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

Lin Zou, Han Leong Goh, Charlene Jin Yee Liew, Jessica Lishan Quah, Gary Tianyu Gu, Jun Jie Chew, Mukundaram Prem Kumar, Christine Gia Lee Ang, and Andy Wee An Ta

***Abstract*—Since the onset of the COVID-19 pandemic in 2019, many clinical prognostic scoring tools have been proposed or developed to aid clinicians in the disposition and severity assessment of pneumonia. However, there is limited work that focuses on explaining techniques that are best suited for clinicians in their decision making. In this article, we present a new image explainability method named ensemble AI explainability (XAI), which is based on the SHAP and Grad-CAM++ methods. It provides a visual explanation for a deep learning prognostic model that predicts the mortality risk of community-acquired pneumonia and COVID-19 respiratory infected patients. In addition, we surveyed the existing literature and compiled prevailing quantitative and qualitative metrics to systematically review the efficacy of ensemble XAI, and to make comparisons with several state-of-the-art explainability methods (LIME, SHAP, saliency map, Grad-CAM, Grad-CAM++). Our quantitative experimental results have shown that ensemble XAI has a comparable absence impact (decision impact: 0.72, confident impact: 0.24). Our qualitative experiment, in which a panel of three radiologists were involved to evaluate the degree of concordance and trust in the algorithms, has showed that ensemble XAI has localization effectiveness (mean set accordance precision: 0.52, mean set accordance recall: 0.57, mean set F_1: 0.50, mean set IOU: 0.36) and is the most trusted method by the panel of radiologists (mean vote: 70.2%). Finally, the deep learning interpretation dashboard used for the radiologist panel voting will be made available to the community. Our code is available at https://github.com/IHIS-HealthInsights/Interpretation-Methods-Voting-dashboard.**

***Impact Statement*—Compared to other sectors that have deployed artificial intelligent (AI), the use of AI in healthcare understandably requires closer scrutiny due to the potential risks to patient safety, especially for clinical AI. As such, AI Explainability (XAI) is a key focus area in regard to the adoption of AI in healthcare. However, most of the current XAI methods for medical imaging revolve around quantitative assessment and there is a lack of systematic qualitative studies that seek to gain trust and concordance with clinicians. In this article, we worked with a panel of clinicians to devise a comprehensive XAI evaluation framework combining quantitative and qualitative metrics to systematically review the efficacy of XAI techniques on deep learning models for pneumonia medical imaging. More importantly, we developed a new image explainability algorithm named Ensemble XAI, which gained the most trust by the panel of radiologists with a mean vote of 70.2%. It is envisioned that with the proposed XAI evaluation framework and ensemble XAI, it will help in proliferating the use of AI in medical imaging.**

***Index Terms*—Explainable artificial intelligent (AI), clinical decision support, pneumonia, COVID-19, chest X-ray, neural network.**

I. Introduction

As of May 17, 2021, 163.71 million cases of COVID-19 infection and 3 393 551 deaths have been reported worldwide. Ethical considerations in scarcity suggest that hospital resources should be prioritized for patients who are most ill. Singapore, with her population of 5.6 million, has faced an unprecedented surge in hospital care, similar to many other countries hit by COVID-19. COVID-19 has pushed Singapore's healthcare systems to the edge and spurred rapid development of AI health informatics solutions to fight against the pandemic.

A number of international studies have been performed and presented in the literature on the importance of deep learning algorithms to facilitate quick diagnosis of COVID-19 detection using medical image datasets [1]–[7]. Most of the work reported good classification performance using deep learning algorithms on computed tomography (CT) images and chest X-ray imaging. For example, in [1], Ozyurt *et al.* proposed a fused feature generator and iterative hybrid feature selector that uses a four-phase image preprocessing technique to extract handcrafted features of CT images. The artificial neural networks and deep neural network models used these features as inputs to classify healthy CT images and Covid-19 CT images and achieved classification accuracies of 94.10% and 95.84% respectively. In [3], Zhu *et al.* highlight the effectiveness of deep learning using pre-trained algorithms for classifying chest X-ray images. However, none of the research papers emphasized model explainability.

In February 2020, Singapore's Changi General Hospital, together with the national HealthTech Agency Integrated Health Information System (IHiS), collaborated to develop an AI predictive model known as the Community Acquired Pneumonia and COVID-19 AI Predictive Engine (CAPE) [8] that

Manuscript received 10 September 2021; revised 18 December 2021; accepted 19 February 2022. Date of publication 25 February 2022; date of current version 24 March 2023. This article was recommended for publication by Associate Editor Yo-Ping Huang upon evaluation of the reviewers' comments. *(Corresponding author: Lin Zou.)*

Lin Zou, Han Leong Goh, Mukundaram Prem Kumar, Christine Gia Lee Ang, and Andy Wee An Ta are with the Integrated Health Information Systems, Singapore 544910 (e-mail: jenny.zou@ihis.com.sg; hanleong.goh@ihis.com.sg; mukund.ram96@gmail.com; christine.ang@ihis.com.sg; andy.ta@ihis.com.sg).

Charlene Jin Yee Liew, Jessica Lishan Quah, Gary Tianyu Gu, and Jun Jie Chew are with the Changi General Hospital, Singapore 529889 (e-mail: charlene.liew.j.y@singhealth.com.sg; jessica.quah.l.s@singhealth.com.sg; gary.gu@mohh.com.sg; junjie.chew@mohh.com.sg).

Digital Object Identifier 10.1109/TAI.2022.3153754

can generate a risk score for pneumonia patients. The CAPE team consisted of senior clinicians with specialties in radiology and respiratory medicine, data scientists, health informatics researchers and system engineers. They set out to design a simple and scalable application that could embed AI into the thoracic imaging workstream.

One of the key challenges with the introduction of AI into a clinical workflow was centered on explainability [9]. Neural networks have been proven to be superior in terms of accuracy in many imaging applications when compared to conventional machine learning approaches such as support vector machines. However, the former is much less explainable. Without clear explainability of how AI algorithms made their predictions, it is difficult for clinicians to be comfortable working with AI and trust the algorithms [10], [11]. While there is extensive work in explaining the decision of an algorithm, for example, commonly used layer-wise relevance propagation [12] and localization, gradients, and perturbations [13], [14]. There are few literatures examining the efficacy of interpretation techniques for deep learning imaging networks [15]–[18], as well as quantifying them using methods, such as overlapping with ground truth bounding boxes and label randomization [19], [20],

As highlighted in [21], Tjoa and Guan there is a lack of standardized and a uniform adoption of interpretability assessment criteria across the medical field. This may create a bias in selection for one method compared to another without justification based on medical practices. It is also found that there is little work involving human studies that evaluate the trustworthiness of interpretation techniques for medical imaging.

The main contributions of this article are as follows.

