

Ensemble explanations in image classification

Weronika Hryniewska-Guzik

Towards Model-Agnostic Ensemble Explanations

Szymon Bobek^{1,2}[0000-0002-6350-8405], Paweł Bałaga¹, and
Grzegorz J. Nalepa^{1,2}[0000-0002-8182-4225]

¹ Jagiellonian Human-Centered Artificial Intelligence Laboratory (JAHCAI) and
Institute of Applied Computer Science, Jagiellonian University, 31-007 Kraków,
Poland

² AGH University of Science and Technology
{szymon.bobek, grzegorz.j.nalepa}@uj.edu.pl

Abstract. Explainable Artificial Intelligence (XAI) methods form a large portfolio of different frameworks and algorithms. Although the main goal of all of explanation methods is to provide an insight into the decision process of AI system, their underlying mechanisms may differ. This may result in very different explanations for the same tasks. In this work, we present an approach that aims at combining several XAI algorithms into one ensemble explanation mechanism via quantitative, automated evaluation framework. We focus on model-agnostic explainers to provide most robustness and we demonstrate our approach on image classification task.

Keywords: explainable artificial intelligence · machine learning · image processing

1 Introduction

Explainable Artificial Intelligence (XAI) has become an inherent component of data mining (DM) and machine learning (ML) pipelines in the areas where the insight into decision process of an automated system is important. Although the explainability (or intelligibility) is not a new concept in AI [16], it has been most extensively developed over the last decade. This is possibly due to the huge successes in black-box ML models such as deep neural networks in sensitive application contexts like medicine, industry 4.0 etc., but also a legal need of providing accountability and transparency to the reasoning process of AI systems [4]. A variety of algorithms for generating justifications for AI decisions and lack of explanations format standards, make it hard to integrate XAI methods into the standard ML/DM pipeline. Moreover, assessing quality of generated explanations is also non trivial task, as there is lack of unified metrics for evaluating

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, 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 paper, we present a new image explainability method named Ensemble 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 an comparable absence impact (decision impact: 0.72, confident impact: 0.24). Our qualitative experiment, in which a panel of 3 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 paper, we worked with a panel of clinicians to devise a comprehensive XAI evaluation framework

framework and Ensemble XAI, it will help in proliferating the use of AI in medical imaging.

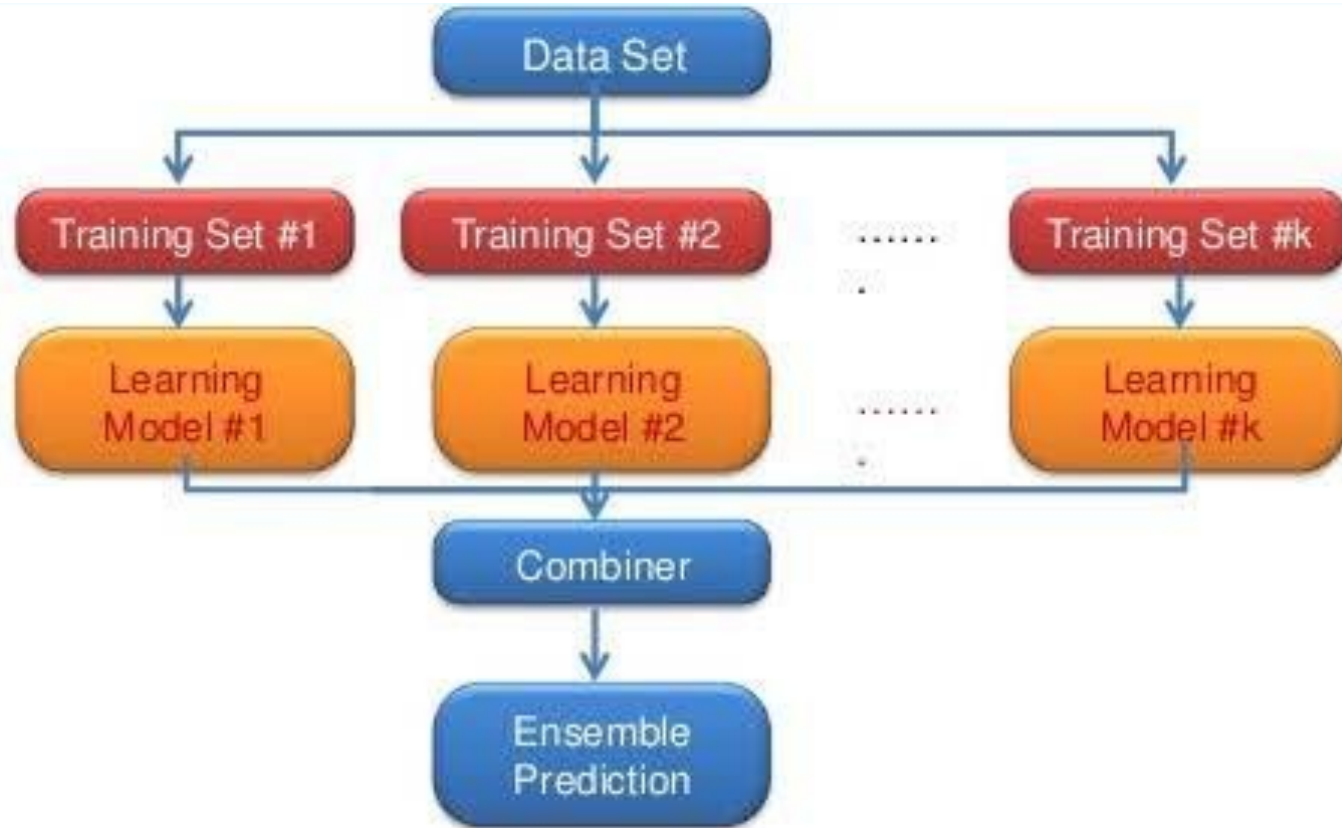
Index Terms— Explainable AI, Clinical Decision Support, pneumonia, COVID-19, Chest X-ray, Neural Network.

