Closing the Neural-Symbolic Cycle: Knowledge Extraction, User Intervention and Distillation from Convolutional Neural Networks

Kwun Ho Ngan[1,*], James Phelan[1], Esma Mansouri-Benssassi[2], Joe Townsend[2] and Artur d'Avila Garcez[1,*]

[1]Data Science Institute, City, University of London, London, EC1V 0HB, UK
[2]Fujitsu Research of Europe Ltd, Slough, SL1 2BE, UK

Abstract

This paper introduces and evaluates a neural-symbolic cycle for Convolutional Neural Networks (CNNs). Knowledge in the form of logic programming rules is extracted from a trained (teacher) CNN model. Domain experts can interact with the rules to assign concepts, intervene and make changes to the model. Distillation is then used to re-train a simplified CNN, closing the neural-symbolic cycle. The approach is evaluated in the classification of medical images (chest x-rays). Experiments indicate that the approach can generate symbolic rules for pleural effusion detection with high accuracy (94.5%) and fidelity (98.2%) in comparison with the original CNN with 96.2% accuracy. Expert intervention produces symbolic rules with clinically relevant concepts while preserving predictive accuracy (94.8%). The approach also enables effective transfer of learning from clinically-relevant rules onto a much simplified (student) CNN that is almost 90% more compact while maintaining accuracy of 93.8%. The goal of this work is to offer an auditable record of network training, elaboration and deployment in the medical domain.

Keywords

Neural-Symbolic System, Knowledge Extraction, Symbolic Reasoning, Human-in-the-loop, Knowledge Distillation

1. Introduction

Deep learning models have shown remarkable success in a range of applications, including image recognition. However, these models are limited by their lack of interpretability and transparency, which hinder their deployment in critical domains (e.g. medical care) where transparency and explainability have become essential.

To address this limitation, explainable AI (XAI) has emerged as a promising research topic aiming to transform black-box models into surrogate models such that humans can understand and interact with the the AI system's predictions [1]. One such approach to XAI is to use a neural-symbolic system to extract knowledge from a neural network and represent it as a set of symbolic rules.

NeSy 2023: 17th International Workshop on Neural-Symbolic Learning and Reasoning
*Corresponding author.
kwun.ngan.3@city.ac.uk (K. H. Ngan); james.phelan1@nhs.net (J. Phelan); esma.mansouri-benssassi@fujitsu.com (E. Mansouri-Benssassi); joseph.townsend@fujitsu.com (J. Townsend); a.garcez@city.ac.uk (A. d. Garcez)

In this paper, we present our development and evaluation of a neural-symbolic cycle that extracts knowledge from Convolutional Neural Networks (CNNs), associates such knowledge with human-understandable concepts, constructs a set of symbolic rules allowing domain experts to interact and intervene in the system to fine-tune the model for specialised applications with known symbolic constraints. Relevant knowledge is then used to re-train a student network such that its response can mimic that knowledge.

The main contribution of this paper is the implementation and evaluation of the proposed neural-symbolic cycle, which includes the process of knowledge extraction, concept association, expert interaction and intervention in the case of a medical diagnosis task, and transfer to a compact student network, closing the cycle. We analyse the effectiveness of this process in a use case for determining the presence of pleural effusion in X-ray images. Pleural effusion can be defined as the abnormal accumulation of fluid in the pleural space, typically observed in the lower lung zones of an X-ray. It can be caused by various underlying diseases, such as pneumonia, congestive heart failure, and malignancy [2]. Our approach is developed modularly such that each of the sub-processes can be reviewed and evaluated or audited to ensure transparency of a model suitable for deployment. Furthermore, we demonstrated the potential of using clinically relevant knowledge to achieve high performance in the student model via a teacher-student network.

The rest of this paper is structured as follows. Section 2 presents an overview of related work in explainable AI, neural-symbolic systems, and knowledge distillation in deep learning. Section 3 describes the proposed neural-symbolic process. Section 4 discusses the experimental results. The paper concludes with a discussion and recommendations for future work in Section 5.

2. Related Work

The Bag-of-Visual-Words (BoVW) method [3, 4] has traditionally been used to classify images based on the frequency distribution of pre-defined visual vocabulary (i.e. image features). Typically, these features are extracted manually at specific regions of interest using various feature descriptors (e.g. Histogram of Oriented Gradients (HOG) [5] or Scale Invariant Feature Transform (SIFT) [6]). In the medical domain, radiomics has also emerged as an important topic of study for the extraction of image features from medical images [7, 8]. However, this extraction process can be laborious, prone to human bias, and limited reproducibility of results due to a lack of standardised feature extraction method [9]. Open-source algorithms such as PyRadiomics [9] and Radiomics [10] aimed to address the reproducibility issue by introducing a broad set of commonly used engineered features for medical imaging.

The development of Convolutional Neural Networks (CNNs) has enabled visual features to be extracted automatically from a large amount of images using a data-driven gradient-based parameter search [11], which can then be used to make effective predictions even in specialised medical domains [12, 13]. Understanding the underlying prediction mechanisms of these complex models, however, has remained challenging as the relationships between the extracted features are embedded in the model's large number of parameters. Model interpretation has helped in the visualisation of the relevant pixels that contribute to the predictions [14, 15, 16, 17], but it is well-known that such interpretation may not constitute a model explanation that is

understood by a user or deemed as acceptable by a domain expert [1].

Previous research discovered that semantic meaning can be embedded in the convolutional kernels [18, 19, 20]. Each kernel (hidden unit) in the convolutional layers outputs activated maps (i.e. extracted feature) from the learnt parameters of preceding layers of a CNN on the corresponding images. In these previous works, such semantic meaning was confined to concepts within the limited lexicon of the Broden dataset rendering any concepts beyond the lexicon as uninterpretable . In specialised fields such as radiology, the relevant concepts for a prediction may be unknown a-priori. Understanding these concepts in isolation is also insufficient to explain how they are related to the predictions. Among the approaches seeking to extract knowledge capable of relating such concepts to produce global explanations from neural networks, layer-wise decompositional methods have been the most effective in the case of CNNs, because of their structure and large number of parameters [21]. In [22], a method for global layer-wise extraction of rules was introduced for CNNs. Outputs from kernels with the highest information gain were translated into literals for the extraction of *M-of-N* rules, where a rule is interpreted as being *true* if any combination of M literals out of a set of N literals is *true*. A heuristic search was used to extract rules that prioritised literals based on the weights of the neurons leading to the target output. This approach worked well for kernels near the output, but became ineffective for larger networks when applied over multiple layers due to information loss. In [23, 24], a post-hoc approach was proposed to decompose a CNN for interpretation. The approach converted a CNN into a decision tree with semantic meaning related to the input image. The method has demonstrated that the decision trees can provide insight into how CNNs make predictions.

The ERIC framework [25], used in this paper, has been shown to derive compact rules expressing global explanations for a CNN's convolutional layer. In ERIC, a quantization is used to binarize kernels into literals and symbolic rules are generated in the form of a decision tree. The rules seek to approximate the behaviour of the convolutional layer with respect to the CNN's output. ERIC achieved good classification accuracy and fidelity, that is, accuracy w.r.t. the CNN's outputs, producing a compact rule set which should in principle be more human comprehensible. We regard the ability to measure fidelity and to apply the extraction method to any CNN, irrespective of the training protocol, as two requirements of XAI. For this reason, this paper is built upon ERIC, a general and efficient global XAI approach for CNNs.

Knowledge distillation has been introduced to transfer knowledge from a complex model called the teacher to a typically simpler model called the student model [26]. The goal is to train a student model that can mimic the teacher's prediction performance. Although the term knowledge is used in the original paper, a knowledge representation is not formalised in the teacher-student approach, differently from the logic programming knowledge rules obtained by ERIC for example. Distillation is adopted in this paper to close the cycle such that a student CNN is trained against the more interpretable logic rules and to evaluate the value of creating an auditable neural-symbolic cycle for the student model prediction performance.

3. A Neural-Symbolic Cycle for Medical Diagnosis

The proposed neural-symbolic cycle is illustrated in Fig 1. A trained CNN is used for pleural effusion classification. Symbolic rules are extracted in the form of a decision tree using ERIC. The rules are investigated by a clinician who can associate concepts from CNN kernels and use only clinically-relevant concepts in the construction of decision tree. Relevant knowledge is then distilled into a student CNN model. In what follows, we discuss each of these steps in turn with a focus on the novel part of closing the neural-symbolic cycle with student training and its performance evaluation.

