# Using Explainable Artificial Intelligence to Locate Pneumonia

Anthony Novokshanov

**Abstract** Artificial intelligence (AI) has already become a vital resource in numerous industries; however, it is often challenging to understand how AI reaches its results. This lack of transparency, combined with potential biases within the machine learning model, prevents professionals in critical fields like healthcare from relying on deep learning models for diagnostic purposes, hindering the widespread use of AI in healthcare. This paper investigates the application of XAI in the medical field, focusing on the detection of pneumonia through the analysis of lung X-ray scans. In this study, we developed an XAI tool, utilizing a Convolutional Neural Network (CNN) constructed with PyTorch and trained on the Pneumonia MNIST dataset. Our model achieves an accuracy of nearly 91% on 28x28 pixel images and highlights the top pixels considered most important by the deep learning model in its decision-making process. The primary aim of this project is to present a proof-of-concept tool for the integration of XAI into healthcare diagnostics, with the goal of assisting medical professionals in making informed decisions and ultimately saving lives. By demonstrating the feasibility and effectiveness of XAI in pneumonia detection, we lay the groundwork for future advancements in healthcare AI, emphasizing the importance of transparency and reliability in AI models.

**Key Words:** Explainable Artificial Intelligence (XAI), Pneumonia, Convolutional Neural Network (CNN), Healthcare Diagnostics, Interpretability, Model-agnostic

## 1. Introduction

### 1.1 Background

Despite artificial intelligence (AI) advancing rapidly, it faces significant limitations that prohibit it from being used in healthcare. The most serious deficiency of AI's use is that it is not able to explain its methodology when producing outputs [1]. This is a very serious drawback, particularly in contexts closely tied to human life [2]. Industries, such as healthcare, are reluctant to use deep learning models due to concealed biases and potential inaccuracies in their outputs [3]. These factors outweigh the increased efficiency that deep models may bring, hindering their widespread adoption in the field. For this reason, getting machine learning to a point where it can cohesively explain what factors determine a specific output has become a very important factor leading to the development of new explainable artificial intelligence (XAI) models.

XAI models work to give a user a level of transparency that deep learning models are unable to by explaining what happens inside the so-called algorithms' "black box." The term "black box model" refers to an algorithm where the inner workings of the model are not easily interpretable by humans, with logical relationships between inputs and outputs being difficult to establish [4]. To address this transparency issue, techniques such as the Local Interpretable Model-agnostic Explanations (LIME) were developed. As indicated by its name, LIME is model-agnostic, which means that it can be applied to virtually any deep learning model regardless of the underlying structure of the framework [5]. This makes it very convenient to use as it can be used alongside already existing machine learning models without rewriting the very core of their algorithms. LIME works by taking a piece of data as an input and using data augmentation to create many perturbations of the same input [6]. The new instances of the datapoint are then inputted into the

model, and prediction values are recorded. A new local deep learning model is subsequently trained using these predictions and input variations, revealing the features with the most significant impact on predictions for different outputs[6]. Unfortunately, despite its convenience, LIME still lacks specificity for particular problems and may incur computational expenses, especially with more complex models and datasets.

### 1.2 Research Goals

This project aims to develop a specialized version of LIME by tailoring it to a specific issue – detecting and identifying the location of pneumonia in lung X-ray scans. Specifically, it achieves this by identifying pneumonia's location based on the pixel display that had the greatest effect on the output prediction. It works similarly to LIME, but instead of training a new AI model on the changes in prediction values, the algorithm will display the pixels that had the greatest net effect on the prediction values of the original data point. This customization makes the algorithm a powerful tool for medical professionals to use supplementally to identify pneumonia quickly and easily.

A refined version of this algorithm could be used by medical professionals to finally introduce artificial intelligence into healthcare. Although this paper uses chest X-ray scans to locate pneumonia, this algorithm would be able to work just as effectively on locating cancers and other medical conditions from computed tomography scans, MRIs, and more. Our XAI model, designed for collaboration with doctors and nurses, can be used as an important tool to significantly improve the identification of life-threatening conditions, potentially saving human lives.

In practical terms, medical practitioners would upload X-ray scans, and in return, receive precise predictions accompanied by visual analyses of the input data. To navigate ethical considerations, our XAI model is positioned as a supportive tool rather than assuming full responsibility for any misclassifications. So, a licensed doctor's presence would be mandatory, emphasizing the program's role as a tool.

## 2. Materials

### 2.1 Dataset

We used the PneumoniaMNIST dataset for both training and testing the deep learning model. It is a binary collection of labeled chest X-ray scans, with positive readings containing pneumonia and negative readings being healthy scans. These images were compressed to a 28 by 28 pixels resolution to make working with the data faster and more efficient. The dataset has a sample size of 5,856 images with 4,708 being training, 524 being validation, and 624 test samples. The images within the collection are gray-scaled X-ray scans of patient’s torsos, with their necks being at the top of the images and their lower abdomen being at the bottom. Figure 1 shows examples of images from the PneumoniaMNIST dataset.

