Symbolic Knowledge Extraction and Distillation into Convolutional Neural Networks to Improve Medical Image Classification

Anonymous Authors

Abstract—**Despite obtaining outstanding performance in radiology applications, the practical uptake of state-of-the-art Convolutional Neural Networks (CNNs) in clinical settings has been limited. This can be attributed to (1) a lack of trust in purely data-driven networks, and (2) the complexities of the medical decision making processes. To address these issues, this paper applies and evaluates a novel neural-symbolic approach. Our approach allows for domain expert intervention following the extraction of meaningful knowledge, such that relevant knowledge can be validated and distilled into the CNN. The neural-symbolic approach is shown to enhance control over conventional CNN training by providing symbolic descriptions that can be combined into simplified CNNs, allowing domain experts to audit the system as part of the decision making process. The kernels in a given CNN layer are mapped onto symbolic knowledge in the form of logic programming rules. Extracted knowledge is evaluated against known radiomics features, allowing doctors to decide based on best practice which kernels to keep or reject. Expert intervention takes place through relevant knowledge distillation back into a more compact CNN. Our results show that a student CNN can learn successfully even from multiple teachers (different knowledge-bases) to replicate the selected relevant kernels and corresponding classification results. The proposed approach delivers a trainable parameter reduction of at least 56.3% while achieving high cosine similarity for kernel replication and a fidelity of 99.3%. Inspection by a domain expert confirms the potential of the neural-symbolic approach to help increase trust by doctors in data-driven AI.**

Index Terms—**Explainable AI, Knowledge Distillation, Image Classification, Neurosymbolic Cycle**

I. Introduction

Recent developments in deep learning and symbolic AI research have drawn significant attention to neural-symbolic approaches [1, 2]. The objective of neural-symbolic AI is to leverage the benefits of both neural networks and symbolic computation, to explore background knowledge, enhance explainability, enable reasoning and facilitate validation of data-driven learning [3, 4]. The ultimate goal is to foster greater trust and better understanding of the processes at play in AI systems.

Neural-symbolic AI methodology calls for translation algorithms to establish a connection between distributed learning models and symbolic representations. When this connection operates effectively in both directions, from the neural to the symbolic representations and vice-versa, it lays the foundation for a virtuous neural-symbolic cycle. In this cycle, features of a trained network can be described symbolically with the benefit of a formal semantics. Additionally, symbolic knowledge can be instilled into neural networks for learning from both data and knowledge, with the promise of improved user interaction, transfer learning and extrapolation beyond the data distribution.

In radiology as a widely analyzed application area for Convolutional Neural Networks (CNNs), the plain chest X-ray (CXR) is an efficient and cost-effective diagnostic tool [5]. Despite its simplicity, CXRs can be used to identify a range of pathologies making interpretation a complex task [6]. The challenge stems from the intricate relationships between features and outcomes, where a single feature can correspond to many outcomes and vice versa. Interpreting a CXR outcome requires accounting for a range of findings that exhibit multi-directional interactions. For example, an enlarged cardiac size alone may be found in different diagnoses, such as pleural effusion, pulmonary edema or congestive cardiac failure. The absence of an enlarged cardiac size does not exclude these diagnoses. Methods solely rely on a location-based interpretation may underperform or utilize irrelevant features, where relevance is assigned through clinical reasoning.

The complexity of radiology classification by CNNs can be illustrated in an example of subtle lymph node enlargement in the context of early infection or malignancy as the labeled diagnosis [7, 8]. Deviation of structures, e.g. mediastinal and tracheal deviations, may be associated with underlying pathologies in the context of pleural effusion [9]. Thus, effective incorporation of clinical knowledge of such multi-factorial and multi-directional pathological process into a CNN is necessary to tackle the complexity of medical decisions, particularly when trying to learn the interactions among the image findings or deriving an explanation for the CNN's outcome.

This paper proposes and evaluates a neural-symbolic cycle for Convolutional Neural Networks applied to the classification of chest X-rays. We use the architecture of CNNs to extract compact logic programming rules from the network's last convolutional layer. This knowledge extraction approach, based on the ERIC framework [10] enhanced with radiomics [11], allows for the extraction of global explanations from CNNs trained to classify X-ray images for pleural effusion [12]. A medical doctor then inspects the extracted rules to assign clinically relevant concepts to the literals in the rules which correspond to CNN kernels. The goal is to empower domain experts to interact with large deep learning models, ask what-if questions, and select relevant features and concepts for re-use. Intervened-upon rules are then re-trained into a simplified CNN using a teacher-student architecture and loss functions introduced in this paper to deal with the case of multiple teachers, that is, multiple sets of rules obtained potentially from different trained CNNs and domain experts. Finally, closing the proposed neural-symbolic cycle allows for knowledge extraction from the student network, yielding a revised set of rules for critical comparative evaluations by experts.

To our best knowledge, this paper is the first to offer:

1) the application of a complete neural-symbolic cycle to CNNs widely used in radiology practice;
2) domain expert validation and intervention capable of recognising when a trained CNN achieves high accuracy but using medically irrelevant features;
3) a teacher-student method for re-training CNNs from data and knowledge given multiple knowledge-bases.

Our experimental results highlight the value of the proposed neural-symbolic cycle at identifying issues with trained networks that are not possible to identify through standard model performance evaluation. Furthermore, the proposed teacher-student architecture and loss functions are shown to align the newly trained student CNN with medically relevant features and kernels. The experimental results illustrate how CNN training can be guided over time to represent more generalizable medical concepts, as normally required by transfer learning, such as in the case of pleural effusion due to trauma and pleural effusion due to infection.

Given that CNNs are also susceptible to variation in data distribution arising from different imaging equipment and settings or in the case of hospitals serving diverse demographic segments [13, 14], the extraction and distillation of logical rules with precise semantics is expected to offer a well-founded alternative to the current efforts towards alignment of deep learning models. The ultimate aspiration is to empower current AI systems with expert evaluation and intervention capabilities at an appropriate level of abstraction to mirror the combined use of data-driven evidence and rule-based protocols in real-world medical decision-making.

The rest of this paper is structured as follows: Section II provides an overview of related work in knowledge extraction, neural-symbolic integration, and knowledge distillation. Section III describes the proposed neural-symbolic approach. Section IV presents the experimental results and discusses the findings in detail, including achieving a student network compression of 56.3% with high cosine similarity for kernel replication, and a network-to-knowledge fidelity of 96.3%. Section V concludes the paper and discusses directions for future work.

II. Related Work

The historical definition of knowledge can be referred to Plato's justified true belief theory of which a conviction for a proposition to be true against an objective reality should be supported by exemplified evidence. In the context of radiology, knowledge extraction is about obtaining visual patterns from medical images for decision-making. Traditional approaches relied on handcrafted features from specific regions of interest. The rise of Convolutional Neural Networks (CNNs) enabled the automation of visual feature extraction to facilitate predictions even in specialized domains. However, understanding the relationships between the extracted features and specialized concepts that make up the prediction remains a challenge. Model interpretation techniques have aided in visualizing relevant pixel-level contributions to the predictions [15, 16, 17, 18], but such interpretation may not provide users with sufficient information about the model to intervene and make changes.

Prior work also managed to achieve some level of disentanglement from convolutional kernels, contributing to the performance of downstream predictions. Each kernel outputs an activation map regarded as extracted features from the learned parameters of preceding layers, down to the input image to which meaning can be associated [19, 20, 21]. Such work, however, rely on a limited lexicon from the Broden data set for the extraction of meaning, making any concept that is not in the lexicon uninterpretable. In specialized field such as radiology, this can become problematic. [22] attempts to utilise clustering and visualize so-called 'kernel fingerprints' to associate meaning to kernels while [23] associate kernels with radiomics features. They were able to derive meaningful model explanations via the extraction of decision trees from CNN kernels.

Numerous approaches have also investigated knowledge extraction from CNNs to produce global explanations in the form of symbolic rules. For example, [24] introduced a global layerwise extraction of rules from CNNs. Kernel outputs are 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 out of a set of N literals is *true*. Knowledge extraction is accomplished using a heuristic search to maximize information gain yet prioritizing literals according to the weights associated with the target output. Although theoretically sound, this approach can become inefficient for large networks. [25, 26] presented a post-hoc approach to decompose a CNN for interpretation. They converted a CNN into a decision tree with meaning associated with the input image. The method has shown that the decision trees can provide insight into how the CNN makes a prediction.

The ERIC framework [10] offers a knowledge extraction method that can produce compact rules expressing global explanations for a convolutional layer. A quantization process is applied to the kernels that converts them into literals such that logic programming rules can be generated. The rules seek to approximate the behavior of the convolutional layer with respect to the CNN's output. The ERIC framework has been shown capable of achieving high classification accuracy and fidelity, that is, accuracy w.r.t. the CNN's output, while producing a compact set of rules which , at least in principle, is expected to be more comprehensible. We regard the ability to measure fidelity and to extract compact rules efficiently as key requirements for the neural-symbolic cycle to work in practice.

