International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Ensemble Image Explainable AI (XAI) Algorithm for Severe Community Acquired Pneumonia and COVID-19 Respiratory Infections

Navaneeth S[1], DR.Dilip R[2]
Department of Electronics and Communication Engineering,DSATM, India

ABSTRACT

Since the 2019 COVID-19 pandemic, clinical prediction score methods have been developed to help physicians assess the severity of pneumonia. However, there are insufficient studies focused on defining the most appropriate decision-making strategies used by physicians.

Having done so. This article introduces XAI, a new image interpretation technique based on SHAP and Grad-CAM++. This video explains how a deep learning prediction model predicts mortality risk for individuals with community-acquired pneumonia and COVID-19 respiratory infections. We conducted a literature review, analyzed quantitative and qualitative parameters, evaluated the effectiveness of Ensemble XAI and compared it with other methods.Since the 2019 COVID-19 pandemic, clinical prediction score methods have been developed to help physicians assess the severity of pneumonia. However, there are not enough studies focused on defining the most appropriate decision-making strategies used by physicians.

Having done so. This article introduces XAI, a new image interpretation technique based on SHAP and Grad-CAM++. This video explains how a deep learning prediction model predicts mortality risk for individuals with community-acquired pneumonia and COVID-19 respiratory infections. We conducted a literature review, analyzed quantitative and qualitative parameters, evaluated the effectiveness of Ensemble XAI and compared it with other methods.

INTRODUCTION

As of May 17, 163.71 million people have been infected. 3 393 551 registered deaths are linked to the virus worldwide. Critically ill patients should be clinically prioritized, as evidenced by the lack of ethical concerns.

Singapore has a population of 5.6 million.

Wonderful hospital treatment, like many others Countries affected by COVID-19 Since then, healthcare systems have gone the extra mile to combat epidemics, to accelerate the development of AI healthcare solutions. Many research projects have been carried out worldwide.

The literature provides evidence of the importance of deep learning.

Rapid diagnosis of COVID-19 is achieved through algorithms using medical image databases. [1]–[7] and respectively. Good classification performance using deep learning algorithms for chest X-ray and

computed tomography (CT) images was reported in several studiesIn February 2020, Changi General Hospital, Singapore, and national health technology company Integrated Health Information Systems (IHiS), collaborated to develop AI.

The prediction model is called the Community Acquired Pneumonia and COVID-19 AI Prediction Engine (CAPE) [8] which can generate risk scores for pneumonia patients. The cape is the team consisted of senior physicians with expertise in radiology and respiratory medicine, data scientists, healthcare journalism

researchers and systems engineers. They set out to be simpleand a customizable application that can put AI inside the chest Imaging workflow.

One of the main obstacles to the application of AI in Explainability was the focal point of a clinical process [9]. It has been demonstrated that neural networks are more accurate as compared to traditional machine learning techniques like support vector machines in several image applications.

The former, however, is far harder to explain. It is challenging for physicians to feel at ease using AI and to have faith in the algorithms in the absence of a detailed explanation of how the algorithms arrived at their predictions [10], [11]. The choice of an algorithm, such as the widely used layer-wise relevance propagation [12] and localization, gradients, and perturbations [13], [14], has been the subject of much research. The effectiveness of interpretation methods for deep learning image networks has not been extensively studied in the literature [15]–[18]

The medical industry does not use consistent or standardised criteria for interpretability assessments, as noted by Tjoa and Guan in [21]. This might lead to a bias in the choosing of one approach over another without any basis in medical procedures. Also, it is discovered that there is not much

research utilising human subjects that assess the reliability of medical imaging interpretation methods

The following are this article's primary contributions

1) We suggested combining the SHAP and Grad-CAM++ approaches to create an enhanced mapping layer that identifies discriminative areas, based on ensemble techniques employed in machine learning. This group is known as XAI.

2) In order to compare different picture explainability methods both numerically and qualitatively, we created a visual explainability assessment checklist.

METHOD

A. Data and Modeling

The model was developed based on data from an isolated acute tertiary hospital. The SingHealth Centralized Institutional Review Board (CIRB 2020/2100) granted ethical approval and obtained permission to use data for the study.

1) Prognostic model development: We built our model based on a retrospective study of adult patients admitted between January 1 and December 31, 2019 at a tertiary acute hospital in Singapore.

The study included individuals admitted to the emergency department with a diagnosis of pneumonia (using ICD-10 codes).

This study included 2235 chest X-ray images from 1966 adult individuals. Patient EMR data were used to develop scores for inpatient mortality.

Data were removed prior to processing.

The data were divided into three phases: training, validation, and testing. Patients admitted between January 1 and October 31, 2019 were divided into training and validation groups.

2) Translational Assessment Data Set: 1475 Adult Patients Prospectively Evaluated Our Methods Of Interpreting The XAI For Those Who A Between January 1, 2020, And June 30, 2020, The Emergency Department Used A Physician-Directed Assessment Of CAPE. When Tested In A Real Clinical Setting, The Predictive Accuracy Of The Model Performance Was 0.811, And The AUC Was 0.803. 76 Really Good Examples Identified By CAPE, An Established Interpretation And Analysis Dataset Was Developed. Figure 1 Shows A Possible Layout For The Semantic Research Data Set.

B. Interpretation Approaches

This article focuses on five state-of-the-art semantic approaches: salience, Grad-CAM, Grad-CAM++, SHAP, and LIME. In addition, this section introduces ensemble XAI.

