IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

EXPLAINING AUTONOMOUS PNEUMONIA DISEASE DETECTION USING GRAD-CAM FOR IMPROVED DECISION SUPPORT

[1]Om Chavan, [2]Yash Bijoor, [3]Chahat Kothari, [4]Arjun Jaiswal

[123]B.Tech Students, [4]Assistant Professor

Department of Information Technology,

Dwarkadas J. Sanghvi College of Engineering, Mumbai, India

Abstract: In recent years, artificial intelligence and machine learning has made major strides in the field of medicine. The medical sector requires a high level of accountability and transparency. Reliable machine decisions and predictions require explanations to justify their outcomes. This necessitates increased interpretability and a deeper understanding of the underlying mechanisms of algorithms. However, the black box nature of deep learning remains a challenge, leading to limited comprehension of many machine decisions. The reason radiologists are weary of using AI is because they do not trust model predictions without any form of explainability. Thus, we aim to create a system that not only focuses on interpretability and explainability but also has a high enough accuracy to make it reliable enough to be trusted and used by radiologists.

Index Terms–artificial intelligence, machine learning, black box, explainability, transparency, interpretablility

I. INTRODUCTION

Explainable AI (XAI) is a field of study and development that focuses on making artificial intelligence (AI) systems interpretable, transparent, and understandable to users. The goal of XAI is to create artificial intelligence systems that can explain their decision-making processes, reasoning and actions in a way that is easily understood by humans. They aim to create more explainable models that still perform at a high level.

Medical imaging is one area where the application of XAI can have a significant impact on patient care. In medical imaging machine learning algorithms are increasingly being used to support diagnosis and treatment decisions. These algorithms, however, can sometimes produce results that are difficult for clinicians to interpret and understand, especially when they involve complex pattern recognition or deep learning. By making AI tools more explainable, healthcare professionals can better understand the reasons behind AI-generated diagnoses, allowing them to make mo re informed decisions.

For example, in radiology, XAI can be used to explain the decisions made by AI algorithms that analyze medical images such as X-rays, MRI scans, and CT scans. Using XAI techniques such as feature attribution or model visualization, these algorithms can provide an explanation for why they identified a particular condition or disease and highlight specific features in the image that led to the diagnosis.

XAI can be used to increase the transparency and accountability of AI systems in medical imaging. By providing explanations for its decisions, it can help address concerns about the potential for artificial intelligence to perpetuate existing biases or inaccuracies in medical diagnoses. [1]

II. LITERATURE REVIEW

Literature Related to Existing Systems

X-ray Classification [Pneumonia/Normal]

It is a system[2] where a user can upload an image of a chest X-ray and receive a diagnosis of it being either Pneumonia or Normal. It is a CNN model for predicting pneumonia. The main issue with the system was that it failed to give the user an explanation for the diagnosis. This does not allow the user to be able to trust the system entirely

Chester: The Radiology AI Assistant

Link: https://mlmed.Org/equipment/xray/

A web-based application called Chester is described in "Chester: A Web Delivered Locally Computed Chest X-Ray Disease Prediction System"[3]. It addresses the privacy and connection issues with cloud-based solutions by using locally computed deep learning models to predict illnesses in chest X-ray pictures. Through the web interface, users upload photos, and the system diagnoses illnesses including pneumonia and tuberculosis. Evaluation reveals competitive sensitivity, specificity, and accuracy. Chester, which runs locally on customers' computers, provides a safe and convenient substitute for viewing chest X-ray images. The main challenge with the system is that it identifies multiple diseases simultaneously often, which is a rare case in the medical industry.

Literature Related to Methodology/ Approaches/ Algorithms

CheXNet

Pranav Rajpurkar and Jeremy Irvin introduced an algorithm - CheXNet - in their paper “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning”[4]. They developed this algorithm to detect pneumonia from chest X-rays at a level that exceeded practising radiologists. ChexNet algorithm is trained on ChestX-ray14, the largest publicly available chest X-ray dataset. It contains over 100,000 frontal view X-ray images of over 30,000 unique patients. CheXNet is a 121-layer CNN that inputs a chest X-ray image and outputs the probability of pneumonia along with a heatmap localizing the areas of the image most indicative of the disease. The dataset was divided into training (70%), testing (20%) and validation (10%) randomly but without any patient overlap between the sets. For comparison against radiologists, they collected a test set of 420 frontal chest X-rays. The four radiologists had 4,7,25, and 28 years of experience and were asked to label each X-ray. The labels for the radiologists and the model were entered into a standardized data entry program. Neither the radiologist nor the model had any access to any patient information. F1 score (with 95% CI) of individual radiologists and CheXNet was computed against each of the other 4 labels as ground truth. As a result, CheXNet achieves an F1 score of 0.435 which is higher than the radiologist average of 0.387 (0.383, 0.356, 0.365, 0.442)[4].