As such, this paper chooses to use the ERIC framework for the extraction of rules from CNN kernels.

To close the neural-symbolic cycle, knowledge extracted and analyzed by experts is instilled back into the CNN. Knowledge distillation was first introduced to compress a complex neural network called the teacher into a simpler network called the student model [27]. Response-based knowledge transfer is achieved if the trained student model is capable of mimicking the teacher's response performance (i.e. if the student has good fidelity to the teacher's prediction output). Feature-based knowledge distillation [28, 29, 30] has also been reported to show model performance improvement by including the learning of features from intermediate layers so as to provide more specific information to the student. However, these feature-based distillation methods tend to transfer entire representations of the intermediate layers to the student. Based on the findings of kernel disentanglement from [21] and [22], feature-based distillation is adopted in this paper, but only specific kernels from (possibly multiple) teachers will be learned. Our goal is to evaluate how distillation can be controlled after expert intervention and the effectiveness of combining relevant concepts from multiple sources.

III. Knowledge Extraction, Expert Intervention and Knowledge Distillation

This paper's proposed approach follows the neural-symbolic cycle illustrated in Fig. 1. CNN models are pre-trained for the detection of pleural effusion. These models are regarded as 'black-boxes'. Symbolic rules can be extracted from each CNN model using the ERIC framework to express the kernels in a convolutional layer in the form of logic programming rules (see bottom left of Fig 1). Each 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 an atom A. Each literal corresponds to a kernel of the CNN's chosen convolutional layer associated with a region of the input image (in our case, a chest X-ray); atom A denotes a given CNN classification output, in our experiments *pleural effusion*. Additionally, each literal is expected to be associated with clinically relevant concepts corresponding to a radiomics feature of the related anatomical region (bottom right of Fig 1). Clinical experts have the option to intervene in the set of rules, presented for user interaction in the form of a decision tree (each node representing a kernel) to which a user can add or remove kernels (Fig 1, top right). Finally, new knowledge obtained from multiple sources, whether directly from the extracted features of trained CNNs or through further expert intervention and validation, can be distilled into a more compact CNN closing the neural-symbolic cycle (Fig 1, top left).

A. Domain of Study and Datasets

Within radiology, this study focused on chest x-rays and used two datasets for training and analysis. The primary dataset, CheXpert [12] was used to train (teacher) CNN models for pleural effusion detection. Only frontal X-rays with labels indicating *pleural effusion* or *no finding* were used. Images with artifacts or supporting aid obstruction were manually removed. Two subsets of 400 images each were randomly selected from the CheXpert dataset to train two teacher models. An additional set of 80 images was used for validation containing equal representation of both classes.

To associate symbols with clinically relevant concepts, a supplementary CNN model using the default YOLOv5x architecture [31] was trained from scratch on the NIH dataset [32]. This model was designed to locate nine anatomical regions in 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). The trained model was then applied to the selected X-ray images from the CheXpert dataset to locate the most relevant anatomical region of the representing kernels from knowledge extraction for clinical concept association. This process is described in Section III-C with further details in Appendix A.

B. Teacher CNN Model Training

Two CNN classification models (M) were trained using standard VGG-16 architecture to simulate two sources of knowledge. The models were trained from scratch (without pre-trained weights) using the Adam optimizer with learning rate 10^{-6}. Training was performed in batches of 32 images from the image subsets of the CheXpert dataset as described in Section III-A. Elite backpropagation (EBP) was applied to increase class-wise activation sparsity during training, as this approach was previously shown to work well with ERIC [33]. EBP works by seeking to associate each class with a smaller set of kernels than would otherwise be obtained from using standard backpropagation. It ranks the kernels based on a penalty function derived from their activation probabilities during training.

C. Rule Extraction and Human Intervention

The extraction of symbolic rules was applied to the last convolutional layer of the VGG-16s. For each input image $\mathbf{x}$ indexed by a subscript i, $1 \leq i \leq n$, and for each kernel k, $1 \leq k \leq k_l$, at a feature extraction layer l, the L1-Norm values $a^l_{i,k}$ were calculated for each kernel given the activation maps $A^l_{i,k}$ at layer l (see Eq. 1 in Fig. 1). Threshold values, θ^l_k, were determined for each kernel by computing the mean L1-Norm over the entire training set (Eq. 2 in Fig. 1). Kernel norm values for each image were then mapped to $\{-1, 1\}$, with -1 denoting *false* and 1 denoting *true* (Eq. 3 in Fig. 1). The resulting set of truth values, $b^l_{i,k}$, is 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 $b^l_{i,k} = -1$ (Eq. 4 in Fig. 1).

Using literals L and the corresponding target output, t, a decision tree, $M*$, is trained in the usual way (using Gini impurity as the criterion for node splitting). The decision tree is expected to offer an interpretable surrogate to the 512 channels of 7x7 kernels in the original CNN model M.

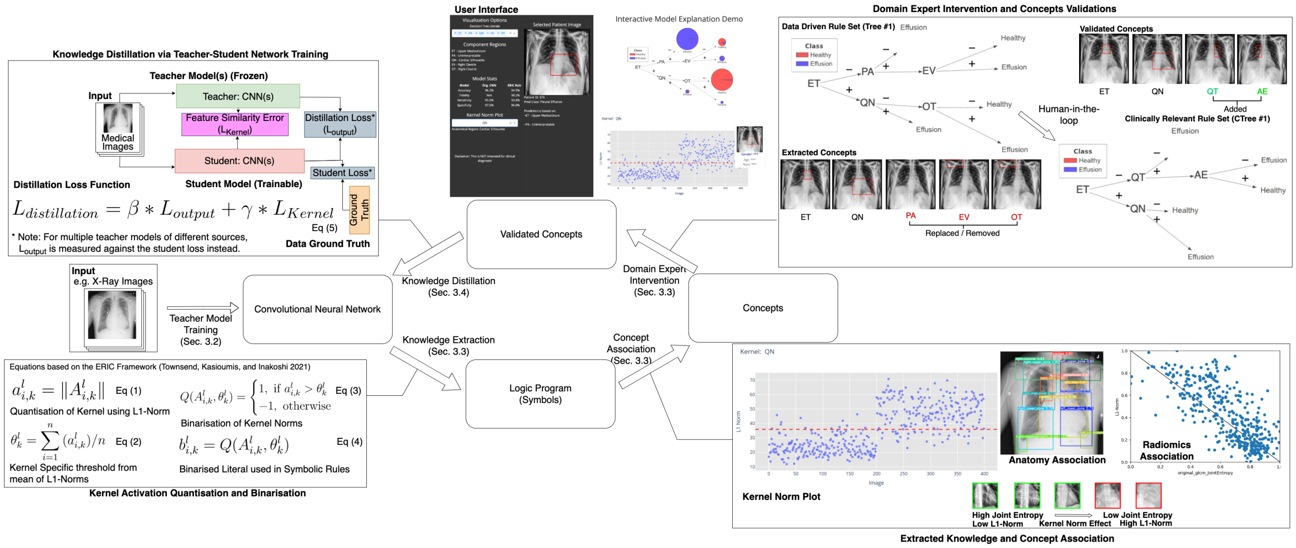


Fig. 1. The proposed neural-symbolic cycle: (a) Logic programming rules are extracted from a trained CNN for medical image analysis, (b) Literals representing 5 kernel are mapped to image regions (literal ET mapped to Upper Mediastinum, QN to Cardiac Silhouette, PA to an uninterpretable region, EV and OT to Left Clavicle); (c) Expert intervention allows for the selection of clinically-relevant knowledge. The system suggests other potentially relevant literals for use. In the example shown here, high fidelity is maintained with the use of 4 clinically-relevant kernels (ET and QN from before plus AE (Right Hilar) and QT (Left Hilar)); (d) Relevant knowledge is distilled into a simplified student CNN through the use of a teacher-student architecture and loss functions.

Fidelity offers a measure of how good this surrogate, $M*$, by comparing the prediction output of the surrogate against the original CNN model, M. An example of extracted decision tree is shown in Fig. 1 (top right inset) as the ”data-driven rule set (Tree #1)”. Previous work has shown that compact trees with a maximum depth of three are capable of approximating the VGG-16’s last convolutional layer with high fidelity [22].