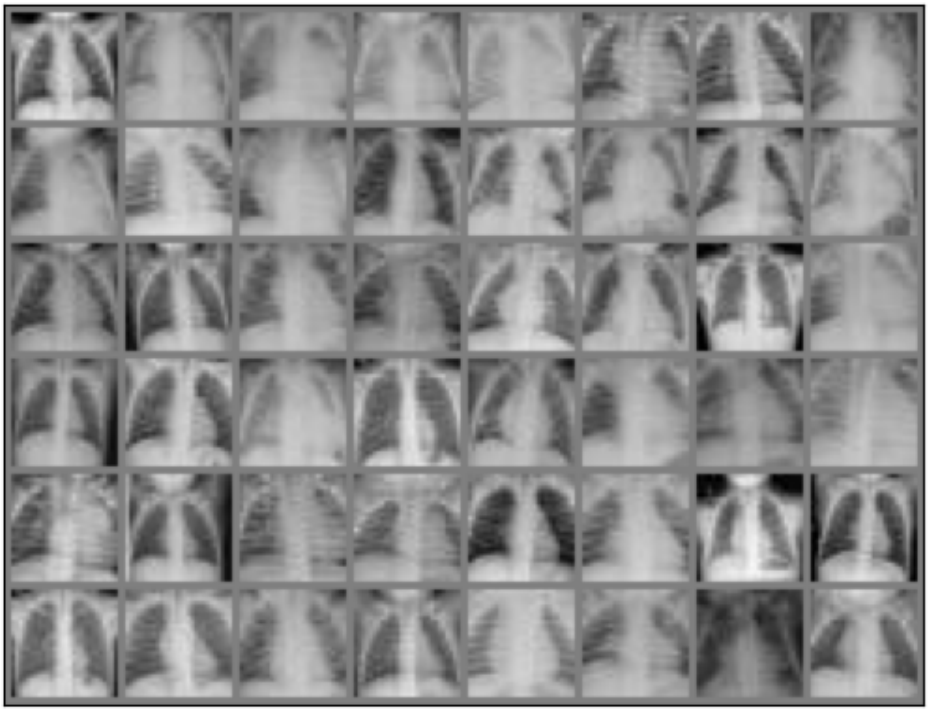

Figure 1: Example of Images from the PneumoniaMNIST Dataset

## 3. Methods

### 3.1 Reproducibility

The study was conducted in Google Collaboratory to ensure the ease of display and access to the code. The program ran using the Intel ® Xeon ® running at 2.3 GHz, but the model was trained with Google's T4 GPU runtime. The model was programmed using PyTorch version 2.1.0 and Cuda 118.

### 3.2 Model Architecture

Our deep learning model is a seven-layer Convolutional Neural Network (CNN) created using the PyTorch library. The network uses Rectified Linear Unit (ReLU) functions to introduce non-linearity into the model, allowing it to learn complex patterns and relationships within the data. The functions used in each layer of the model contain implicit regularization to ensure the network does not overfit the data, which may lead to diminished testing results. We trained the neural network with a 15-epoch training loop, using Stochastic Gradient Descent to update the parameters of the neural network and backpropagation to ensure efficient training of an accurate algorithm. Figure 2 shows the Convolutional Neural Network used in this study.

