LISA : Enhance the explainability of medical images unifying current XAI techniques

I. Abstract

This work proposed a unified approach to increase the explainability of the predictions made by Convolution Neural Networks (CNNs) on medical images using currently available Explainable Artificial Intelligent (XAI) techniques. This method in-cooperates multiple techniques such as LISA aka Local Interpretable Model Agnostic Explanations (LIME), integrated gradients, Anchors and Shapley Additive Explanations (SHAP) which is Shapley values-based approach to provide explanations for the predictions provided by Blackbox models. This unified method increases the confidence in the black-box model's decision to be employed in crucial applications under the supervision of human specialists. In this work, a Chest X-ray (CXR) classification model for identifying Covid-19 patients is trained using transfer learning to illustrate the applicability of XAI techniques and the unified method (LISA) to explain model predictions. To derive predictions, an image-net based Inception V2 model is utilized as the transfer learning model.

II. Introduction

With the rapid spread of COVID-19 and its variants, accurate and highly available alternative methods to identify patients should be Incorporated with reliability and accountability. Currently, available explanation methods for image classification have their inherent drawbacks and strategies when providing explanations by using current XAI techniques. The black box models are machine learning models whose predictions cannot be explained solely by the data that feeds the algorithm. The collaborative approach of exciting methods with experts intervention for the predictions made by the black box model could increase the reliability and trustworthiness of the predictions.

Chest X rays are commonly used to diagnose coronavirus patients with a precise approach and can be used as an alternative to PCR (Polymerise Chain Reaction) [1, 2]. Although the model's accuracy is substantial, it is very significant that the predictions are explained in the context of health-related predictions. CNNs are black-box models, and the characteristics that influence predictions must be consistent with human intuition when making judgments in critical systems.

LIME is a commonly used approach that is presently integrated into explainable AI. However, despite its simplicity of use in image classification explanation, it cannot offer a generalized explanation of the predictions. Anchors are an if-then rule-based technique of explaining things. The Anchors, on the other hand, are not always guaranteed to generate explanations for every input. On the other hand, integrated gradients satisfied the following axioms namely implementation invariance and completeness which assure fair explanations and mitigates the drawbacks associated with the previous attribution based methods [3, 4]. Shapley values follow four basic properties namely Efficiency, Symmetry, Dummy and Additivity which guarantee the fair representation of a given feature associated with the decision [5, 6]. The calculation of shapely values for a given feature is time-consuming and resource-intensive compared to other methods.

In this study, currently, accessible methodologies are used; nevertheless, the explanations offered by the aforementioned approaches are not always in agreement with one another. LISA considers the final explanations to be the union or intersection of explanations supplied for the picture by the explanation techniques to a particular prediction.

III. Related Work

A. Overview

The resemblance of attributions to an observer's expectations, on the other end, is a more practical measure of an attribution system. It requires the use of a human expert for a specific task and involves observer bias, since approaches closer to the observer's expectation may be chosen on behalf of those who explain the model behaviour. The majority of the medical imaging research that explored the interpretability of deep learning algorithms employed attribution-based approaches due to their simplicity [3]. JCS (Joint Classification and Segmentation) uses segmentation and activation maps to create decision explainability [1]. It has become important to screen a significant number of people to identify those who are infected and reduce the spread of the COVID-19 virus. A real-time PCR is a common diagnostic tool for pathological testing. There are failure cases for this tool because it produces more false test results, demanding the search for an alternative tool. For COVID-19 screening, chest x-rays are a preferable option to PCR. However, in this case, the accuracy of results is critical [2]. BS-Net is an explainable CNN that uses lung segmentation and Brixia score [7] to explain COVID-19 chest X-rays [8]. There are several additional models for the detection of patients of COVID-19 without incorporating frameworks of explanation [9, 10] that are based on transference learning models.