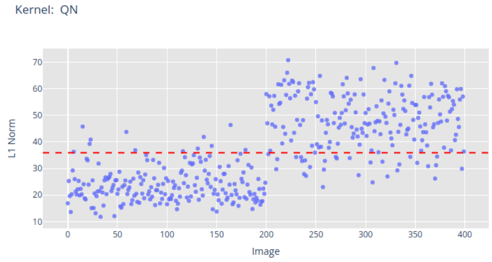
The literals L are assigned meaning based on the anatomical region, inferred from the YOLOv5x model, that is most activated for the corresponding kernel across the entire training set. L1-Norm values, $a^l_{i,k}$, are compared with the radiomics features. For instance, a literal corresponding to a kernel labelled as QN was associated with the cardiac silhouette and the Grey Level Cluster Matrix (GLCM) Joint Entropy at this region was the most correlated by mutual information with the L1-Norm values across 93 radiomics features applied. The GLCM Joint Entropy provides a quantifiable measure of pixel intensity randomness in relation to its spatial vicinity, where low entropy values indicate a more homogeneous visual texture and vice-versa (see Fig. 2). This association of images to concepts resembles the visual findings sought by medical clinicians when inspecting X-rays. In the case of kernel QN, the commonly known ”silhouette sign” would show an obscured lung space along with the border of the left ventricle of the heart due to the presence of pleural effusion. Further details on the radiomics analysis can be found in Appendix B. A medical doctor was involved in the analysis removing 3 uninterpretable/irrelevant kernels, PA, EV and OT, and added 2 relevant kernels, AE and QT, to form the ”clinically relevant rule set” having 4 relevant kernels (CTree #1 in Fig. 1, top right inset).

D. Relevant Knowledge Distillation

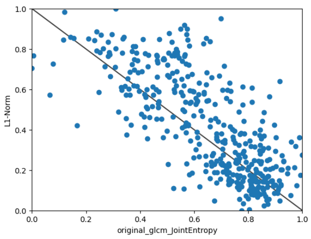
The task here is to take a set of rules, obtained potentially from expert intervention on multiple decision trees, along with the original CNNs and produce a simplified CNN that is expected to account for the knowledge in the set of rules. This was accomplished through knowledge distillation using a teacher-student model, where the teachers were derived from the models trained in Section III-B. A modified VGG-16 model (see Fig. 3) was chosen as the student model, sharing the same feature extraction layers as the teacher models but, crucially, with an additional bottleneck convolutional layer having the same number of kernels as determined by the number of kernels deemed relevant by an expert following rule extraction and intervention described earlier. To maintain consistency with the original VGG-16 configuration, a reduction of the number of neurons in the fully-connected layers was made at a ratio of 6.125 based on the length of the flattened kernels in the bottleneck. The student model with 4 relevant kernels at the bottleneck, as an example, has 14.7 million trainable parameters, a reduction of 56.3% in the number of parameters in comparison with one of the teacher models with 33.6 million trainable parameters.

Knowledge distillation is guided through a training loss function, $L_{disillation}$, comprising of two components: L_{output} that calculates the difference between the softmax outputs of the teacher and the student using the Kulback-Leibler (KL) divergence, and L_{Kernel} that seeks to minimise the negative of the cosine similarity between the relevant kernels in the teacher(s) and those at the convolutional bottleneck layer of the student.

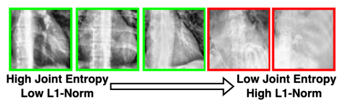
During multi-teacher distillation, a slight modification to the



(a) Kernel norm plot



(b) Kernel radiomics correlation



(c) Joint entropy (GLCM) for cardiac silhouette

Fig. 2. (a) Kernel norm plot displaying the L1-Norm values for kernel QN on a training set with 400 images. The first 200 images have label *no finding* and the next 200 images have label *pleural effusion*. The threshold value denoting the average L1-Norm (dotted red line) separates images for which literal QN is assigned $true$ (on or above the line) or $false$ (below the line); (b) The highest negative correlation between radiomics features measured by GLCM joint entropy and L1-Norms for kernel QN is observed at the cardiac silhouette region; (c) Images of the cardiac silhouette region in descending order of joint entropy and increasing L1-Norms (from left to right) shows the gradual obscurity of the cardiac silhouette known as "silhouette sign". Images labeled *no finding* are shown with a green outline and images labeled *pleural effusion* with a red outline.

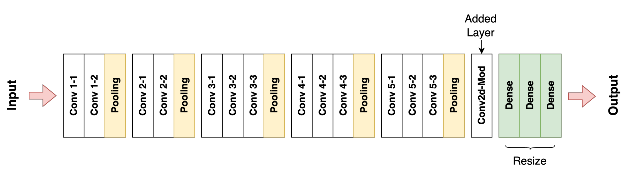


Fig. 3. A schematic of the student network model where a bottleneck convolutional layer is included (see arrow) with a much reduced number of kernels. The number of neurons in the subsequent fully-connected layer is reduced accordingly to maintain the same ratio as the teacher network.

distillation loss was introduced so that L_{output} was measured against the ground-truth instead unless otherwise stated. This is because it would not be appropriate to apply any aggregation method to the outputs of the two teachers trained from different subsets of the CheXpert X-ray dataset as defined in Section III-A. Using the ground-truth in this case will allow us to evaluate predictive accuracy and the efficacy of the selected learned kernels against that ground-truth. Eq. 5 in Fig. 1 shows the training loss function and its constituent terms with weighted hyper-parameters β and γ for controlling the contribution of each term.

IV. Experimental Results and Findings

Two CNN teacher models (Model #1 & #2) were trained to detect pleural effusion in frontal chest X-rays as described in Section III-B. Symbolic rules in the form of decision trees (Tree #1 & #2) were generated respectively from Model #1 & #2 following the ERIC framework. From the radiomics analysis, Tree #1 contained uninterpretable kernels (kernel PA, to which a specific anatomical region or clinically relevant concept could not be assigned) and irrelevant kernels (kernels EV and OT pertaining to the right clavicle, which should be unrelated in a pleural effusion diagnosis). Intervention by a domain expert then produced a clinically-relevant tree (CTree #1), built from Tree #1 by selecting kernels AE, ET, QN and QT, validated against clinical concepts as described in Section III-C while maintaining accuracy and fidelity. Tree #2, on the other hand, did not require any expert intervention. Its kernels were already both interpretable and relevant to pleural effusion. For this reason, Tree #2 is referred to as the "data-driven tree". Fig.4 contrasts the rules sets from CTree #1 and Tree #2 which form the basis for our experiments to close the neural-symbolic cycle with knowledge distillation. A performance comparison of the teacher CNN models and the derived trees is also presented in Table I. All trees exhibit high fidelity with respect to the corresponding CNN models. It should be noted that the training accuracy is nearly indistinguishable between Tree #1 and CTree #1. Despite a small reduction in fidelity, intervention achieves a comparable increase in validation accuracy. Tree #2, despite not requiring intervention, has a loss of information noticeable in the drop in validation accuracy.

$$\neg ET \wedge \neg QT \implies Effusion$$
$$\neg ET \wedge QT \wedge \neg AE \implies Effusion$$
$$\neg ET \wedge QT \wedge AE \implies Healthy$$
$$ET \wedge \neg QN \implies Healthy$$
$$ET \wedge QN \implies Effusion$$

(a) Clinically relevant rules

$$\neg DH \implies Effusion$$
$$DH \wedge \neg IB \wedge \neg N \implies Healthy$$
$$DH \wedge \neg IB \wedge N \implies Effusion$$
$$DH \wedge IB \implies Effusion$$

(b) Data driven rules

Fig. 4. (a) Set of rules for pleural effusion obtained following expert intervention (CTree #1) and (b) set of rules extracted from CNN without expert intervention (Tree #2), each generated from CNNs trained on different subsets of the CheXpert dataset.

CTree #1 has four clinically relevant kernels, namely ET (Upper Mediastinum), QN (Cardiac Silhouette), QT (Left Hilar) and AE (Right Hilar). Tree #2 has three clinically relevant kernels, namely DH (Trachea), along with IB and N both associated with the Cardiac Silhouette. A rough comparison of the trees indicates that Tree #2 seems to be more specialized, concentrating only on kernels relating to the Cardiac Silhouette and Trachea, with the Trachea typically associated with pleural effusion related to tuberculosis or other respiratory infections [34]. By contrast, CTree #1 with its four kernels, each associated with a different clinically-relevant region, seems to offer a more generalist explanation, which is expected to transfer better onto a related but new image.

The following experiments focus on evaluating the effectiveness of knowledge distillation onto a student network using

TABLE I
CNN MODEL PERFORMANCE AND CORRESPONDING KNOWLEDGE EXTRACTION PERFORMANCE WITH AND WITHOUT EXPERT INTERVENTION. ALTHOUGH THE TRAINING ACCURACY OF TREE #1 IS SIMILAR TO THAT OF CTREE #1, AN IMPROVEMENT IN VALIDATION ACCURACY IS SEEN FOR CTREE #1, WHICH USES ONLY CLINICALLY-RELEVANT KERNELS.