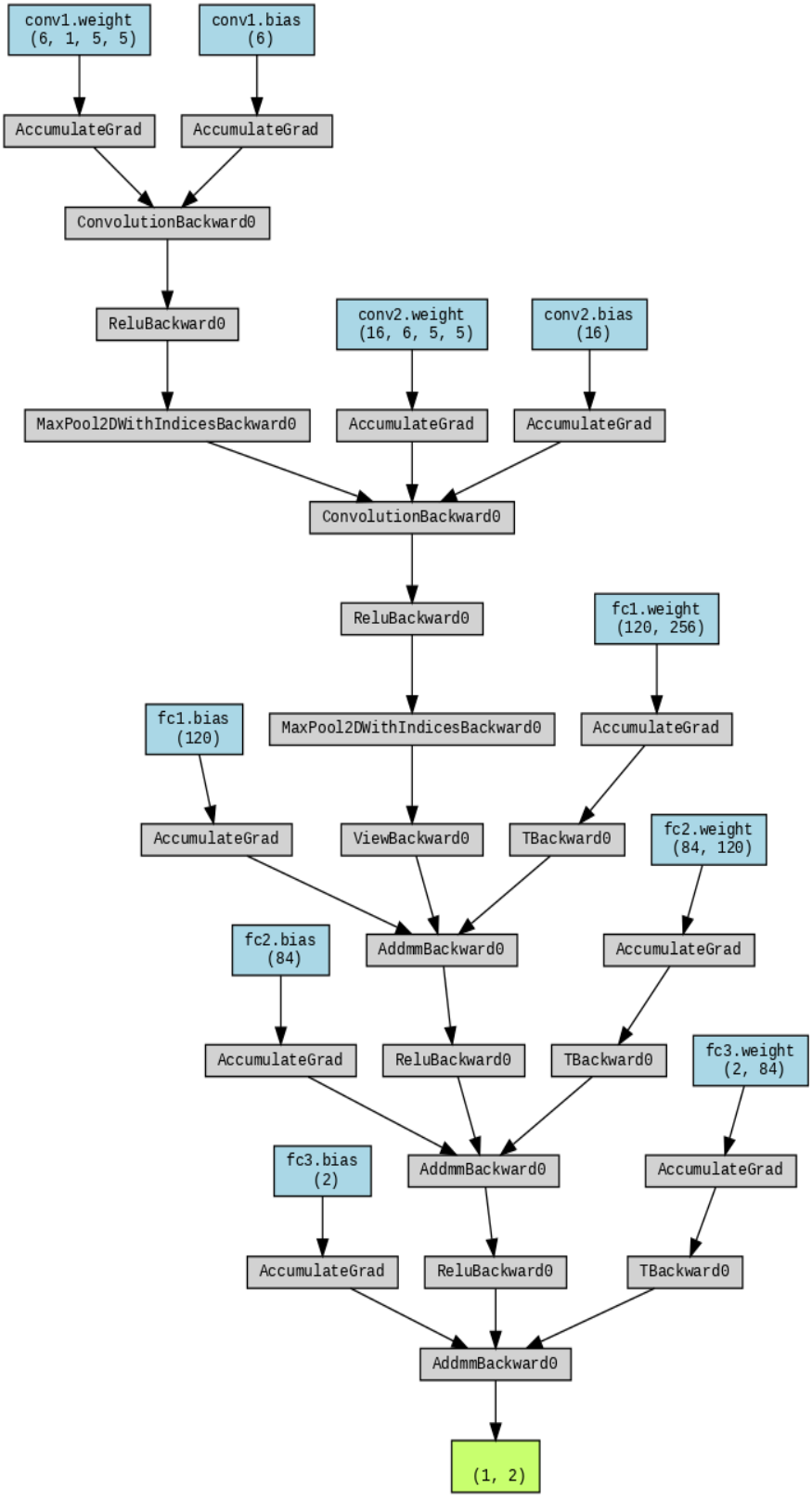


Figure 2: Convolutional Neural Network Diagram

### 3.3 XAI Algorithm

The custom-developed explainable artificial intelligence algorithm works similarly to the well-known XAI: Local Interpretable Model-agnostic Explanations (LIME). True to the LIME model, our algorithm is model-agnostic, meaning that this algorithm is added on top of an already developed AI model instead of being written deep inside its layers. However, instead of integrating the transparency code in the machine learning model's main body, a new algorithm is created that takes the AI model as an input parameter. This makes the technique convenient to apply to many different classification algorithms without making noticeable changes to the code. Our deep learning model outputs a vector of positive and negative predictions based on how likely it believes each output to be true. Changing an important pixel in an image will have noticeable effects on these values, which is the foundation of the whole algorithm.

### 3.4 XAI Implementation Details

The function takes an image and AI model as parameters. Its first step then is to run the original image through the convolutional neural network and calculate the initial prediction vector of the image. In the next step, the function takes the first pixel in the image, sets its color value to black, and runs that slightly altered image through the AI model again. The prediction values of the image are stored in a matrix. The pixel color is then reverted to its original value, and the image is restored to its starting picture. This is repeated for every single pixel in the image, with the prediction values stored in the matrix. Figure 3 shows the pseudocode for the XAI used in the study.

"model" function returns a number from 0-1 corresponding to the likely-hood of pneumonia being present. "input image" is a gray-scaled image with pixel values from 0-1 corresponding to the darkness of each pixel (0 is black)

```
originalPrediction ← model(inputImage)                    ▷ get AI prediction
for i in x-pixels do
    for j in y-pixels do
        inputImage[i, j] ← 0                               ▷ set pixel color to black
        newPrediction ← model(inputImage)
        predictionChange ← originalPrediction − newPrediction
        pixelImportanceImage[i, j] ← predictionChange
        inputImage[i, j] ← originalPixelValue              ▷ set pixel to initial color
    end for
end for
display pixelImportanceImage;
```

Figure 3: Pseudocode for the XAI Used in the Study

The difference between the augmented prediction vectors in the matrix and the original prediction vector is then calculated. This difference signifies the "effect" of each pixel on the conclusion of the AI model as it shows the magnitude of change it had on the model's output. To make these values easy to understand, they can be graphically represented as shown in Figure

4. As seen in Figure 4, the magnitude of the effect of the pixel is shown from the least impactful (dark purple) to the most impactful (bright yellow) in the corresponding pixel in the original image.

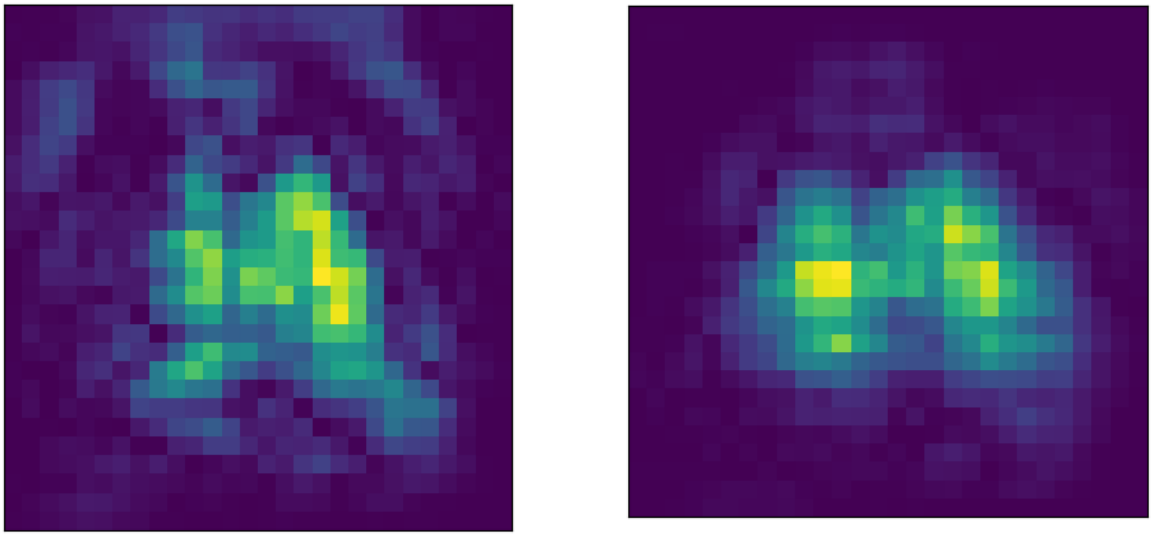

Figure 4: Visual Representation of the Effect Each Pixel Has on the Deep Model Output

In the next step, the top pixels with the largest effect on the image can be found and displayed on top of the original image to produce a visual aid for medical professionals to identify the exact location of pneumonia in patients' lungs. Figure 5 illustrates the top one hundred pixels with the largest effect on the output of the model, overlaid on the original data image. Within the algorithm, the number of overlayed pixels can be changed so that if the highlighted areas create zones that are too broad to be useful, their sizes can be configured to match the user's needs.