B. LIME

LIME explanations work by learning which lines the best match the model with a certain local weighting [11]. Local

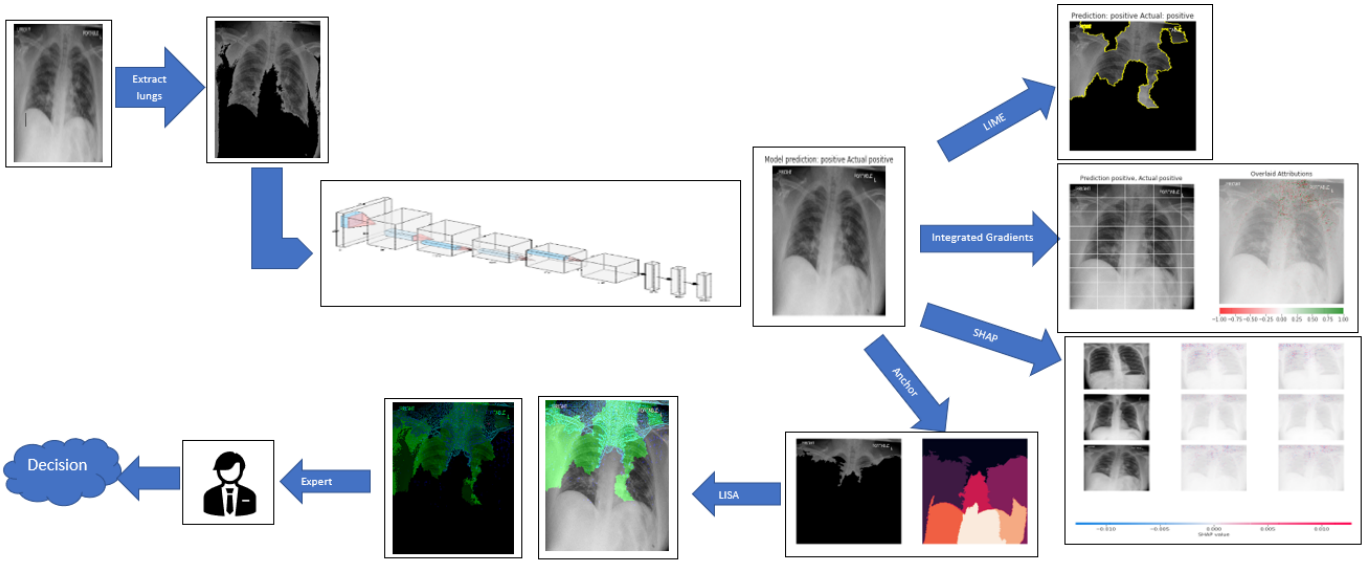


Fig. 1: Proposed Framework

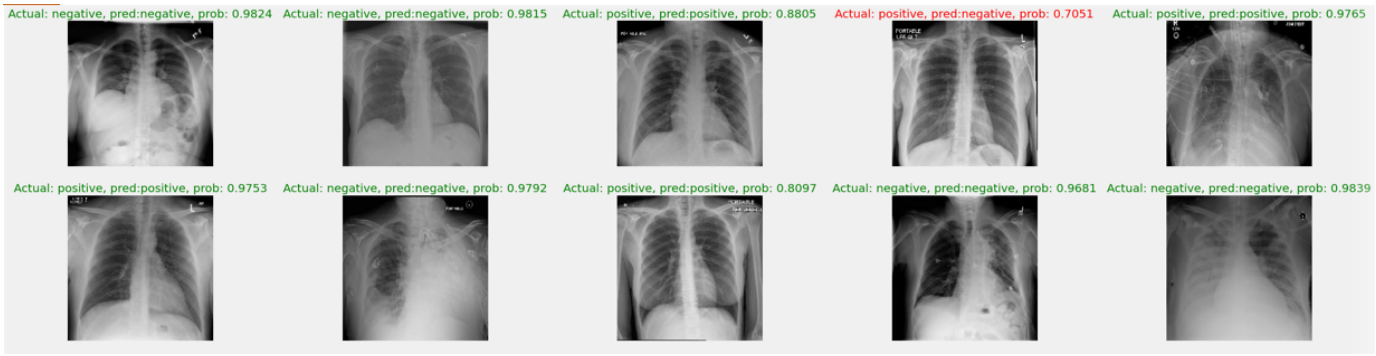
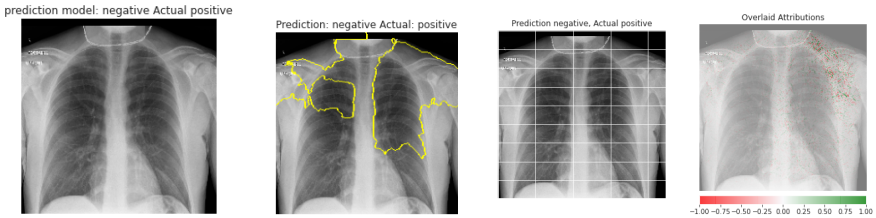