Model	Train Acc.	Train Fid.	Val. Acc.
Model #1 CNN	96.2%	-	92.5%
Data Driven Tree (Tree #1)	94.5%	98.2%	92.5%
Clinically Relevant Tree (CTree #1)	94.8%	97.0%	93.8%
Model #2 CNN	96.8%	-	91.3%
Data Driven Tree (Tree #2)	95.0 %	97.3%	86.3%

relevant knowledge extracted either directly from data or with human intervention, as described in Section III-D. This is expected to produce a more compact network with minimal information loss, with the option to systematically combine different knowledge, e.g. from a specialist and generalist.

Experiment #1 compares the distillation process of learning only the teacher's output (response-based learning) and learning both the output and the relevant kernels' activation maps (feature-based learning). Fig. 5(a) shows that response-based learning activates regions near to the cardiac silhouette and the hilars primarily via the first and fourth kernels at the bottleneck (arbitrarily labeled as kernel A & D). Introducing the additional loss term L_{Kernel} ensures that the activation maps are learned from kernels derived from Model #1. Both learning approaches achieved high training accuracy of 94.0% and fidelity of 99.3%. The response-based trained student however performs poorly on the validation set (87.5% accuracy) against the feature-based trained student (92.5% accuracy). Additionally, interpretable rules were extracted from the feature-based trained student (see Fig. 5(c)) which appear to deviate from the rules obtained from the teacher (Fig. 4(a)). It should be noted that no specific constraint was imposed on the loss function to seek to maintain the tree structure. Nonetheless, the kernels remained associated with the cardiac silhouette (QN) and upper mediastinum (ET), which is consistent with clinical research on the appearance of "silhouette sign" [35]. Additional experiments (Appendix C) with the replacement of kernel QN with kernel OD (cardiac silhouette) both extracted from Model #1 showed similar results. Furthermore, feature-based learning was extended to a multi-teacher configuration where each teacher contributed a subset of the four relevant kernels from CTree #1. The student also learned the teachers' responses with a similar preference for cardiac silhouette and upper mediastinum kernels (Table II). These results confirm the possibility of learning kernel activations with high fidelity, closing the neural-symbolic cycle with a teacher-student approach capable of combining knowledge from multiple teachers.

Experiment #2 and Experiment #3 evaluate feature-based learning from multiple teachers, each represented by a different model for detecting pleural effusion and evaluated against

TABLE II
STUDENT MODEL PERFORMANCE COMBINING KERNELS FROM MULTIPLE TEACHERS USING DIFFERENT TRAINING DATA; $x \rightsquigarrow y$ DENOTES REPLACING KERNEL x WITH y.

Model	Train Acc.	Val. Acc.
Student CNN (QN $\rightsquigarrow$ IB)	100.0%	93.8%
Student CNN (QN $\rightsquigarrow$ N)	100.0%	93.8%
Student CNN (DH added)	100.0%	90.0%

the ground truth. In Fig.6), the kernel QN (Cardiac Silhouette) was replaced with a similar Cardiac Silhouette kernel from Tree #2, namely IB or N, yielding similar learning profiles and prediction performance. In Fig.7, the inclusion of a new literal (kernel DH) maintains successful learning against the ground truth while expanding the student model to account for DH. Taken together, these experiments offer a proof-of-concept for effective pleural effusion detection through the use of a small number of clinically relevant kernels that can be described as a knowledge-base, manipulated, combined and re-evaluated using the proposed neural-symbolic cycle.

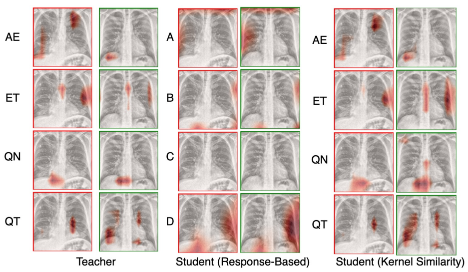
The extracted rules with kernel replacement (i.e. from kernel QN to kernels IB and N) continued to highlight the significance of the cardiac silhouette with a similar outcome as Experiment #1. The inclusion of kernel DH (trachea) also reaffirms the relative importance of cardiac silhouette (QN) over trachea (DH), as shown in the rule set.

V. Conclusion and Future Work

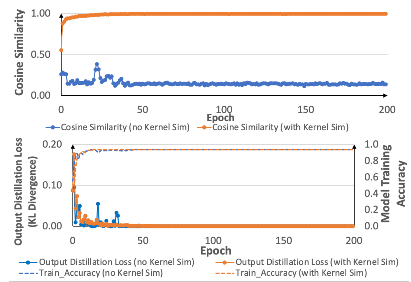
This paper confirmed that measuring the performance of CNN classification in radiology using accuracy alone is inadequate, as CNNs can achieve high accuracy for the wrong reasons. A proposed neural-symbolic cycle has been shown to offer a better control of the learning process as evaluated by domain experts. The paper also highlights the importance of using clinically-relevant kernels, allowing intervention to combine learning from data and possibly multiple knowledge-bases, so as to guide a CNN's model training. Through this approach, relevant knowledge is distilled into a more compact student model utilizing only very few kernels. The experimental results validated the proposed approach showing that high fidelity can be achieved with a small subset of clinically relevant kernels. This should offer the potential to tailor models for specific demographics, blend concepts coming from different sources into more robust models, or transfer models to use with new disease detection and evaluation such as COVID-19. As future work, we shall carry out further knowledge and data validation and optimization, integrate evolving knowledge from different modalities [36] to enhance system performance, trust and applicability in practice in the case of critical decision-making.

VI. Supplementary Material

Appendix with details on the anatomical region localization, concept association and additional experimental results can be accessible at https://bit.ly/IJCNN2024appendix.



(a) Teacher and Student Kernel Visualizations

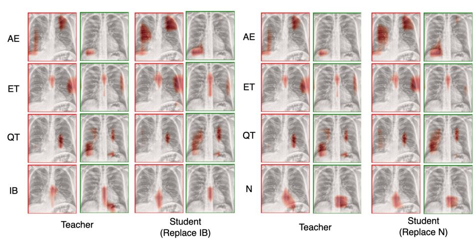


(b) Knowledge distillation performance

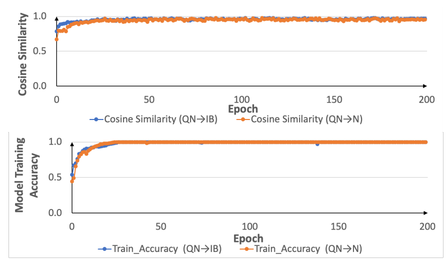
$\neg QN \wedge \neg ET \wedge \neg QT \implies Effusion$
$\neg QN \wedge \neg ET \wedge QT \implies Healthy$
$\neg QN \wedge ET \implies Healthy$
$QN \wedge \neg ET \implies Effusion$
$QN \wedge ET \implies Healthy$

(c) Extracted rules from student (Kernel Similarity)

Fig. 5. (a) Learning from 4 clinically relevant kernels (AE, ET, QN and QT) during knowledge distillation. Arbitrary labels (A,B,C,D) are given to the four kernels in the student model for response-based learning; γ is set to 10 when kernel similarity loss is included. (b) Loss curves: (top plot) kernel similarity is needed to replicate the relevant kernels with a high cosine similarity, (bottom plot) high model accuracy of 94.0% can be achieved. (c) Extracted rules from student model using kernel similarity maintains reliance on kernels QN (Cardiac Silhouette) and ET (Upper Mediastinum).



(a) Teacher and Student Kernel Visualizations



(b) Knowledge distillation performance

Replacing QN with IB

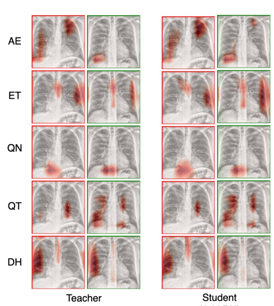
$\neg IB \wedge \neg AE \wedge \neg ET \implies Effusion$
$\neg IB \wedge \neg AE \wedge ET \implies Healthy$
$\neg IB \wedge AE \implies Healthy$
$IB \wedge \neg QT \implies Effusion$
$IB \wedge QT \wedge \neg AE \implies Effusion$
$IB \wedge QT \wedge AE \implies Healthy$

Replacing QN with N

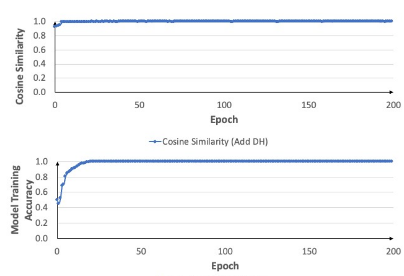
$\neg N \implies Healthy$
$N \wedge \neg QT \implies Effusion$
$N \wedge QT \wedge \neg ET \implies Effusion$
$N \wedge QT \wedge ET \implies Healthy$

(c) Student rules with kernel replacement (IB & N)

Fig. 6. (a) Kernel learning when kernel QN (CTree #1) is substituted with kernel IB or N (both representing the Cardiac Silhouette region) from another teacher model (Tree #2). Good replication of kernels can be seen during training. (b) The learning profile achieves similar performance results given that the substitutions are very similar. (c) Both sets of extracted rules also show strong dependence on kernels for the Cardiac Silhouette (kernel IB and N respectively).



