-norm values) for kernel 'LC'. This plot demonstrated the appearance variations around the brow ridge. Higher values indicated an arched brow or protruding brow ridge, whereas lower values indicated with straighter brow or less pronounced brow ridge. Images for the top 5 and bottom 5 kernel norm values are shown above and below the kernel norm plot respectively.

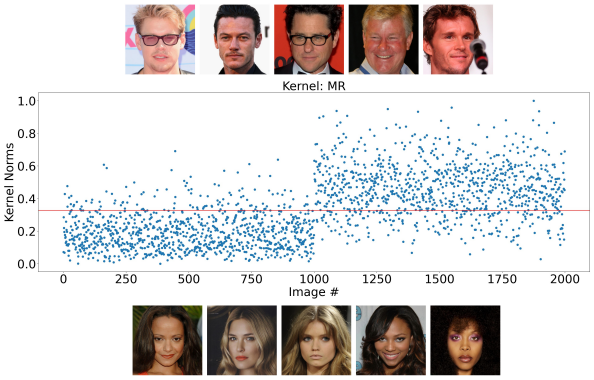
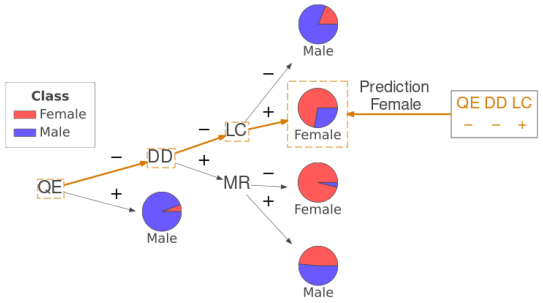


Fig. A.8: The corresponding kernel norm plot (L1-norm values) for kernel 'MR'. This plot demonstrated the appearance variations on the squareness of the lower face. Higher values correlated with wider chins, while lower values correlated with narrower V-shaped chins. Images for the top 5 and bottom 5 kernel norm values are shown above and below the kernel norm plot respectively.

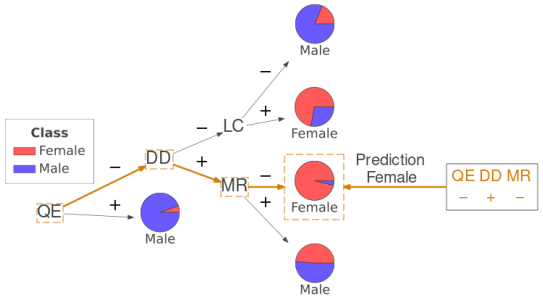
As shown in Figs. A.9, A.10 and A.11, the identified facial features can be used to provide the appropriate descriptive explanations for the prediction of the illustrated images defined by each of the symbolic rules. For example, Fig. A.9 categorised faces with rounder jawbone, sunken eyes and/or shallow cheeks, and arched eyebrow/prominent brow ridge. The rule in Fig. A.10 provides an alternative definition of the female gender where faces have fuller cheeks and fewer traits of sunken eyes. This is supplemented with the appearance of V-shaped lower faces and narrower chins. In the case of male appearance, Fig. A.11 illustrated an alternative decision route to using only kernel 'QE'. This rule generates a decision that differed from the rule in Fig. A.9 due to the straighter brows and less pronounced brow ridge in the representative images.



(a)

(b)

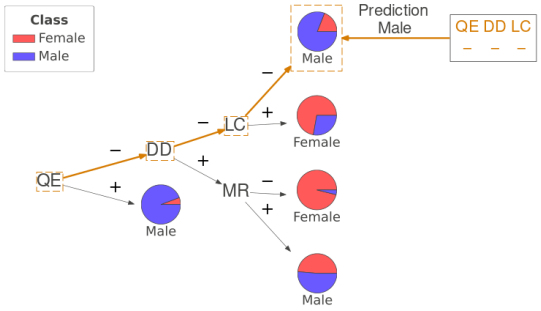
Fig. A.9: The decision path for one of the symbolic rules extracted from a tree-based method within ERIC [7]. Representative images in (b) indicate that the prediction is based on facial features - (1) rounded jawbone (¬QE), (2) sunken eyes and/or shallow cheeks (¬DD), and (3) arched eyebrow/prominent brow ridge (LC).