I. INTRODUCTION

As of 17 May 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] Fatih Ozyurt et al. proposed a fused feature generator and iterative hybrid feature selector that uses a four-phase image pre-processing technique to extract handcrafted features of CT images. The artificial neural networks (ANN) and deep neural network (DNN) 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], Ningbo Zhu et al. highlight the effectiveness of deep learning using pre-trained algorithms for classifying chest X-ray images (CXR). However, none of the research

Idea of ensembling



Explanation methods for images

Gradient-based

Integrated Gradients

DeepLift

Guided GradCAM

Saliency

Gradient SHAP

DeepLift SHAP

Guided Backprop / Deconvolution

LRP

Input x Gradient

NoiseTunnel (SmoothGrad, VarGrad, SmoothGrad Square)

Perturbation-based

FeatureAblation

FeaturePermutation

LIME

Occlusion

Kernel SHAP

SHAP Methods

Shapley Value Sampling

<https://github.com/pytorch/captum>

Bobek, S., Bałaga, P., & Nalepa, G. J. (2021). Towards Model-Agnostic Ensemble Explanations. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12745 LNCS, 39–51.
https://doi.org/10.1007/978-3-030-77970-2_4

Metrics-based ensemble XAI

Main contributions

- Presented an approach that aims at **combining several XAI algorithms into one ensemble explanation** mechanism via **quantitative, automated evaluation framework**.

Why metric-based?

3.2 Ensemble score

Calculating ensemble score is not limited to the metrics defined above and can be easily extended and modified as we will show in Section 4. The main goal of ES score is to capture the weighted importance of different metrics into one value. The definition of ES for a set of metrics M and weights w , was given in Eq. (4).

$$ES(M, w) = \sum w_i \cdot M_i \quad (4)$$

Having the ensemble score, we calculate a new, combined vector of explanation Φ^{ens} as a weighted sum of ensemble scores and associated with them original explanations $\Phi^{e_1}, \Phi^{e_2}, \dots, \Phi^{e_n}$. The weights are assigned arbitrary depending on the desired influence of a particular metric to the ensemble explanation.