1) Based on ensemble techniques used in machine learning, we proposed integrating the SHAP and Grad-CAM++ methods to produce an augmented mapping layer identifying discriminative regions. We named this ensemble XAI.
2) We compiled a visual explainability evaluation checklist that aims to benchmark various image explainability techniques quantitatively and qualitatively. For the qualitative studies, a panel of expert radiologists was involved in the localization effectiveness and subjective voting assessment to determine which image explainability techniques were best suited for thoracic medical images.
3) Finally, we provide an in-depth discussion on the impact of visual explainability on clinical pathways.

II. Method

In this section, we introduce the data and deep learning model used for mortality identification, review five state of the art interpretation methods, proposes ensemble XAI, and elaborate on evaluation metrics and experiments for the visual explainability evaluation checklist.

A. Data and Modeling

Model development was based on a single acute tertiary hospital's data. Ethics approval was given by the SingHealth Centralized Institutional Review Board (CIRB 2020/2100), and a study consent waiver for the use of data was obtained.

1) Predictive Model Development: Our predictive model was developed by using a retrospective study of a cohort that encompasses adult patients admitted to a tertiary acute hospital in Singapore from January 1, 2019 to 31st December 2019. The inclusion criterion was based on patients admitted through the emergency department with pneumonia diagnosis (using ICD-10 coding).

The study cohort consisted of 2235 chest X-ray images from 1966 unique adult patients. The EMR data of patients were also collected to generate labels for the inpatient mortality indicator. The data was deidentified prior to processing.

The data were categorized into three different sets named "training," "validation," and "test," Patients admitted from January 1, 2019 to October 31, 2019 were split into the training and validation sets with a ratio of 9:1. Patients admitted from November 1, 2019 to December 31, 2019 were used to create the "test" set to ensure temporal generalizability of the model.

Our objective was to predict the demise of patients with pneumonia during inpatient episodes. The starting point of prediction was on the day of admission. A binary classification was used with the label 1 defined as inpatient mortality and 0 for non-mortality.

A deep learning classifier was developed that combined a pretrained image classification network—Xception—with a fully-connected network. Xception is an extension of the Inception architecture that replaces the standard Inception modules with depthwise separable convolutions [22]. A transfer learning approach, which uses a predefined model, has the benefit of taking advantage of the knowledge gained while learning generic features from large-scale image datasets. The models were implemented in Keras (version 2.3.0), Scikit-learn (version 0.19.1) and Python (version 3.7).

The final model developed termed as CAPE has an AUC of 0.890 and accuracy of 0.899 when tested on this retrospective cohort data. To date, CAPE has been implemented as a computer application, where independent chest radiographs can be analyzed for determination of an image-based mortality risk score.

2) Interpretation Evaluation Dataset: To evaluate our approaches for XAI interpretation, a prospective cohort consisting of 1475 adult patients who required inpatient admission for a physician-determined diagnosis of CAPE via the emergency department, over the period of January 1, 2020 to June 30, 2020 was employed. The model performance on the prospective data has an AUC of 0.803 and accuracy of 0.811 when tested in real world clinical setting. The interpretation evaluation dataset established was 76 true positive cases identified by CAPE. Fig. 1 illustrates the prospective cohort formation of the interpretation evaluation dataset.

B. Interpretation Approaches

There are five state of the art interpretation methods that are targeted in this article, namely Grad-CAM, Grad-CAM++, SHAP, LIME, and saliency. In addition, ensemble XAI is introduced in this section.

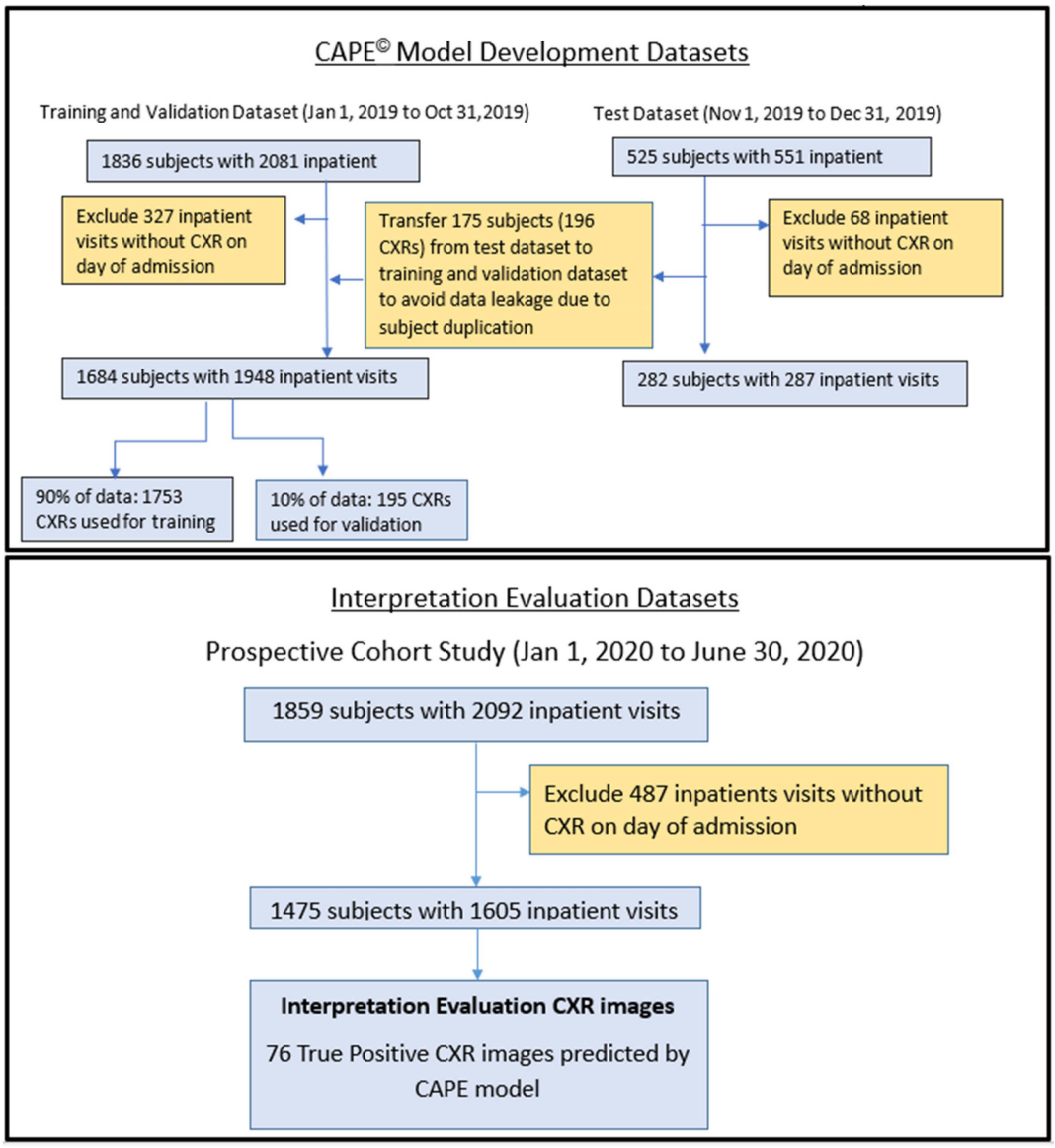

Fig. 1. Datasets for CAPE model development and prospective cohort study.