(a) Teacher and Student Kernel Visualizations



(b) Knowledge distillation performance

$\neg QN \wedge \neg QT \wedge \neg ET \implies Effusion$
$\neg QN \wedge \neg QT \wedge ET \implies Healthy$
$\neg QN \wedge QT \wedge \neg ET \wedge \neg AE \implies Effusion$
$\neg QN \wedge QT \wedge \neg ET \wedge AE \implies Healthy$
$\neg QN \wedge QT \wedge ET \implies Healthy$
$QN \wedge \neg ET \wedge \neg DH \implies Effusion$
$QN \wedge \neg ET \wedge DH \wedge \neg QT \implies Effusion$
$QN \wedge \neg ET \wedge DH \wedge QT \implies Healthy$
$QN \wedge ET \wedge \neg AE \wedge \neg DH \implies Healthy$
$QN \wedge ET \wedge \neg AE \wedge DH \implies Effusion$
$QN \wedge ET \wedge AE \wedge \neg QT \implies Effusion$
$QN \wedge ET \wedge AE \wedge QT \implies Healthy$

(c) Student rules with additional kernel DH

Fig. 7. (a) Kernel learning when kernel DH representing the trachea obtained from a separate teacher model (Tree #2) is added to the set of kernels from the clinically-validated tree (CTree #1). Good replication of kernels can be seen during training. (b) The learning profile with the addition of DH shows successful learning of the kernels w.r.t. the corresponding output against the ground-truth. (c) The rule set extracted from the student model shows that kernels for Cardiac Silhouette (QN) and Upper Mediastinum (ET) are fundamental, with the kernel for the trachea (DH) being used in only 5 rules.

References

[1] S. Yu *et al.*, "Understanding convolutional neural networks with information theory: An initial exploration," *IEEE Trans Neural Netw Learn Syst*, vol. 32 (1), pp. 435–442, 2021.

[2] A. D. Garcez *et al.*, *Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning*. CEUR Workshop Proceedings, 2023, vol. 3432.

[3] ——, *Neural-Symbolic Learning Systems: Foundations and Applications*, ser. Perspectives in neural computing. Springer, 2002.

[4] A. D. Garcez and L. C. Lamb, "Neurosymbolic AI: the 3rd wave," *Artificial Intelligence Review*, 2023.

[5] F. A. Mettler *et al.*, "Radiologic and nuclear medicine studies in the united states and worldwide: Frequency, radiation dose, and comparison with other radiation sources—1950–2007," *Radiology*, vol. 253 (2), pp. 520–531, 2009.

[6] S. Raoof *et al.*, "Interpretation of plain chest roentgenogram," *Chest*, vol. 141 (2), pp. 545–558, 2012.

[7] S. Andronikou *et al.*, "CT scanning for the detection of tuberculous mediastinal and hilar lymphadenopathy in children," *Pediatr. Radiol.*, vol. 34 (3), pp. 232–236, 2004.

[8] A. Fritscher-Ravens *et al.*, "Mediastinal lymphadenopathy in patients with or without previous malignancy: EUS-FNA-based differential cytodiagnosis in 153 patients," *Am. J. Gastroenterol.*, vol. 95 (9), pp. 2278–2284, 2000.

[9] C. R. Whitten *et al.*, "A diagnostic approach to mediastinal abnormalities," *Radiographics*, vol. 27 (3), pp. 657–671, 2007.

[10] J. Townsend *et al.*, "ERIC: Extracting relations inferred from convolutions," in *Computer Vision – ACCV 2020*. Springer, 2021, pp. 206–222.

[11] van Griethuysen *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer Res.*, vol. 77 (21), pp. e104–e107, 2017.

[12] J. Irvin *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *arXiv*, 2019.

[13] A. J. DeGrave *et al.*, "AI for radiographic COVID-19 detection selects shortcuts over signal," *medRxiv*, 2020.

[14] E. Tartaglione *et al.*, "Unveiling COVID-19 from CHEST X-Ray with deep learning: A hurdles race with small data," *Int. J. Env. Res. Public Health*, vol. 17 (18), 2020.

[15] B. Zhou *et al.*, "Learning deep features for discriminative localization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.

[16] R. R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep networks via Gradient-Based localization," *Int. J. Comput. Vis.*, vol. 128 (2), pp. 336–359, 2020.

[17] J. T. Springenberg *et al.*, "Striving for simplicity: The all convolutional net," *arXiv*, 2014.

[18] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *arXiv*, 2013.

[19] B. Zhou *et al.*, "Interpreting deep visual representations via network dissection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41 (9), pp. 2131–2145, 2019.

[20] ——, "Comparing the interpretability of deep networks via network dissection," pp. 243–252, 2019.

[21] 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.

[22] K. H. Ngan *et al.*, "Extracting meaningful High-Fidelity knowledge from convolutional neural networks," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–17.

[23] ——, "Closing the Neural-Symbolic cycle: Knowledge extraction, user intervention and distillation from convolutional neural networks," in *NeSy*, 2023.

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

[25] Q. Zhang *et al.*, "Interpreting CNN knowledge via an explanatory graph," *AAAI*, vol. 32 (1), 2018.

[26] ——, "Interpreting CNNs via decision trees," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 6261–6270.

[27] G. Hinton *et al.*, "Distilling the knowledge in a neural network," *arXiv*, 2015.

[28] R. Adriana *et al.*, "Fitnets: Hints for thin deep nets," *Proc. ICLR*, vol. 2, p. 3, 2015.

[29] S. Park and N. Kwak, "Feature-level ensemble knowledge distillation for aggregating knowledge from multiple networks," in *ECAI 2020*. Amsterdam, NY: IOS Press, 2020, pp. 1411–1418.

[30] M. Ji *et al.*, "Show, attend and distill: Knowledge distillation via attention-based feature matching," *AAAI*, vol. 35 (9), pp. 7945–7952, 2021.

[31] G. Jocher *et al.*, "YOLOv5 SOTA realtime instance segmentation," 2022.

[32] X. Wang and othersM, "ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2097–2106.

[33] T. Kasioumis *et al.*, "Elite BackProp: Training sparse interpretable neurons," in *NeSy*, 2021, pp. 82–93.

[34] R. Khajotia, "Respiratory clinics: Mediastinal shift: A sign of significant clinical and radiological importance in diagnosis of malignant pleural effusion." *Malays Fam Physician*, vol. 7 (1), pp. 34–36, 2012.

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

[36] T. Tu *et al.*, "Towards generalist biomedical AI," *arXiv*, 2023.

[37] A. Karargyris *et al.*, "Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for AI development," *Sci Data*, vol. 8 (1), p. 92, 2021.