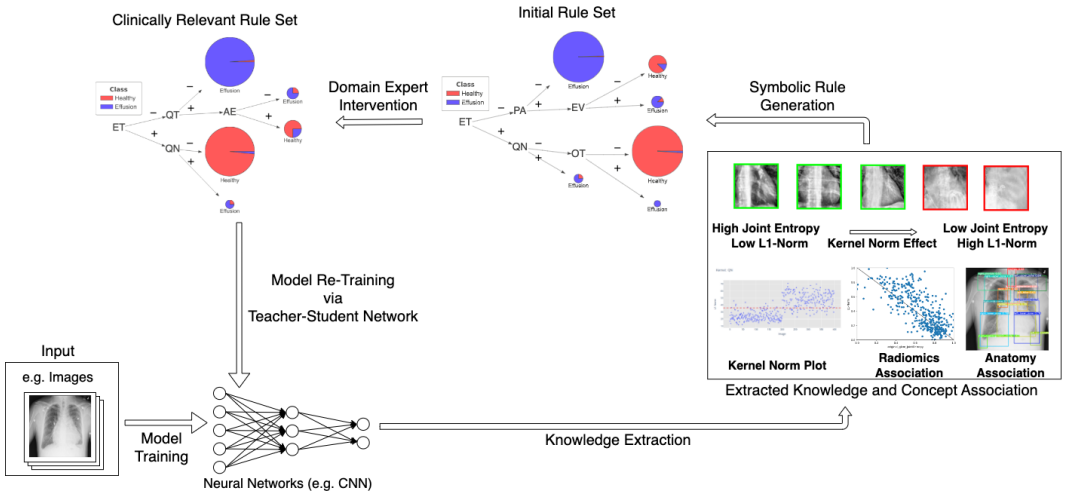


Figure 1: An overview of a neural-symbolic cycle illustrating the process of (a) extracting knowledge from a trained CNN for medical image diagnosis, (b) generating symbolic rules based on the extracted knowledge, (c) expert rule interaction and intervention to produce clinically-relevant knowledge, (d) transferring of relevant knowledge from the rules to a student CNN, closing the neuro-symbolic cycle.

3.1. Datasets

Two datasets were employed in this work. The first dataset, CheXpert [27], was used to train a CNN for pleural effusion detection. Only frontal X-rays with labels *pleural effusion* or *no finding* were used. After images with artefacts or supporting aid obstruction were removed, 400 images were randomly selected for training and 80 images for validation, with both classes equally represented.

The second dataset, NIH dataset [28] with reference to the study metadata in [29], was used to train a supplementary CNN model to locate nine anatomical regions within frontal chest X-rays, namely (a) Trachea (T), (b) Upper Mediastinum (UM), (c) Cardiac Silhouette (CS), (d) Left Clavicle (LC), (e) Right Clavicle (RC), (f) Left Hilar (LH), (g) Right Hilar (RH), (h) Left Costophrenic Angle (LCA), and (i) Right Costophrenic Angle (RCA). This model was then applied to the CheXpert X-rays to assign the associated anatomical concepts to image regions.

3.2. CNN Classification: Model Training

A CNN classifier (M) was trained using the VGG-16 architecture and the Adam optimiser with a learning rate of 10^{-6}. The model was trained from scratch with no pre-trained weights in batches of 32 images using the CheXpert dataset. Elite backpropagation (EBP) was used to improve *class-wise activation sparsity* [30]. This was achieved by associating each class with a small number of kernels that are activated rarely but strongly for related images. The kernels were ranked according to a penalty function based on the activation probabilities of the kernels during training. EBP was previously shown to produce a more distinct separation of kernel concepts and, arguably, more interpretable representations. When seeking to assign semantic meaning to kernels, the above separation will become useful.

3.3. Symbolic Rule Extraction and Human Intervention

The ERIC framework [25, 31] was used to extract rules from the last convolutional layer of the trained VGG-16. Kernels were binarized and the CNN from the last convolutional layer replaced by the derived rules M^* evaluated for its fidelity as a metric of how well M^* approximates M. Earlier work found that rules with a maximum of three literals in the body were sufficient to approximate the trained CNN model for pleural effusion well [32]. Literals associated with CNN kernels were assigned semantic meaning based on the anatomical region localisation as part of a radiomics concept association process. The details of the rule extraction and concept association processes are described in Appendix A and Appendix B, respectively. An experienced medical professional reviewed the generated set of symbolic rules to analyse the clinical relevance of the kernels used, and made modifications to use only anatomically-relevant kernels for the construction of decision tree.

3.4. Model Re-Training via Teacher-Student Network

Model re-training was based on a teacher-student network [26] (see Fig.10 in Appendix E for training schematics). A VGG-16 architecture was chosen for the student CNN model as it has the same feature extraction layers as the teacher model. The training loss, L_{total}, was governed by a loss function which typically consists of a student loss, $L_{student}$, and a distillation loss, $L_{distillation}$ calculated using the categorical cross-entropy and Kullback-Leibler (KL) divergence respectively as shown in Eq. 1 [26, 33]. In this work, the α term in the loss function was set to zero to mimic only the behaviour from the teacher model. The re-training was aimed to investigate how well the CNN model can respond based on the captured clinically-relevant knowledge as measured by the student model's accuracy and losses.

$$L_{total} = \alpha * L_{student} + (1 - \alpha) * L_{distillation} \tag{1}$$

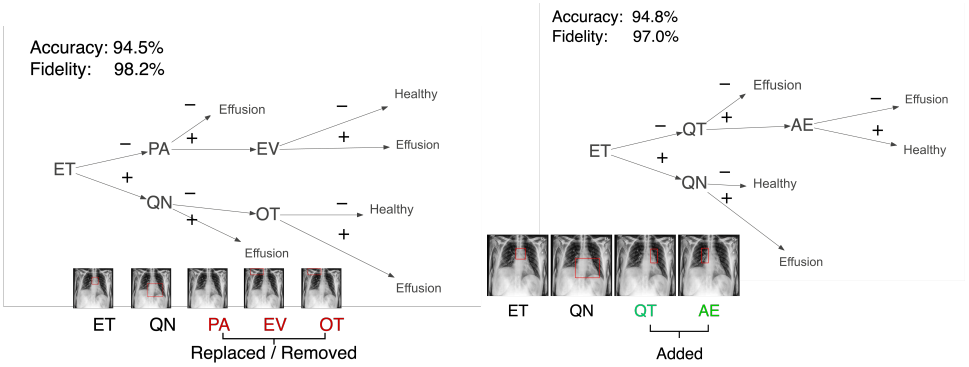
4. Experimental Results

As described in Section 3.2, a CNN model was trained to detect pleural effusion from frontal chest X-rays using the CheXpert dataset and achieved an accuracy of 96.2%. Symbolic rules

were extracted from the trained CNN using the ERIC framework (see Fig 2 (a)) with an accuracy of 94.5% and a fidelity of 98.2%. Five kernels were used (labelled here with associated anatomical regions): ET (Upper Mediastinum), QN (Cardiac Silhouette), PA (Uninterpretable), EV and OT (both relating to the Right Clavicle).

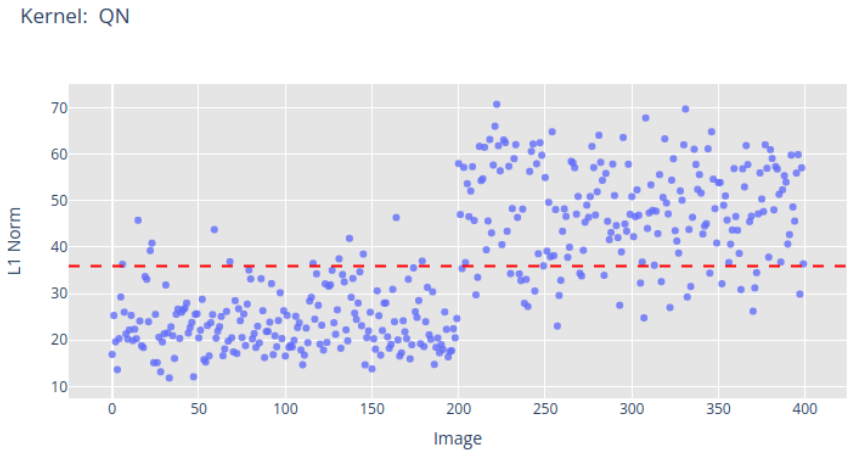
Expert intervention discarded several uninterpretable and other kernels deemed as irrelevant, namely PA, EV and OT. Kernels QT and AE, relating to the left and right hilar, were added instead to reconstruct the decision tree according to the corresponding L1-Norm values. These were deemed as more relevant rules by the expert as being plausibly linked to pleural effusion. This new set of kernels forming a clinically relevant rule set is shown in Fig 2 (b) in the form of a decision tree, where $\neg ET \wedge QT \wedge \neg AE \rightarrow Effusion$ is an example of a rule.