1) Grad-CAM : Grad-CAM is now a well developed technique. By revealing input areas with finer details that are critical for prediction, this increased the clarity of CNN-based models [23 final feature map visualization Since the final transition level can be thought of as part of the classification model, Ak refers to areas of discrimination in the image Grad-CAM recommends using average gradient scores as weights for feature map the, which is,

$$\partial_k = \frac{1}{uv} \sum_{i=1}^{u} \sum_{j=1}^{v} \frac{dy}{dA_{i,j}^{k}}$$

Is given by , where the kth element of height u and width v from the last transition point is represented by Ak ∈ Ru∗v . However, this method may have some shortcomings, such as not being detected if it occurs more than once in an image or some features reaching the entire data set due to the exclusion of only partially produced features.

2) Grad-CAM++: Grad-CAM++ is a Grad-CAM correction method for Grad-CAM boundaries. This method creates a feature map that helps CNN make all its decisions by calculating the relevance of each pixel.

Furthermore, the overall location capture of the embedded image has been well demonstrated that the entire object moves to any location where there are multiple instances of the same object [24].

3) SHAP: SHAP calculates each feature's contribution to the prediction in order to explain the prediction of instance x.

The SHAP gradient explainer is a variation on the integrated gradients method, a feature attribution technique created for differentiable models using Aumann-Shapley values, an extension of Shapley values to infinite player games [25], [26].The anticipated gradients compute approximate SHAP values if we assume that the input characteristics are independent and we approximate the model using a linear function between each background data sample and the current input to be explained. In order to operate, the gradient explainer integrates the gradients of each interpolation made between the background sample—which is the sample being compared to—and the foreground sample, which is the sample being explained.

4) Local interpretation model-agnostic interpretation (LIME): Locally complex models are estimated using LIME [27].A reasonable model that can provide an explanation for the prediction of a particular case of interest. The following is a summary of the LIME program.

A. Be sure to explain the context logically. Superpixels, or the continuum of the relevant pixels, are used to describe images so that their interpretable representations are two vectors, where 0 denotes the bleached superpixel and 1 denotes the original superpixel
B. Constrain an interpretable position to obtain a model. The two vectors of the sample image contain a zero, representing bleached superpixels, as opposed to all of them in the binary vector for the real image
C. Apply the original image to the disturbed images and make a prediction.
D. Fit interpretable models to phase iii predictions and proximity-weighted sampled images.
E. Use a definable example to illustrate the significance of each definable object.Due to the complexity of the aforementioned method, the LIME calculation takes a very long time. Furthermore, the rendering by this simple superpixel-based rendering method is unstable when there are small amounts of noise in the input [29]

5) Saliency Map: First proposed in 2014 [30], saliency map is a basic simple semantic technique. High gradient numbers are predicted to highlight input regions that cause large changes in the output, because the gradient of the output with respect to the input image shows how the output value varies with slight changes in the input with the resulting pixels contributing significantly to the results They are the displayed. However, due to the absolute values of partial derivatives, this method cannot distinguish between positive and negative evidence [31] .

6) Ensemble XAI: Since ensemble methods can reduce bias and variance and increase reliability, they are widely used in deep learning [32], [33 stacking-based ensemble methods are widely used in medicine in painting, and in many in-depth studies Having shown promising results experimentally [8], [34 we propose a stacking-based ensemble method for image rendering, which uses the results of the basic rendering method. It is interesting to know whether the simultaneous use of Grad-CAM++ and SHAP gradient explainer methods offers complementary advantages since they are both gradient-based algorithms with different approaches and related advantages We noted it is recommended that a team approach be adopted to verify this assumption: Kernel Ridge on generalized positive SHAP and generalized Grad-CAM++ use regression.

C. Interpretation Evaluation Metrics

1) Decision impact coefficient: Percentage change in alternatives from the critical point determined by the semantic approach.D(x) represents a deep learning decision function that, given image x as input, yields a classification decision. Let 1logic be a pointer function that returns one in case the argument is true. The following assumptions can be used to calculate decision impact coefficients.

$$\text{Decision impact ratio} = \sum_{i}^{N} \frac{1_{\mathrm{D}(x_i)\neq \mathrm{D}(x_i - c_i)}}{N} \qquad (2)$$

where xi denotes the ith original image, and ci denotes the critical area identified by the deep learning model for the ith image.

Confidence Impact Ratio:

The proportion of confidence lost as a result of leaving out the crucial region that the interpretation technique found.

Given a picture x as input, let C(x) be the deep learning confidence function that yields the classification confidence probability. The confidence impact ratio m

ay be computed using the following formula:

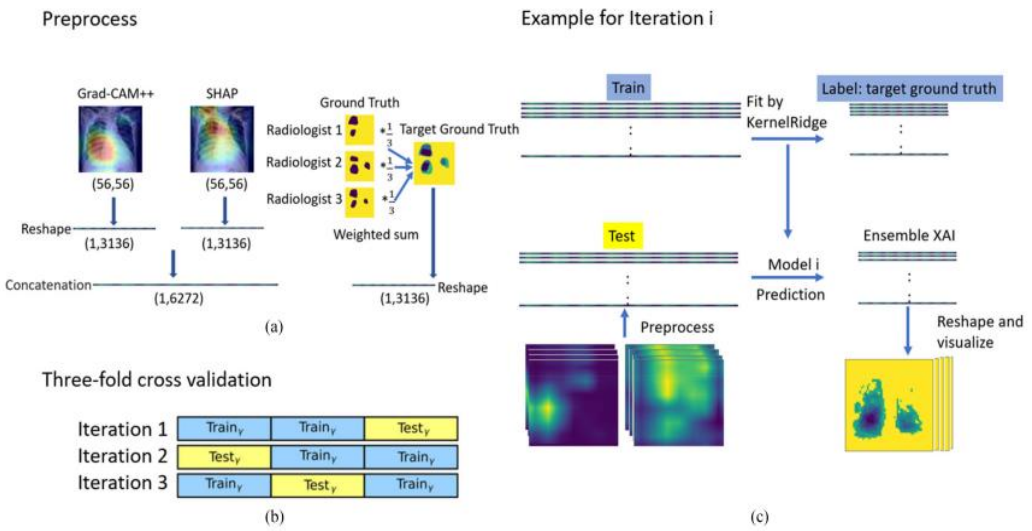

Fig. 2. Advanced ensemble XAI. (a) Preprocessing of Grad-CAM++, SHAP, and ground truth for each image. (b) Three-fold cross validation is applied to generate ensemble XAI. (c) Workflow for iteration *i*.

$$\text{Confidence impact ratio} = \sum_{i}^{N} \frac{\max\left(C\left(x_i\right) - C\left(x_i - c_i\right), 0\right)}{N} \tag{3}$$

where xi denotes the ith original image, and ci denotes the critical area identified by deep learning model for the ith image.

The core region identified by deep learning was compared with the area annotated by experienced physicians with residual consistency and accuracy analyzes for each image

1) Concordance recall: the proportion of the total annotated field that was correctly recognized by the translation method.

2) Concordance accuracy: The percentage of the total area of interest presented by the interpretation method that is identified with accuracy.

3) F1 score: harmonic means accuracy of matching and recall.

4) Intersection Over Union (IOU): That portion of the union area found exactly between the critical point determined by the interpretive line and the radiologist's notes.

Let F(x) represent the important region found by the interpretation technique, and let S(x) represent the suspicious pneumonia area noted by the clinician for picture x. The following definitions apply to the accordance recall and precision, set accordance recall, set accordance precision, set F1, and set IOU formulas:

$$\text{Accordance recall}\left(x_i\right) = \frac{S\left(x_i\right) \cap F\left(x_i\right)}{S\left(x_i\right)} \tag{4}$$

$$\text{Accordance precision}\left(x_i\right) = \frac{S\left(x_i\right) \cap F\left(x_i\right)}{F\left(x_i\right)} \tag{5}$$

$$\text{Set Accordance recall} = \sum_{i}^{N} \frac{1}{N} \times \text{Accordance recall}\,(x_i) \tag{6}$$

$$\text{Set Accordance precision} = \sum_{i}^{N} \frac{1}{N} \times \text{Accordance precision}\,(x_i) \tag{7}$$

$$\text{Set } F_1 = \sum_{i}^{N} \frac{1}{N} \left(2 \times \frac{\text{Accordance recall}\,(x_i) + \text{Accordance precision}\,(x_i)}{\text{Accordance recall}\,(x_i) \times \text{Accordance precision}\,(x_i)}\right) \tag{8}$$

$$\text{Set IOU} = \sum_{i}^{N} \frac{1}{N} \times \frac{S(x_i) \cap F(x_i)}{S(x_i) \cup F(x_i)} \tag{9}$$

where x_i denotes the ith original image.

D. Visual Explainability Evaluation Checklist

The quantitative and qualitative evaluation of the performance of each translation method consists of a comprehensive analysis of translation visibility Three multi-objective experiments were planned to obtain these measures.

1) To assess the impact of each translation method in the absence of critical areas.

2) To observe the effective localization of each translation method

3) To gauge radiologists' confidence in every technique of interpretation Of these, experiment 1 evaluates each interpretation method's quantitative performance, and experiments 2 and 3 evaluate each method's qualitative performance. The framework's structure is displayed in Fig. 3. This framework may be used for various imaging modalities, such as magnetic resonance imaging (MRI), computed tomography (CT), or ultrasound imaging, in addition to radiography pictures.

1) First Experiment: Impact of Absence: This study evaluated the decision impact and confidence impact ratios of many widely used interpretation techniques, including ensemble XAI, Grad-CAM, Grad-CAM ++, SHAP, Saliency, and LIME. The same deep learning model, created by fusing a fully connected network with a pretrained image categorization network (Xception), was used for each technique.

Radiographic images were obtained from individuals who died between January and June 2020. A pool of 76 images from this sample After successfully generating images recognized by the model, heat maps of significant regions were generated using cluster XAI and five interpretation methods and then the deep learning model was used to generate prediction scores for these images though for, where the prediction scores and associated methods are shown in each of the above figures. This test was designed to assess how well different methods can be described in the general situation where the same network makes decisions based on the same information The next two tests were used to examine the three description methods with surface and into the XAI system qualitatively.