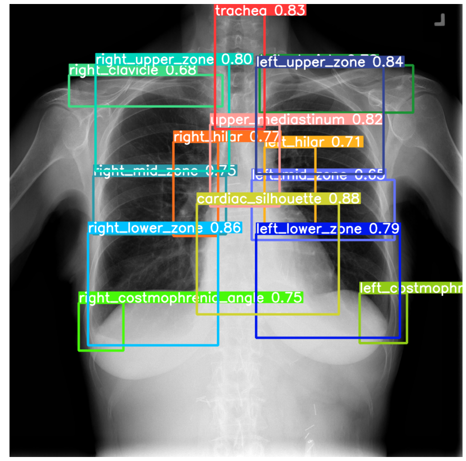
APPENDIX

SUPPORTING INFORMATION

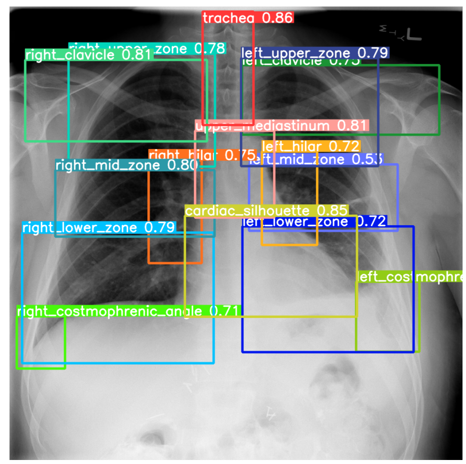
A. Anatomical Region Localization

An object localization model, based on the YOLOv5x architecture [31], was independently trained using X-ray images from the NIH dataset [32] to detect nine specific anatomical regions in individual frontal chest X-rays - namely, Trachea (T), Upper Mediastinum (UM), Cardiac Silhouette (CS), Left Hilar (LH), Right Hilar (RH), Left Clavicle (LC), Right Clavicle (RC), Left Costophrenic Angle (LCA), and Right Costophrenic Angle (RCA). Annotations for anatomical regions were referenced to [37]. The identified anatomical regions were superimposed on the activated image regions for each convolutional neural network (CNN) kernel from each of the trained classification models (i.e. 512 kernels at the last convolutional layer of the VGG16 model in this work) to assess the region of intersection. For each image, a hit was recorded if the intersection over union (IoU) score exceeded 0.5. The most representative anatomical region for each individual kernel was identified based on the highest aggregated hit rate across all images in the training dataset, with a minimum empirical defined hit threshold over 60% of the dataset. To target the kernel to specific anatomical region, anatomical regions most activated by kernel responses were limited to a maximum of two regions. This anatomical association offered a more comprehensible representation of clinical concepts than the kernel fingerprints used in prior work [22]. Consequently, 184 interpretable kernels were identified for Model #1, and 192 from Model #2., while the rest were considered non-specific to any anatomical region.

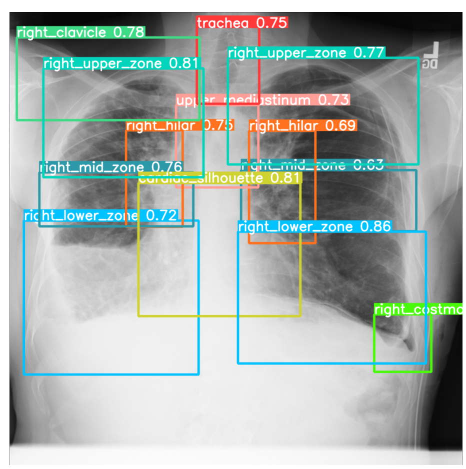
The utilization of representative anatomical regions as guiding points in the kernel norm plots facilitated the interpretation of semantic concepts expressed by the kernels with greater clarity. This also allowed filtering any uninterpretable kernel for future investigation. The use of only frontal X-ray images ensured the consistency of anatomical region positioning, facilitating the inspection and manual correction of any mislocated regions. Notably, the hilar regions and costophrenic angles posed the greatest challenges due to their similarity between the left and right regions. The object localization model also provided the correct labels for the located regions, although it showed weaker performance on the hilar and costophrenic regions. Representative X-ray images with annotated bounding boxes of the located anatomical regions were presented in Fig. 8, including *no finding* and *pleural effusion* cases of varying severity. This trained model enabled the annotation of representative regions for the kernels for each CNN classification models, thus enabling further evaluation analysis by correlating with radiomics features.



(a) No finding



(b) Pleural effusion



(c) Pleural effusion

Fig. 8. Samples of anatomical region localization using YOLOv5x model for plain chest X-rays of patients with different severity in pleural effusion. Any missing/wrongly labeled regions (e.g. costophrenic angles, duplicated right lung labelling) have been manually amended prior to the radiomics analysis.

B. Concept Association through Radiomics feature on anatomical regions

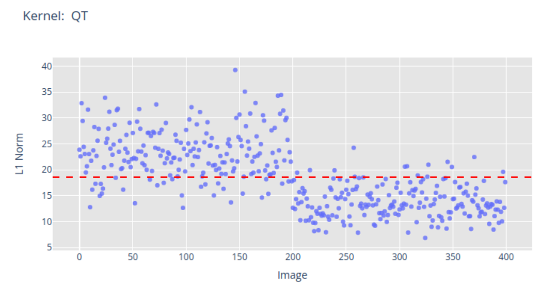
In this section, additional representative examples were presented to explore the association of concepts with radiomics features specific to the represented anatomical regions. For instance in the left hilar region (see Fig. 9, there was a positive correlation between L1-Norm values for kernel QT and the Run Length Gray Level Non-Uniformity (GLRLM) feature. This positive correlation corresponds well with the visual observation that the left hilar region becomes opaque (i.e. low L1-Norm values and low GLRLM value) as the presence of pleural effusion and vice versa.

Similarly for the right hilar region (see Fig. 10, kernel AE exhibited a positive correlation with the associated radiomics feature, Run Length Gray Level Non-Uniformity (GLRLM). In particular, cases labeled as *no finding* showed high L1-Norm values and high Gray Level Non-Uniformity, while pleural effusion cases showed the opposite values. By observing Fig 10 (c), the change in L1-Norm values (refer to Fig 10 (a)) also corresponded to the visual observation of the right hilar region becoming more opaque in the presence of pleural effusion.

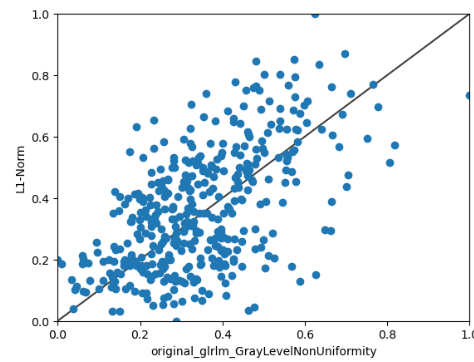
Fig. 11 presents the concept association between the L1-Norm values for kernel DH (Model #2) for the Trachea region and the corresponding association with the best fitted radiomics feature - First Order Entropy (FOE) which specifies the uncertainty/randomness of the pixel intensity values. High FOE implies more randomness within the region and more homogeneous otherwise. From the figure, cases of pleural effusion (in red border at Fig. 11(c)) will tend to have whiteness at the Trachea region and a more clear image of the Trachea if the image is labeled as *no finding* (in green).

Fig. 12 and Fig. 13 present the concept association between the L1-Norm values for both kernel IB and N (Model #2) on the Cardiac Silhouette region and the corresponding association with the best fitted radiomics feature - Dependence Matrix Gray Level Non-Uniformity (GLDM) which measures the similarity of gray level pixel intensity values related to the central pixel of the image. A lower value indicates greater similarity (i.e. more homogeneous) and higher difference/randomness otherwise. From figures, both kernels show very similar L1-Norm values to radiomics feature correlation (see Fig. 12(b) and Fig. 13(b)). The correlation implies that high L1-values (cases of pleural effusion) shows clear indication of "silhouette sign" where the border at the left ventricle is obscured while images labeled as *no finding* will have low L1-Norm values and corresponding clear left ventricle border.

These examples demonstrated that the L1-Norm values displayed in the kernel norm plot can effectively approximate the changes in visual changes in a specific anatomical region and effectively simulate how a clinician would examine an X-ray image. The correlation between L1-Norm values and radiomics features provides valuable insights into the radiological features of representing anatomical regions extracted by the CNN models, contributing to better interpretability and clinical relevance of the features (i.e. symbols) used for making the prediction.



(a) Kernel norm plot

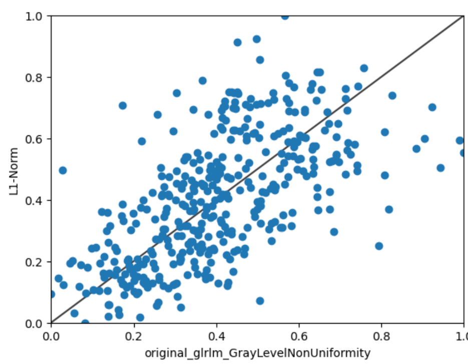


(b) Kernel radiomics correlation

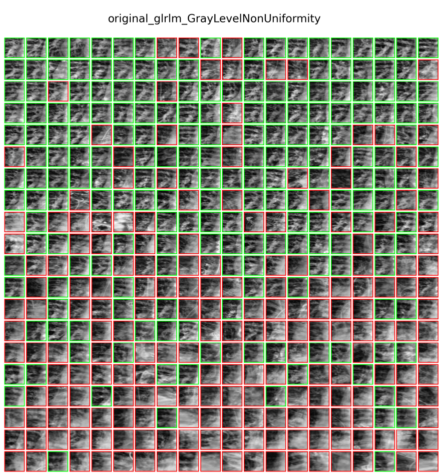
(c) Gray Level Non-Uniformity for left hilar

Fig. 9. (a) A kernel norm plot for kernel QT (Model #1) generated from a trained CNN's convolutional kernel output (indexed as 'QT') which represent the left hilar. The first 200 data points from the training dataset are labeled as *no finding* 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 *no finding* as ground truth label are outlined in green while those with *pleural effusion* are outlined in red.