(a)

(b)

Fig. A.10: The decision path for one of the symbolic rules extracted from a tree-based method within ERIC [7]. Representative images in (b) indicate that the prediction is based on facial features - (1) rounded jawbone (¬QE), (2) no sunken eyes and/or fuller cheeks (DD), and (3) sharper chin (¬MR).



(a)

(b)

Fig. A.11: The decision path for one of the symbolic rules extracted from a tree-based method within ERIC [7]. Representative images in (b) indicate that the prediction is based on facial features - (1) rounded jawbone (¬QE), (2) sunken eyes and/or shallow cheeks (¬DD), and (3) straight eyebrow/shallow brow ridge (¬LC).

APPENDIX B
SUPPLEMENTARY MATERIALS FOR PLEURAL EFFUSION AND COVID-19 DETECTION

A. Representative Images from CheXpert and IEEE8023 datasets

Fig. B.1 presents some of the representative images of frontal chest X-rays used in this work to detect pleural effusion and COVID-19 respectively. Images were pre-processed as described in Section IV.

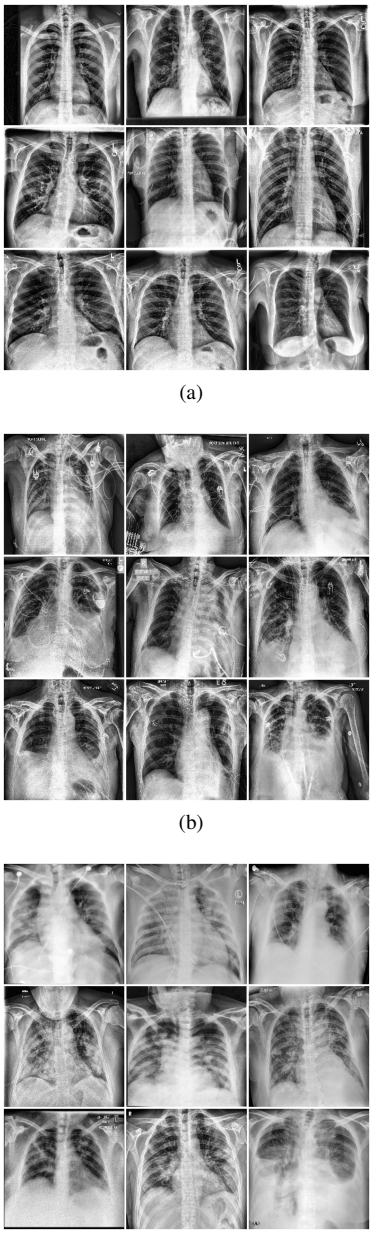


Fig. B.1: Representative frontal chest X-ray images of patients defined as (a) healthy, (b) having pleural effusion from the CheXpert [25] dataset, and (c) having COVID-19 from the IEEE8023 [26] dataset.

B. Relevant work on Pleural Effusion Detection from Chest X-Ray

Pleural effusion is defined as the abnormal accumulation of fluid in the pleural space, which is typically caused by an imbalance in fluid formation and absorption. Pneumonia, congestive heart failure and malignancy account for being the cause in the majority of cases [33]. Chest X-ray is generally taken during an initial diagnosis as standard practice. Typically, pleural effusion can be characterised by homogeneous opacity, obliteration of the costophrenic angle, and a curved lower boundary commonly referred to as the Ellis S-shaped curve [33].

As a matter of good practice, anatomical locations of abnormalities should always be specified in a radiological report [39]. Fig. B.2 shows a normal chest X-ray with annotations of anatomical regions identified through concept clustering in this work. A normal chest should appear like this figure with clear lung air space and a normal-sized heart (e.g. as defined by a cardiothoracic ratio), and no fluid accumulation at both costophrenic angles. Labels of anatomical regions will provide valuable references for describing abnormalities and correlating relevant medical concepts with the findings from activated kernels in a CNN model reported in this work.