1) Grad-CAM: Grad-CAM is one of the methods that have gained popularity in recent years. It has made CNN-based models more transparent by visualizing input regions with high resolution details that are important for making predictions [23]. Visualization of the final feature map A^k shows the discriminative regions of the image, as the last convolutional layer can be considered features of a classification model. Grad-CAM proposes to use the averaged gradient score as weights for the feature map which is defined as

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

where $A^k \in R^{u*v}$ is the kth feature map from the last convolutional layer with height u and width v.

However, this approach may have certain shortcomings, such as the failure to localize an object in the image if there are multiple occurrences of the same object, or the localization of only a portion of the objects due to the unweighted average of partial derivatives [23].

2) Grad-CAM++: Grad-CAM++ is a generalized method of Grad-CAM that improves upon Grad-CAM's limitations. This approach provides a measure of importance to each pixel in a feature map that contributes to the overall decision of the CNN. As such, all the spatially relevant regions of the input image are equally highlighted such that the entire object is localized in instances where there are multiple occurrences of the same object [24].

3) SHAP: SHAP is used to explain prediction of instance x by computing the contribution of each feature to the prediction. The SHAP gradient explainer is an extension of the integrated gradients method—a feature attribution method designed for differentiable models based on an extension of Shapley values to infinite player games (Aumann–Shapley values) [25], [26].

If we approximate the model with a linear function between each background data sample and the current input to be explained, and we assume that the input features are independent, then expected gradients compute approximate SHAP values. The gradient explainer works by integrating the gradients of all interpolations between a foreground sample (the sample being explained) and a background sample (the sample being compared to). As an adaptation to make them approximate SHAP values, expected gradients reformulate the integral as an expectation and combine that expectation with sampling reference values from the background dataset. This leads to a single combined expectation of gradients that converge to attributions that sum to the difference between the expected model output and the current output [25], [26].

4) Local Interpretable Model-Agnostic Explanation (LIME): LIME [27] is used to approximate a complex model locally by an interpretable model that can explain prediction of a particular instance of interest. The LIME procedure can be summarized as follows.

1) Determine an interpretable representation of the instance of interest. For an image, superpixels (contiguous patches of similar pixels) are used, such that the interpretable representation of an image is a binary vector where 1 indicates the original superpixel and 0 indicates a grayed out superpixel [28].
2) Draw a sample by disturbing the interpretable representation. Instead of having all ones in the binary vector for the original image, the sample image has some zeros in the binary vector that indicate grayed out superpixels.
3) Apply the original model to the perturbed images and generate predictions.
4) Fit the interpretable model to the proximity-weighted sampled images and the predictions in step iii.
5) Use the interpretable model to draw conclusions about the relevance of each interpretable component.

The complexity of the above process makes the computation of LIME very time consuming. Furthermore, this sparse superpixel-based explanation method is sensitive to small amounts of noise in the input [29], resulting in instability of the explanations.

5) Saliency Map: Saliency map is a simple and straightforward interpretation method that was first introduced in 2014 [30]. Since the gradient of output with respect to the input image represents how the output value changes with respect to a small change in input, high magnitudes of gradients are expected to highlight input regions that cause the most change in the output. Thus, the pixels that are highlighted are the ones that contribute most to the output. However, this approach is unable to distinguish between positive and negative evidence due to absolute values of the partial derivative [31].

6) Ensemble XAI: Ensemble methods are widely applied in deep learning due to their ability to minimize bias and variance, which can result in improved reliability [32], [33]. For medical imaging, the stacking-based ensemble method is often used and has shown success in many deep learning studies [8], [34]. For image interpretation, we propose a stacking-based ensemble method that works on the output of the base interpretation method.

Since the Grad-CAM++ and SHAP gradient explainer methods are both gradient-based algorithms with different mechanisms but corresponding advantages, it is of interest to know whether the combined use of both methods results in complementary effects. To test this hypothesis, we proposed an ensemble method that applies Kernel Ridge regression to the normalized Grad-CAM++ and the normalized positive SHAP values to generate the mapping layer identifying discriminative regions. We have termed this method as ensemble XAI. Fig. 2 illustrates the workflow of ensemble XAI algorithm.

For each image, a pair of Grad-CAM++ and SHAP heat maps is generated by the base model, as well as the corresponding ground truth annotated by radiologists. Preprocessing is applied before the Kernel Ridge as shown in Fig. 2(a). First, as the image is annotated by three different radiologists, to generate the *y* function for Kernel Ridge, we calculate the weighted sum of three annotations to produce the target label. The target label shows three different intensity colors with the darkest area representing the concordance area of three radiologists, the green area representing the concordance area of two radiologists and light green representing the area annotated by only one radiologist. Second, the Grad-CAM++, SHAP heat map and ground truth two-dimensional (2-D) images are resized and converted to 1-D pixel features array. Then Grad-CAM++ and SHAP 1-D pixel arrays are concatenated as one combined pixel array.

Finally, the experiment data are split into three folds. For each iteration in Fig. 2(b), two folds of data are used in kernel ridge to fit the corresponding target ground truth, and then predict on the other fold. After three iterations, ensemble XAI for all folds is generated without information leakage.

C. Interpretation Evaluation Metrics

To determine the interpretation performance, the following evaluation metrics are used.

1) *Decision Impact Ratio*: The percentage change in decisions as a result of omitting the critical area identified by interpretation method.

Let $D(x)$ be the deep learning decision function which returns classification decision when the input is image x. Let 1_{logic} be the indicator function which returns one when the logic is true. The formula decision impact ratio can be calculated as follows:

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

where x_i denotes the *i*th original image, and c_i denotes the critical area identified by the deep learning model for the *i*th image.

2) *Confidence Impact Ratio*: The percentage drop in confidence as a result of omitting the critical area identified by the interpretation method.