(a) Kernel norm plot

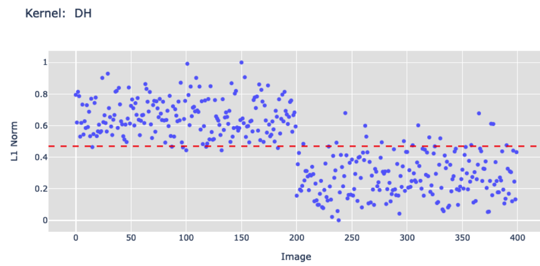


(b) Kernel radiomics correlation

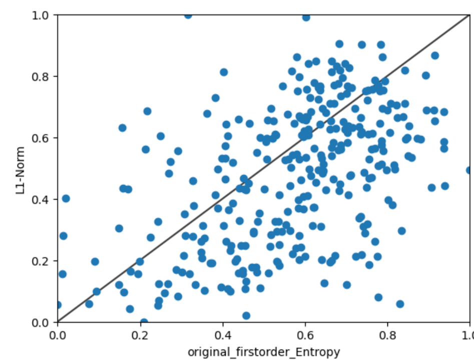


(c) Gray Level Non-Uniformity for right hilar

Fig. 10. (a) A kernel norm plot for kernel AE (Model #1) generated from a trained CNN's convolutional kernel output (indexed as 'AE') which represent the right hilar. The first 200 data points from the training dataset are labeled as *no finding* 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 (¬AE). (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 *no finding* as ground truth label are outlined in green while those with *pleural effusion* are outlined in red.



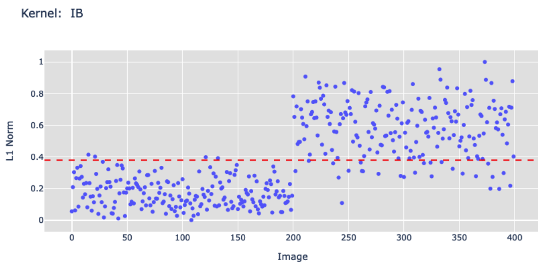
(a) Kernel norm plot



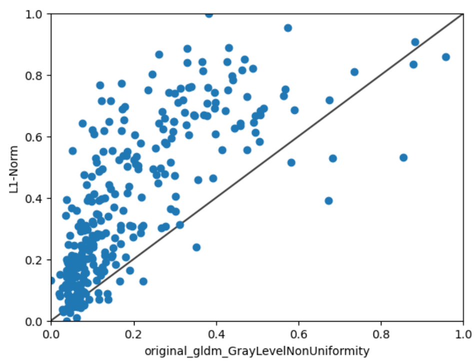
(b) Kernel radiomics correlation

(c) First order entropy for trachea

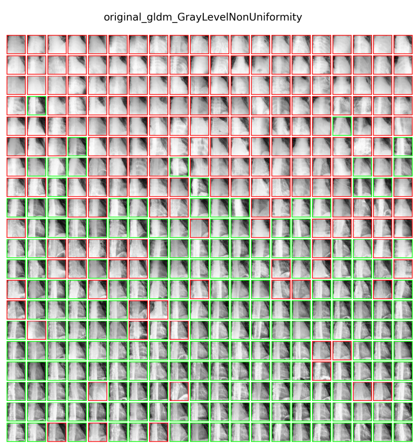
Fig. 11. (a) A kernel norm plot for kernel DH (Tree #2) generated from a trained CNN's convolutional kernel output (indexed as 'DH') which represent the Trachea. The first 200 data points from the training dataset are labeled as *no finding* and the next 200 as *pleural effusion* according to the ground truth. A threshold value (red line) separates positive literals (e.g. DH) (above the line) and negative literals (¬DH). (b) A positive correlation is found between the First Order Entropy (FOE) with L1-Norms for kernel DH. Sub-figure (c) shows images of the Trachea region sorted row-wise from highest entropy (top left) to lowest entropy (bottom right). Those images with *no finding* as ground truth label are outlined in green while those with *pleural effusion* are outlined in red.



(a) Kernel Norm Plot

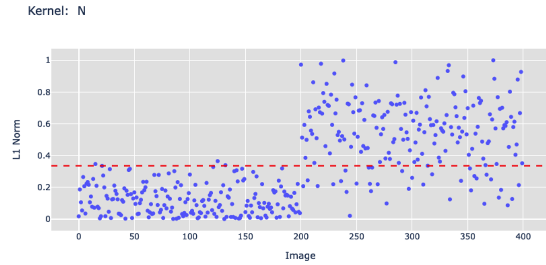


(b) Kernel Radiomics Correlation

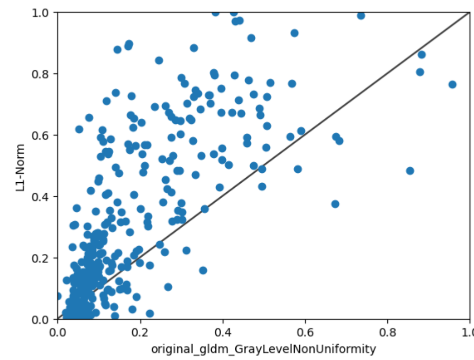


(c) Gray Level Non-Uniformity for Cardiac Silhouette

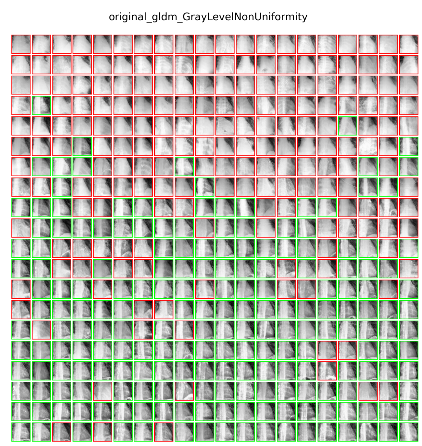
Fig. 12. (a) A kernel norm plot for kernel IB (Tree #2) generated from a trained CNN's convolutional kernel output (indexed as 'IB') which represent the Cardiac Silhouette. The first 200 data points from the training dataset are labeled as *no finding* and the next 200 as *pleural effusion* according to the ground truth. A threshold value (red line) separates positive literals (e.g. IB) (above the line) and negative literals (¬IB). (b) A positive correlation is found between Dependent Matrix Gray Level Non-Uniformity (GLDM) with L1-Norms for kernel IB. Sub-figure (c) shows images of the cardiac silhouette region sorted row-wise from highest gray level non-uniformity (top left) to lowest gray level non-uniformity (bottom right). Those images with *no finding* 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



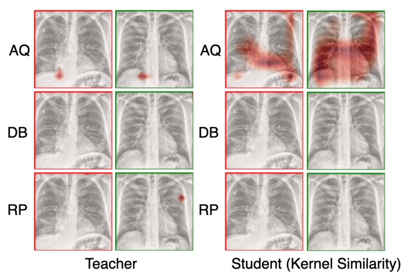
(c) Gray Level Non-Uniformity for cardiac silhouette

Fig. 13. (a) A kernel norm plot for kernel N (Tree #2) generated from a trained CNN's convolutional kernel output (indexed as 'N') which represent the Cardiac Silhouette. The first 200 data points from the training dataset are labeled as *no finding* and the next 200 as *pleural effusion* according to the ground truth. A threshold value (red line) separates positive literals (e.g. N) (above the line) and negative literals (¬N). (b) A positive correlation is found between Dependent Matrix Gray Level Non-Uniformity (GLDM) with L1-Norms for kernel N. Sub-figure (c) shows images of the cardiac silhouette region sorted row-wise from highest gray level non-uniformity (top left) to lowest gray level non-uniformity (bottom right). This sub-figure is the same as Fig. 12(c) as both kernels are trained from the same model and images. Those images with *no finding* as ground truth label are outlined in green while those with *pleural effusion* are outlined in red.

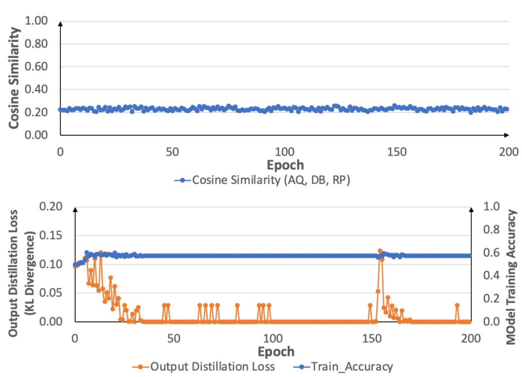
C. Supplementary Experimental Finding

This section provides further supporting experimental results to those that had been presented in the main paper. It aims to substantiate that relevant knowledge can be extracted and effectively distilled into a student model.