Therefore, the final ensemble explanation is given by the Eq. (5).

$$\Phi^{ens} = \frac{ES(M, w) \cdot [\gamma_1 \Phi^{e_1}, \gamma_2 \Phi^{e_2}, \dots, \gamma_n \Phi^{e_n}]}{\sum_{i=1}^n ES_i(M, w)} \quad (5)$$

Where $\gamma_1, \gamma_2, \dots, \gamma_n$ are scaling factors that make it possible to compare and combine explanations obtained from different XAI frameworks. Note that

Scaling factor

Scaled feature importance values fit in a normalized range $[-1, +1]$.

Feature with the strongest contribution is assigned importance of ± 1 .

$$\gamma_k = \frac{1}{||\Phi^{e \rightarrow k}||_{max}}$$

$$\Phi' = \frac{\Phi - \min(\Phi)}{\max(\Phi) - \min(\Phi)}$$

Metrics

Consistency

$$C(\Phi^{e \rightarrow m_1}, \Phi^{e \rightarrow m_2}, \dots, \Phi^{e \rightarrow m_n}) = \frac{1}{\max_{a,b \in m_1, m_2, \dots, m_n} \|\Phi_j^{e \rightarrow m_a} - \Phi_j^{e \rightarrow m_b}\|_2 + 1}$$

a complete explanation matrix for every feature and every instance generated by explanation \mathbf{e} for model \mathbf{m}

Stability

image

$$\hat{L}(\Phi^{e \rightarrow m}, X) = \max_{x_j \in N_\epsilon(x_i)} \frac{\|x_i - x_j\|_2}{\|\Phi_i^{e \rightarrow m} - \Phi_j^{e \rightarrow m}\|_2 + 1}$$

the importance of feature i and instance j delivered by explanation model \mathbf{e} for machine learning model \mathbf{m}

where $N_\epsilon(x_i)$ is a set such as:

$$N_\epsilon(x_i) = \{x_j \in X \mid \|x_i - x_j\| < \epsilon\}$$

Area under the
loss curve (AUCx)

depicts the loss in accuracy (or other selected metric) when features are perturbed gradually according to their inverse importance returned by explanation algorithm.

Zou, L., Goh, H. L., Liew, C. J. Y., Quah, J. L., Gu, G. T., Chew, J. J., Prem Kumar, M., Ang, C. G. L., & Ta, A. (2022). Ensemble image explainable AI (XAI) algorithm for severe community-acquired pneumonia and COVID-19 respiratory infections. *IEEE Transactions on Artificial Intelligence*, 1–1.
<https://doi.org/10.1109/TAI.2022.3153754>

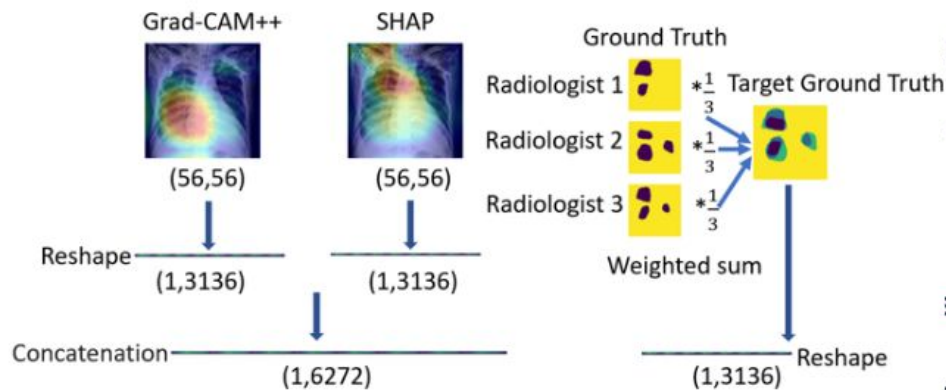
Supervised, mask-based ensemble XAI

Main contributions

- 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.
- 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 were involved in the localization effectiveness and subjective voting assessment to determine which image explainability techniques were best suited for thoracic medical images.
- Finally, we provide an in-depth discussion on the **impact of visual explainability on clinical pathways**.