Fig. 2: Transfer Learning model results

weighting in images indicates the influence of superpixels on the decision. Lime generates changes by splitting the picture into ”superpixels” and toggling the superpixels on and off. Superpixels are linked colour pixels that may be turned off by replacing each pixel with a user-defined colour. In each permutation, the user may also choose the likelihood of turning off a superpixel. Although the LIME interpretations are simple to use and can use interpretable features derived from data instances besides those on which the original model was trained to provide explanations [5], the authors of an article [12] demonstrated that the explanations of two very close points varied greatly in a simulated setting. The data scientist can change LIME explanations to hide biases, as seen in [13]. The possibility of tampering makes accepting LIME-generated explanations more challenging.

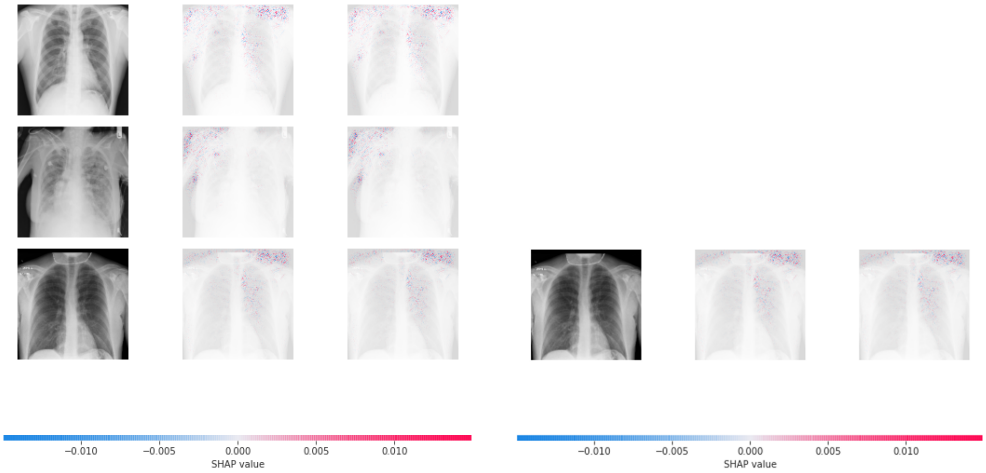
C. Integrated Gradients

This method is based on two fundamental axioms: sensitivity and implementation invariance, and it does not require network instrumentation and can be computed rapidly with a few calls to the gradient function in CNNs[4]. This approach addresses the sensitivity issue, which is an indicator of the influence of the size of the input on the prediction even though it begins higher than the decision function [4]. Attribution techniques should be implementation invariant, which implies that attributions for two functionally equivalent networks should always be the same [4] even widely different implementations. If the outputs of two networks are the same for all inputs, they are functionally equivalent. Because attribution methods that fail to meet Implementation Invariance may be sensitive to minor elements of the models [4]. Integrated gradients satisfy the completeness axiom, which states that the attributions add up to the difference between the output of function F(x) at the input x and the baseline [4] where x and x's are superpixel value and baseline respectively. The baseline is employed in the explanation to keep the counterfactual understanding intact. Choosing a solid baseline is an important step in using integrated gradients [4, 14, 15]. Even though in general gradient-based models are faster than model-agnostic methods it is difficult to know whether an explanation is a correct and huge part of the evaluation is only qualitative [5] i.e. may not comply with human knowledge.