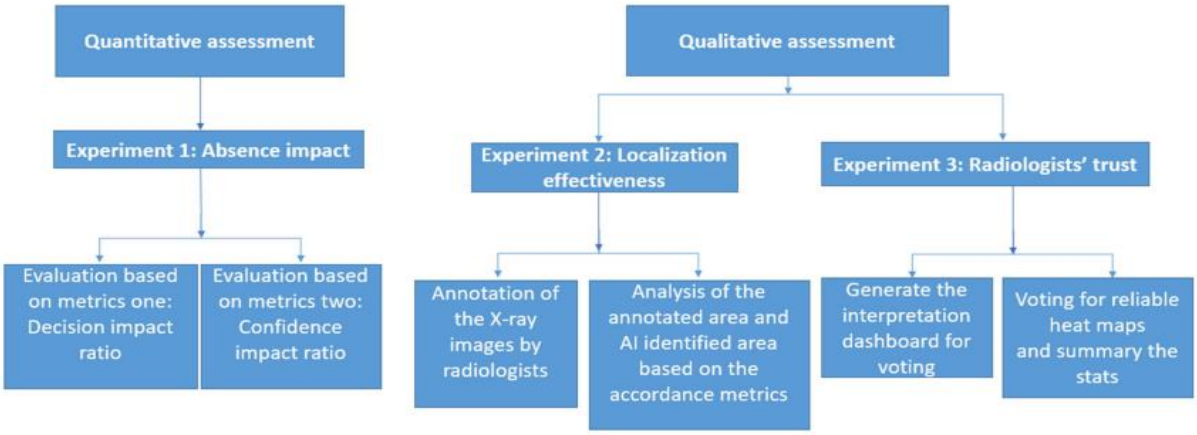

Fig. 3. Visual explainability framework.

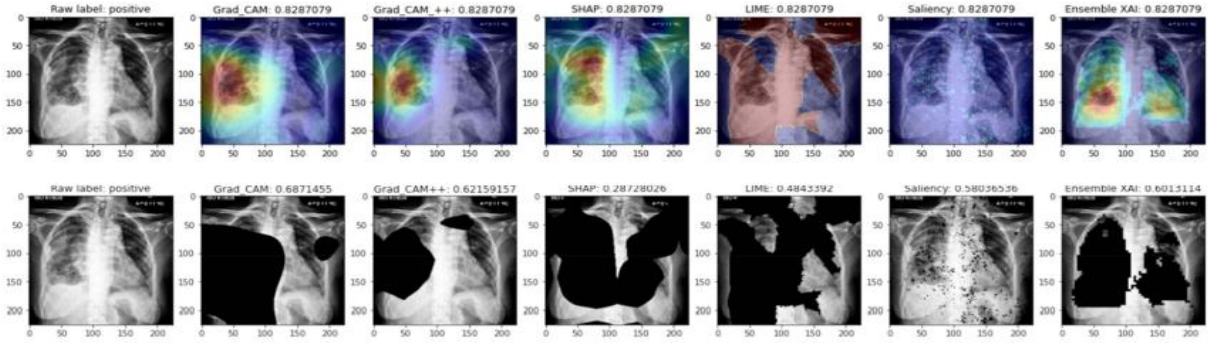

Fig. 4. Heat map identified by six interpretation methods with mortality risk score of original images in first row; images in absence of critical area of corresponding interpretation methods with new mortality risk score in second row.

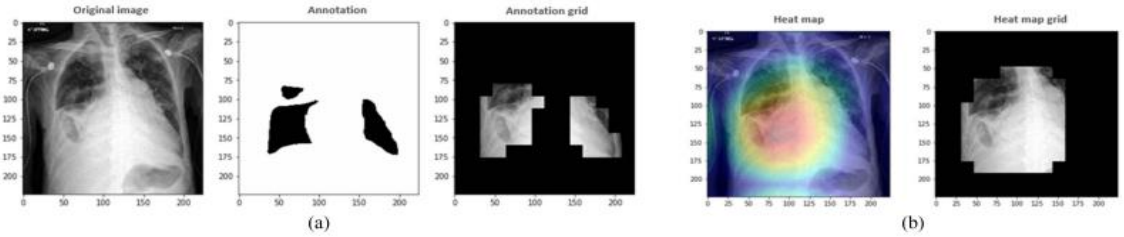

Fig. 5. Annotation and heat map in grid form. (a). From left to right are the original image, annotated area recognized by experienced clinicians and annotation in grid form. (b). Critical area identified by Grad-CAM++ in grid form.

2) Experiment 2. Localization Effectiveness: This experiment evaluated how well each interpretation technique could locate the possible areas of severity. By comparing the crucial regions found by the interpretation technique with the ground truth of severity areas marked by skilled radiologists, the set tagged regions in the pictures that they believed might lead to the patient's death using the accordance recall, set accordance precision, set F1, and set IOU values were calculated.

Three radiologists with at least five years of experience reviewed the set of 76 pictures for severity area irritation. The radiologists identified and PixelAnnotationTool_x64_v1.4.0 annotation tool. The centre picture in Fig. 4(a) displays the result of the annotated

watershed mask image. Since (77,1024) is the form of the output generated by the final convolutional layer as evaluated by the interpretation techniques, (77) may be understood as the shape of the image feature. The annotated region was approximated into 14∗14 grids with a greater
tolerance, as the picture on the right of Fig. 5(a) illustrates.
The crucial regions determined by the interpretation techniques were also formed into 14 x 14 grids for comparison in order to preserve uniformity. Fig. 5(b) provides an example, where the critical area acquired by the AI interpretation approach is displayed on the left, and the matching critical area in grid form is displayed on the right.
Experiment 3. Radiologists' Trust: In this experiment, we used voting to gauge radiologists' confidence in each technique of interpretation. The purpose of this experiment was to get radiologists' subjective and qualitative opinions on how reliable AI techniques are in identifying crucial regions. In this experiment, the top three interpretation techniques from experiment 1 and the ensemble XAI were applied to the same collection of 76 photos.
a) Creation of an interpretation dashboard: Using streamlit, an open-source Python toolkit for building and sharing web apps for machine learning and data science, a web application was created to record and examine the decisions made by the seasoned radiologists. The web was used to present the chosen photos and the heat map graphs that went with them use as seen in Figure 6. There was also a checkbox available for every heat map type. Each radiologist then utilised the online programme to carry out any of the following tasks.