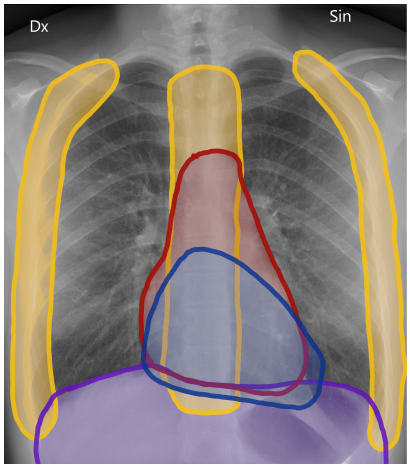


Fig. B.2: An annotated approximation of the anatomical regions identified from clustering. It shows four key regions: (1) Central & Peripheral (yellow), (2) Mediastinum (red), (3) Cardiophrenic (blue), and Diaphragm & Subprenic (purple).

C. Additional Experimental Results on Pleural Effusion Detection

As with the illustrative case for gender identification in Appendix A, the semantic meaningful concepts for the relevant kernels used in the generated rule set for model explanation were evaluated through the kernel norm plots and concept assignment via clustering. Fig. B.3 presents the plot of kernel norm values for kernel 'DH'. It was associated with the central & peripheral region (P) of the chest. The plot indicated variations in this region where high values appeared in images with a clear periphery around both lung air spaces or haziness around the same region otherwise. Fig. B.4 presents the plot of

kernel norm values for kernel 'IB'. Kernel 'IB' was related to the mediastinum (M) region, including the heart. High kernel norm values were found in images with significant opacity near the left side of the heart and/or visible heart enlargement. Fig. B.5 presents the plot of kernel norm values for kernel 'N'. Kernel 'N' was likewise related to the mediastinum (M) region. The representative examples further demonstrated the effect of changing the kernel norm value on the appearance of images. At a high value, significant fluid accumulation(i.e. shown in whiteness) could be observed in the lower-left lobe (near the left side of the heart and the left hemidiaphragm).

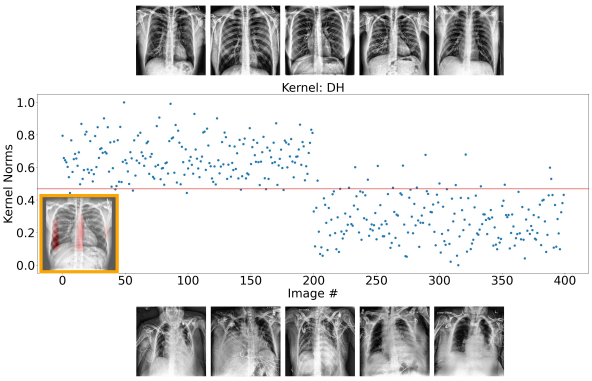


Fig. B.3: The corresponding kernel norm plot (L1-norm values) for kernel 'DH'. This plot demonstrated the variation in clarity around the central & peripheral of the lung air space. Higher values show images with clear peripheries around the lung air spaces, while lower values show haziness in the same regions. Images for the top 5 and bottom 5 kernel norm values are shown above and below the kernel norm plot respectively.

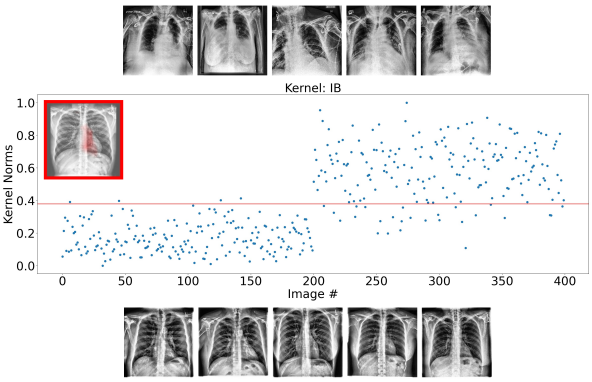


Fig. B.4: The corresponding kernel norm plot (L1-norm values) for kernel 'IB'. This plot presented the appearance variations at the mediastinum (including the heart). Higher values indicate significant opacity near to the left side of the heart with possible left ventricle enlargement. Images for the top 5 and bottom 5 kernel norm values are shown above and below the kernel norm plot respectively.