The selected kernels were evaluated using the kernel norm values (L1 norms calculated from the activation maps using Eq. 4 in Appendix A) and associated with a human understandable concept through the localisation of anatomical regions and the correlation with radiomics features. For example in Fig 3, kernel QN was related to the region of Cardiac Silhouette and the change in L1-norms were highly correlated with the Joint Entropy in the Gray Level Cluster Matrix (GLCM). With joint entropy viewed as a quantifiable measure of randomness/variability of pixel intensity in relation to its spatial neighbourhood, low entropy values were taken to indicate a more homogeneous texture and vice-versa (see Fig 3 (c)). This resembles the visual change that a medical doctor will observe in the presence of pleural effusion which will obscure the lung space and the border of the left ventricle of the heart, commonly understood as the "silhouette sign". Additional results on the radiomics analysis for the left and right hilars are presented in Appendix C.

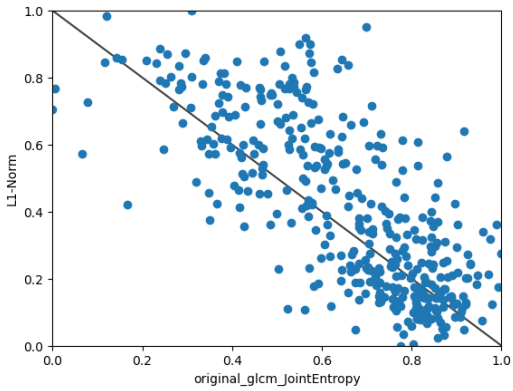


(a) Initial Rule Set (b) Clinically Relevant Rule Set

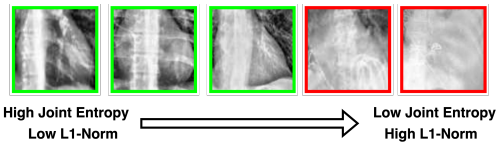
Figure 2: (a) An initial set of rules depicted as a decision tree where ET, PA, QN, etc are literals, + denotes a positive literal, − denotes a negative literal, generated from the trained CNN with an accuracy of 96.2% using the ERIC framework. (b) A clinically-relevant rule set following expert intervention, achieving a comparable accuracy including for example the rule $\neg ET \wedge QT \wedge \neg AE \rightarrow Effusion$. Human intervention produced a clinically-relevant rule set with constituent kernels all relating to anatomical regions: Upper Mediastinum (ET), Left Hilar (QT), Right Hilar (AE) and Cardiac Silhouette (QN). Uninterpretable or irrelevant kernels PA, EV and OT were discarded.



(a) Kernel Norm Plot



(b) Kernel Radiomics Correlation



(c) Joint Entropy (GLCM) for Cardiac Silhouette

Figure 3: (a) A kernel norm plot (L1-norm) for literal QN generated from one of the trained CNN's convolutional layer's output with kernel indexed as 'QN' representing the cardiac silhouette. The first 200 data points from the training set are labelled as *healthy* and the next 200 as *pleural effusion* according to the ground truths. A threshold value (red line) separates positive literals (e.g. QN) (above the line) and negative literals (¬QN), otherwise. (b) A negative correlation between Joint Entropy (GLCM) and L1-norms for Kernel QN. (c) images of the cardiac silhouette region sorted from highest (left) to lowest (right) joint entropy with green outlines indicating *healthy* ground-truth labels and with red outlines indicating *pleural effusion*.

Knowledge distillation to a student network was evaluated using the teacher-student network architecture as described in Section 3.4. Fig 4 presents an investigation of results by removing specific kernels from the last convolutional layer, l, and evaluating the impact on the student model's accuracy. Four runs were conducted including a base case where none of the 512 kernels were removed, a case using 99 interpretable and anatomically-relevant kernels, a case with only the kernels associated with the anatomic regions present in the relevant decision tree (i.e. CS_LH_RH_UM), and a case including the kernels regarded as less anatomically relevant (i.e.

LC_LCA_RC_RCA). In all cases, the student model was trained completely from the teacher model (i.e. alpha = 0). For the first three cases (no kernel muting, interpretable, CS_LH_RH_UM), all the student networks approached similar accuracy as the teacher model. For the case with less relevant kernels (LC_LCA_RC_RCA), the student model performed poorly (accuracy: 53.8%) given that the teacher model also had an accuracy of only 54.8%. When evaluating the change in student loss, it can be seen that the control case with no kernel muting has the highest drop in loss as it has utilised all the kernels to generate contrasting values in the softmax output for the calculation of categorical cross-entropy. The student loss for the case of LC_LCA_RC_RCA remained nearly unchanged, indicating no relevant learning gained against the ground-truth. For both the interpretable and the CS_LH_RH_UM cases, the drop in student loss has fallen in between the other two cases.

Figure 4: Experimental results: model accuracy, student loss and distillation loss comparing student models learned from different teacher models (with different combinations of kernel groups). A teacher model made up of combinations of interpretable and anatomically-relevant kernels (CS,LH,RH,UM) can produce predictive accuracy close to the teacher model without any modification. However, the student models' losses given a teacher model with only interpretable kernels or anatomically-relevant kernels have plateaued at a higher level than the student loss with the unmodified teacher. The teacher model using the least relevant kernels (LC,LCA,RC,RCA) has the lowest model accuracy and the student loss in this case does not drop significantly.

Fig 5 evaluates the use of the extracted and intervened-upon rule sets as the teacher model against the baseline of the originally trained CNN as teacher. Since decision trees are deterministic, a one-hot step was introduced when training. This can be attributed to the initial high distillation loss compared to the trained CNN as teacher, as the discrepancy in value would be significant. The results show that both rule sets performed comparably well on model accuracy and training loss. This tells us that the symbolic rules could be a more interpretable form of knowledge representation replacing complex CNN models as the teacher in knowledge distillation. Based on these results, the intervened-upon and clinically-relevant decision tree should be the preferred choice to train a simpler student model for deployment or further training and analysis because the cycle of knowledge extraction and distillation can make an otherwise obscure system highly transparent and auditable, which is key in the medical field.

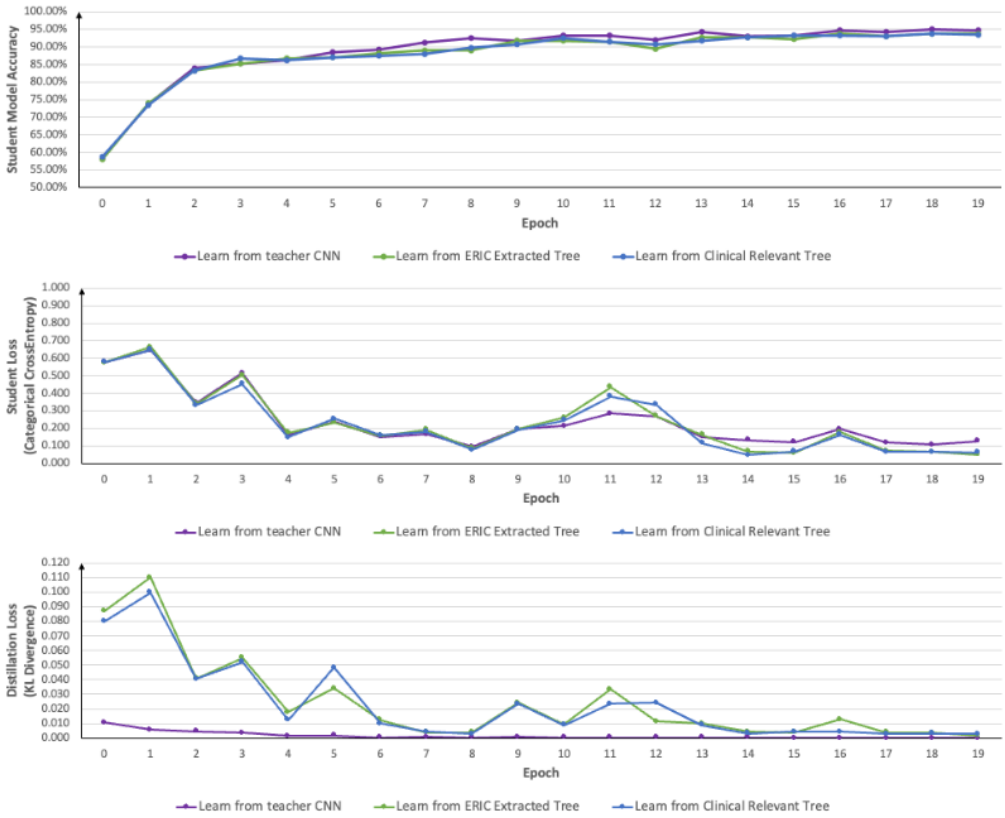


Figure 5: Experimental results: model accuracy, student loss and distillation loss by comparing a student model learning with three teachers: original CNN teacher, ERIC decision tree and intervened-upon (i.e. expert validated) decision tree. Using an ERIC-extracted decision tree or a clinically-relevant, intervened-upon tree as teacher both produce predictive accuracy close to the accuracy produced using the original CNN as teacher. Student losses obtained using both trees are also similar to that obtained using the original CNN as teacher. Results on effect from new and out-of-distribution data are yet to be carried out and may indicate relevant variations from the results obtained thus far.