1) Select the best interpretation technique for an image from among those available (the interpretation winners from experiment 1).
2) Select the two interpretation techniques that work best for a given image.
3) Select the most appropriate interpretation approach for each of the three images.
4) Select each interpretation technique that seems to fit an image the best.
5) If none of the three interpretation strategies seem acceptable enough, choose none of them.
Each of the 76 distinct photos' selections made by the various radiologists were automatically recorded and stored in a CSV file. The interpretation approach that was determined to be a reasonable interpretation will subsequently be given a score of one, or zero if it was not selected.

V. CLINICAL IMPACT

Chest X-rays are commonly used to diagnose and predict illness severity due to their widespread availability and inexpensive cost. The COVID-19 epidemic has increased the use of chest X-rays for clinical diagnosis and treatment. Interpretation models can help speed up workflow and prioritise important X-ray reports. Abnormal chest X-rays might be flagged for priority assessment by a radiologist. Early and correct diagnosis allows for more effective treatment of patients. Furthermore, these models can perform pattern recognition and search algorithms far quicker than radiologists.
The visual explainability framework helps radiologists analyse X-rays by spotting abnormalities and evaluating their concordance. It improves X-ray detection by providing clear visual clues and facilitating quick evaluation. While not yet capable of independent clinical diagnosis, these models can enhance the accuracy and reliability of X-ray reporting. This is very useful for marking certain regions on photographs to double-check or examine more closely.

DISCUSSION

Reliable interpretation of heat maps generated by gradient-based techniques like GradCAM++ and SHAP requires a stable model due to the influence of changing models on performance. Compared to Grad-CAM++ and SHAP, Ensemble XAI provides more reliable interpretation by automatically assigning weights to pixel attributes based on a small collection of annotation. Ensemble XAI develops reliable interpretations by combining high-contributed pixel information from Grad-CAM++ and SHAP. Grad-CAM++ and SHAP heat maps may emphasise non-lung regions owing to text, catheters, or lines in the X-ray picture.
Precision and recall metrics are important to measure business decisions. A high accuracy metric indicates the ability to identify specific regions, while a high recall rate indicates the ability to identify all doubtful regions as a wider core area will benefit more from recall,
Action Grad-CAM++ outperformed two other major methods, SHAP and LIME, with an accuracy of 0.46. Ensemble XAI outperformed Grad-CAM++ (0.45) with an accuracy of 0.52 and a recall of 0.57.F1 and IOU is used as the main matrix for analyzing the data. The analysis showed that Ensemble XAI performed better than other translation methods in terms of localization effectiveness (medium group F1: 0.50, average group IOU: 0.36).
F1 and IOU are used as significance matrices to analyze the findings. The analysis showed that Ensemble XAI performed better than other translation methods in terms of localization effectiveness (medium group F1: 0.50, average group IOU: 0.36).
The purpose of the radiologist reliability test is to determine the reliability of a high residual-based F1 score. Radiologists voted for the method with the most published sites based on their comments, so larger, better-remembered regions would not receive more votes Overall, Ensemble XAI received the highest votes with 70.2% of the votes of all methods tested.

Since the overall mean group F1, mean group IOU score, and mean results for ensemble XAI are higher than any interpretation method, it is clear that SHAP and Grad-CAM++ complement each other.As for localization about effectiveness (mean group F1: 0.46 and mean set IOU: 0.32) and about radiologist -Trust (mean agreement: 67.1%), SHAP was the second best descriptive method. Compared to other methods Grad-CAM++ method provides faster rendering time, acceptable vote rate, higher understanding of increased average conformity accuracy Due to superpixel-based rendering, which has more variability and with external space is consistently related therefore, LIME performed inconsistently in quantitative and qualitative assessment consequently radiologists in it They were not convinced It did not score negatively in the localization-effectiveness assessment. In addition to the conventional IOU measure, we specified the accordance accuracy and recall during the localization evaluation in the visual explainability checklist. These metrics are readily comprehended by doctors.The ensemble XAI approach yields a maximum of 0.52 and a minimum of 0.33 using the LIME technique. The real positive regions found using the explainability approaches are effectively represented by the mean set precision values. Although ensemble XAL has reached 0.52 percent, this indicates that 48% of false positive regions have been discovered. In a clinical environment, these areas would need to be further interpreted and confirmed by a qualified radiologist before they could be considered true disease locations. As a result, it restricts the use of explainability approaches by nonradiologists as independent on-the-ground detection tools.

Conversely, although adequate, the mean set recall values are still not good enough for clinical application at this time. Using the SHAP, the greatest mean set recall value of 72% was found among the four explainability approaches examined. Although this is a positive and high score, it also indicates that 28% of the disease-affected sites that radiologists have discovered are not included in the explainability technique. Therefore, radiologists prefer to employ an adjunctive tool rather than an independent detection tool for the nonradiologist physician.

To further characterise these false positive rates, more study utilising the explainability techniques on normal X-rays will need to be done as part of future development. Specifically, how frequently do the explainability methods indicate abnormal regions that a radiologist judged normal? By providing an answer, you will be able to better describe how well explainability techniques work to distinguish between radiographs that are normal and those that may be abnormal.