Explainability

In recent years, deep learning models have made a lot of progress, giving us a lot of insights. The reliability in these models has also increased, due to high levels of accuracy. But a few sectors, like healthcare, need transparency as well as reliability. Unfortunately, the black box mechanism, while it ensures high reliability and accuracy, fails to be transparent towards the user. As these models grow more and more complex, understanding the working of the model becomes more and more difficult-almost impossible. To combat this issue, explainability is implemented to make the model more transparent in its working. Many papers have examined various elements of XAI, such as its history, study fields, methodologies, obstacles, and potential [5].

The study by Erico Tjoa and Cuntai Guan provides a summary of the current condition of XAI in the medical industry. The authors evaluate the benefits and drawbacks of various XAI methodologies. They also discuss the opportunities and hurdles of XAI in medical applications, such as transparency and interpretability in decision-making[6].

The paper by Feiyu Xu, Hans Uszkoreit, Yangzhou Du, and Wei Fan offers a thorough overview of the background, study areas, methodologies, and difficulties in the field of XAI. The authors investigate the benefits and drawbacks of several XAI methodologies; including rule-based systems, decision trees, and model-based methods. They also talk about the limitations of XAI, including the necessity for uniform evaluation measures as well as the trade-off between interpretability and accuracy[1].

In the study by Shaw-Hwa Lo and Yiqiao Yin, they provide a novel way for producing understandable justifications for pneumonia chest X-ray images. Experts and non-experts collaborate with the model to create as well as improve the explanations using the authors' "human in the loop" methodology. The study shows how well this method produces precise and understandable interpretations for chest X-ray images of pneumonia [7].

Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever's study offers a summary of the state of XAI in deep learning-based medical image processing at the moment. The authors examine the advantages and disadvantages of several XAI strategies, including saliency maps and attention processes. They also draw attention to the difficulties and possibilities of XAI in medical picture analysis, such as the necessity of interpretability and openness in decision-making[8].

In summary, these publications illustrate the obstacles and prospects in the field of XAI while giving a thorough review of its state in numerous fields, including medical image analysis. They also describe several methodologies and examine their advantages and disadvantages, and lay the groundwork for more in-depth investigations in the future. The field of XAI undergone massive breakthroughs and has the potential to increase the accountability and transparency of AI systems in the coming decades.

LIME

Local Interpretable Model-Agnostic Explanations, or LIME, is a set of rules that can explain the predictions of any classifier, by using approximating it regionally with an interpretable version. It is a python library based totally on the paper “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”[9]. The LIME model tries to locate the explanation of the black box machine getting to know model with the aid of approximating the nearby linear conduct of the model. It is a popular library that may be applied to any black-container version to generate evidence. LIME tries to interpret the model thru these 4 steps:

1. Input information permutation: - The first step LIME does is it creates numerous artificial data near the information that is to be explained. If the input is a photograph, then LIME will generate numerous samples which might be comparable with our enter photograph via turning on and off a number of the excellent-pixels of the photo.