Additional experimental results are presented in Appendix E. As shown in Figs 13 and 14, student model training requires relevant knowledge from our proposed concept association approach resulting in reducing the student loss and high model accuracy for classifying the ground truth. Fig 15 also shows that a smaller student CNN model (even with 4 kernels or almost 90% more compact) can store the relevant knowledge with a transparent knowledge transfer trail.

5. Conclusion and Future Work

The neural-symbolic cycle was applied successfully to extract, intervene and transfer relevant knowledge back into a CNN model. A simpler student CNN model was trained on extracted knowledge from a more complex teacher CNN model. Extracted knowledge in symbolic form was validated by a domain expert resulting in an interpretable model with a transparent knowledge representation and transfer back into a CNN model while maintaining predictive accuracy. This approach can serve as a blueprint for CNN model audits to validate the flow of knowledge transfer prior to operational deployment particularly in critical fields such as in medicine.

Future work will seek to incorporate kernel feature similarity as an additional training loss component to facilitate better feature learning from the teacher models as well as rule construction methods for human-defined knowledge, thereby strengthening the whole knowledge distillation process. In addition, it is intended that the proposed teacher-student network will expand to include learning from multiple teachers so that a single streamlined student model is needed for training to capture only the relevant knowledge from each of the teachers. Repetition of the neural-symbolic cycle can also be applied to investigate the effect of new knowledge introduction and data drift over time. Kernels will be modified over the cycles as more appropriate knowledge will be discovered. Alternative teacher/student network architectures, different chest X-ray datasets on various respiratory disease types may be investigated to enhance the generalisability of the proposed neural-symbolic cycle. The goal of these future work is to enhance the prediction performance and interpretability of the student models for a variety of respiratory diseases, as well as to facilitate the customisation of these models to meet specific hospital requirements, such as different demographic compositions. Overall, the revised neural-symbolic cycle is expected to result in a student model that captures relevant knowledge with enhanced model explanation and transparency that are essential in critical fields.

References

[1] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv (2017).

[2] V. S. Karkhanis, J. M. Joshi, Pleural effusion: diagnosis, treatment, and management, Open Access Emerg. Med. 4 (2012) 31–52.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: ECCV Workshop on statistical learning in computer vision, vol.1, 2004, pp. 1–2.

[4] Sivic, Zisserman, Video google: a text retrieval approach to object matching in videos, in: 9th IEEE International Conference on Computer Vision, vol.2, 2003, pp. 1470–1477.

[5] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol.1, 2005, pp. 886–893.

[6] D. G. Lowe, Object recognition from local scale-invariant features, in: 7th IEEE International Conference on Computer Vision, vol.2, 1999, pp. 1150–1157.

[7] A.-N. Frix et al., Radiomics in lung diseases imaging: State-of-the-Art for clinicians, J Pers Med 11 (2021).

[8] P. Lambin et al., Radiomics: extracting more information from medical images using advanced feature analysis, Eur. J. Cancer 48 (2012) 441–446.

[9] J. J. M. van Griethuysen et al., Computational radiomics system to decode the radiographic phenotype, Cancer Res. 77 (2017) e104–e107.

[10] M. Vallières, radiomics: MATLAB programming tools for radiomics analysis, 2015.

[11] R. Geirhos et al., ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, arXiv (2018).

[12] P. Rajpurkar et al., Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists, PLoS Med. 15 (2018) e1002686.

[13] J. T. Wu et al., Comparison of chest radiograph interpretations by artificial intelligence algorithm vs radiology residents, JAMA Netw Open 3 (2020) e2022779.

[14] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929.

[15] R. R. Selvaraju et al., Grad-CAM: Visual explanations from deep networks via Gradient-Based localization, Int. J. Comput. Vis. 128 (2020) 336–359.

[16] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, arXiv (2014).

[17] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, arXiv (2013).

[18] B. Zhou, D. Bau, A. Oliva, A. Torralba, Interpreting deep visual representations via network dissection, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2019) 2131–2145.

[19] B. Zhou, D. Bau, A. Oliva, A. Torralba, Comparing the interpretability of deep networks via network dissection, 2019.

[20] D. Bau et al., Understanding the role of individual units in a deep neural network, Proc. Natl. Acad. Sci. U. S. A. 117 (2020) 30071–30078.

[21] J. Mu, J. Andreas, Compositional explanations of neurons, Adv. Neural Inf. Process. Syst. 33 (2020) 17153–17163.
[22] S. Odense, A. D. Garcez, Layerwise knowledge extraction from deep convolutional networks, arXiv (2020).
[23] Q. Zhang, R. Cao, F. Shi, Y. N. Wu, S.-C. Zhu, Interpreting CNN knowledge via an explanatory graph, AAAI 32 (2018).
[24] Q. Zhang, Y. Yang, H. Ma, Y. N. Wu, Interpreting cnns via decision trees, in: IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6261–6270.
[25] J. Townsend, T. Kasioumis, H. Inakoshi, ERIC: Extracting relations inferred from convolutions, ACCV Lecture notes in computer science, Springer, 2021, pp. 206–222.
[26] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv (2015).
[27] J. Irvin et al., CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison, arXiv (2019).
[28] X. Wang et al., ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2097–2106.
[29] A. Karargyris et al., Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for AI development, Sci Data 8 (2021) 92.
[30] T. Kasioumis, J. Townsend, H. Inakoshi, Elite BackProp: Training sparse interpretable neurons, in: NeSy, 2021, pp. 82–93.
[31] J. Townsend, M. Kudla, A. Raszkowska, T. Kasiousmis, On the explainability of convolutional layers for Multi-Class problems, 1st International Workshop on Combining Learning and Reasoning, 2022.
[32] K. H. Ngan, A. D. Garcez, J. Townsend, Extracting meaningful High-Fidelity knowledge from convolutional neural networks, in: International Joint Conference on Neural Networks (IJCNN), 2022, pp. 1–17.
[33] J. Gou, B. Yu, S. J. Maybank, D. Tao, Knowledge distillation: A survey, arXiv (2020).
[34] J. Ross Quinlan, C4.5: Programs for Machine Learning, Elsevier, 2014.
[35] G. Jocher et al., YOLOv5 SOTA realtime instance segmentation, 2022.

Supporting information

A. Symbolic Rule Extraction using the ERIC framework

Let **x** denote a set of input images and **t** denote a set of target outputs, each indexed by the subscript i, where $1 \leq i \leq n$. A convolutional neural network, M, is trained on examples $\{x_i, t_i\}$ and consists of two parts: $g(\cdot)$ 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 the feature extraction layer l, where $1 \leq k \leq k_l$ denotes a *kernel* of the CNN, represented by a square matrix of vectorized real numbers. 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 (kernel norms) in $A^l_{i,k}$ (see Eq.3), and θ^l_k is an user-defined threshold value calculated for each kernel. In this work, the mean L1-norm value for the entire training set was used (see Eq.4).