RESULTS

A. Experiment 1: Absence Impact

This test is based on the selection effect coefficient and the reliability effect coefficient of each visual presentation method to measure its explanatory power.Table I shows the performance of the methods in the first stage. LIME is the most efficient measure with a decision effect of 0.96 and a reliability effect of 0.43. This means that 96% of positive images will be placed in the negative category and if the main area detected by LIME is removed, the associated confidence will decrease to 43%. Besides LIME, the other two that perform well are Grad-CAM++ and SHAP. Critical areas with significant impact on decision-making and reliability were also identified. XAI by Grad-CAM yielded comparable findings (decision effect 0.72, reliability effect 0.24). Thus, in trials 2 and 3, the top three base methods—Grad-CAM++, LIME, and SHAP—as well as ensemble XAI are assessed in more detail.

Fig. 6. Dashboard for CNN interpretation voting.

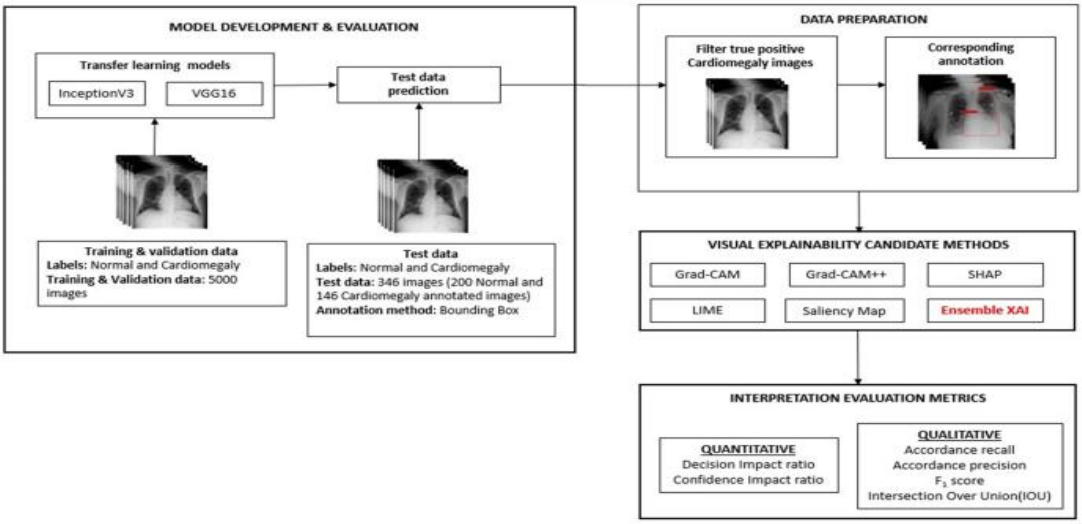

Fig. 7. Workflow for methods comparison on InceptionV3 and VGG16 using public data.

TABLE I
VISUAL EXPLAINABILITY EVALUATION CHECKLIST FOR DIFFERENT INTERPRETATION METHODS BASED ON XCEPTION MODEL (AUC: 0.803)

		Visual explainability methods					
Evaluation Measures		Ensemble XAI	SHAP	Saliency Map	Grad-CAM	Grad-CAM++	LIME
Quantitative	**Absence impact**						
	Decision impact	0.72	0.84	0.65	0.78	0.89	0.96
	Confident impact	0.24	0.30	0.18	0.23	0.33	0.43
	Representative Paper(s): (Chattopadhay et al.,2018; Lin Zhong Qiu et al.,2019)						
Qualitative	**Localization effectiveness**						
	Mean set accordance precision	0.52(0.08)	0.39(0.06)	--	--	0.46(0.06)	0.33(0.07)
	Mean set accordance recall	0.57(0.05)	0.72(0.05)	--	--	0.45(0.03)	0.61(0.02)
	Mean set F_1 score	0.50(0.03)	0.46(0.04)	--	--	0.41(0.02)	0.40(0.05)
	Mean set IOU	0.36(0.03)	0.32(0.03)	--	--	0.28(0.02)	0.26(0.04)
	Representative Paper(s): (Chattopadhay et al.,2018; Padilla et al.,2020)						
	Radiologists' trust						
	Mean vote for reliable interpretation methods by radiologists	70.18% (0.03)	67.10% (0.12)	--	--	49.60% (0.06)	26.30% (0.06)
	Representative Paper(s): (Selvaraju et al.,2019)						
	Overall assessment	In the quantitative assessment, LIME had the highest decision impact and confidence impact, followed by Grad-CAM++, SHAP, Grad-CAM and ensemble XAI. In the qualitative assessment, we take the mean value (standard deviation) from three radiologists. The Ensemble XAI achieved the best performance in both localization effectiveness (mean set F_1: 0.50, mean set IOU: 0.36), and reliability votes from the panel of radiologists (mean vote: 70.2%). SHAP followed in second place in reliability votes (mean vote: 67.1%) and localization effectiveness (mean set F_1: 0.46, mean set IOU: 0.32). Grad-CAM++ and LIME did not achieve good performance in this round.					

B. Experiment 2: Localization Effectiveness

In this experiment, concordance recall, concordance accuracy, and IOU are used to determine the localization effectiveness of each method.

The second column of Table I compares the critical area obtained by deep learning with the area identified by experienced meteorologists in order to assess the effectiveness of each method.

The Ensemble XAI method outperformed the other three methods in terms of imputed F1 scores, with an average set concordance precision of 0.52 and an average set concordance recall of 0.57 This indicates that location significant findings in the translational process were 52% annotated XAI, . After SHAP and Grad-CAM++, he has likewise achieved the highest group IOU of 0.36 for IOU analysis.