In the first supporting experiment (see Fig. 14), three clinically irrelevant/uninterpretable kernels were identified, namely AQ, DB and RP. During knowledge distillation, it is observed that it takes more epochs in order for the training to learn the response of the teacher (i.e. minimising the output distillation loss, L_{output}). It is noted that the distillation met with challenges learning the kernels as the activated regions are either random or non-existed (reflected by the low cosine similarity). In addition given that the performance of the teacher itself is poor (i.e. accuracy of 58.8%), it is also reflected from the model accuracy of the student model to achieve only 57.5% despite a high fidelity of 95.8% against the teacher upon training completion. This outcome reflects clearly on the poor learning outcome of using irrelevant kernels as knowledge.



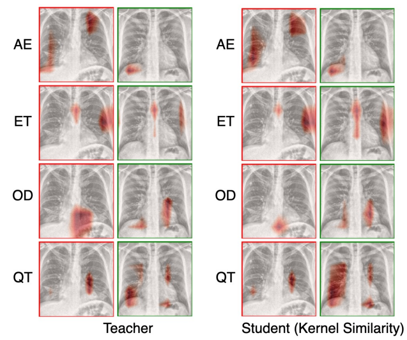
(a) Kernel learning during training



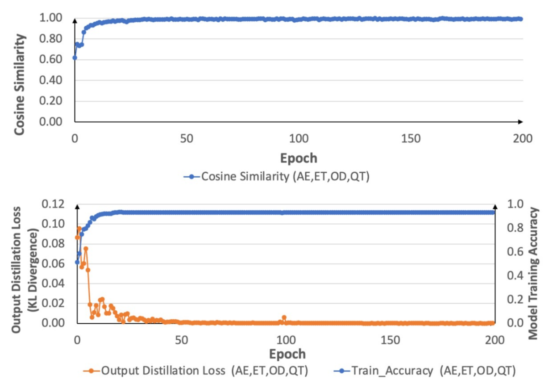
(b) Knowledge distillation performance

Fig. 14. (a) Kernel Learning for the 3 irrelevant kernels (AQ, DB, RP) from Model #1 during knowledge distillation. It can be seen that poor learning was achieved to learn the random activated patches. (b) Cosine Similarity for learning the kernels remained low at approximately 0.2 despite that the output distillation loss has plateaued to near zero. It is also shown that the learned knowledge were irrelevant against ground truth based on the low model accuracy.

The next experiment extends the investigation of feature-based learning (shown in the main paper) with an alternative kernel related to the Cardiac Silhouette region (i.e. kernel QN to kernel OD) derived from a tree of Model #1. Similar to Fig. 5, the proposed feature-learning approach (see Fig. 15 continued to successfully replicate the relevant kernels (i.e. high cosine similarity). The student model achieves a high training accuracy of 93.3% and fidelity of 98.8%. This model also yielded good generalisability with a validation accuracy of 93.8%. The extracted rules from this student model continued to exhibit dominance of the Cardiac Silhouette and Upper Mediastinum regions represented by kernel OD and ET respectively.



(a) Kernel learning during training



(b) Knowledge distillation performance

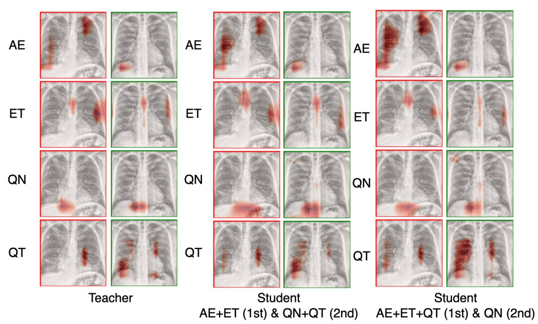
$$\neg OD \land \neg ET \land \neg AE \implies Effusion$$
$$\neg OD \land \neg ET \land AE \implies Healthy$$
$$\neg OD \land ET \implies Healthy$$
$$OD \land \neg AE \implies Effusion$$
$$OD \land AE \land \neg QT \implies Effusion$$
$$OD \land AE \land QT \implies Healthy$$

(c) Extracted rules from student (replaced with kernel OD)

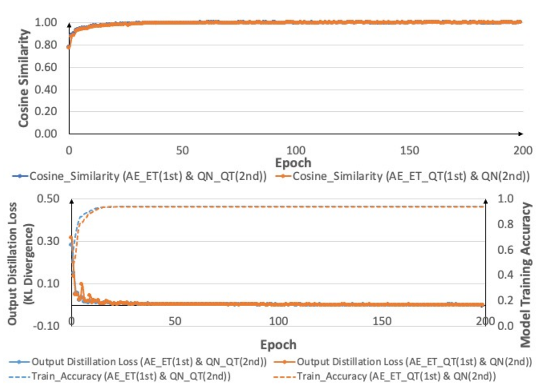
Fig. 15. (a) Kernel Learning for the 4 clinical relevant kernels (AE, ET, OD and QT) during knowledge distillation. It continued to show good kernel replications across the selected relevant kernels (b) The loss minimisation plot shows good Cosine Similarity indicating successful kernel replication. In addition, good model accuracy of 93.3% is achieved once output distillation loss is minimised to near zero. (c) The extracted tree rules from the student models exhibits dominance in the Cardiac Silhouette and Upper Mediastinum region represented by kernel OD and ET.

The last supplementary experiment investigates feature-based learning with a multi-teachers configuration. In this experiment, two teacher models were employed with the same teacher from CTree #1 and Model #1. Each of these teacher models taught a subset of relevant kernels as specified in Fig. 16. The student also learned the output from CTree

#1. In the representative examples (see Fig. 16), successful learning can be demonstrated using a mix of kernels in this multi-teachers configuration into student models with similar learning profiles (i.e. the gradual change in cosine similarity and training accuracy over the training epochs). The activation maps for the learned kernels were also successfully replicated. In addition for both cases, the student models yields good training accuracy of 94.0% against the ground truth and fidelity of 99.3% against the teacher tree (CTree #1). Good generalisation with validation accuracy of above 90.0% was achieved. In both cases, the extracted rules from the student models continued to show dominance in the Cardiac Silhouette and Upper Mediastinum regions represented by kernel QN and ET respectively.



(a) Kernel Learning during Training



(b) Knowledge Distillation Performance

$\neg QN \wedge \neg ET \wedge \neg AE \implies Effusion$
$\neg QN \wedge \neg ET \wedge AE \implies Healthy$
$\neg QN \wedge ET \implies Healthy$
$QN \wedge \neg ET \implies Effusion$
$QN \wedge ET \wedge \neg AE \implies Healthy$
$QN \wedge ET \wedge AE \implies Effusion$

(c) Extracted rules from student (AE_ET(1st) & QN_QT(2nd))

$\neg QN \wedge \neg ET \wedge \neg AE \implies Effusion$
$\neg QN \wedge \neg ET \wedge AE \implies Healthy$
$\neg QN \wedge ET \implies Healthy$
$QN \wedge \neg ET \implies Effusion$
$QN \wedge ET \wedge \neg AE \implies Healthy$
$QN \wedge ET \wedge AE \implies Effusion$

(d) Extracted rules from student (AE_ET_QT(1st) & QN(2nd))

Fig. 16. (a) Kernel Learning for the 4 clinical relevant kernels (AE, ET, QN and QT) during knowledge distillation in a multi-teacher configuration. Each teacher contributed a subset of relevant kernels as shown while the student learned the output of the teacher tree (CTree #1). Good kernel replication can be observed. (b) In the loss minimisation plots, the successful kernel replication is shown by the high cosine similarity. In addition, both cases show similar learning profile with model accuracy achieving 94.0% upon training completion. (c) The extracted rules from both cases show dominance of kernels from the Cardiac Silhouette and Upper Mediastinum region (Kernel QN and ET) respectively.

TABLE III
STUDENT MODEL PERFORMANCE ON SPLITTING RELEVANT KERNELS FOR LEARNING FROM TWO TEACHERS COMPOSED OF THE SAME MODEL. FOR MIX #1, THE 1ST TEACHER CONTRIBUTES KERNELS AE AND ET WHILE THE 2ND TEACHER CONTRIBUTES KERNELS QN AND QT. FOR MIX #2, THE 1ST TEACHER CONTRIBUTES KERNELS ET, QN AND QT WHILE THE 2ND TEACHER CONTRIBUTES ONLY KERNEL AE.

Model	Train Acc.	Train Fidelity	Val. Acc.
Student CNN (Mix #1)	94.0%	99.3%	90.0%
Student CNN (Mix #2)	94.0%	99.3%	92.5%