$$b^l_{i,k} = Q(A^l_{i,k}, \theta^l_k) \tag{2}$$

$$Q(A^l_{i,k}, \theta^l_k) = \begin{cases} 1, \text{ if } a^l_{i,k} > \theta^l_k \\ -1, \text{ otherwise} \end{cases} \tag{3}$$

$$a^l_{i,k} = \|A^l_{i,k}\| \tag{4}$$

$$\theta^l_k = \sum_{i=1}^{n} (a^l_{i,k})/n \tag{5}$$

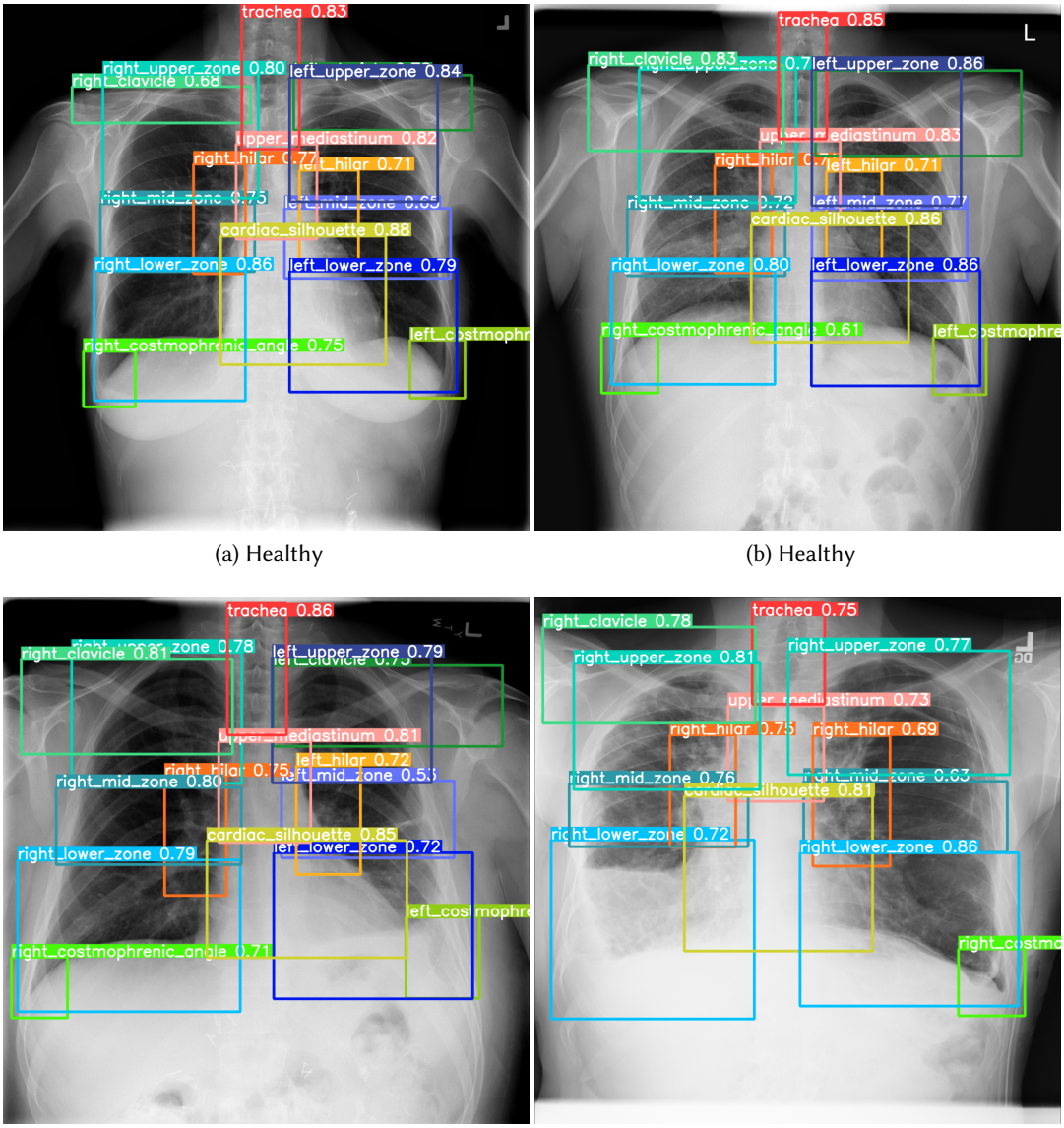
In ERIC, a set of symbolic rules R is generated as an approximation M^* of M using a decision tree-based rule extraction algorithm similar to the C4.5 algorithm [34] trained on instances $\{b^l_{i,k}, h(g(x_i))\}$. Each rule R_r takes the form of a conjunction of literals $L_1 \wedge L_2 \wedge ... \wedge L_{k_l}$, obtained from the feature extraction layer, which implies a CNN classification target output t_i, that is, $L_1 \wedge L_2 \wedge ... \wedge L_{k_l} \rightarrow t_i$. A rule defines a path from the root node to a leaf node in the extracted decision tree. Tree pruning is applied to prevent overfitting. The Gini index is used to determine the branching of tree nodes. If a leaf node has multiple outcomes following pruning, the majority class is selected as the prediction. 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 by up-sampling and inspecting of literals in the rules against the input images.

B. Anatomical Region Localisation

A segmentation model based on the YOLOv5x architecture [35] was trained independently using X-ray images from the NIH dataset [28] to locate nine specific anatomical regions from individual frontal chest X-rays applied in the original trained CNN classification model. Anatomical regions were annotated with reference to [29]. The identified anatomical regions were superimposed on the activated image regions for each CNN kernel (i.e. 512 kernels at the last convolutional layer of a VGG16 model) to evaluate the region of interception. A hit was empirically considered if the interception over union (IoU) score was above 0.7 for each image. The region with the highest aggregated hit rate across the entire training dataset (and hits in at least more than 70% of the dataset) was regarded as the most frequently activated and representative anatomical region for the respective kernel. Additional criteria were implemented, including the restriction of anatomical regions highly hit by kernel activation to no more than two regions, to ensure the kernels were targeted to specific anatomical regions. This anatomical association offered a more comprehensible representation of clinical concepts than the kernel fingerprints used in previous work [32].

With the new representative anatomical regions serving as guiding points in the kernel norm plots, the concept expressed by the kernels could be interpreted with greater clarity. This process also allowed the filtering of uninterpretable kernels (i.e. those not associated with a specific anatomical region) for future research when more knowledge from medical image analysis is made known. To ensure the explainability of the extracted rules, only interpretable kernels (99 in total) were applied to generate the final clinically relevant rule set.

As only frontal X-ray images were used in this work, the relative positioning of the anatomical regions was consistent. This helped with the inspection and manual correction of any missing/incorrectly localised regions needed in this work. The hilar regions and the costophrenic angles were among the more challenging regions, as the left and right regions were very similar. It should also be highlighted that the segmentation model was also capable of labelling the identified regions, albeit with weaker performance on the hilar and costophrenic regions as discussed earlier. Fig 6 displays sample X-ray images with annotated bounding boxes of the anatomical regions in healthy and pleural effusion cases of varying severity. This trained model was separately used to infer on the CheXpert dataset. The annotation were reviewed and manually corrected prior to further evaluation analysis by correlating with radiomics features.



(a) Healthy (b) Healthy

(c) Pleural Effusion (d) Pleural Effusion

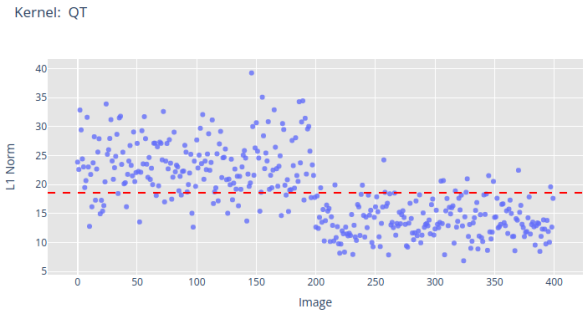
Figure 6: Samples of anatomical region localisation using YOLOv5x model for plain chest X-rays of patients with different severity in pleural effusion. Any missing/wrongly labelled regions (e.g. costophrenic angles, duplicated right lung labelling) have been manually amended prior to the radiomics analysis.

C. Concept Association through Radiomics feature on anatomical regions

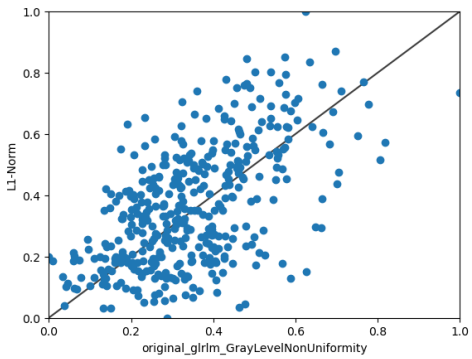
In this section, further representative examples on concept association with radiomics feature on specific anatomical regions are presented. For example in the left hilar region, the correlation between L1-Norm values for kernel QT with Run Length Gray Level Non-Uniformity (GLRLM) was positive. As seen in Fig 7 (a), the L1-norm values were high in the healthy cases (i.e. the first 200) and low for the pleural effusion cases (i.e. remaining 200). As a result, it explained the positive correlation observed when compared to the Gray Level Non-Uniformity feature (i.e high for the healthy cases and low for pleural effusion cases). This corresponded to the visual observation that the left hilar region becomes opaque as the presence of pleural effusion increases (denoted by the change in L1-norm values).

For the right hilar, kernel AE had a positive correlation with the radiomics features to which it was most closely fitted, namely Gray Level Non-Uniformity (GLRLM) (i.e. high L1 norm values and high Gray Level Non-Uniformity for healthy cases and vice versa). By observing Fig 8 (c), the change in L1-norm values (see Fig 8 (a)) could again be translated to the visual observation of the right hilar, which became more opaque in the presence of pleural effusion.