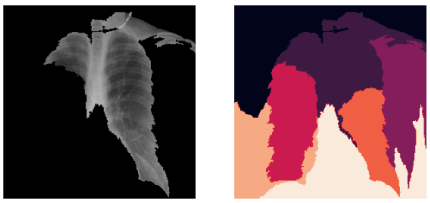
(a) False Negative Result (b) False Negative Result LIME explanation (c) False Negative Result Integrated gradients explanation

Fig. 3: False negative results explanation



(a) False Negative Result, SHAP Integrated gradients explanations with background samples (b) False Negative Result, SHAP Integrated gradients explanations

Fig. 4: False negative results explanation



(a) False Negative Result Anchor explanation

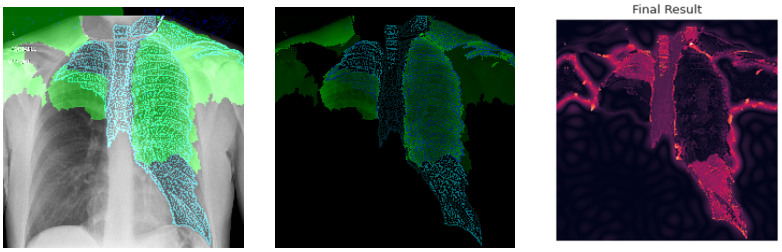
Fig. 5: False negative results explanation

D. SHAP integrated gradients

Explains a model that uses expected gradients, an extension of the integrated gradients technique, and a method for differentiable models based on Shapley values that are extended into infinite player games [16]. The integrated gradient values are somewhat different from SHAP values and need the integration of a single reference value. As an adaptation to approximate Shapley values, the expected gradients recreate the integral as expectations and combine this expectation with the sampling reference values from the background data set [16]. This results in a single combined gradient expectation that converges to attributions that amount to the difference between the expected model output and the predicted output [16].

E. Anchors

Anchors are built using if-then logic [17]. Superpixels will be generated using the segmentation function. To provide a helpful explanation, meaningful superpixels are required. The pixels in the segmentation that are not in the proposed anchor will take the average value of their superpixel. Another possibility is to superimpose pixel values from another image. Images, like LIME, are first divided into superpixels to pre-

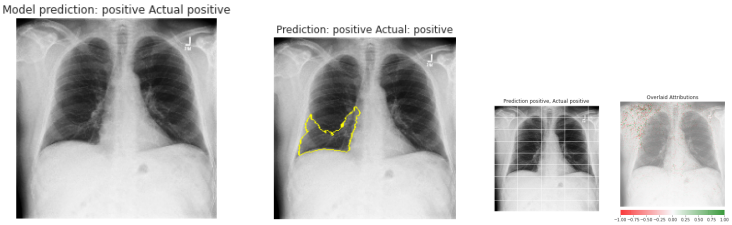


(a) False Negative LISA union explanation

(b) False Negative LISA intersection explanation

(c) False Negative LISA explanation decision boundary

Fig. 6: False negative results explanation



(a) Model prediction on failed Anchor explanation

(b) Lime explanation on failed Anchor explanation

(c) Integrated Gradient explanation on failed Anchor explanation

Fig. 7: Positive result anchor failed to explain

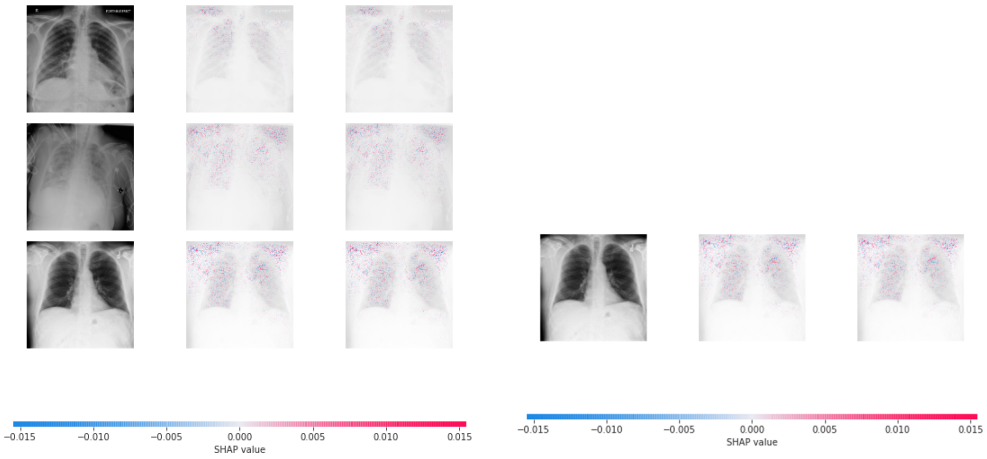
serve local picture structure. The existence or absence of each superpixel in the anchor then constitutes the interpretable representation. To arrive at interpretable explanations, meaningful superpixels must be generated. The algorithm supports various basic picture segmentation algorithms such as felzenszwalb, slice, and rapid shift. It also allows the user to provide their segmentation algorithm. Individual predictions are explained by local model-agnostic explanations instead of the whole model at once. These methods include a trade-off: each explanation is simple to grasp even for complex models and tasks, but only captures the model's behaviour on a particular portion of the input space. Anchor approach also belongs to the prior mentioned category of local explanations[17]. An anchor explanation is a rule that sufficiently "anchors" the prediction locally. So that changes to the instance's other feature values are insignificant. In other words, the prediction is almost always the same while the anchor holds [17]. The anchor approach combines the advantages of local model-agnostic explanations with the interpretability of rules, which are built in such a way that they best facilitate human understanding [17]. The fundamental benefit of employing anchors is that they are simple to interpret, model-agnostic, and highly efficient because they can be parallelized. The significant disadvantages are that coverage is unclear in some domains, for example, there is no simple or universal description of how superpixels in one image compared to those in other images. In rare circumstances, anchors are unable to generate explanations [16].