Let $C(x)$ be the deep learning confidence function that returns the classification confidence probability when the input is

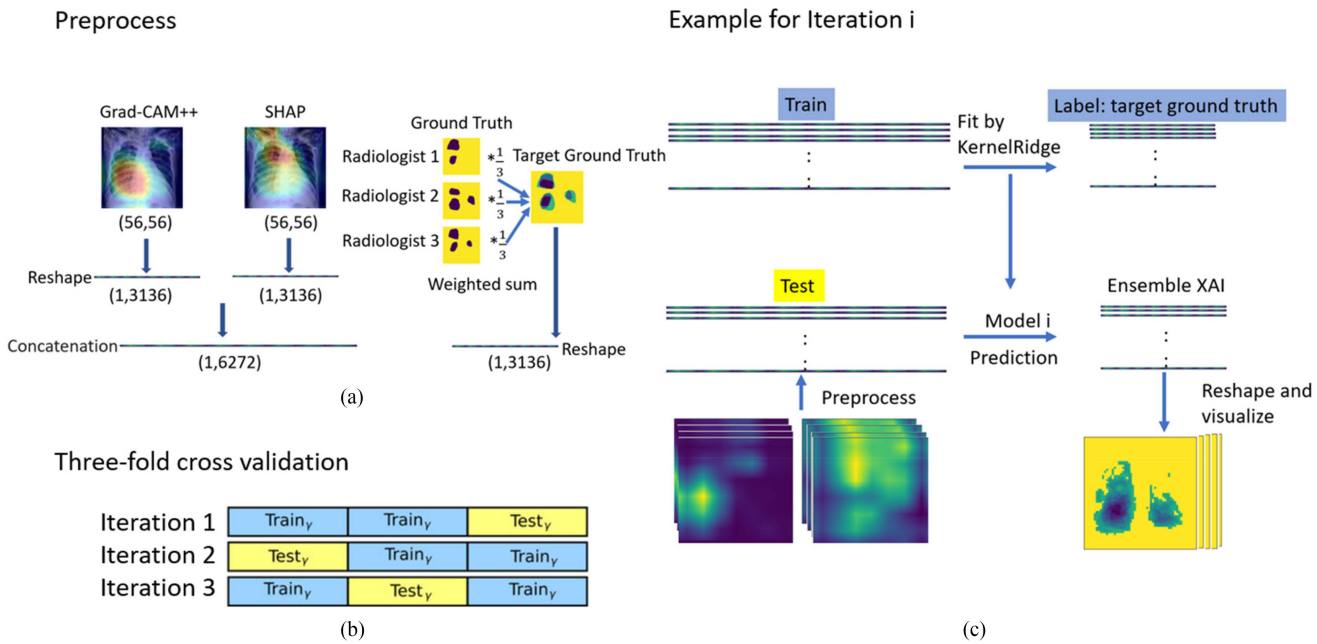

Fig. 2. Advanced ensemble XAI. (a) Preprocessing of Grad-CAM++, SHAP, and ground truth for each image. (b) Three-fold cross validation is applied to generate ensemble XAI. (c) Workflow for iteration *i*.

image x. The formula, confidence impact ratio, can be calculated as

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

where x_i denotes the *i*th original image, and c_i denotes the critical area identified by deep learning model for the *i*th image.

When comparing the critical area recognized by deep learning and the annotated area by experienced clinicians, there was a pair of accordance recall and precision measurements for each image.

1) *Accordance Recall*: fraction of the total annotated area that was correctly recognized by the interpretation method.
4) *Accordance Precision*: fraction of correctly recognized area among the entire critical area identified by the interpretation method.
5) F_1 *Score*: harmonic mean of the accordance recall and precision.
6) *Intersection Over Union (IOU):* fraction of the correctly recognized area among the union area encompassed by both the radiologist's annotation and the critical area identified by interpretation method.

Let $S(x)$ be the suspicious pneumonia area that is annotated by the clinician for image x and let $F(x)$ be the critical area that is identified by the interpretation method. The accordance recall and precision, set accordance recall, set accordance precision, set F_1 and set IOU formulas are as defined follows:

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

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

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

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

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

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

where x_i denotes the *i*th original image.

D. Visual Explainability Evaluation Checklist

The comprehensive visual explainability evaluation checklist consists of measures that assess both quantitative and qualitative performance of each interpretation method. Three experiments with different objectives were designed to obtain these measures.

1) To assess the critical area absence impact of each interpretation method.
2) To assess localization effectiveness of each interpretation method.

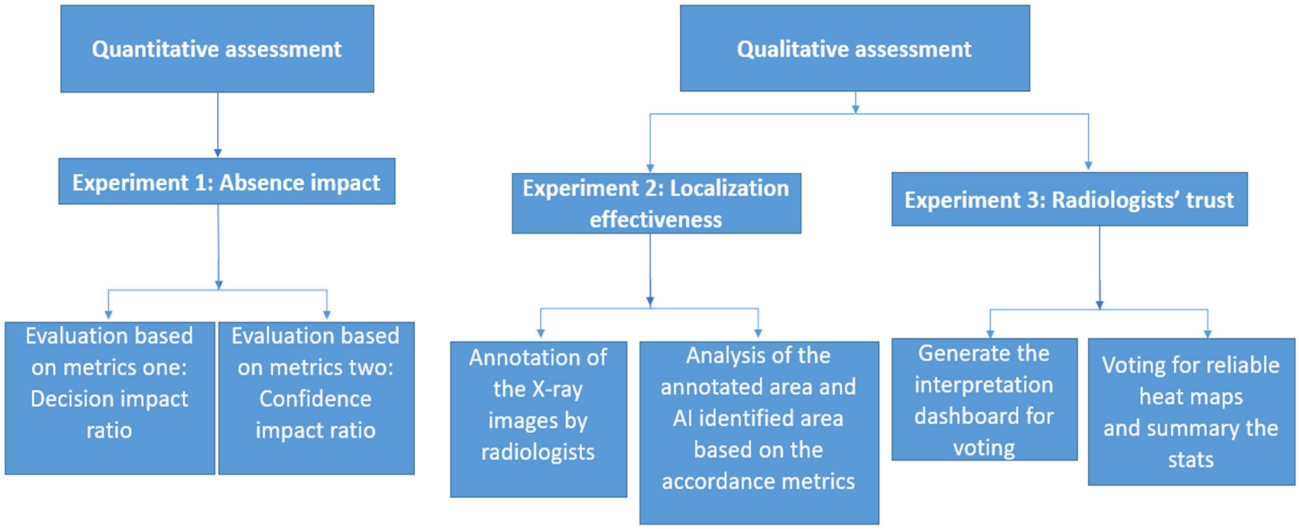

Fig. 3. Visual explainability framework.