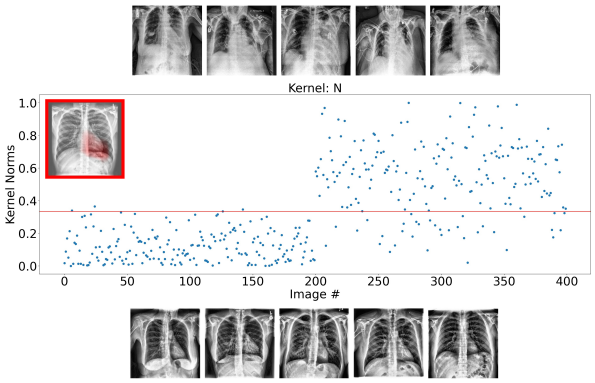
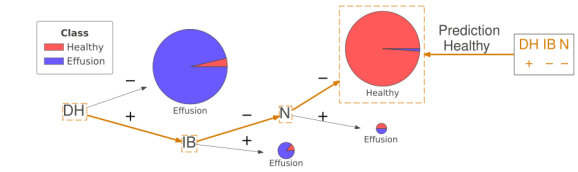


Fig. B.5: The corresponding kernel norm plot (L1-norm values) for kernel 'N'. This plot presents the appearance variations in the lower-left lobe, where the left ventricle located. High values show images with significant consolidation (i.e. shown in whiteness) near the left ventricle and the left hemidiagphragm. Images for the top 5 and bottom 5 kernel norm values are shown above and below the kernel norm plot respectively.

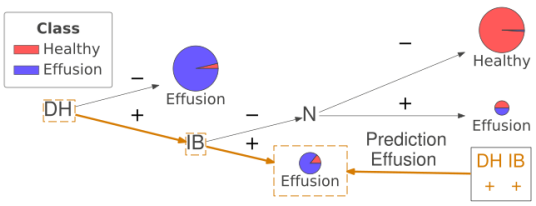
These identified conceptual descriptions were applied to the symbolic rules to make the associated explanations more human-comprehensible. For example, the rule generated in Fig. B.6a can be used to represent the healthy cases. All of the images show clear peripheries of the lung air spaces, normal-shaped heart and clear lower left lobes. On the other hand, the rule at B.7a defines images of pleural effusion as a result of haziness and/or enlargement at the left ventricle (i.e. positive kernel 'IB'). Lastly, the rule in Fig. B.8 shows that haziness around the periphery of the lung air spaces is sufficient for the model to detect pleural effusion in the example images.



(a)

(b)

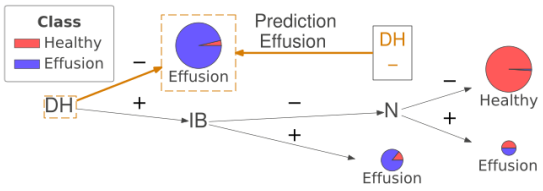
Fig. B.6: The decision path for one of the symbolic rules extracted from a tree-based method within ERIC [7]. Representative images in (b) indicate that the prediction is based on anatomical features - (1) clear lung periphery (DH), (2) normal-sized heart (¬IB), and (3) clear lower left lobe around the left side of the heart (¬N).



(a)

(b)

Fig. B.7: The decision path for one of the symbolic rules extracted from a tree-based method within ERIC [7]. Representative images in (b) indicate that the prediction is based on anatomical features - (1) clear lung periphery (DH), and (2) haziness near the heart with possible left ventricle enlargement (IB)



(a)

(b)

Fig. B.8: The decision path for one of the symbolic rules extracted from a tree-based method within ERIC [7]. Representative images in (b) indicate that the prediction is based on one anatomical feature - (1) haziness at the lung periphery (¬DH).

D. Additional Experimental Results on COVID-19 Detection

This section includes further examples of using the conceptual clusters to adopt the conceptual descriptions on the symbolic rules for COVID-19 detection. While these examples are not intended for clinical diagnosis, the descriptive explanation will aid current clinical research addressing the present pandemic.