C. Experiment 3: Radiologists' Trust

It contradicts the confidence of radiologists in the use of any technique.

This was achieved through an assessment by our psychological research performed by our experienced team of radiologists. They voted for the most credible interpretations of each image.

The results of the interpretation methods are shown in the third column of Table I: Grad-CAM++, ensemble XAI, SHAP, and LIME. The group of radiologists determined that Ensemble XAI was the most reliable measure, with an average vote of 70.2%.

Before developing translation methods, we first developed models. In this step, two sets of images from the National Institutes of Health (NIH) chest X-ray dataset—absent/normal and cardiac enlargement (pathology class)—were used to generate VGG and onset binary classification model no. The flowchart in Figure 7 using the NIH chest x-ray dataset illustrates the sequential steps required in building models and developing semantic analysis metrics for each model. 2500 random images from each category in the original data set were used for model training and validation. The experimental dataset consists of 146 fully annotated images for the cardiomegaly group and 200 random samples for the no finding/normal category. This data set was used to develop binary classification models using two pre-trained

image classification models: VGG16 and InceptionV3. The test data set yielded accuracy and AUC scores of 0.817 and 0.917, respectively, for the InceptionV3 model. On the test data set, the VGG16 model achieved accuracy and AUC scores of 0.855 and 0.948, respectively. We extracted really good predictions (correctly predicted cardiac expansion images from a total of 146 cardiac expansion test images) for the two models on the baseline test according to basic methodology analysis criteria is the right. A default model threshold of 0.5 was used to eliminate true positives. There were 128 really good predictions for Inception V3 and 137 really good predictions for VGG16.

TABLE II
VISUAL EXPLAINABILITY EVALUATION CHECKLIST FOR DIFFERENT INTERPRETATION METHODS BASED ON INCEPTION MODEL (AUC: 0.917)

	Evaluation Measures	Visual explainability methods					
		Ensemble XAI	SHAP	Saliency Map	Grad-CAM	Grad-CAM++	LIME
Quantitative	**Absence impact**						
	Decision impact	0.22	0.21	0.10	0.12	0.06	0.42
	Confidence impact	0.19	0.19	0.11	0.15	0.11	0.29
	Representative Paper(s): (Chattopadhay et al.,2018; Lin Zhong Qiu et al.,2019)						
Qualitative	**Localization effectiveness**						
	Mean set accordance precision	0.66(0.13)	0.42(0.15)	--	--	0.36(0.07)	0.30(0.07)
	Mean set accordance recall	0.87(0.13)	0.81(0.24)	--	--	0.95(0.08)	0.87(0.11)
	Mean set F_1 score	0.74(0.10)	0.54(0.17)	--	--	0.52(0.08)	0.45(0.07)
	Mean set IOU	0.60(0.12)	0.39(0.14)	--	--	0.36(0.07)	0.29(0.06)
	Representative Paper(s): (Chattopadhay et al.,2018; Lin Zhong Qiu et al.,2019)						
	Overall Assessment	In the quantitative assessment, LIME had the highest decision and confidence impact, followed by Ensemble XAI and SHAP. In the qualitative assessment, Ensemble XAI achieved the best performance with mean set F_1: 0.74 and mean set IOU: 0.60. The second-best results were obtained using SHAP (mean set F_1: 0.54, mean set IOU: 0.39) followed by Grad-CAM++ (mean set F_1: 0.52, mean set IOU: 0.36).					

Semantic analysis metrics for VGG16 and InceptionV3 were generated using these two images.

The InceptionV3 semantic analysis metrics are given in Table II . With a decision effect of 0.42 and a reliability effect of 0.29, LIME is the best performing method in the absence effect test. Besides LIME, the two best performing methods were ensemble XAI and SHAP. In the localization effectiveness experiment, the ensemble XAI method obtained the highest score among the four measurements. The composite F1 scores, average set recall, set IOU, and set IOU set for XAI were all, in that order, 0.74, 0.87, 0.66, and 0.60. Translational analysis criteria for VGG16 are shown in Table III . Ensemble XAI gives the best results in the absence effect test, with a decision effect of 0.59 and a reliability effect of 0.46. Apart from ensemble XAI, LIME and Grad-CAM were the other best performing methods. The Ensemble XAI method produced the best overall metrics in the localization effectiveness experiment. The average set IOU, average set F1 score, average set conformity accuracy, and average set recall for ensemble XAI were 0.77, 0.88, 0.72, and 0.64, respectively. According to the interpretive evaluation criteria for InceptionV3 and VGG16 using public datasets, ensemble XAI outperforms other visual interpretation methods and gives better results than the original dataset used in this work.While LIME takes significantly longer computational time to generate visual interpretations, Grad-CAM, Grad-CAM++, SHAP, salience map, and ensemble XAI require comparatively much less time.

TABLE IV
COMPUTATIONAL COMPLEXITY COMPARISONS FOR DIFFERENT INTERPRETATION METHODS

Methods	Grad-CAM	Grad-CAM++	SHAP	Saliency Map	LIME	Ensemble XAI
Average Time per image (ms)	1760	1880	3450	3099.5	133320	5152
Hardware Processor: Intel® Core™ i7-8700K Processor CPU @ 3.70GHz, RAM: 64GB, GPU: Dual NVIDIA GeForce GTX 1080 @ 8GB memory **Image Property** Average Size: 400KB ~ 500KB, Resolution: 1024 * 1024, Format: PNG						

CONCLUSION

Our ensemble XAI was created using Grad-CAM++ and SHAP as its foundation. In comparison to other interpretation techniques, it performed better in terms of radiologists' trust and the accuracy of localization. The article provided assurance of the possible application of ensemble approaches in the interpretation of imaging.

domain. Additionally, the radiologists on our panel recommended using ensemble XAI, which performed the best in the explainability test, as an additional tool to aid in the interpretation of the X-ray.