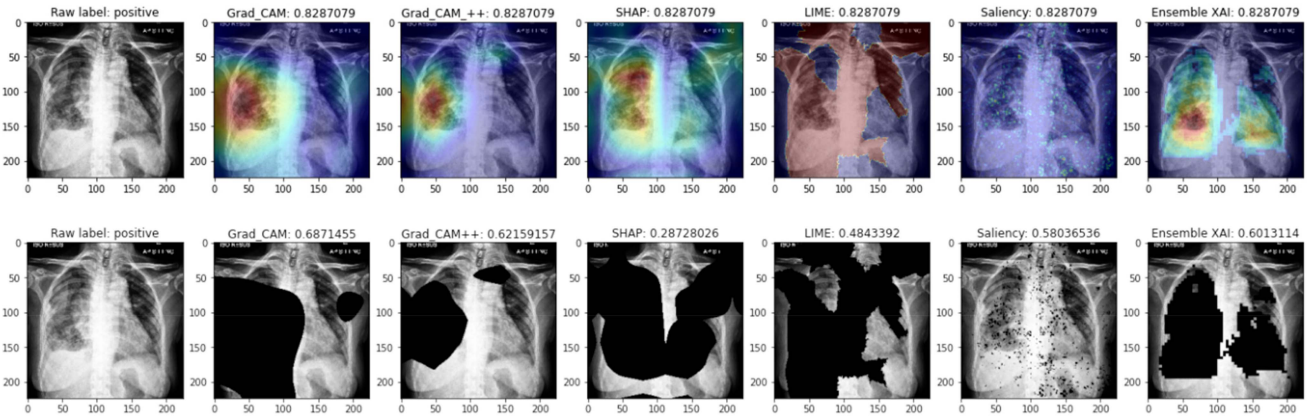

Fig. 4. Heat map identified by six interpretation methods with mortality risk score of original images in first row; images in absence of critical area of corresponding interpretation methods with new mortality risk score in second row.

3) To assess radiologists' trust for each interpretation method.

Of these, experiment 1 assesses the quantitative performance, while experiments 2 and 3 assess the qualitative performance of each interpretation method. Fig. 3 shows the structure of the framework. The use of this framework is not limited to radiographic images and can be applied to other imaging modalities, such as magnetic resonance imaging (MRI), CT, or ultrasound imaging.

1) Experiment 1. Absence Impact: In this article, the decision impact and confidence impact ratios of different popular interpretation methods (Grad-CAM, Grad-CAM ++, SHAP, LIME, saliency, and ensemble XAI) were assessed. Each method was applied using the same deep learning model, which was developed by combining a pretrained image classification network (Xception) with a fully connected network. The same last convolutional layer was used as the target for each method.

Radiographic images from patients who died from the January to June 2020 were obtained. From these images, a set of 76 images correctly recognized by the model was generated, to which heat maps of critical areas were added using the five interpretation methods and ensemble XAI. Prediction scores were subsequently derived for these images using the deep learning model in two different groups: one for the original image and one in an altered image (in which the part of the image corresponding to the abovementioned critical areas were removed). An example of this is shown in Fig. 4, where the interpretation methods used and the corresponding prediction scores are shown at the top of each image. As expected, the prediction score for each image was significantly changed when the corresponding critical area was removed.

The purpose of this experiment was to quantify the interpretation capabilities of different methods under a general scenario where decisions were made by the same network on the same data. The top three interpretation methods and ensemble

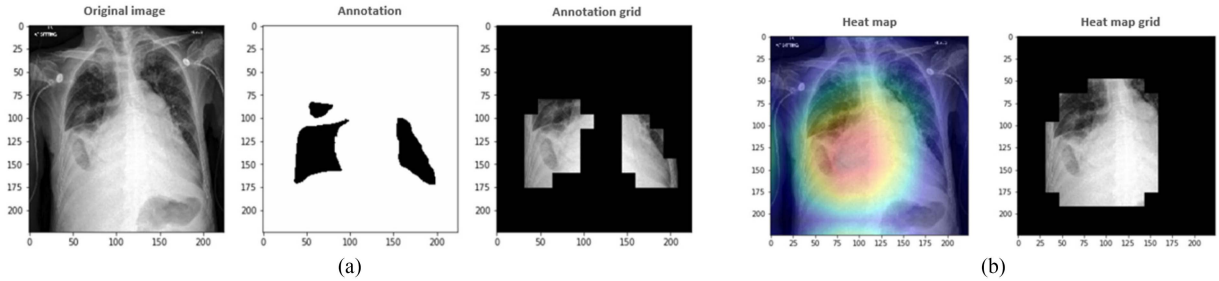

Fig. 5. Annotation and heat map in grid form. (a). From left to right are the original image, annotated area recognized by experienced clinicians and annotation in grid form. (b). Critical area identified by Grad-CAM++ in grid form.

XAI were then qualitatively assessed via the subsequent two experiments.

2) Experiment 2. Localization Effectiveness: In this experiment, the ability of each interpretation method to correctly localize the potential severity areas was assessed. This was measured using the set accordance recall, set accordance precision, set F_1 and set IOU values, which were obtained by comparing the critical areas identified by the interpretation method with the ground truth of severity areas annotated by experienced radiologists.

The set of 76 images was sent to a panel of three radiologists who had been working for at least five years for severity area annotation. Using the PixelAnnotationTool_x64_v1.4.0 annotation tool, areas in the images that were expected to contribute to the patient's mortality were identified and annotated by the radiologists. The output of the annotated watershed mask image is shown in the middle image in Fig. 4(a).

Since the interpretation methods evaluate the last convolutional layer generating output with a shape (77,1024), (77) can be interpreted as the shape of the image feature. We approximated the annotated area into 14∗14 grids with wider tolerance as shown in the rightmost image in Fig. 5(a).

To maintain consistency, the critical areas identified by the interpretation methods were also shaped into 14∗14 grids for comparison. An example is shown in Fig. 5(b), where the left image shows the critical area obtained by the AI interpretation method while the right image shows the corresponding critical area in grid form.

3) Experiment 3. Radiologists' Trust: In this experiment, we evaluated the radiologists' trust in each interpretation method through voting. The aim of this experiment was to obtain a qualitative and subjective assessment from radiologists with regard to the reliability of AI methods in identifying critical areas.

The same set of 76 images was again used for this experiment, to which the top three interpretation methods from experiment 1 and the ensemble XAI were applied.

a) Interpretation dashboard generation: To capture and analyze the choices of the experienced radiologists, a web application was developed using streamlit—an open-source Python library for creating and sharing web applications for machine learning and data science. The selected images and their corresponding heat map graphs were displayed using the web application as shown in Fig. 6. A checkbox was also provided for each of the different heat maps. The web application was then used by each radiologist to perform any of the following actions.

1) Choose any one of the interpretation methods (the interpretation winners from experiment 1) as the best choice for an image.
2) Choose any two of the interpretation methods as the best choice for an image.
3) Choose any three of the interpretation methods as the best choice for an image.
4) Choose all the interpretation methods as the best choice for an image.
5) Select none of the three interpretation methods if they are not reasonable enough.

The choices made by the different radiologists for each of the 76 different images were captured and saved automatedly in a CSV file. The interpretation method that was chosen as a reasonable interpretation will then be counted as one, or zero if not chosen. The aggregated trust percentages score for the different methods was then compared.

III. Results

In this section, the experimental results are summarized.

A. Experiment 1: Absence Impact

This experiment quantifies the interpretation capability of each visual explainability method based on their decision impact ratio and confidence impact ratio.