These examples had illustrated that the L1-norm values displayed in the kernel norm plot could be used to approximate the change in visual texture in a particular region, simulating how a clinician would examine an X-ray image.



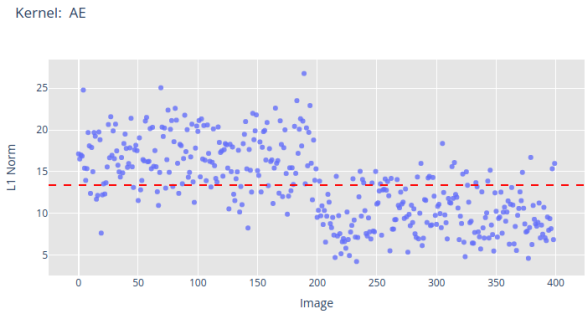
(a) Kernel Norm Plot



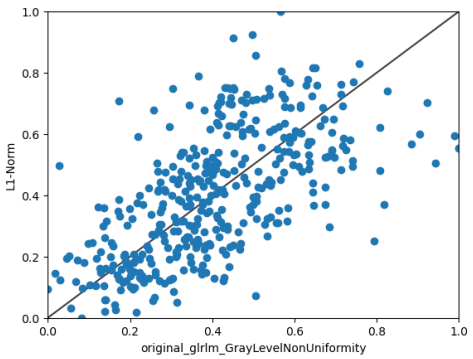
(b) Kernel Radiomics Correlation

(c) Gray Level Non-Uniformity for Left Hilar

Figure 7: (a) A kernel norm plot for kernel QT generated from one of the trained CNN's convolutional kernel output (indexed as 'QT') which represent the left hilar. The first 200 data points from the training dataset are labelled as *healthy* and the next 200 as *pleural effusion* according to the ground truth. A threshold value (red line) separates positive literals (e.g. QT) (above the line) and negative literals (¬QT). (b) A positive correlation is found between Run Length Gray Level Non-Uniformity (GLRLM) with L1-Norms for Kernel QT. Sub-figure (c) shows images of the left hilar region sorted row-wise from highest gray level non-uniformity (top left) to lowest gray level non-uniformity (bottom right). Those images with *healthy* as ground truth label are outlined in green while those with *pleural effusion* are outlined in red.



(a) Kernel Norm Plot



(b) Kernel Radiomics Correlation

(c) Gray Level Non-Uniformity

Figure 8: (a) A kernel norm plot for kernel AE generated from one of the trained CNN's convolutional kernel output (indexed as 'AE') which represent the right hilar. The first 200 data points from the training dataset are labelled as *healthy* and the next 200 as *pleural effusion* according to the ground truth. A threshold value (red line) separates positive literals (e.g. AE) (above the line) and negative literals (¬QT). (b) A positive correlation is found between Run Length Gray Level Non-Uniformity (GLRLM) with L1-Norms for Kernel AE. Sub-figure (c) shows images of the right hilar region sorted row-wise from highest gray level non-uniformity (top left) to lowest gray level non-uniformity (bottom right). Those images with *healthy* as ground truth label are outlined in green while those with *pleural effusion* are outlined in red.

D. Interactive User Interface

A graphical user interface was built to present the decision tree generated from the extracted symbolic rules, with each kernel's corresponding anatomical region displayed. In addition, the kernel norm plot (positioned at the bottom right) can be used to facilitate the assignment of concept descriptions to the respective kernels as demonstrated in [32]. A clinical user can access additional patient information by hovering over the data points, and display the corresponding X-ray image superimposed with the representing anatomical region of the kernel. The interface enables the user to evaluate and modify the decision tree by selecting different kernels, allowing the user to interact with the system by posing questions such as "what happens if a kernel is replaced by another representing the same or another anatomical region?".

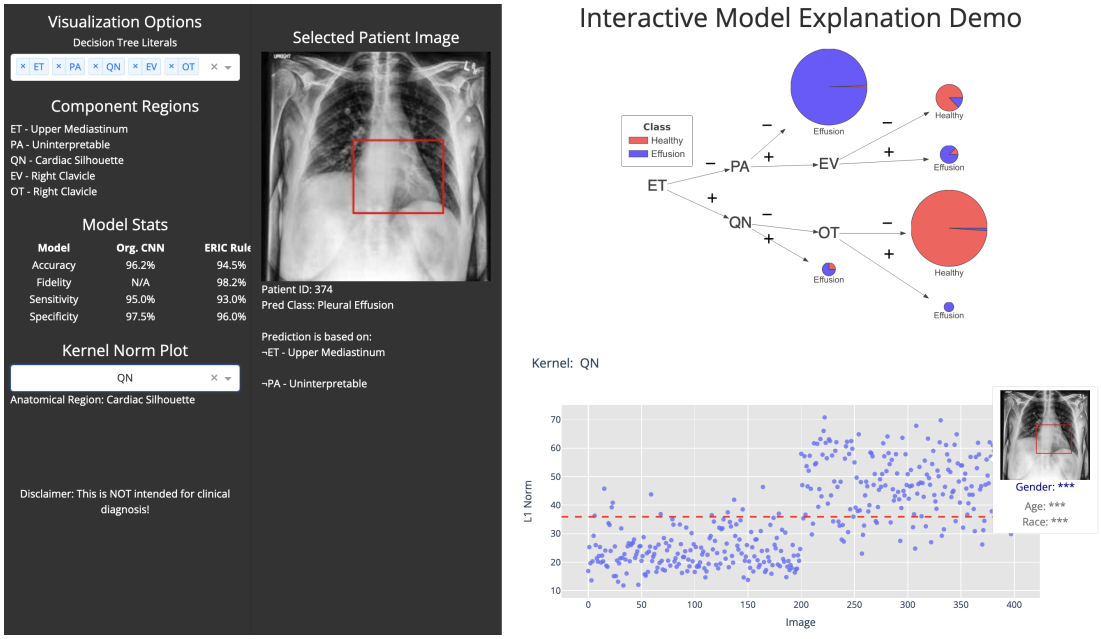


Figure 9: An interactive user interface that enables clinical users to analyse and intervene with the model for improved clinical relevance. The grey sidebar (left) displays the user-defined kernel selection with the corresponding model output. The derived decision tree (from Fig 2a) is displayed on the top right, and the kernel norm plot (from Fig 3a) for a selected kernel is displayed on the lower right for deeper analysis.

E. Supplementary Material on Model Re-Training

This section provides further materials regarding training student models from selected teacher models as described in Section 3.4. Fig 10 presents a schematic of the teacher-student network where a selected teacher model can be used to train a student model based on the minimisation of distillation loss.

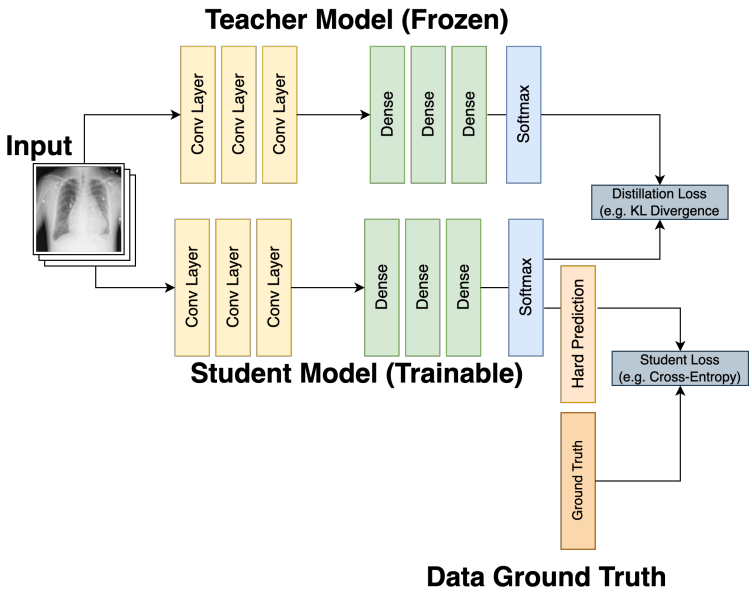


Figure 10: A schematic of the teacher-student network where a selected teacher model can be trained to distill the knowledge necessary to provide similar response on a student model of choice.