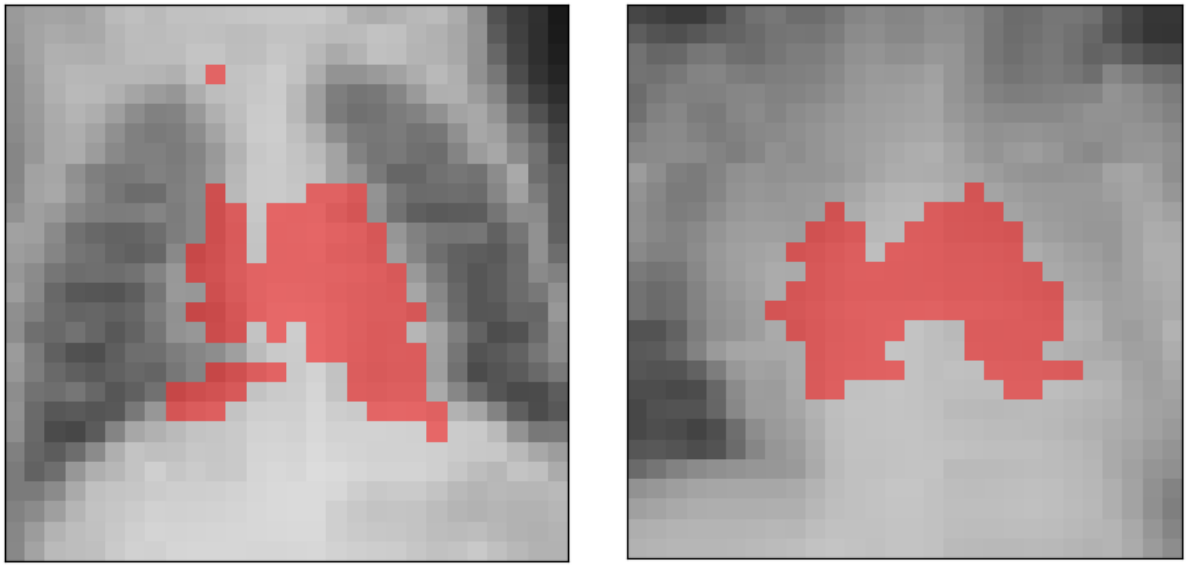

Figure 5: Top 100 Most Important Pixels on the Output of the Model Overlaid on the Original Data Image

### 3.5 Further Experimentation

To assess the robustness of the XAI algorithm in detecting pneumonia-related patterns under different pixel perturbations, we conducted some further tests with our XAI model. By modifying the pixel color to white and then its complement, we explored how the algorithm responds to changes in pixel values and whether it maintains its interpretability across different color transformations.
In these experiments, the algorithm was slightly changed while the test data point was held constant. As opposed to the original algorithm where pixels were changed to be black, the pixel color was first modified to be white and then modified to be the complement of the current color. As the shades of each pixel were determined on a 0 to 1 scale, the complement was found by taking the absolute value of 1 minus the current pixel value. The results of these experiments were recorded.
The results of these experiments revealed that changing the pixel colors to black when creating the augmented images highlighted the most notable pixels in the image in which pneumonia would be found. The analysis of these results sheds light on the algorithm's sensitivity to pixel alterations and its ability to identify pneumonia regions despite varying pixel representations accurately. Furthermore, it provides insights into potential enhancements or adjustments needed to improve the algorithm's resilience and interpretability in real-world clinical settings.

## 4. Results

The trained deep model was able to reach an accuracy of approximately 90% after the 15-epoch training and testing phase. Although not the primary purpose of the project, these outcomes show the adaptability and strength of deep learning models and how they can be applied to solve complex world problems.

### 4.1 Model Performance and Interpretability

The positive results we achieved with our XAI model can be attributed to the large amounts of similar data found inside the PneumoniaMNIST dataset, which when paired with regularization and multiple ReLU functions created a fast and accurate model to classify the x-ray scans. The ReLU functions played a crucial role in this specific model by introducing non-linearity to an otherwise linear algorithm. This enabled the neural network to analyze a broader range of intricate variables that would have been overlooked in the absence of these functions. As a result, the XAI was able to construct useful diagrams to explain how it reached its outputs.

#### *4.1.1 Interpretation of XAI Visualizations*

Figures 6 and 7 illustrate how our XAI model provides visual aids not only for the location of pneumonia in lung X-ray scans but also for particular hotspots that doctors should be looking at when reading scans. Specifically, Figure 6 shows a collage of the original X-ray images with red highlights going along certain features in the image, highlighting the pixels the CNN deemed important in its output.