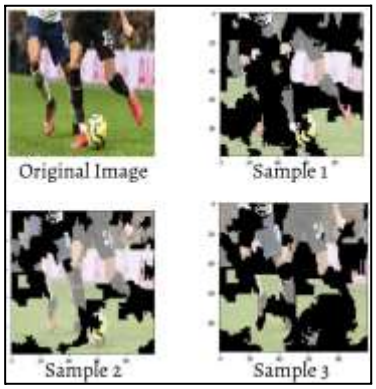

Fig. 2.1 Steps in LIME

2. Predict the class of every synthetic records point: Next, LIME predicts the elegance of each of the synthetic data factors which have been generated the usage of the trained model. If the input information is an image, then the prediction of each perturbed photograph may be generated at this stage.
3. Calculate the load of every artificial records point: The next step is to calculate the weight of each artificial statistics to measure its importance. The larger the weight, the bigger the significance of a positive artificial statistics factor. If the enter information is a photo, then the cosine distance among every perturbed image and the original photograph is computed. The extra the similarity between a perturbed image to the unique photo, the larger its weight and importance.
4. Fit a linear classifier to explain the maximum critical features: The remaining step is fitting a linear regression version the usage of the weighted synthetic information factors. After this step, we need to get the fitted coefficient of every characteristic. Now if we kind the coefficient, the capabilities that have large coefficients are those that play a large role in figuring out the prediction of the version.

SHAP

Shapley Additive Explanations (SHAP) is based on Shapley values, a method used to calculate the contributions of each player to the outcome of a game. The Shapley values are calculated with all possible combinations of players. For N players, it has to calculate 2^N possibilities. In the case of Machine Learning, the "players" are the features (or pixels for an image) and the "outcome of the game" is the prediction. Calculation of Shapley values for all the pixels of an image is not optimal due to the large numbers of N. One of the biggest advantages of SHAP is that it can provide both global and local explainability.[10]

GRAD-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) is a visualization method that was developed by Selvaraju et al.[11] to explain the predictions made by deep neural networks and to pinpoint the areas of an image that are most responsible for the prediction. Grad-CAM has been widely used and adopted since it was first introduced.

Grad-CAM has been used to produce visual justifications for the judgments made by deep neural networks in the area of image categorization. Also, it has been used to pinpoint the areas of a picture that are most pertinent to particular classes, such as pinpointing the characteristics of an image that indicate the presence of a bird.

Grad-CAM has been used to locate areas of interest in medical pictures, such as tumors in MRI scans and aberrant areas in mammograms. In order to assess the interpretability and robustness of deep learning models for medical image processing, this method has also been applied.[12]

Overall, Grad-CAM has shown to be a helpful tool for illustrating and elucidating the choices made by deep neural networks. Its versatility and ability to provide visual explanations for complex models make it a valuable asset in various fields, including image classification, object detection, and medical imaging.

Table 2.1 Comparative Analysis of Grad-CAM, LIME and SHAP

Parameters	Grad-CAM	LIME	SHAP
Focus	Visualizing important regions in an input image that contribute to model predictions.	Providing local, interpretable explanations for individual model predictions.	Assigning feature importance to understand the contribution of individual features in the model's predictions.
Methodology	Utilizes gradient information from the model's final convolution layer to compute class activation maps.	Generates local surrogate models to approximate the behavior of the black-box model within a local neighborhood around the prediction.	Based on cooperative game theory, it computes Shapley values that represent the contribution of each feature in different combinations.
Benefits	Provides intuitive heat maps highlighting regions of importance.	Model-agnostic, applicable to a wide range of machine learning algorithms.	Provides a unified framework for feature importance across different model types.
Limitations	Limited to convolution neural networks (CNNs) and does not capture fine-grained pixel-level importance.	May not capture global model behavior and can be sensitive to perturbations in input data.	Computationally expensive for large feature sets and may require approximation techniques.

After comparing these three algorithms for explainability, we found Grad-CAM to be the most ideal for our project on the basis of these parameters.

First, Grad-CAM may be used to apply different neural networks and can show deep neural networks in a visual manner. This makes it a great tool for figuring out how complicated models used in medical imaging make decisions. LIME, on the other hand, has a limited scope for use in medical imaging because it is primarily intended for explaining linear models. Similar to SHAP, it might be challenging to apply to medical imaging datasets due to processing costs, especially for big models.

Second, Grad-CAM has been extensively applied in the field of medical imaging to pinpoint regions of interest, such as tumors in MRI scans or aberrant areas in mammograms. It has been shown to be quite useful in identifying these locations due to its capacity to produce heatmaps for different layers of a neural network. Although SHAP can likewise produce feature importance scores, its visual heatmaps might not be as user-friendly as Grad- CAM's. LIME, on the other hand, is more suited for deriving regions of interest than for explaining specific forecasts.