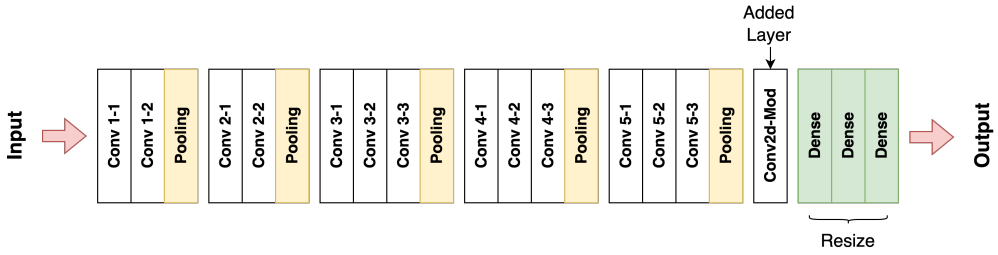


Figure 11: A schematic of the modified student model where an additional bottleneck convolutional layer (see arrow) is included with specific number of kernels. The number of neurons in the subsequent fully connected layers are reduced at a ratio of 6.125 with the flattened kernels to maintain the same ratio as the original VGG16 network.

Fig 12 compares the training loss by learning from ground truth (alpha = 1) and learning from the teacher CNN model (alpha = 0) described in Section 3.2. For the case of learning from teacher, it can be observed that the distillation loss plateaued beyond epoch 6 indicating that training has completed. It is also found that the corresponding student loss has dropped to a

level similar to that learned from the ground truth. This indicates that the student model has learned relevant information from the teacher that can be used to detect pleural effusion.

Table 1 and Fig 13 provide an representative sample comparison analysis of the student model trained using the clinical relevant tree as the teacher and a selected choice of single kernels. It can be shown that the distillation loss for learning from the clinical relevant tree, or relevant kernels alone (ET and QN respectively) decreases close to zero beyond epoch 6. For the case of irrelevant and anatomically unspecific kernel (ie. DB), it has difficulty in mimicking the response of the teacher leading to a slow decline in distillation loss. In addition while the student loss declines significantly for the clinical relevant tree and for the cases of relevant kernels, the student loss for the case of DB (anatomically unspecific) remains high and fluctuating. This indicates that the information learned from the teacher (DB) is not relevant to improving pleural effusion detection in contrast with information learned from the other teacher models. This is correspondingly reflected in the difference in student model accuracy.

Table 2 and Fig 14 provide an alternative comparison between learning from the presented clinical relevant tree and selected cases when a single kernel in the tree is replaced with an alternative kernel of the same anatomical region (e.g. QN to OD (Cardiac Silhouette), QT to DM (left hilar) and AE to MH (right hilar)). This further exemplifies that the training losses (student loss and distillation loss) will decline comparatively similar when relevant teacher models are used to train the student for the detection of pleural effusion.

Lastly, Table 3 and Fig 15 compare learning from the presented clinical relevant tree to a modified student network with varying number of kernels (ranging from 4 to 512 kernels (base case)) at the modified last convolutional layer (see Fig 11). The experimental results show that there is no noticeable performance drop when the modified last convolutional layer is compressed to only four kernels (i.e. almost 90% reduction in trainable parameters). This is consistent with prior work [32] that a rule set consisting of very limited number of kernels is sufficient for the binary classification of pleural effusion.

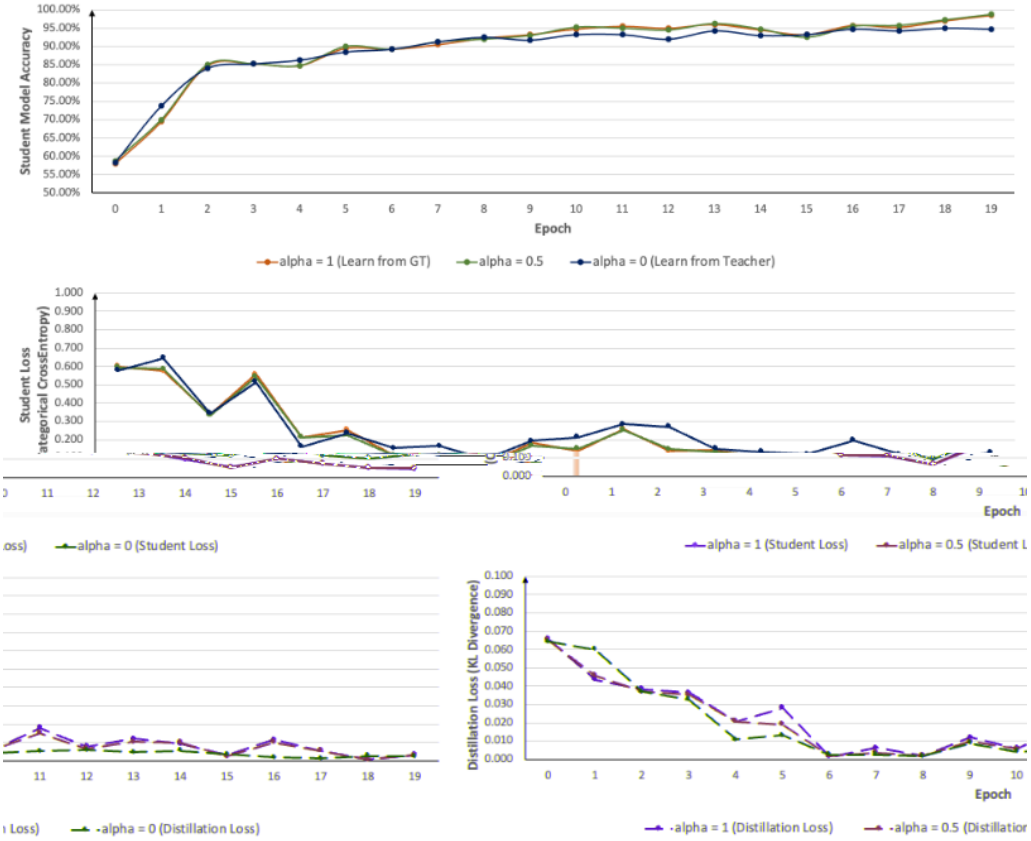


Figure 12: Experimental results: model accuracy, student loss and distillation loss by adjusting the weightage for ground truth (GT) learning (alpha = 1) and teacher learning (alpha = 0). Given that the teacher is a well-trained model against ground truth, learning from the teacher model will yield a similar decline in student loss as distillation loss is plateaued beyond epoch 6. Similarly, training from the ground truth yields a corresponding low distillation loss.

Teacher Model	Teacher Model Train Acc. (%)	Student Model Train Acc. (%)	Student Model Val. Acc. (%)
Clinical Relevant Tree	94.8	93.5	93.8
ET (Upper Mediastinum)	91.8	90.5	92.5
QN (Cardiac Silhouette)	89.5	89.5	91.3
DB (Anatomically Unspecific)	55.8	54.3	55.0

Table 1
Comparison of Accuracy and Fidelity trained from a clinically relevant tree, single relevant kernel only (ET and QN) and single irrelevant kernel (DB).

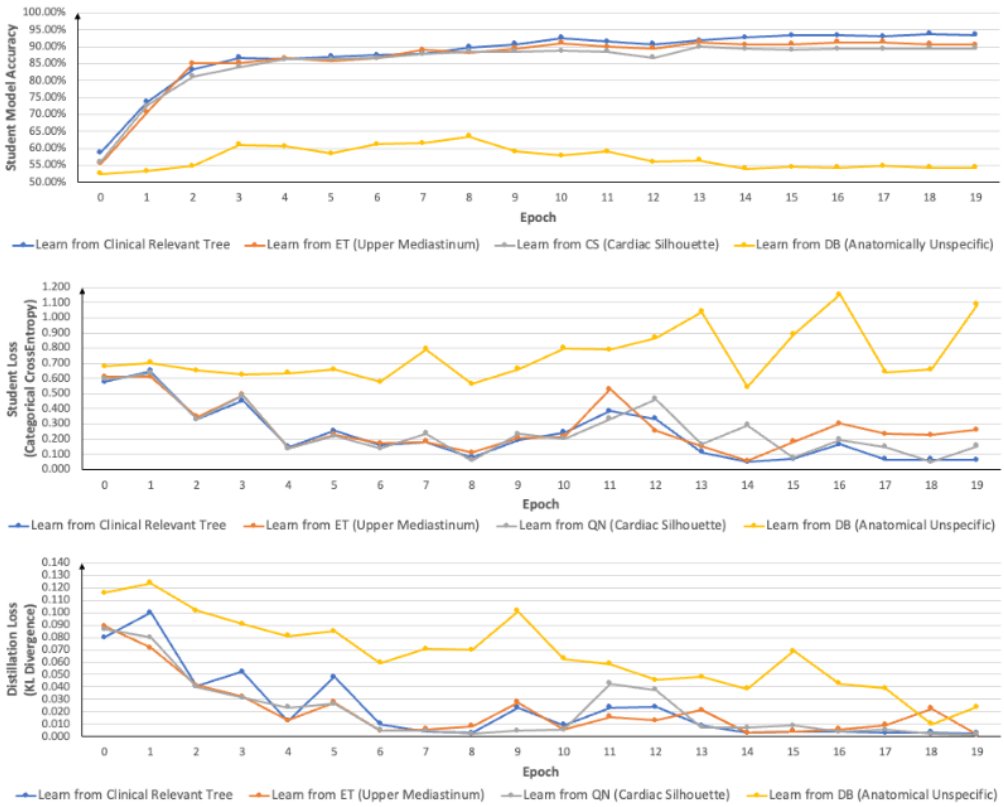


Figure 13: Experimental results: model accuracy, student loss and distillation loss by comparing the learning from a single kernel(ET, QN and DB) with the presented clinical relevant tree. It is observed that it is challenging to learn from an anatomically unspecific kernel (i.e. DB) as shown in the distillation loss. The high student loss from kernel DB is shown to remain high comparing to the use of relevant tree and kernel indicating that the information provided by the relevant tree and kernel helps to improve the detection of pleural effusion.