a

Preprocess

**b**

Three-fold cross validation

Iteration 1	Train _y	Train _y	Test _y
Iteration 2	Test _y	Train _y	Train _y
Iteration 3	Train _y	Test _y	Train _y

c

Example for Iteration i

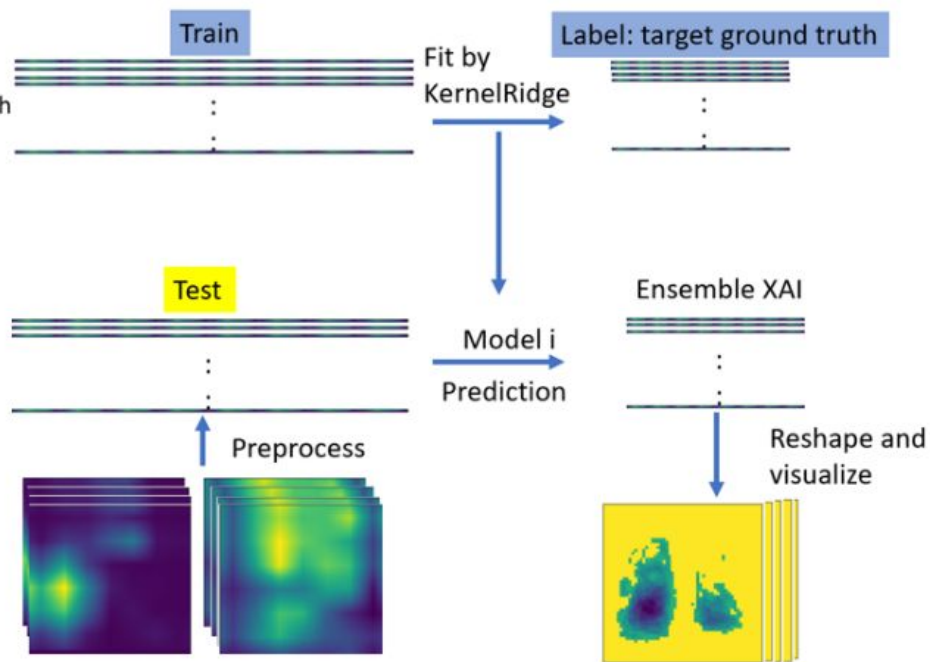


Fig. 2. Advanced ensemble XAI. a, Preprocessing of Grad-CAM++, SHAP and ground truth for each image. b, Three-fold cross validation is applied to generate Ensemble XAI. c. Workflow for iteration i .

Quantitative and qualitative metrics

the indicator function which returns one when the logic is true

the deep learning decision function which returns classification decision when the input is i^{th} original image x_i

the critical area identified by the deep learning model for the i^{th} image

$$\text{Decision impact ratio} = \sum_i^N \frac{1_{D(x_i) \neq D(x_i - c_i)}}{N}$$

The percentage change in decisions as a result of omitting the critical area identified by interpretation method.

the deep learning confidence function that returns the classification confidence probability when the input is image x

$$\text{Confidence impact ratio} = \sum_i^N \frac{\max(C(x_i) - C(x_i - c_i), 0)}{N}$$

The percentage drop in confidence as a result of omitting the critical area identified by the interpretation method.

Quantitative and qualitative metrics

$$\text{Accordance recall}(x_i) = \frac{S(x_i) \cap F(x_i)}{S(x_i)}$$

the suspicious pneumonia area that is annotated by the clinician for image x

$$\text{Accordance precision}(x_i) = \frac{S(x_i) \cap F(x_i)}{F(x_i)}$$

the critical area that is identified by the interpretation method

$$\text{Set Accordance recall} = \sum_i^N \frac{1}{N} \times \text{Accordance recall}(x_i)$$

$$\text{Set Accordance precision} = \sum_i^N \frac{1}{N} \times \text{Accordance precision}(x_i)$$

$$\text{Set } F_1 = \sum_i^N \frac{1}{N} \left(2 \times \frac{\text{Accordance recall}(x_i) + \text{Accordance precision}(x_i)}{\text{Accordance recall}(x_i) + \text{Accordance precision}(x_i)} \right)$$

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

Quantitative assessment

Experiment 1: Absence impact

Evaluation based
on metrics one:
Decision impact
ratio

Evaluation based
on metrics two:
Confidence
impact ratio

Qualitative assessment

Experiment 2: Localization effectiveness

Annotation of
the X-ray
images by
radiologists

Analysis of the
annotated area and
AI identified area
based on the
accordance metrics