IV. Methodology

Because publicly available COVID-19 chest x-ray data sets are hard to come by, and the volume of data relevant to corona patients is limited in datasets, this model was created using transfer learning. In the image preprocessing phase of the ML model training, a custom design mask is used to extract the sections of lungs from the X-ray imagery. The mask is created using a Gaussian filter and Otsu thresholding. Here the Gaussian filter is used to smooth the CXR image and Otsu thresholding is used to separate the lungs from the CXR image. The trained model and CXRs are then passed to explanation generation algorithms to create explanations. Integrated gradients technique, LIME based image explanation method, Shapley values-based Integrated gradient-based method, and Anchor rule-based method are among the explanation approaches. These techniques give explanations that are distinct from one another. For the above purpose, the LISA method proposed in this paper can be used. It unified the explanation into one single explanation. The architecture of the explanation providing neural network and its components are as follows.1

And the algorithm used to combine the explanations are as follows 1. This algorithm uses union and intersection of the explanations provided by the existing methods. Other than providing union and intersection of the explanations by utilizing Fourier transformation and high pass filtering it generates the decision boundary of the LISA explanations.

High time consumption to provide explanations and the in-



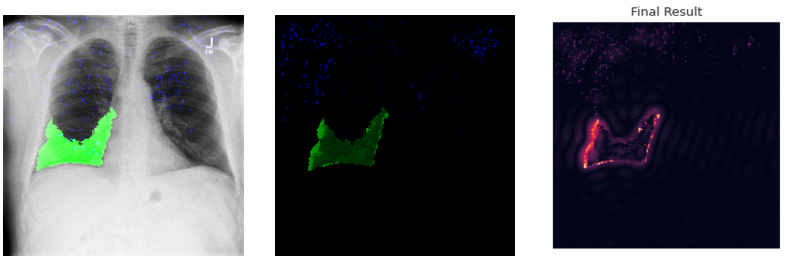
(a) SHAP explanation with background samples on failed Anchor explanation

(b) SHAP explanation on failed Anchor explanation

Fig. 8: Positive result anchor failed to explain

(a) Failed Anchor explanation

Fig. 9: Positive result anchor failed to explain



(a) Positive result LISA union explanation

(b) Positive result LISA intersection explanation

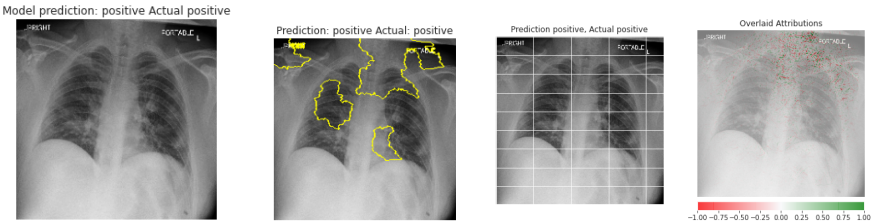
(c) Positive result explanation LISA passed through high pass filter