The performance of the different methods is shown in the first section of Table I. The top performing method is LIME with a decision impact of 0.96 and a confidence impact of 0.43. This means that if the critical area identified by LIME is removed, 96% of the positive images will be classified into the negative category and the corresponding confidence will drop to 43%. In addition to LIME, the other top performing methods are Grad-CAM++ and SHAP. The critical areas recognized by them are also associated with high decision impact and confidence impact. The ensemble XAI has achieved comparable results (decision impact: 0.72, confident impact: 0.24) with Grad-CAM. Thus, the top three base methods Grad-CAM++,

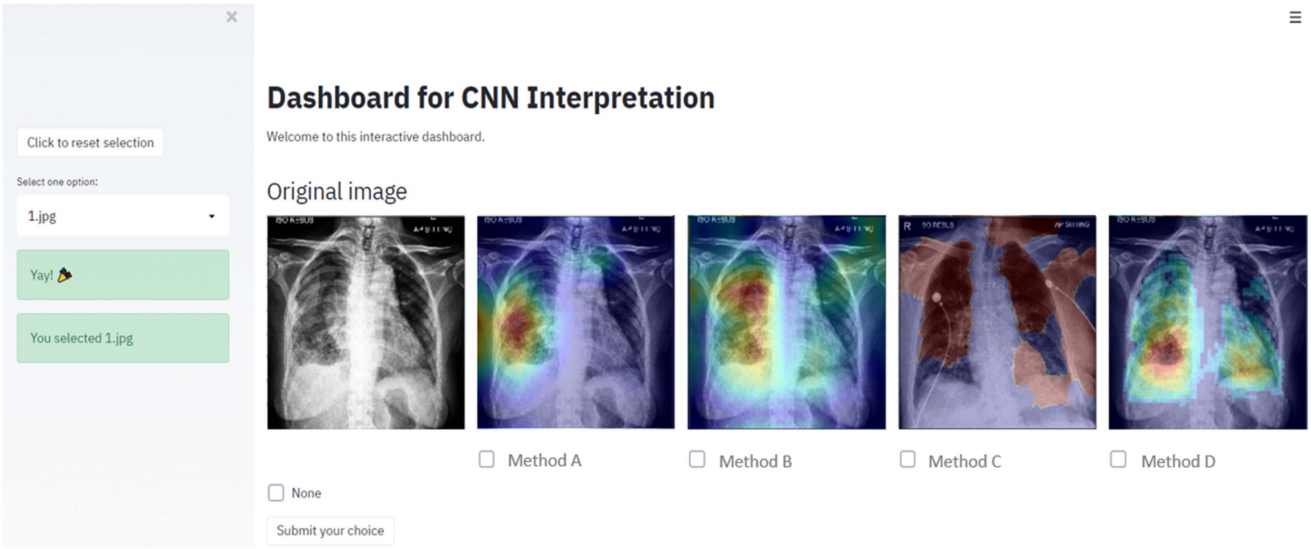

Fig. 6. Dashboard for CNN interpretation voting.

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.					

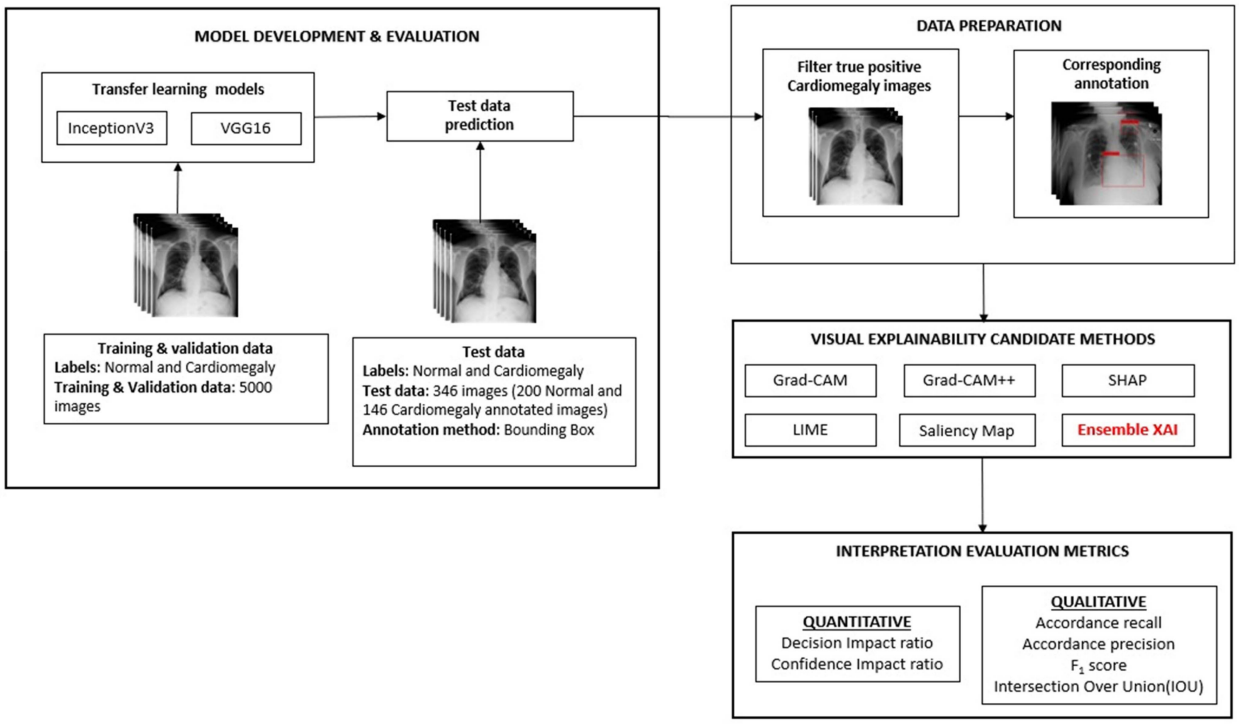

Fig. 7. Workflow for methods comparison on InceptionV3 and VGG16 using public data.

LIME and SHAP as well as ensemble XAI are evaluated further in experiments 2 and 3.

B. Experiment 2: Localization Effectiveness

This experiment quantifies the localization effectiveness of each method based on accordance recall, accordance precision and IOU.

The second section of Table I evaluates the performance of each method by comparing the critical area identified by deep learning against the area annotated by experienced radiologists. Among the four methods, the ensemble XAI achieved the highest mean set F_1 score with mean set accordance recall of 0.57 and mean set accordance precision of 0.52. This means that 57% of the annotated area was correctly identified by interpretation method and 52% of the critical area identified by interpretation method was consistent with the annotation. For the IOU evaluation, ensemble XAI has also achieved the highest mean set IOU of 0.36, followed by SHAP and Grad-CAM++.