Lastly, in the paper, 'Evaluating the performance of the LIME and Grad-CAM explanation methods on a LEGO multi-label image classification task', they found that Grad-CAM provides a more detailed insight than LIME from the view of core performance and people seemed to trust Grad-CAM more than LIME.

In conclusion, Grad-CAM is a strong tool for comprehending and analyzing deep neural networks utilized in medical imaging because of its adaptability and capacity to produce visual heatmaps.

III. OUR APPROACH

Dataset

The provided Kaggle dataset, titled "Chest X-ray COVID-19, Pneumonia," contains chest X-ray images pertaining to COVID-19, pneumonia, and normal cases.The dataset is a useful tool for creating and assessing machine learning algorithms for classifying respiratory diseases. The dataset is a set of PNG-formatted grayscale chest X-ray images. It comprises pictures from many places, like medical facilities, online archives, and scholarly publications. COVID-19, pneumonia, and normal are the three categories under which the photos fall.

Images of sufferers with COVID-19 diagnoses are included within the COVID-19 class. These photos document the abnormalities of the lungs delivered on via the viral infection, including consolidations, opacities, or different COVID-19-precise capabilities. The pneumonia class consists of X-rays of people who have diverse forms of pneumonia, whether they had been added on through bacterial, viral, or fungal ailments. These pix exhibit numerous pneumonia-related signs and symptoms of inflammation, like infiltrates or opacities. X-ray scans of humans without any obvious breathing disorders or illnesses make up the regular organization. These images display a healthful lung appearance devoid of any obvious anomalies and serve as a baseline for evaluation. Dataset is organized into 2 folders (educate, check) and both teach and check incorporate three subfolders (COVID19, PNEUMONIA, NORMAL). The Dataset consists of a total of 6432 x-ray pix and look at facts has 20% of general snap shots.

Training Procedure: -

1. Data Preprocessing: The chest X-ray photos are first preprocessed by means of the gadget to create a standardised layout this is appropriate for evaluation. This may also involve scaling, normalising pixel values, and the usage of any required picture filters or improvements.

2. Model Prediction: The trained deep studying version for pneumonia class is then fed the preprocessed photographs. Convolutional neural networks (CNNs) are used by the version to become aware of vital factors inside the photos and offer predictions.

3. Grad-CAM Visualization: The Grad-CAM method is then used to system the selected photos after getting the version's predictions. Grad-CAM creates heatmaps by emphasizing the regions of the chest X-ray that have the maximum effect at the version's judgment. These heatmaps visually imply the regions that make a contribution substantially to the expected classification.

4. LIME Explanation: The LIME approach is used in addition to Grad-CAM to provide neighborhood interpretability for specific predictions. Within small photo patches, LIME develops surrogate models that resemble the behaviors of the CNN. This makes it less complicated to supply extra targeted justifications for the version's selections, enabling medical employees to realize the version's thinking locally.

5. User Interaction: The system's consumer-friendly interface shows the unique chest X-ray pix alongside the Grad-CAM heatmaps and LIME explanations. Users may additionally add X-ray snap shots for analysis and interpretation. They can interact with the system, see the Grad-CAM-highlighted areas of hobby, and appearance over the LIME reasons to learn greater approximately how the model makes selections.

6. Interpretation and Diagnosis: Based at the visualized Grad-CAM heatmaps and LIME explanations, users can interpret the version's predictions and make informed diagnostic choices. They can utilize their domain know-how and knowledge to validate or question the model's findings, imparting an extra layer of clinical judgement to the diagnostic method.

7. By using this method of operation, the system allows clinical employees to make properly-knowledgeable judgments primarily based at the model's predictions while facilitating the detection of pneumonia in chest X-rays.

IV. RESULTS

The validation accuracy of the Grad-Cam model is evaluated to be 81.57 %.
Below are the plots of model accuracy and model loss over 25 epochs.