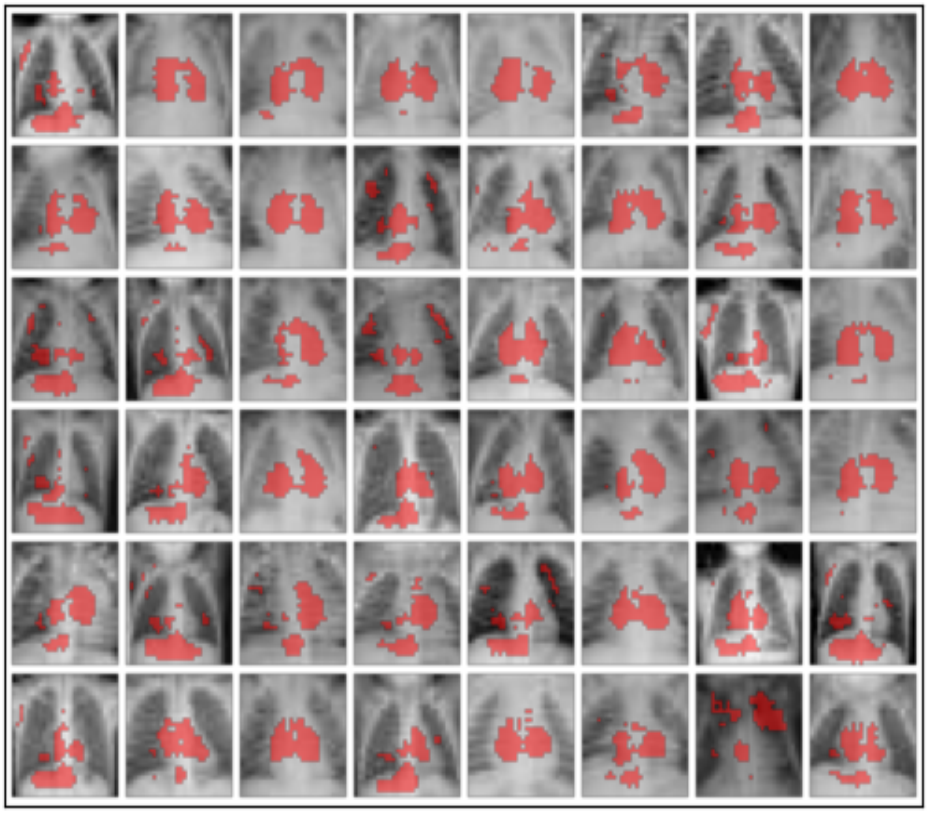

Figure 6: Most Important Pixels Displayed on 48 Images

If medical professionals analyze the images found in Figure 6 and still require additional information to make an accurate diagnosis, additional assistance like the images shown in Figure 7 could be displayed.

### *4.1.2 Analysis of Pixel Effects*

Figure 7 plots how much each pixel affected the AI model's output, which can be thought of as the chance that each pixel is pneumonia-infected. The dark purple pixels which can be found on the outside edges of the image signify a low or negligible chance of pneumonia. In contrast, the green and yellow pixels in the center of the image demonstrate medium and high chances respectively. The dark green pixels generally trace the shape of the lungs in the original image, while bright yellow spots signify hotspots that should be closely examined for traces of pneumonia.

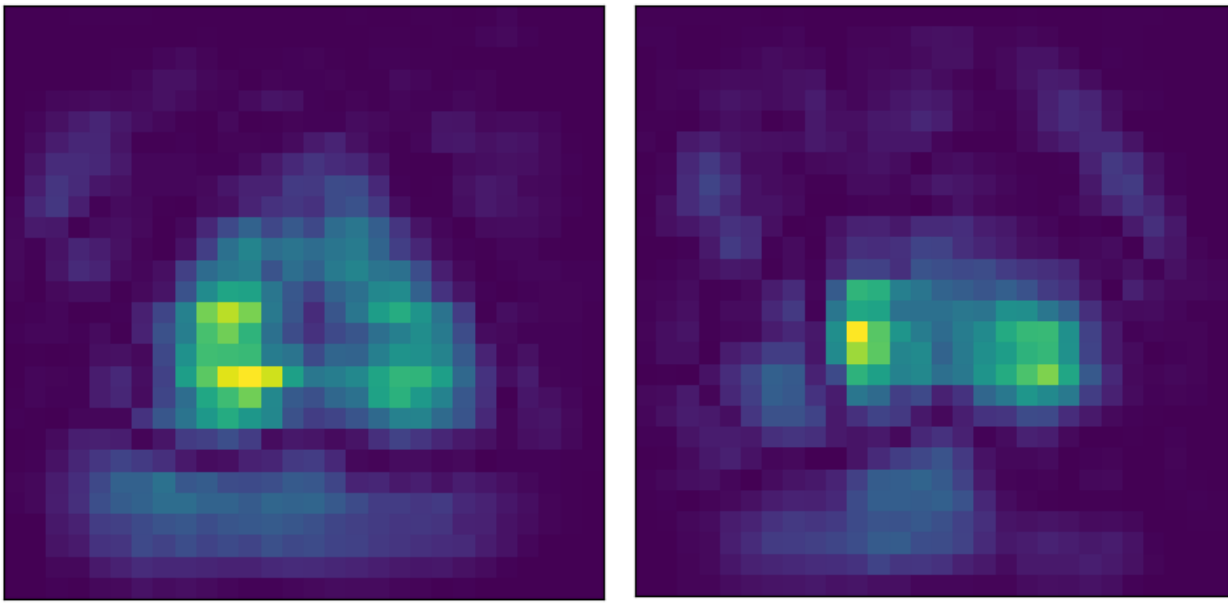

Figure 7: Relative Pixel Effect on Model Output

Combinations of the visual diagrams such as shown in Figures 6 and 7 can be used in hospitals around the world to help with the diagnoses of patients with life-threatening conditions. The algorithm can quickly and visually describe what areas of the image doctors should pay more attention to, with the highlighted spots aligning with the likely location of the pneumonia.

### 4.2 XAI Function Performance Evaluation

The XAI function of the project produced results that could be useful to radiologists and other healthcare professionals. As expected, pixels in the lungs were on average brighter than pixels in other parts of the image, but still had hotspots to pinpoint the exact location of the pneumonia.