The first literal in the rule is based on kernel 'MC'. It is associated with the semantically non-meaningful 'redundant (R)' cluster. Fig. B.9 presents the plot of kernel norm values for this kernel. As observed, this kernel made a critical separation between healthy images (first 200) and those with COVID-19 (remaining 200). Based on the representative images at both extremes, it was observed that high kernel norm values appeared in sharper images with higher contrast with the blacker background. On the other hand, low values appeared in images that are more grey. Despite the manual pre-training screening of images, this variation in image appearance remained. This revealed that the original CNN model had retained this appearance feature as a sign of 'healthy' during training. This needs to be rectified for better model performance.

Kernel 'NY' and 'ET' were both related to the central & peripheral regions. Kernel 'NY' identified haziness around the lower spine (see Fig. B.10), whereas kernel 'ET' identified clarity across the lung peripheral with high kernel norm values (see Fig. B.12). Kernel 'E' was associated with the opacity at the cardiophrenic region.

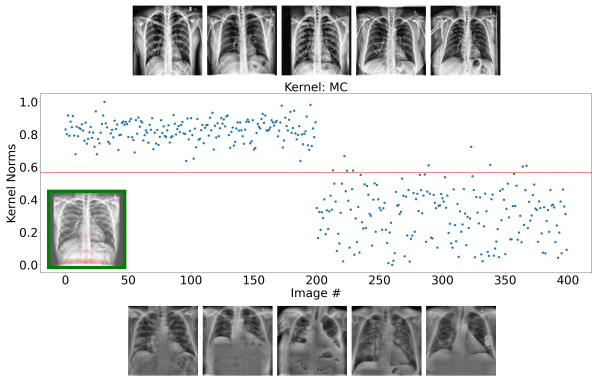


Fig. B.9: The corresponding kernel norm plot (L1-norm values) for kernel 'MC'. This plot presents the variations in image appearance. High values show up as sharper chest X-ray images with good contrast against the black background, while images at lower values typically appear more grey. Images for the top 5 and bottom 5 kernel norm values are shown above and below the kernel norm plot respectively.

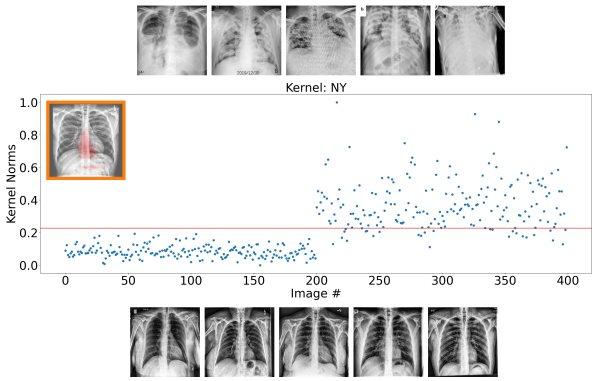


Fig. B.10: The corresponding kernel norm plot (L1-norm values) for kernel 'NY'. This plot differentiates images based on haziness around the lower spine. Higher values appear on images with haziness at the lower spine, while lower values indicate with clear lower spine. Images for the top 5 and bottom 5 kernel norm values are shown above and below the kernel norm plot respectively.

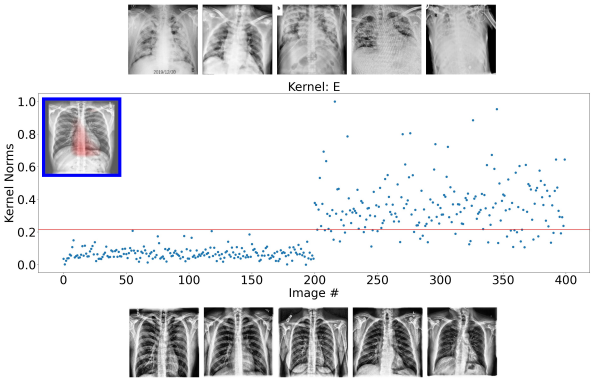


Fig. B.11: The corresponding kernel norm plot (L1-norm values) for kernel 'E'. This plot identifies the opacity at the cardiophrenic region (A). Higher values indicate with opacity around the cardiophrenic region and clear region at lower values. Images for the top 5 and bottom 5 kernel norm values are shown above and below the kernel norm plot respectively.