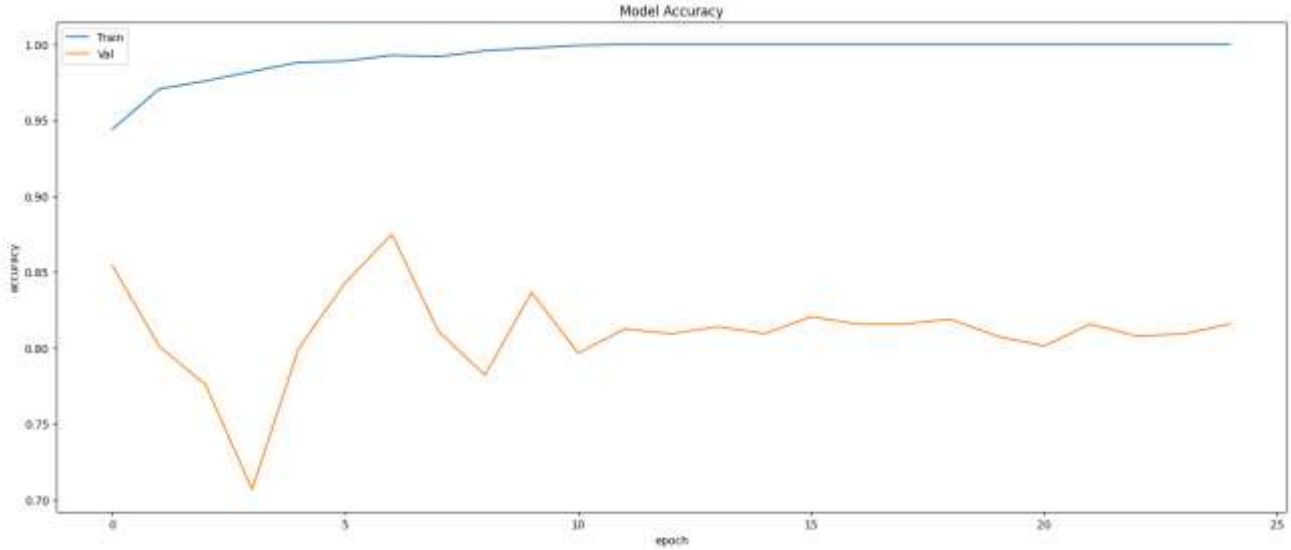

Fig. 4.1 Plot of model accuracy

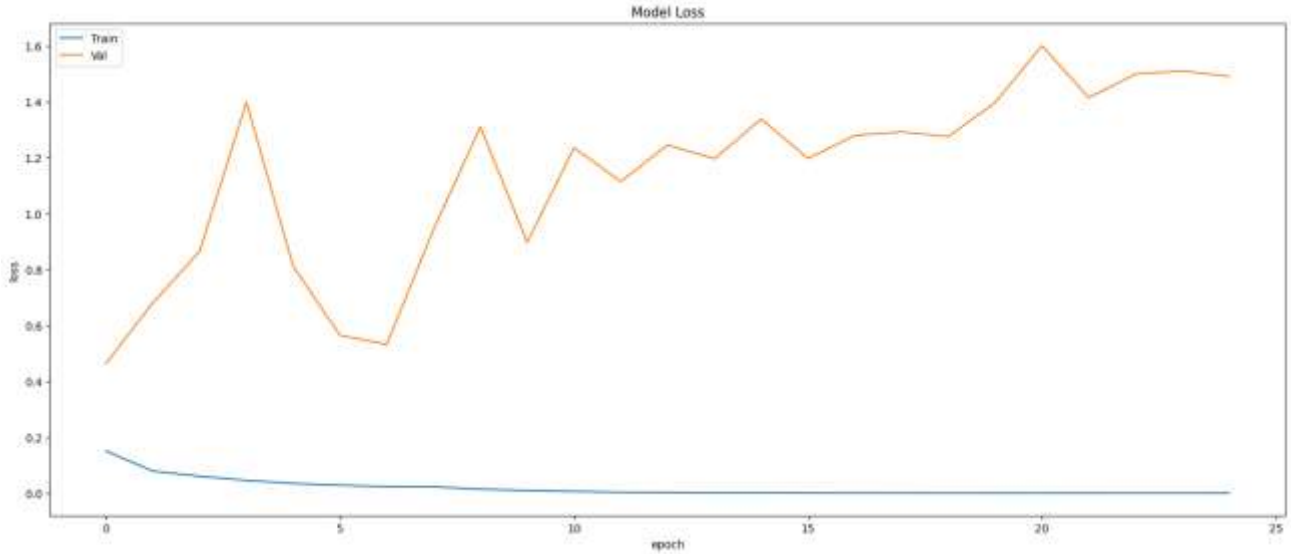

Fig. 4.2 Plot of model loss

1.1 Experimental Results

1. Severe Pneumonia (Bacterial):

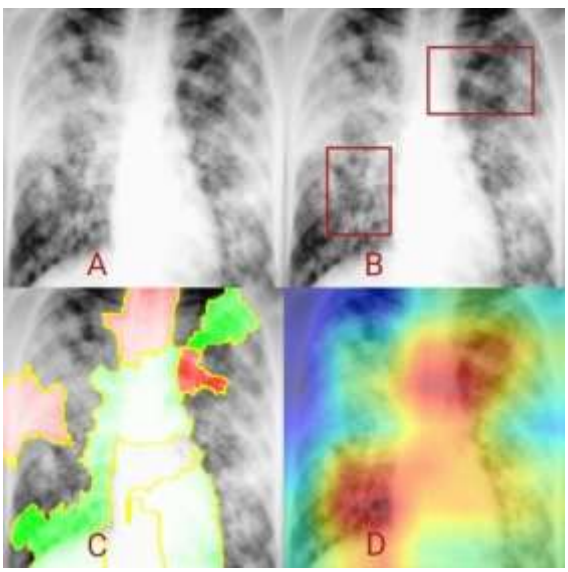

Fig. 4.3

Images classified as severe pneumonia exhibit extensive lung abnormalities, consolidation, and inflammation. These cases represent advanced stages of pneumonia, indicating a severe infection and potential complications. The model's ability to identify and classify these severe cases demonstrates its effectiveness in detecting critical pneumonia conditions. Fig 4.3 A: The original chest X-ray image Fig 4.3 B: Pneumonia region marked and validated by a medical professional Fig 4.3 C: Generated LIME image (Green represents area infected by pneumonia and red represents non-infected regions of the chest) Fig 4.3 D: Generated Grad-CAM heatmap (Red represents area infected by pneumonia and blue represents non-infected regions of the chest. Yellow and green represent mild infection)