C. Experiment 3: Radiologists' Trust

This experiment compares radiologists' trust in each method. This was done through a subjective vote by our panel of experienced radiologists. For each image, they voted for the interpretation methods that were deemed reliable.

The third section of Table I presents the voting results for the interpretation methods: Grad-CAM++; SHAP; LIME; and ensemble XAI. Ensemble XAI was chosen as the most trusted method by the panel of radiologists with a mean vote of 70.2%.

IV. Experiments Using Public Dataset

First, we developed models as prerequisites for interpretation methods. During this step, VGG and Inception binary classification models are developed based two classes of images—no finding/normal and cardiomegaly (pathology class) from National Institutes of Health (NIH) chest X-ray dataset [35]. Second, we conducted quantitative and qualitative experiments based on true positive cases which identified by respective models, and corresponding bounding box. Fig. 7 is a flowchart depicting the step-by-step process involved in developing the models and generating the interpretation evaluation metrics for each model using the NIH chest X-ray dataset.

For model training and validation, 2500 images from each class were randomly sampled from the original dataset. All the annotated images (146 images) available for the Cardiomegaly class along with 200 random samples for no finding/normal class were used as the testing dataset. Two pretrained image classification models – VGG16 and InceptionV3 were used to develop binary classification models using this dataset. The InceptionV3 model achieved an accuracy and AUC score of 0.817 and 0.917, respectively, on the test dataset. The VGG16 model achieved an accuracy and AUC score of 0.855 and 0.948, respectively, on the test dataset.

In accordance with the original approach for interpretation evaluation metrics, we filtered the true positive predictions (correctly predicted Cardiomegaly images from a total of 146 Cardiomegaly test images) for both the models on test set. The default model threshold of 0.5 was used to filter out the true positives. Inception V3 had a total of 128 true positive predictions and VGG16 had a total of 137 true positive predictions.

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

		Visual explainability methods					
Evaluation Measures		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).					

TABLE III
VISUAL EXPLAINABILITY EVALUATION CHECKLIST FOR DIFFERENT INTERPRETATION METHODS BASED ON VGG MODEL (AUC: 0.948)

		Visual explainability methods					
Evaluation Measures		Ensemble XAI	SHAP	Saliency Map	Grad-CAM	Grad-CAM++	LIME
Quantitative	**Absence impact**						
	Decision impact	0.59	0.44	0.15	0.53	0.35	0.59
	Confidence impact	0.46	0.39	0.12	0.42	0.32	0.43
	Representative Paper(s): (Chattopadhay et al.,2018; Lin Zhong Qiu et al.,2019)						
Qualitative	**Localization effectiveness**						
	Mean set accordance precision	0.72(0.14)	0.86(0.20)	--	--	0.72(0.18)	0.33(0.07)
	Mean set accordance recall	0.88(0.14)	0.39(0.15)	--	--	0.60(0.14)	0.81(0.12)
	Mean set F_1 score	0.77(0.10)	0.51(0.15)	--	--	0.63(0.09)	0.47(0.08)
	Mean set IOU	0.64(0.13)	0.36(0.13)	--	--	0.47(0.10)	0.31(0.07)
	Representative Paper(s): (Chattopadhay et al.,2018; Lin Zhong Qiu et al.,2019)						
	Overall Assessment	In the quantitative assessment, Ensemble XAI had the highest decision and confidence impact score of 0.59 and 0.46, followed by LIME and Grad-CAM. In the qualitative assessment, Ensemble XAI achieved the best performance with mean set F_1: 0.77 and mean set IOU: 0.64. The second-best results were obtained using Grad-CAM++ (mean set F_1: 0.63, mean set IOU: 0.47) followed by SHAP (mean set F_1: 0.51, mean set IOU: 0.36).					

These two sets of images were used for generating the interpretation evaluation metrics for InceptionV3 and VGG16, respectively.

Table II gives the interpretation evaluation metrics for InceptionV3. The top performing method in absence impact experiment is LIME with a decision impact of 0.42 and a confidence impact of 0.29. Besides LIME, the other top performing methods were ensemble XAI and SHAP. In localization effectiveness experiment, ensemble XAI method achieved the highest scores for all the four metrics. The mean set F_1 score, mean set accordance recall, mean set accordance precision, mean set IOU for ensemble XAI are 0.74, 0.87, 0.66, and 0.60, respectively.

Table III gives the interpretation evaluation metrics for VGG16. The top performing method in absence impact experiment is ensemble XAI with a decision impact of 0.59 and a confidence impact of 0.46. Besides ensemble XAI, the other top performing methods were LIME and Grad-CAM. In localization effectiveness experiment, ensemble XAI method achieved the

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

best metrics overall. The mean set F_1 score, mean set accordance recall, mean set accordance precision, mean set IOU of ensemble XAI was 0.77, 0.88, 0.72, and 0.64, respectively.

The interpretation evaluation metrics for both InceptionV3 and VGG16 using the public dataset is in accordance with the results generated using the original dataset used in this article and shows that ensemble XAI produces better results in comparison to other visual explainability methods. Table IV gives the computation time results for the visual interpretation methods. The methods Grad-CAM, Grad-CAM++, SHAP, saliency map, and ensemble XAI require comparatively much lesser time for general visual explanations while LIME requires much higher computation time to generate the visual explanations.

V. CLINICAL IMPACT

Chest X-ray is widely used to obtain initial diagnosis and prognosticate disease severity due to their broad availability, accessibility and low cost. The COVID-19 pandemic has brought along a further rise in the usage of chest X-ray to assist clinical decision making and treatment.

There is a role for these interpretation models in expediting workflow and prioritizing more urgent X-ray for reporting. This can be done by flagging abnormal chest X-ray for priority review by a radiologist. By facilitating earlier and more accurate diagnosis, definitive treatment of patients can be initiated earlier. Furthermore, these models can complete their pattern recognition and search algorithms far faster than radiologists.

The visual explainability framework aids the interpretation of the X-ray by a radiologist through identifying areas of abnormality to be evaluated for concordance. It augments the detection of abnormal areas on X-ray through easy interpretation and intuitive visual cues, while also allowing for rapid review. Although these models are yet to be capable of independent clinical diagnosis, they can improve the accuracy and confidence of reporting X-ray. This can be especially helpful by flagging specific areas on images to double check or take a closer look.

VI. DISCUSSION

Since different models and performances will impact the heat map generated by gradient-based methods, such as Grad-CAM++ and SHAP, generating reliable interpretation based on fixed model is important. Ensemble XAI has the advantage of stable interpretation compared to individual Grad-CAM++ and SHAP, as it automatically assigns weights to respective pixel features by learning from a small set of annotation. Ensemble XAI generates the stable interpretation by extracting and combine the high contributed pixel features from Grad-CAM++ and SHAP. In addition, it is inevitable that the base heat maps generated by Grad-CAM++ and SHAP sometimes highlight areas outside the lungs due to presence of text, catheters or lines in the X-ray image. Even though this special and distracted area may indicate a sign of severe disease in the lungs, it is not helpful for decision making. The ensemble XAI outperformed individual Grad-CAM++ and SHAP by assigning low weight to the special area outside the lungs.