Fig. 10: Positive result anchor failed to explain

ability to generate anchors for certain predictions is the major downside of the approach for evaluating models. It might be due to X-ray images and in anchor explanation creation and the usage of the segmentation function. Even though the felzenszwalb graph-based approach is widely utilized in the field of medical picture segmentation [18] this approach was not possible due to a shortage of memory resources and used slic segmentation with Anchors to experiment. The novel coronavirus can infect either the upper or lower respiratory tract. It makes its way through your airways. The lining might become swollen and irritated. The infection can go all the way down into your alveoli in some circumstances. If the model forecasts understand these anomalies, then the model, which is based on human knowledge, may be utilized in production contexts for prediction. The domain expert within the loop plays a critical role in the system by ensuring the model predictions and the framework explanations are rely upon with expert knowledge.

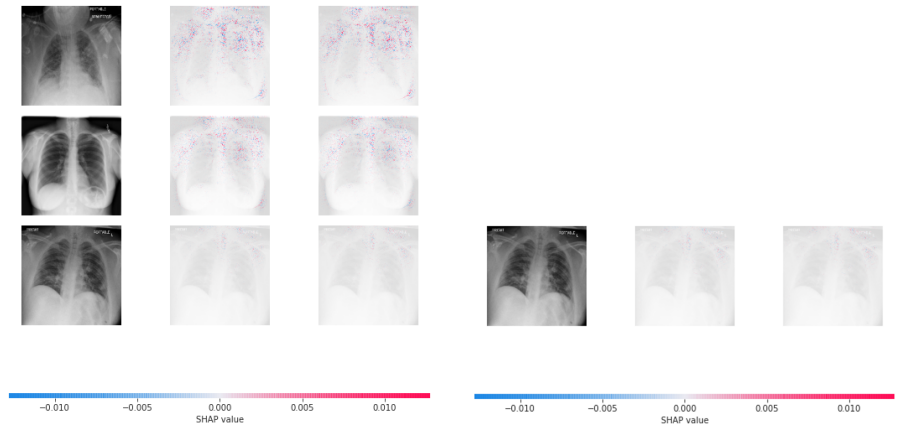
V. Results

With the limited availability of data, we were able to create fairly accurate CNN with transfer learning to check the explainable framework. The model able to achieve modest testing accuracy of more than 90 % with the dataset click



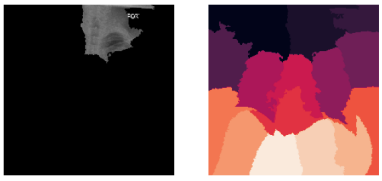
(a) Positive result (b) Positive result explanation by LIME (c) Positive result explanation by Integrated Gradients

Fig. 11: Positive results explanation



(a) Positive result explanation by SHAP with background data points (b) Positive result explanation by SHAP

Fig. 12: Positive results explanation



(a) Positive result explanation by Anchor

Fig. 13: Positive results explanation

[19]. The following are some predictions made by the model on the random testing data.

The major shortcoming when predicting patients, as seen in the image, is the model's false-positive and false-negative results. Using the framework, the domain expert is able to identify the parts of the CXR that are responsible for the prediction. By assessing those regions, the model can be fine-tuned with new data, changed learning parameters, or introduced a more sophisticated image segmentation approach in preprocessing stage.

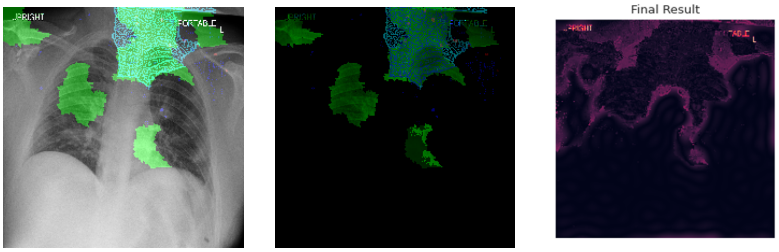
For a particular model prediction, the explanations offered by the framework can be depicted as follows.

A. *False Negative Result Analysis*

The interpretations provided to false negative result by the model as follows 3a.

LIME explanations are capable of producing intuitive explanations. Even though the prediction is inaccurate in this case, if the chosen region is comparable to the CXR of a healthy patient, the model is consistent with human understanding. Otherwise, the model selected incorrect features throughout the training phase and will need to be retrained. Lime indicates the left region and right upper region of the lungs as the critical region for the decision 3a.