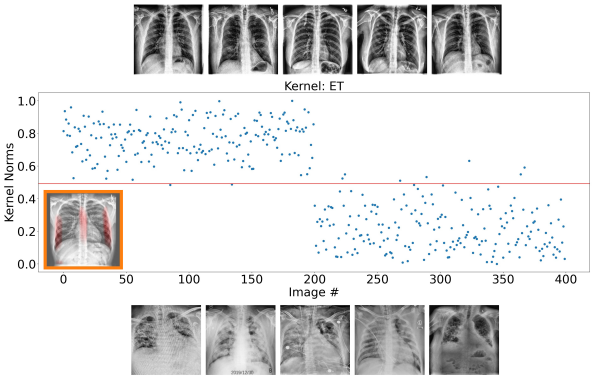
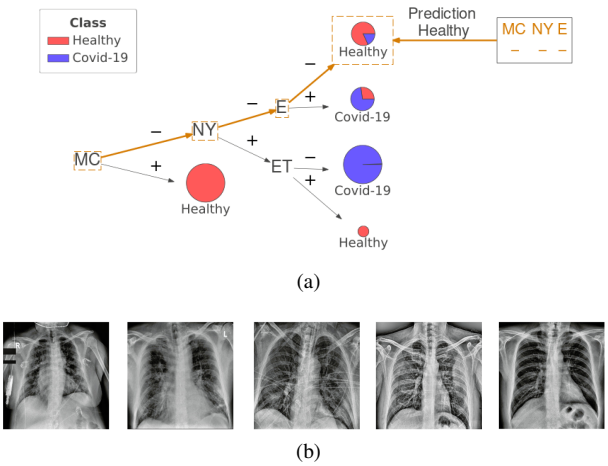


Fig. B.12: The corresponding kernel norm plot (L1-norm values) for kernel 'ET'. This plot differentiates images with clear peripheral regions. High values appear in images with clear peripheries of the lungs, while haziness observed in lower values. Images for the top 5 and bottom 5 kernel norm values are shown above and below the kernel norm plot respectively.

Besides determining healthy images solely on image greyness (Kernel 'MC'), the rule from Fig. B.13 identified a subset of false negatives predicted by the CNN model against the labelled ground truth as having COVID-19. The representative examples indicate that the lower spine appears relatively clear with mild haziness at the cardiophrenic region. The healthy traits in the anatomical regions were identified to cause the false prediction. This conceptual explanation will enable clinicians to re-evaluate these images for more accurate diagnosis.



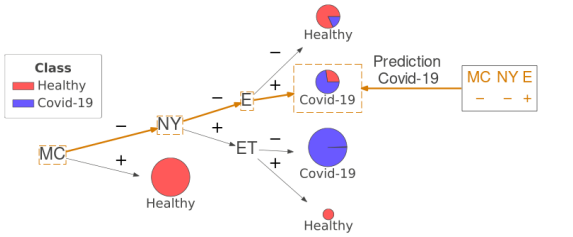
(a)

(b)

Fig. B.13: The decision path for one of the symbolic rules extracted from a tree-based method within ERIC [7]. Representative images in (b) indicate that the prediction is based on anatomical features - (1) image greyness (¬MC), (2) clear lower spine (¬NY), and (3) clear cardiophrenic region (¬E).

Figs. B.14 and B.15 shows rules for determining images with COVID-19 with varying degrees of 'severity' (i.e. based on haziness at anatomical regions). Fig. B.14 presents cases where haziness was only observed at the cardiophrenic region. In contrast, haziness was observed around the lung periphery for cases governed by the rule in Fig. B.15. These examples demonstrated clear evidence of enhancement in model explanation power using symbolic rules and concept assignment via clustering, as described in the main text.