#### *4.2.1 Experimentation to Identify Pneumonia Features*

Additionally, an experiment was performed to deduce exactly what features in the dataset the XAI was identifying as pneumonia. In general, pneumonia is identified in X-rays by locating cloudy, milky-white spots inside the lungs, so setting a pixel inside of the lungs to black creates the exact opposite of what pneumonia symptoms would appear to be. Hence, if previously light pixels that drastically increased the pneumonia prediction in the AI are changed to black, the switch will have a larger effect on the prediction vector than changing already dark pixels to be darker, as dark pixels wouldn't be interpreted as pneumonia initially in the first place.

#### *4.2.2 Impact of Pixel Alteration on Prediction*

To test this, the pixels were changed to white instead of black during the recalculation of prediction values, and the difference between predicted pneumonia locations was stark. Figure 8 shows the predicted pneumonia locations of the algorithm when pixels were changed to white instead of black.

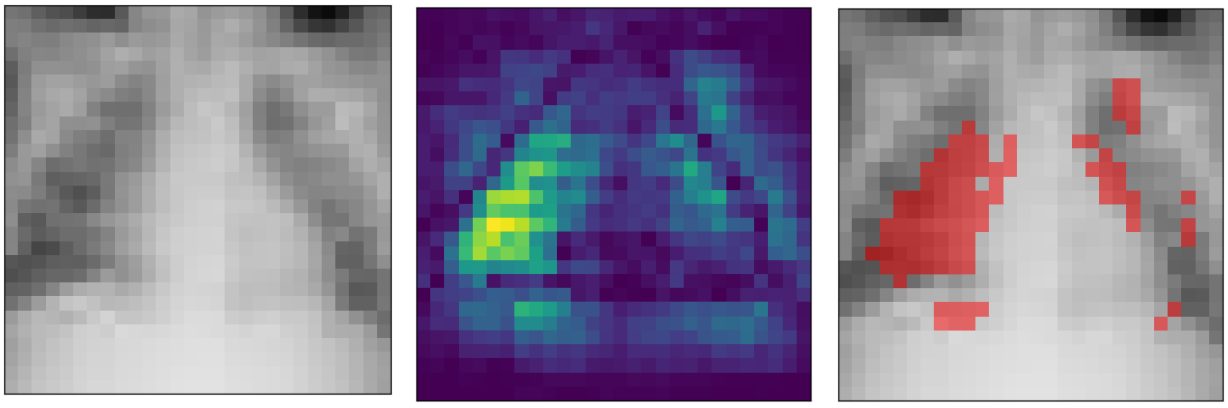

Figure 8: Original Image, Relative Pixel Importance, Top 100 Overlaid Pixels Changed to "White" in Algorithm

In comparison, the original algorithm which substituted pixel colors for black produced a result that better visualizes the location of pneumonia. Figure 9 shows the original algorithm where pixels are changed to be black.

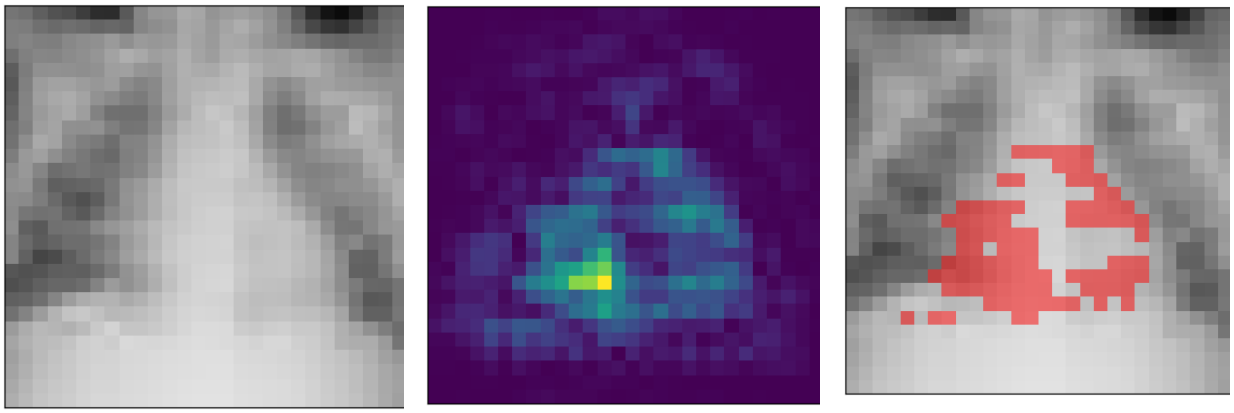

Figure 9: Original Image, Relative Pixel Importance, Top 100 Overlaid Pixels Changed to "Black" in Algorithm

As can be seen in Figures 8 and 9, when the pixels were changed to white (Figure 8), the model's predictions were more affected by the changing of darker pixels to lighter pixels. This is demonstrated by the clustering of highlights along the dark sections of the lungs rather than the light sections as seen in Figure 9.

### *4.2.3 Algorithm Optimization for Desired Output*

These results show that the most effective version of the algorithm to reach the desired output substitutes every pixel in the picture with black pixels, one by one, rather than any other gray-scaled color.
This fact makes logical sense, as when dark pixels in the lungs – which are initially seen to be pneumonia-free – become white, or "pneumonia infested," the prediction vector changes by a larger amount than when already white or "pneumonia infested" pixels are kept white.

This change in the algorithm produces a result that highlights the dark parts of the image red, the parts of the image that are pneumonia-free. However, when changing light-colored pixels to dark-colored pixels (as seen in Figure 9), the highlights grouped up around the light areas of the lungs, correctly locating the true pneumonia-infected spots.