1) Integrated Gradient: Even though the explanation region is distributed within the left side of the X-ray it isn't



(a) Positive result explanation by LISA union

(b) Positive result explanation by LISA intersection

(c) Positive result explanation LISA passed through high pass filter

Fig. 14: Positive results explanation

Algorithm 1 Extract Explanation from Available Methods

if ANCHOR **then**
Add to Explanations
end if
if SHAP **then**
Add to Explanations
end if
if LIME **then**
Add to Explanations
end if
if INTEGRATED GRADIENTS **then**
Add to Explanations
end if

Take Union of the Explanations
$LISA$ = $ANCHOR$ $\cup$ SHAP $\cup$ LIME $\cup$ INTEGRATED GRADIENTS

Image Blending
$G_1(x) = (1-\alpha) f_{(0)}(x) + \alpha f_{(1)}(x)$

where α *is* $Transparency$,
$f_{(0)}(x)$ *is* $LISA$ *and* $f_{(1)}(x)$ *is* $Predicted\ Image$

Bit-wise and of Image and LISA
$G_{(2)}(x) = f_{(0)}(x)\ \&\ \alpha f_{(1)}$

LISA Fourier Transformation and High Pass Filtering
$G_3(x) = HPF(DFT(LISA))$
where DFT is Discrete Fourier Transformation,

where HPF is High Pass Filtering,

Return $G_1(x), G_2(x), G_3(x)$
=0

intuitive as the LIME explanation 3c. The explanations tend to concentrate towards the left upper region of the lungs.

2) SHAP: For this image, the SHAP integrated gradients method won't able to produce meaningful results along with sample known inputs 4a. SHAP results also somewhat comply with the results of the Integrated gradient 4b.

3) Anchor: In this instance, Anchor explanations are able to generate a fairly intuitive explanation for the prediction 5a by providing most of the left side of the lungs as the explanation for the prediction.

4) LISA: The newly proposed unified method is able to extract the essence of the explanations provided by the above methods and produce a Blended image explanation as follows 6a,6b,6c

B. Positive Result Analysis Where Anchor Explanations Failed

In this particular positive result analysis Anchor explanation is unable to create anchors on this given input 7a.

LIME is able to produce some intuitive explanation regions for the instance 7b. LIME explanations capture the right lower regions of the lungs which comply with human knowledge when diagnosing the disease.

1) Integrated Gradient: For this instance Integrated gradients unable to produce intuitive results 7c. The explanations are intense in the upper and middle portions of the right lung.

2) SHAP: For this instance, SHAP is unable to produce intuitive results 8a. The explanations of the predictions are distributed over the whole CXR 8b.

3) Anchor: The failed Anchor explanations 9a.

4) LISA: The proposed method is able to extract the essence of the explanations provided by the above methods and produce a blended image explanation and 10a intersection of the explanations 10b and LISA decision boundary 10c as follows. The explanations appear to be based on human domain expertise.

C. Positive prediction

In this instance, all the explanation methods except SHAP are able to generate intuitive results 11a.

1) LIME: The lime explanations are able to provide the upper left corner of the lungs as the explanation 11b. It also gives an explanation for the lower right corner of the right lung that is compatible with human comprehension.

2) Integrated Gradient: The Integrated Gradients are also able to provide condensed explanations which somewhat comply with the explanations provided by the LIME 11c.

3) SHAP: For this instance, SHAP is unable to produce intuitive explanations 12a,12b.

4) Anchor: The anchor explanations coincide with the explanations provided by the LIME 13a. Both the explanations seem to comply with human intuition.

5) LISA: The LISA method is able to unify the explanations for the above instance and provide a blended image as the explanation for the prediction 14a,14b and the decision boundary 14c.

Aside from utilizing the technique as the only explanation providing method, the four explanation providing ways can be combined to guarantee that the forecast region is comply with human experts' expertise.

VI. Conclusion

Since the trustworthiness of the CNN in medical diagnosis is essential, this framework can be used to evaluate model performance and determine whether the features included in the prediction are consistent with human knowledge and intuition. As a result, this framework is suitable for model debugging in mission-critical domains.