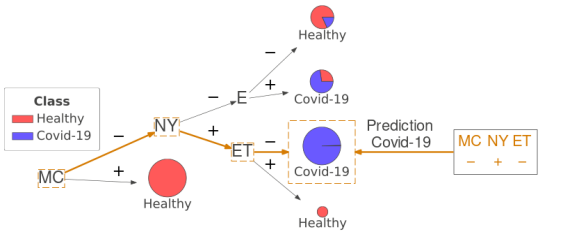
Finally, a comparable study on cluster significance for COVID-19 is presented in Figs. B.16 and B.17. Fig. B.16 shows a similar finding that model accuracy could be maintained when up to approximately 60-70% of the kernels were muted. Cluster combinations, MA and MAP, were found in Fig. B.17a to produce the largest loss in accuracy when muted, further supporting the evidence that these clusters contain relevant information for the classification from these anatomical regions that other cluster kernels could not compensate. Notably, APD was found in Fig. B.17b to be a cluster combination capable of maintaining high model accuracy on its own, possibly because COVID-19 X-rays consisted of more cases with significant fluid accumulation at the lower lobes.



(a)

(b)

Fig. B.14: The decision path for one of the symbolic rules extracted from a tree-based method within ERIC [7]. Representative images in (b) indicate that the prediction is based on anatomical features - (1) image greyness (¬MC), (2) clear lower spine (¬NY), but (3) haziness is observed at the cardiophrenic region (E).



(a)

(b)

Fig. B.15: The decision path for one of the symbolic rules extracted from a tree-based method within ERIC [7]. Representative images in (b) indicate that the prediction is based on anatomical features - (1) image greyness (¬MC), (2) haziness at the lower spine (NY) as well as (3) other peripheral regions of the lungs (¬ET).

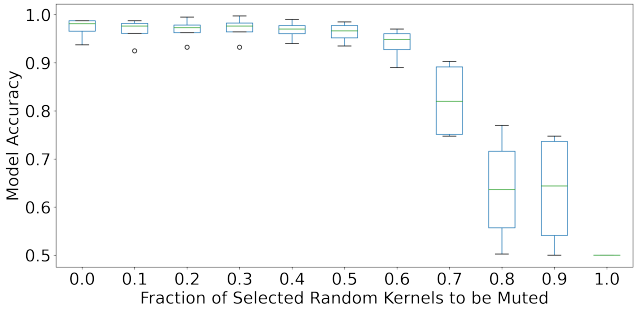
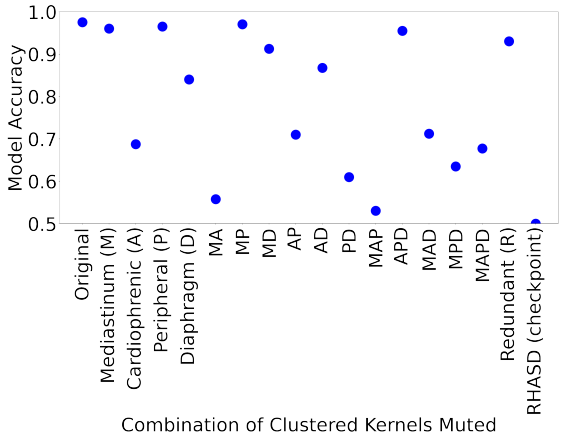
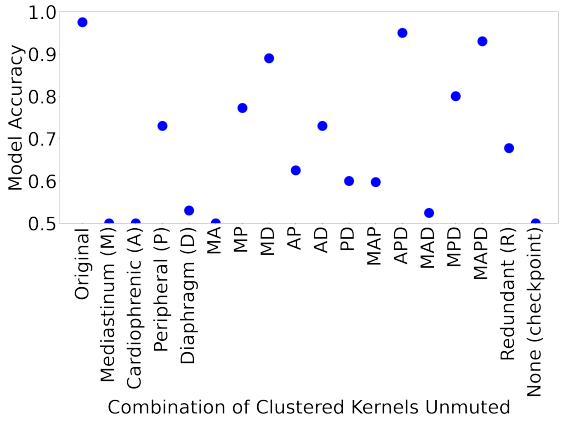


Fig. B.16: Evaluation of model accuracy degradation through muting a random selection of kernels at 10% intervals when applied to trained CNN models for COVID-19.



(a)



(b)

Fig. B.17: Evaluation of average model accuracy through (a) muting and (b) preserving activations of a subset of anatomical relevant clusters or their combination in five repeated CNN models for COVID-19.