#### *4.2.4 Clinical Relevance and Trust in Predictions*

Further on, Figure 9 shows that the deep learning model is indeed considering the lighter pixels in the lungs to be the tell-tale sign of pneumonia, just as doctors do. Doctors will now be able to use this tool and look at the augmented X-ray scans and see that the model was looking at the same details that they were when coming to their conclusion about the patient.

### 4.3 Evaluation of Prediction Patterns

When setting pixels to be their complements (as outlined in the methods section) as seen in Figure 10, the augmented image resembles the effect of just setting dark pixels to be light. This variation to the algorithm also gives incorrect locations of pneumonia, as it is effectively pointing out the pixels that haven't yet been infected by pneumonia. Figure 10 shows images that are similar to the ones in Figure 8, as the dark parts of the lungs are highlighted rather than the light parts.

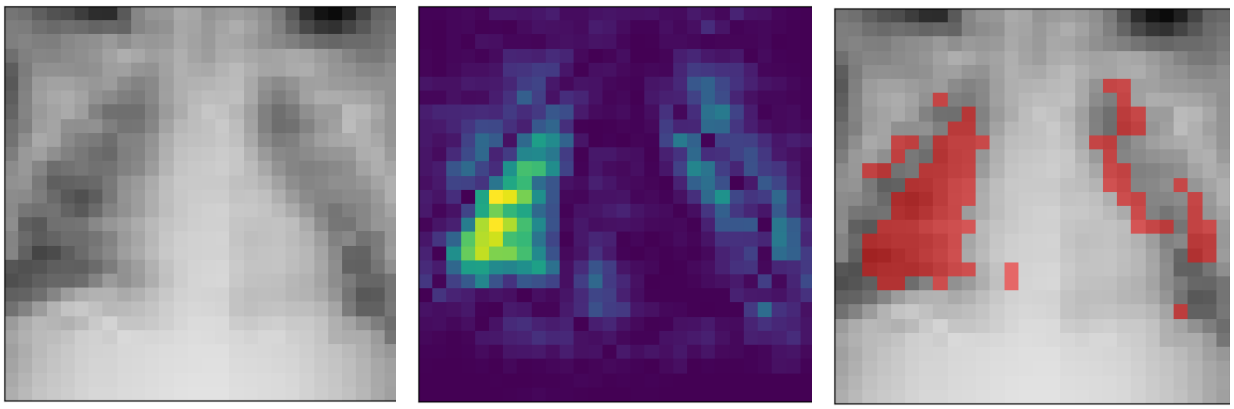

Figure 10: Original Image, Relative Pixel Importance, Top 100 Overlaid Pixels Changed to "Complement Pixel"

This aligns with the researcher's expectations because, in medical contexts, doctors identifying a white splotch in a pair of lungs is a stronger indicator of pneumonia compared to a random dark spot in a lighter area. This variation to the algorithm also gives incorrect locations of pneumonia, as it is effectively pointing out the pixels that haven't yet been infected by pneumonia. These results show that the most effective version of the algorithm to reach the desired output substitutes every pixel in the picture with black pixels, one by one, rather than any other color.

### 4.4 Assessment of False Predictions

In cases where the deep model predicts that there is no pneumonia, the highlights on the image look completely different than when it predicts that there is pneumonia. Figure 11 shows examples of what the XAI will output when it predicts "False" – that there is no pneumonia.

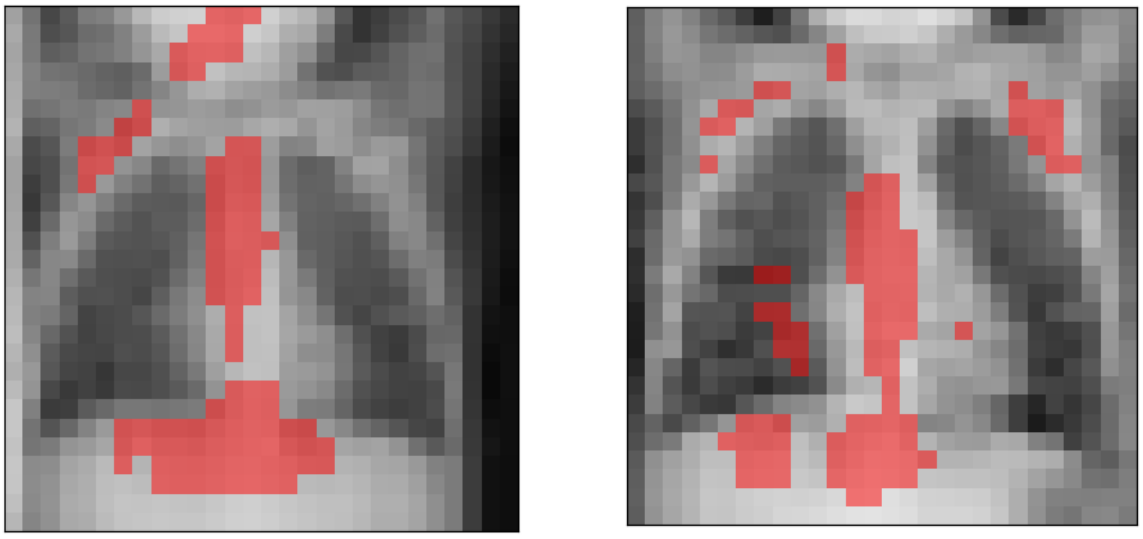

Figure 11: Top 100 Overlaid Pixels When Prediction Was Negative ("No Pneumonia")

Although the AI still focuses on the lighter pixels, it highlights more of the spine and the bones rather than specific locations in the lungs. The images in Figure 11 have clear highlights going up along the spine, ones that are not present when it predicts that there is pneumonia; the highlights are grouped up along the bottom of the images where other organs are found. Either one or both of these features are present in all of the false predictions and can be easily noticed by doctors. The XAI likely behaves this way due to the lack of lighter pixels found in the lungs when there is no pneumonia present in the image, so the pixels that look the most like pneumonia are the bones surrounding the lungs.