For the reproducibility of the results the code is available in the following repository link.

References

[1] Yu-Huan Wu et al. "JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation". In: *IEEE Transactions on Image Processing* 30.1 (2021), pp. 3113–3126. DOI: 10.1109/TIP.2021.3058783.

[2] Bhukya Jabber et al. "Detection of Covid-19 Patients using Chest X-ray images with Convolution Neural Network and Mobile Net". In: 2 (Dec. 2020), pp. 1032–1035. DOI: 10.1109/ICISS49785.2020.9316100.

[3] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. "Explainable Deep Learning Models in Medical Image Analysis". In: *Journal of Imaging* 6.3 (2020). ISSN: 2313-433X. DOI: 10.3390/jimaging6060052. URL: https://www.mdpi.com/2313-433X/6/6/52.

[4] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: Proceedings of Machine Learning Research 70.4 (Aug. 2017). Ed. by Doina Precup and Yee Whye Teh, pp. 3319–3328. URL: https://proceedings.mlr.press/v70/sundararajan17a.html.

[5] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 5. https://christophm.github.io/interpretable-ml-book/. 2019.

[6] Scott Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *CoRR* abs/1705.07874.6 (2017). arXiv: 1705.07874. URL: http://arxiv.org/abs/1705.07874.

[7] Andrea Borghesi and Roberto Maroldi. "COVID-19 outbreak in Italy: experimental chest X-ray scoring system for quantifying and monitoring disease progression". In: *La radiologia medica* 125.7 (2020), pp. 509–513. DOI: 10.1007/s11547-020-01200-3.

[8] Alberto Signoroni et al. "BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset". In: *Medical Image Analysis* 71.8 (2021), p. 102046. DOI: 10.1016/j.media.2021.102046.

[9] Narayana Darapaneni et al. "Inception C-Net(IC-Net): Altered Inception Module for Detection of Covid-19 and Pneumonia using Chest X-rays". In: *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)* 9 (2020). DOI: 10.1109/iciis51140.2020.9342741.

[10] Xin Li, Chengyin Li, and Dongxiao Zhu. "COVID-MobileXpert: On-Device COVID-19 Patient Triage and Follow-up using Chest X-rays". In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 10 (2020). DOI: 10.1109/bibm49941.2020.9313217.

[11] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* 11 (2016). DOI: 10.18653/v1/n16-3020.

[12] David Alvarez - Melis and Tommi S. Jaakkola. "On the Robustness of Interpretability Methods". In: *CoRR* abs/1806.08049.12 (2018). arXiv: 1806.08049. URL: http://arxiv.org/abs/1806.08049.

[13] Dylan Slack et al. "Fooling LIME and SHAP". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 13 (2020). DOI: 10.1145/3375627.3375830.

[14] Avanti Shrikumar et al. "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences". In: *CoRR* abs/1605.01713.14 (2016). arXiv: 1605.01713. URL: http://arxiv.org/abs/1605.01713.

[15] Alexander Binder et al. "Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers". In: *Artificial Neural Networks and Machine Learning – ICANN 2016 Lecture Notes in Computer Science* 15 (2016), pp. 63–71. DOI: 10.1007/978-3-319-44781-0_8.

[16] Mukund Sundararajan and Amir Najmi. "The Many Shapley Values for Model Explanation". In: Proceedings of Machine Learning Research 119.16 (July 2020). Ed. by Hal Daumé III and Aarti Singh, pp. 9269–9278.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Anchors: High-Precision Model-Agnostic Explanations". In: *Proceedings of the AAAI Conference*

on Artificial Intelligence 32.17 (Apr. 2018). URL: https://ojs.aaai.org/index.php/AAAI/article/view/11491.

[18] Luqman Subki et al. “A REVIEW ON MEDICAL IMAGE SEGMENTATION: TECHNIQUES AND ITS EFFICIENCY”. In: 7 (Jan. 2017), pp. 59–82.

[19] Linda Wang, Zhong Qiu Lin, and Alexander Wong. “COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images”. In: *Scientific Reports* 10.1 (Nov. 2020), p. 19549. ISSN: 2045-2322. DOI: 10.1038/s41598-020-76550-z. URL: https://doi.org/10.1038/s41598-020-76550-z.