Teacher Model	Teacher Model Train Acc. (%)	Student Model Train Acc. (%)	Student Model Val. Acc. (%)
ET, AE, QN, QT	94.8	93.5	93.8
ET, AE, QN, DM	93.0	92.5	93.8
ET, MH, QN, QT	94.3	93.0	92.5
ET, AE, OD, QT	94.5	93.5	92.5

Table 2
Comparison of Accuracy and Fidelity trained from variations of the presented clinically relevant tree by replacing a single kernel from the same anatomical region.

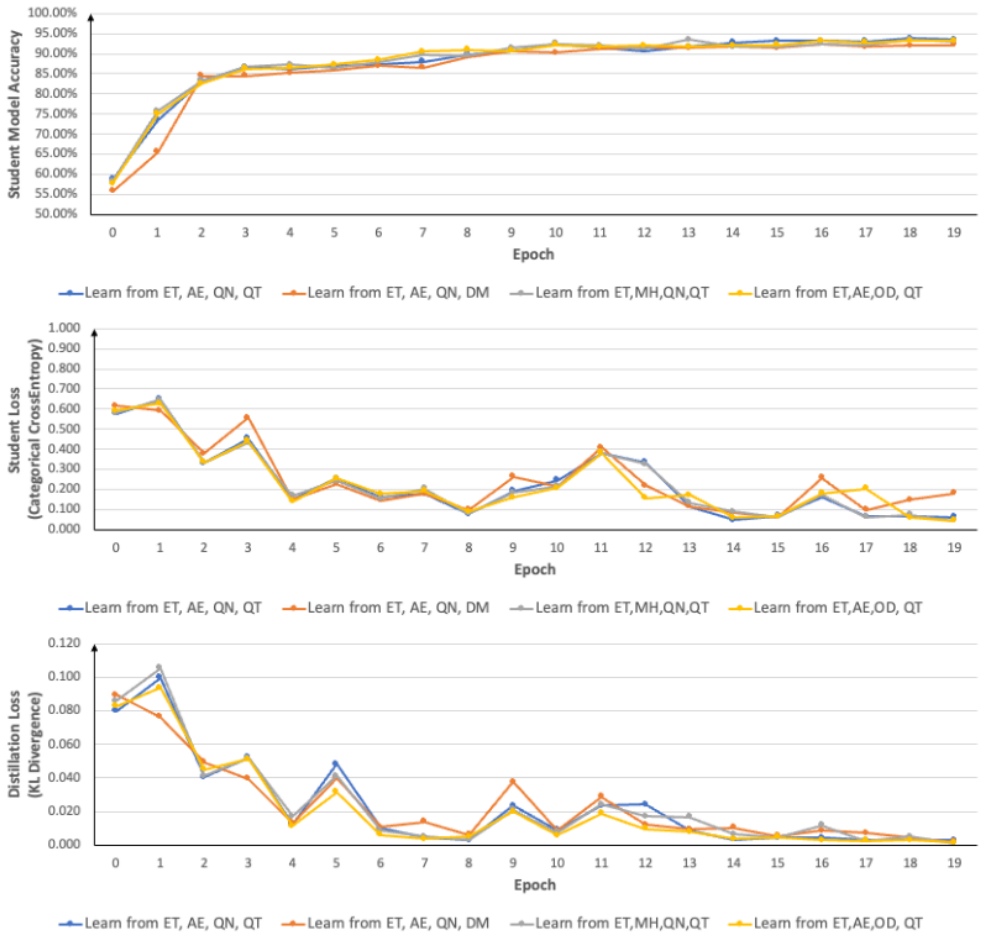


Figure 14: Experimental results: model accuracy, student loss and distillation loss when comparing between learning from variations of the presented clinical tree (ET (Upper Mediastinum), AE (Right Hilar), QN (Cardiac Silhouette), QT (Left Hilar)). It has shown that the learning of the student model appears to be similar where the student and distillation losses are declining in a similar trend between the different teacher models. This indicates that the suggested concepts at the anatomical region have provided similar relevant information to the detection of pleural effusion.

Teacher Model	Teacher Model Train Acc. (%)	Student Model Train Acc. (%)	Student Model Val. Acc. (%)
VGG16_4	94.8	93.8	92.5
VGG16_12	94.8	94.0	92.5
VGG16_50	94.8	93.3	93.8
VGG16_99	94.8	93.3	93.8
VGG16_512 (control)	94.8	93.5	92.5

Table 3
Comparison of Accuracy and Fidelity trained from a clinically relevant tree (ET, AE, QN, QT) across modified student models with a bottleneck layer with varying number of kernels (4, 12, 50, 99, 512 (control)).

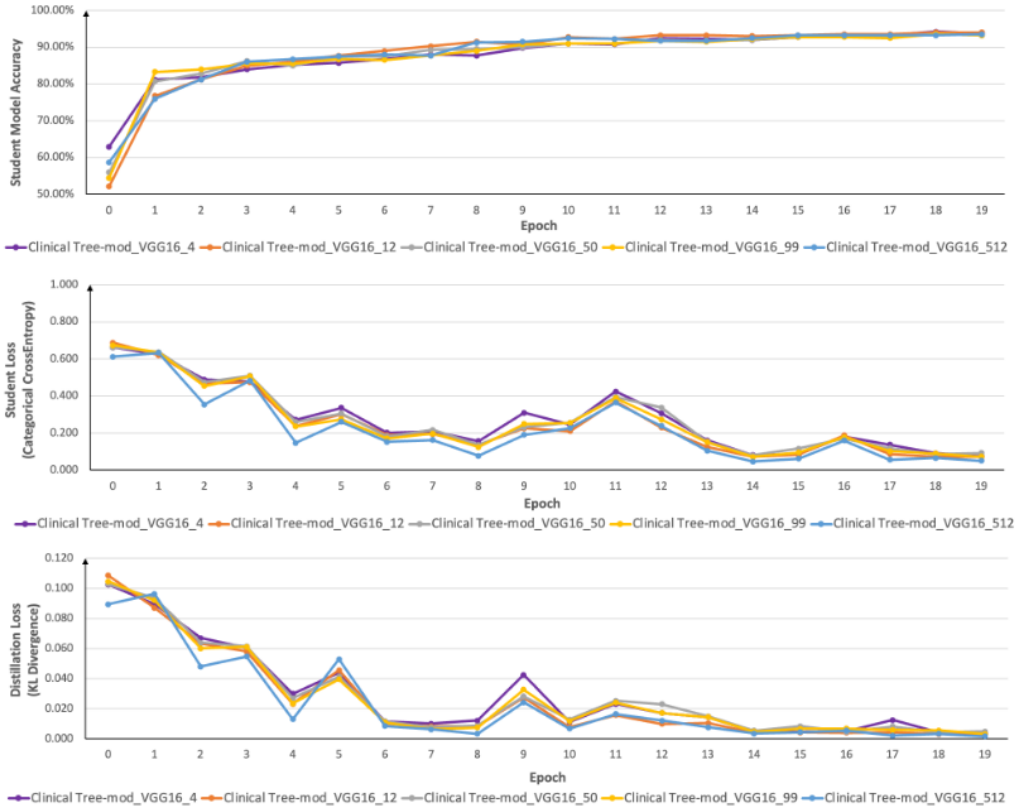


Figure 15: Experimental results: model accuracy, student loss and distillation loss when comparing between learning from va clinical tree (ET (Upper Mediastinum), AE (Right Hilar), QN (Cardiac Silhouette), QT (Left Hilar)) to modified student models with a bottleneck layer of varying number of kernels (4, 12, 50, 99, 512 (control)). It has shown that the learning of the student model appears to be similar where the student and distillation losses are declining in a similar trend between different modified student models. This indicates that the suggested concepts at the anatomical region can be learned in smaller number of kernels to generate accuracy binary prediction.