When measuring the performance metrics, both concordance precision and recall metrics are important since high precision metric shows the capability of identifying the correct areas and high recall metric shows the capability of discovering all suspicious regions. Since large critical area will benefit a lot of recall, maintaining a relatively high precision without sacrificing much recall is the expected direction. Among three base methods: SHAP, LIME and Grad-CAM++, Grad-CAM++ has achieved the best precision of 0.46. Ensemble XAI achieved a better precision of 0.52 and recall of 0.57 compared to Grad-CAM++, which was 0.45.

In addition, F_1 and IOU are used as key matrices to discuss the results. The results of the study have demonstrated that ensemble XAI outperformed the other interpretation methods in both localization effectiveness (mean set F_1: 0.50 and mean set IOU: 0.36).

By conducting radiologists' trust experiment, we can also address the question that if high F_1 score pulled up by high recall is reliable. Since the radiologists voted for the method where the highest density of highlighted area is in accordance with their own annotation, larger region (higher recall) does not necessarily get more votes. As a result, among all the methods, ensemble XAI achieved best mean vote of 70.2%.

The SHAP and Grad-CAM++ is shown to complement each other since all mean set F_1, mean set IOU score and mean vote for ensemble XAI is higher than those from each interpretation method.

SHAP is the second-best interpretation method in terms of localization effectiveness (mean set F_1: 0.46 and mean set IOU: 0.32) and radiologists' trust (mean vote: 67.1%). The Grad-CAM++ is a fast interpretation method with acceptable votes and relatively high mean set accordance precision among the other methods. LIME showed inconsistent performance in

quantitative and qualitative assessment due to its superpixel-based explanations which have large variance and are always linked to some area outside the lung. Hence it was less competent in the localization effectiveness assessment and did not score well on radiologists' trust.

During the assessment of localization in the visual explainability checklist, in addition to the traditional metric of IOU, we defined the accordance precision and recall, which can be easily interpreted by clinicians. The mean set precision across all four explainability methods tested ranged from between 0.33 to 0.52. The lowest is 0.33 with the LIME method and the highest is 0.52 with the ensemble XAI method. The mean set precision values essentially represent the true positive areas identified via the explainability methods. While ensemble XAL has achieved 0.52, this means 48% percentage of false positive areas are identified, of which if used in a clinical setting, would require further interpretation and confirmation by a trained radiologist before they can be deemed actual areas of disease. It, therefore, limits the usage of the explainability methods as independent on-the-ground detection tools for the nonradiologist clinician.

The mean set recall values, on the other hand, while acceptable, are still not optimal enough for clinical use at the current stage. The highest mean set recall value across the four explainability methods studied is 72% via the SHAP. While such a value is high and promising, it also means that 28% of disease-affected areas identified by radiologists are missed when using the explainability method. Thus rather than used as an independent detection tool for the nonradiologist clinician, an adjunctive tool is more preferred for radiologists.

One key limitation faced in this article, is the interpretation evaluation dataset is only based on the true positive cases. This is due to two reasons. First, enormous laborious effort is needed in annotating the medical images for all four categories (true positive, true negative, false positive, and false negative) of the confusion matrix. Second, for true negative cases which mostly are mild cases, critical area is usually not appeared when radiologists annotate lung area during the assessment of risk of patient's mortality. Also heat maps generated by interpretation methods on true negative cases do not show much critical area. Thus, there is not enough information for accordance comparison. For false positive and false negative cases, as both are the incorrect prediction, it's less meaningful to conduct accordance comparison in this article comparing to True Positive cases.

As part of future enhancement, more work will have to be performed using the explainability methods on normal X-ray to further define these false positive rates—how often do the explainability methods identify abnormal areas that a radiologist deemed normal? Answering this question will further help characterize the efficacy of explainability methods in their use to dichotomize normal and potentially abnormal radiographs.

VII. Conclusion

We developed ensemble XAI, which was based on SHAP and Grad-CAM++. It had better performance than other interpretation methods in terms of localization effectiveness and radiologists' trust. This article gave confidence for the potential use of the ensemble techniques in the imaging interpretation field. Our panel of radiologists also suggested that ensemble XAI, which had the best performance in the explainability evaluation, can be used as an adjunctive tool to support the interpretation of the X-ray.

The visual explainability evaluation checklist (with input from a panel of radiologists) proved to be an effective and comprehensive assessment framework in determining the best image explainability techniques for thoracic medical images. This will aid researchers in generating appropriate image interpretations that align with clinical assessment. In turn, this process can be scaled up and applied to different clinical scenarios rapidly and easily. The dashboard used for our radiologist panel voting will be made available to the community.

To ascertain the clinical impact, we underwent in-depth discussions with radiologists on the impact of visual explainability on clinical pathways. This provides important feedback and guidelines for future development and improvement of the AI interpretation algorithm.

Finally, practical use of the proposed visual explainability Evaluation framework and ensemble XAI will potentially be used in AI-enabled medical imaging platform by IHiS [36] Singapore. In the near future, ensemble XAI is also foreseen to be easily extensible to other medical imaging interpretations, such as CT and MRI which are rapidly gaining adoption in 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-of-saliency-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
[25] M. Sundararajan *et al.*, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328. [Online]. Available: https://dl.acm.org/doi/10.5555/3305890.3306024
[26] S. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," 2017. [Online]. Available: https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
[27] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," 2020. [Online]. Available: https://arxiv.org/abs/2006.11371
[28] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144. [Online]. Available: https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf
[29] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," 2018. [Online]. Available: https://export.arxiv.org/pdf/1806.08049
[30] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2014. [Online]. Available: https://arxiv.org/abs/1312.6034v2
[31] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *J. Imag.*, vol. 6, no. 6, p. 52, 2020.
[32] M. A. Ganaie and M. Hu, "Ensemble deep learning: A review," 2021. [Online]. Available: https://arxiv.org/abs/2104.02395
[33] T. C. Lin and H. C. Lee, "Skin cancer dermoscopy images classification with meta data via deep learning ensemble," in *Proc. Int. Comput. Symp.*, 2020, pp. 237–241.
[34] R. Liu *et al.*, "Ensemble Learning with multiclassifiers on pediatric hand radiograph segmentation for bone age assessment," *Int. J. Biomed. Imag.*, vol. 2020, pp. 1–12, 2020.
[35] X. Wang *et al.*, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," 2017. [Online]. Available: https://arxiv.org/abs/1705.02315
[36] "Call for collaboration (CFC) on AI-enabled medical imaging platform," 2020. [Online]. Available: https://www.ihis.com.sg/HealthLab/Pages/AIMedicalImagingCFC.aspx