Experiment 3: Radiologists' trust

Generate the
interpretation
dashboard for
voting

Voting for reliable
heat maps
and summary the
stats

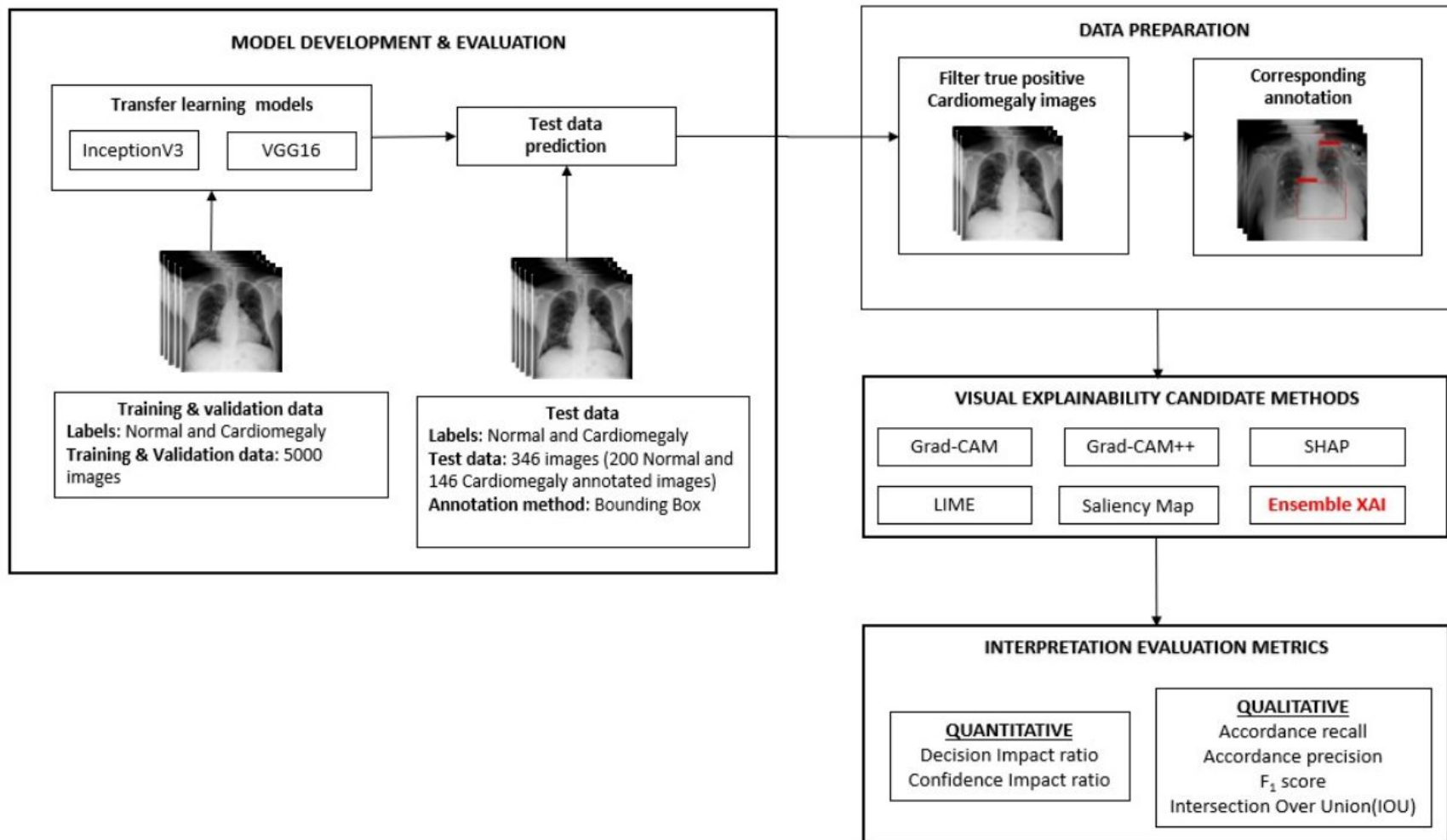


Fig. 7. Workflow for methods comparison on InceptionV3 and VGG16 using public data