### 4.5 Conclusion and Clinical Implications

When the AI produces an inaccurate prediction, it is easily identifiable through visual inspection. As seen in Figure 12, if the prediction is unreliable, the highlights have no general structure and aren't found in the lighter regions of the lungs. For experienced medical professionals, this would immediately raise red flags, and the AI prediction would be nullified. This adds an extra layer of security to the XAI, making sure that any incorrect readings are obvious to detect.

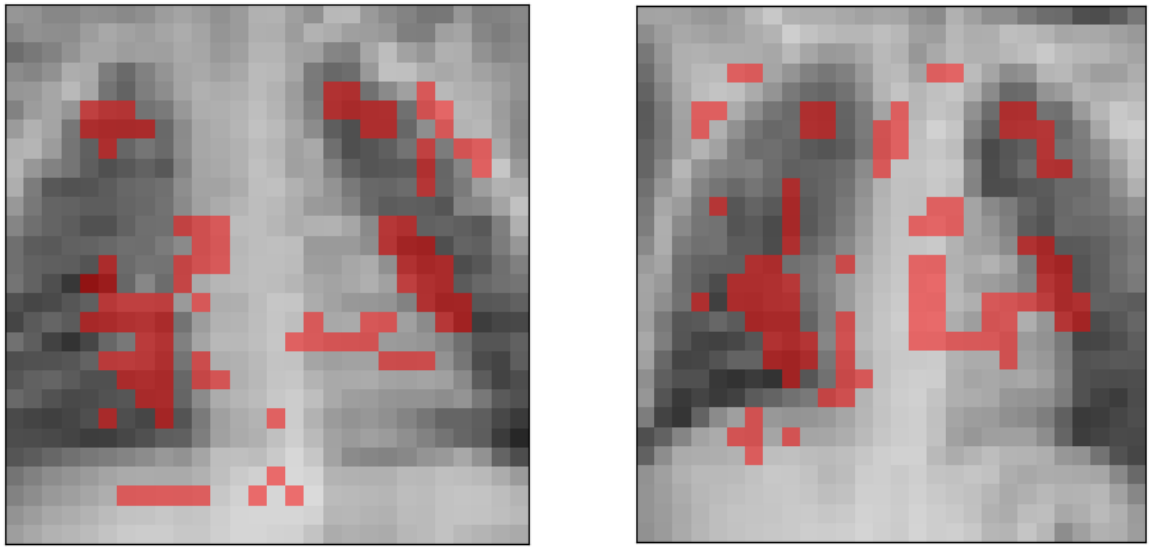

Figure 12: Top 100 Overlaid Pixels When Prediction Was Incorrect

## 5. Discussion & Acknowledgements

### 5.1 Effectiveness and Adaptability of the XAI Model

The developed learning model demonstrated a 90% accuracy in diagnosing lung conditions. Our explainable algorithm (XAI) was able to identify important pixels in the image that most influenced the output of the machine learning model, aiding as a tool for researchers to use to identify pneumonia location or existence. The substitution of different colors of pixels allowed us to understand the methodology the learning model takes when determining an output, with substituting black values for the pixels achieving the best result in terms of highlighting points of interest for pneumonia on the x-rays. The created XAI is model-agnostic and can be easily implemented in other fields of healthcare, not being limited to convolutional neural networks. A shortcoming that this algorithm has is that it runs in $O(n^2)$, which means it has to complete many computations to achieve its results and will take a long time to process the image as the resolution increases.

### 5.2 Future Enhancements and Explorations

This XAI model can still be improved in the future, and the exact accuracy of the pneumonia location prediction could be further explored. Trying this experiment on the same dataset but with higher resolution could provide better data about algorithm speed, efficiency, and precision of the predicted location. Although the red highlights were in the correct pneumonia-infected areas, the pixel density was so small that it was more showing an area of interest than exact pinpoint locations. With a larger and higher-resolution image, though, it would be interesting to see if the model highlighted the exact, miniscule white spots in the lungs that are symptoms of pneumonia, or if it would behave in the same way and show general areas.
If the speed of detection dramatically decreased, it would be intriguing to examine a recursive function that segments the image into large sections and then decreases the size of those sections depending on the effect their masking had on the prediction. This would allow large,

unimportant sections of the image to be ignored, and the time spent computing to be decreased. While these new avenues have not yet been explored, I would like to continue this study and see exactly how impressive the results could be when using new and interesting techniques to analyze the behavior of deep learning models.

## References

1. J. M. Durán, K. R. Jongsma. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*. (2021). doi: 10.1136/medethics-2020-106820. Epub ahead of print. PMID: 33737318.
2. Gramegna, P. Giudici. SHAP and LIME: An evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*. 1-6 (2021).
3. Khan, H. Fatima, A. Qureshi, S. Kumar, A. Hanan, J. Hussain, S. Abdullah. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. *Biomedical Materials & Devices*, (2023).
4. Kiseleva, D. Kotzinos, P. De Hert. Transparency of AI in healthcare as a multilayered system of accountabilities: Between legal requirements and technical limitations. *Frontiers in Artificial Intelligence*. **5** (2022).
5. Rai. Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, **48**, 137–141 (2019).
6. M. T. Ribeiro, S. Singh, C. Guestrin. "Why Should I Trust You?": Explaining the predictions of any classifier. arXiv:1602.04938 (2016).