2. Mild Pneumonia

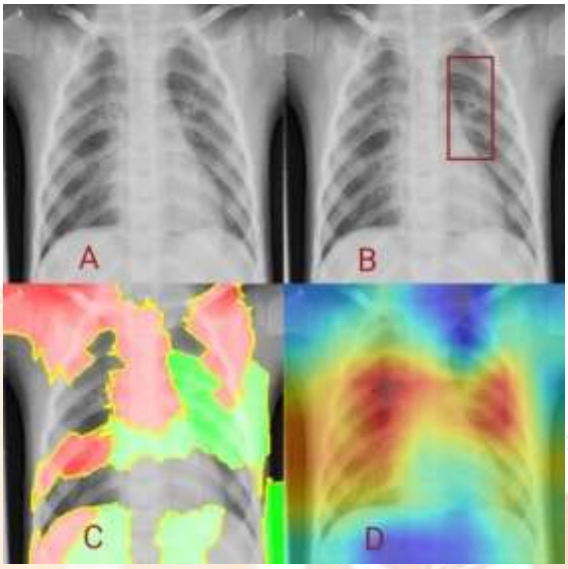

Fig. 4.4

This category comprises images with more pronounced lung abnormalities and infiltrates compared to mild pneumonia cases. The presence of moderate pneumonia suggests a more significant infection and inflammation in the lungs. The model's accurate classification of these cases indicates its capability to detect and distinguish between varying levels of pneumonia severity. Fig 4.4 A: The original chest X-ray image Fig 4.4 B: Pneumonia region marked and validated by a medical professional Fig 4.4 C: Generated LIME image (Green represents area infected by pneumonia and red represents non-infected regions of the chest) Fig 4.4 D: Generated Grad-CAM heatmap (Red represents area infected by pneumonia and blue represents non-infected regions of the chest. Yellow and green represent mild infection)