Experiment 1 - absence impact

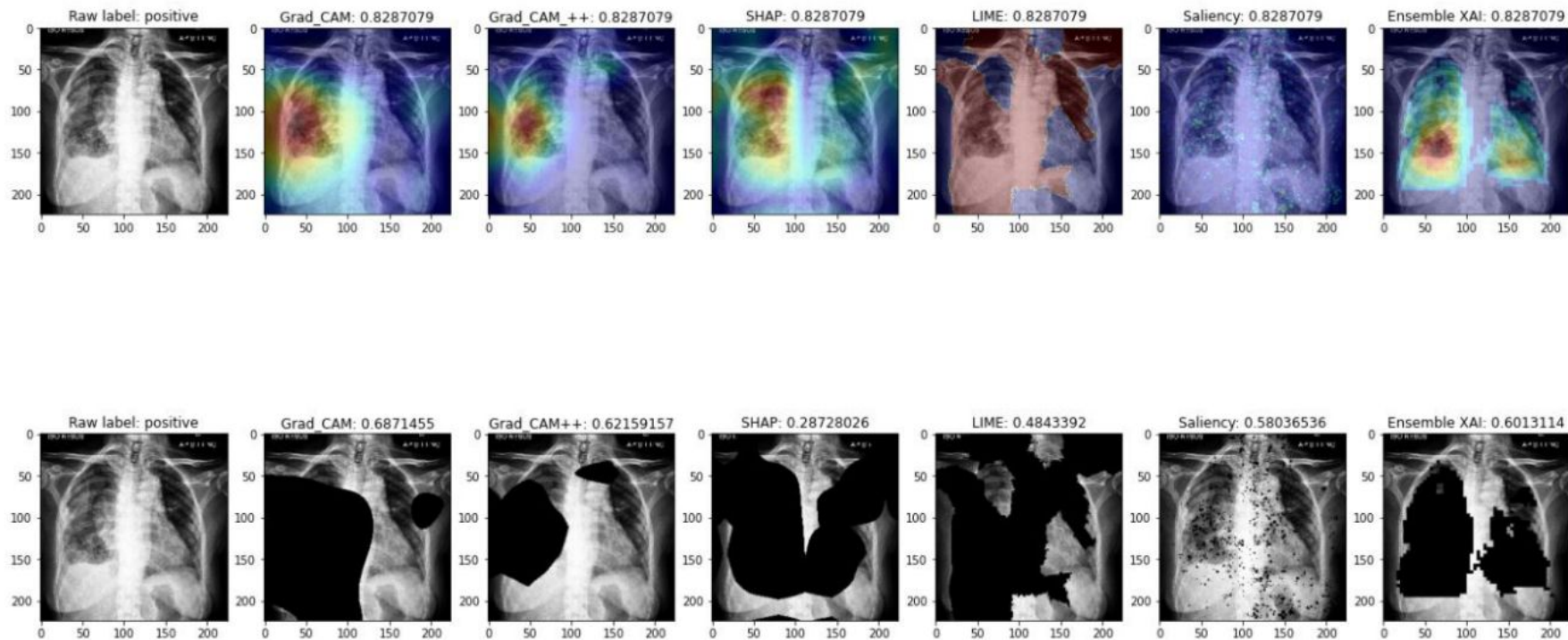


Fig. 4. Heat map identified by six interpretation methods with mortality risk score of original images in first row; images in absence of critical area of corresponding interpretation methods with new mortality risk score in second row.