In order to determine the most effective and complete assessment framework for establishing the best image explainability procedures for thoracic medical imaging, a panel of radiologists provided feedback for the visual explainability evaluation checklist. This will make it easier for researchers to analyse images correctly and in line with clinical evaluation. Consequently, this procedure may be quickly and simply expanded to accommodate various therapeutic settings. The public will have access to the dashboard that was utilised for our radiologist panel vote.

We conducted extensive interviews with radiologists about the influence of visual explainability on clinical pathways in order to determine the clinical impact. This offers crucial input and recommendations for further research and advancement of the AI interpretation algorithm. Lastly, IHiS [36] Singapore's AI-enabled medical imaging platform may make use of the suggested visual explainability evaluation methodology and ensemble XAI in practice. It is also anticipated that ensemble XAI would be readily expandable in the near future to accommodate additional medical imaging interpretations, such CT and MRI, which are quickly gaining traction in the field of health care AI.

REFERENCES

[1] F. Ozyurt et al., "An automated COVID-19 detection based on fused dynamic exemplar pyramid feature extraction and hybrid feature selection using deep learning," Comput. Biol. Med., vol. 132, May 2021, Art. no. 104356.

[2] T. Tuncer et al., "A novel Covid-19 and pneumonia classification method based on F-transform," Chemometrics Intell. Lab. Syst., vol. 210, Mar. 2021, Art. no. 104256.

[3] M. M. Taresh et al., "Transfer learning to detect COVID-19 automatically from X-Ray images using convolutional neural networks," Int. J. Biomed. Imag., vol. 2021, May 2021, Art. no. 8828404.

[4] S. Thakur et al., "X-ray and CT-scan-based automated detection and classification of Covid-19 using convolutional neural networks (CNN)," Biomed. Signal Process. Control, vol. 69, Aug. 2021, Art. no. 102920.

[5] P. G. Vaz et al., "Evaluation of COVID-19 chest computed tomography: A texture analysis based on three-dimensional entropy," Biomed. Signal Process. Control, vol. 68, Jul. 2021, Art. no. 102582.

[6] J. Zhang et al., "Dense GAN and multi-layer attention-based lesion segmentation method for COVID-19 CT images," Biomed. Signal Process. Control, vol. 69, Aug. 2021, Art. no. 102901.

[7] H. Golamalinejad et al., "A novel deep learning based method for COVID-19 detection from CT image," Biomed. Signal Process. Control, vol. 70, Sep. 2021, Art. no. 102987.

[8] J. Quah et al., "Chest radiograph-based artificial intelligence predictive model for mortality in community-acquired pneumonia," BMJ Open Respirotary Res., vol. 8, no. 1, 2021, Art. e001045. [9] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?," 2017. [Online]. Available: https://arxiv.org/abs/1712.09923 .

[10] K. Paranjape, M. Schinkel, and P. Nanayakkara, "Short keynote paper: Mainstreaming personalized healthcare–transforming healthcare through new era of artificial intelligence," IEEE J. Biomed. Health Informat., vol. 24, no. 7, pp. 1860–1863, Jul. 2020.

[11] S. Thomas, "Artificial intelligence, medical malpractice, and the end of defensive medicine," Bill Health, 2017. [Online]. Available: https: //blog.petrieflom.law.harvard.edu/2017/01/26/artificial-intelligence -medical-malpractice-and-the-end-of-defensive-medicine/

[12] S. Bach et al., "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," PLoS One, vol. 10, 2015, Art. no. e0130140.

[13] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," 2017. [Online]. Available: https://arxiv.org/abs/ 1704.03296
[14] A. Shrikumar et al., "Learning important features through propagating activation differences," 2017. [Online]. Available: https://arxiv.org/abs/ 1704.02685
[15] S. Lapuschkin et al., "Unmasking clever Hans predictors and assessing what machines really learn," Nature Commun., vol. 10, Mar. 2019, Art. no. 1096. [16] D. Wang et al., "Designing theory-driven user-centric explainable AI," in: Proc. CHI Conf. Hum. Factors Comput. Syst., 2019, pp. 1–15, doi: 10.1145/3290605.3300831.
[17] D. Bau et al., "Network dissection: Quantifying interpretability of deep visual representations," 2017. [Online]. Available: https://arxiv.org/abs/ 1704.05796
[18] C. Olah et al., "The building blocks of interpretability," ResearchGate, Berlin, Germany, 2018.
[19] N. T. Arun et al., "Assessing the validity of saliency maps for abnormality localization in medical imaging," DeepAI, 2020. [Online]. Available: https://deepai.org/publication/assessing-the-validity-ofsaliency-maps-for-abnormality-localization-in-medical-imaging
[20] Z. Q. Lin et al., "Do explanations reflect decisions? A machine-centric strategy to quantify the performance of explainability algorithms," 2019. [Online]. Available: https://arxiv.org/abs/1910.07387 [21] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.
[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp.–1800–1807. [Online]. Available: https://arxiv.org/abs/1610.02357
[23] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via Gradient-based localization," 2019. [Online]. Available: https: //arxiv.org/abs/1610.02391
[24] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Improved visual explanations for deep convolutional networks," 2018. [Online]. Available: https://arxiv.org/abs/1710.11063