3. Normal Case

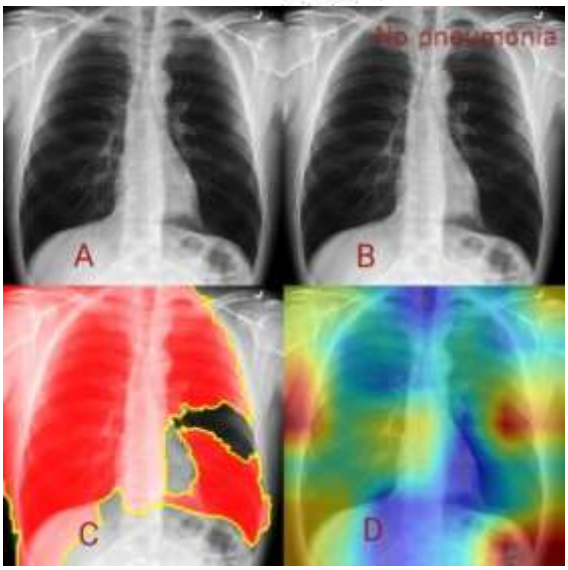

Fig 4.5

This class contains photos of instances in which pneumonia is absent. The model accurately classifies these instances as normal. Fig 4.4 A: The original chest X-ray image Fig 4.4 B: Marked as 'No pneumonia' by a medical professional Fig 4.4 C: Generated LIME image shows that there is no infection (Green represents area infected by pneumonia and red represents non-infected regions of the chest) Fig 4.4 D: Generated Grad-CAM heatmap shows little to no pneumonia in the chest region (Red represents area infected by pneumonia and blue represents non-infected regions of the chest. Yellow and green represent mild infection)

V. CONCLUSION

We present a model that provides suitable explainability with a high accuracy that can be beneficial in healthcare and increase access to chest radiograph expertise. This system has a very wide future scope. It could be used for diagnoses of various other diseases and fractures and many useful features can also be implemented.

VI. ACKNOWLEDGEMENT

We would like to express our deepest gratitude to the many people who supported us throughout the development of this paper. First and foremost, we want to thank our project guide, Assistant Prof. Arjun Jaiswal, for his guidance, advice, and unwavering support throughout this project. His insights and expertise were invaluable in shaping the direction of this research. We are also grateful to our colleagues, who provided valuable feedback and encouragement throughout this project. Their diverse perspectives and expertise helped to shape the ideas presented in this paper

VII. REFERENCES

[1] S.-H. Lo and Y. Yin, "A novel interaction-based methodology towards explainable AI with better understanding of pneumonia chest X-ray images," Discover Artificial Intelligence, vol. 1, no. 1, 2021. doi:10.1007/s44163-021-00015-z

[2] Hardik, "X-Ray Classification [Pneumonia/Normal]" Share.streamlit.io, https://share.streamlit.io/smarthardik10/xrayclassifier/main/webapp.py

[3] P. Bertin and V. Frappier, "Chester: A web delivered locally computed chest X-ray disease prediction system," arXiv.org, https://doi.org/10.48550/arXiv.1901.11210

[4] P. Rajpurkar et al., "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," PLOS Medicine, vol. 15, no. 11, 2018. doi:10.1371/journal.pmed.1002686

[5] E. Tjoa and C. Guan, "A survey on Explainable Artificial Intelligence (XAI): Toward medical xai," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, 2021. doi:10.1109/tnnls.2020.3027314

[6] F. Xu et al., "Explainable AI: A brief survey on history, research areas, approaches and challenges," Natural Language Processing and Chinese Computing, pp. 563–574, 2019. doi:10.1007/978-3-030-32236-6_51

[7] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," Medical Image Analysis, vol. 79, p. 102470, 2022. doi:10.1016/j.media.2022.102470

[8] M. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 2016. doi:10.18653/v1/n16-3020

[9] F. López, "Shap: Shapley additive explanations," Medium, https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3 (accessed May 19, 2023).

[10] R. R. Selvaraju et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization," International Journal of Computer Vision, vol. 128, no. 2, pp. 336–359, 2019. doi:10.1007/s11263-019-01228-7

[11] H. Panwar et al., "A deep learning and grad-cam based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-scan images," Chaos, Solitons & Fractals, vol. 140, p. 110190, 2020. doi:10.1016/j.chaos.2020.110190

[12] P. Patel, "Chest X-ray (covid-19 & pneumonia)," Kaggle, https://www.kaggle.com/datasets/prashant268/chest-xray-covid19-pneumonia (accessed May 19, 2023).