Experiment 2 - localization effectiveness

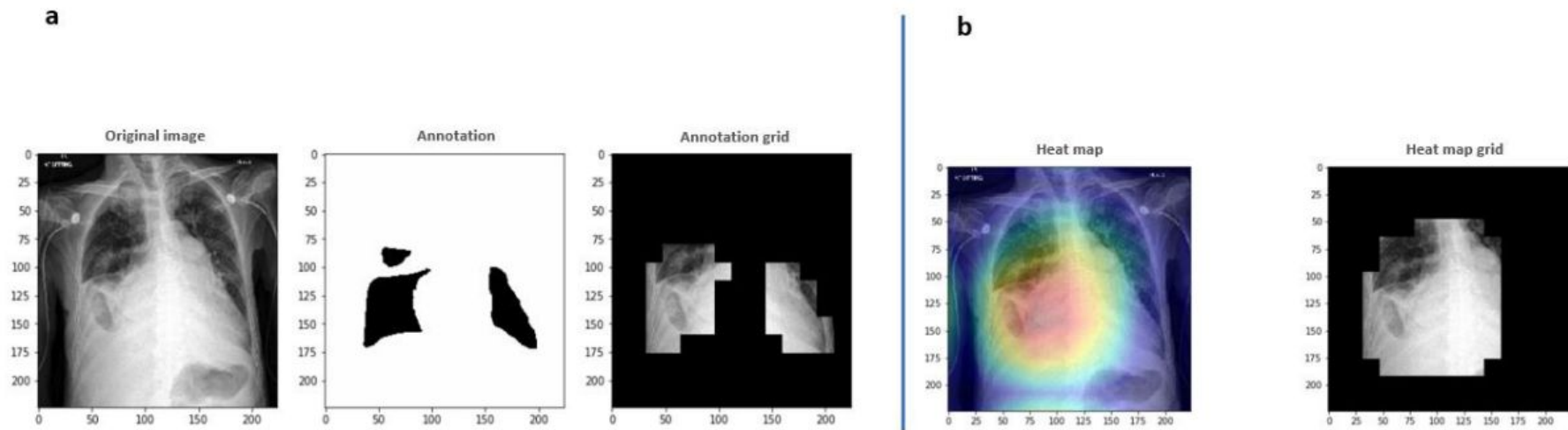


Fig. 5. Annotation and heat map in grid form. a, From left to right are the original image, annotated area recognized by experienced clinicians and annotation in grid form. b, Critical area identified by Grad-CAM++ in grid form.

Experiment 3

×

Click to reset selection

Select one option:

1.jpg

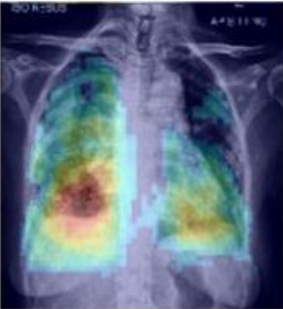




Yay! 🎉

You selected 1.jpg

Dashboard for CNN Interpretation

Welcome to this interactive dashboard.

Original image



☐ None

☐ Method A

☐ Method B

☐ Method C

☐ Method D

Submit your choice

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): (Chattopadhyay 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): (Chattopadhyay 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.				

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): (Chattopadhyay 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): (Chattopadhyay et al.,2018; Lin Zhong Qiu et al.,2019)						
Overall Assessment		<p>In the quantitative assessment, LIME had the highest decision and confidence impact, followed by Ensemble XAI and SHAP.</p> <p>In the qualitative assessment, Ensemble XAI achieved the best performance with mean set F₁: 0.74 and mean set IOU: 0.60. The second-best results were obtained using SHAP (mean set F₁: 0.54, mean set IOU: 0.39) followed by Grad-CAM++ (mean set F₁: 0.52, mean set IOU: 0.36).</p>					

TABLE III

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

Evaluation Measures		Visual explainability methods					
		Ensemble XAI	SHAP	Saliency Map	Grad-CAM	Grad-CAM++	LIME
Quantitative	Absence impact	0.59 0.46	0.44 0.39	0.15 0.12	0.53 0.42	0.35 0.32	0.59 0.43
	Decision impact						
	Confidence impact						
	Representative Paper(s): (Chattopadhyay et al.,2018; Lin Zhong Qiu et al.,2019)						
Qualitative	Localization effectiveness	0.72(0.14) 0.88(0.14) 0.77(0.10) 0.64(0.13)	0.86(0.20) 0.39(0.15) 0.51(0.15) 0.36(0.13)	-- -- -- --	-- -- -- --	0.72(0.18) 0.60(0.14) 0.63(0.09) 0.47(0.10)	0.33(0.07) 0.81(0.12) 0.47(0.08) 0.31(0.07)
	Mean set accordance precision						
	Mean set accordance recall						
	Mean set F_1 score						
	Mean set IOU						
	Representative Paper(s): (Chattopadhyay et al.,2018; Lin Zhong Qiu et al.,2019)						
Overall Assessment		<p>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.</p> <p>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).</p>					

Thank you for attention!