Pneumonia Detection using chest X-Ray and Deep Learning Technologies

Kavana Anil
DATA 240 Department of Applied Data Science
San Jose State University

Veena Ramesh Beknal
DATA 240 Department of Applied Data Science
San Jose State University

Rutuja Kokate
DATA 240 Department of Applied Data Science
San Jose State University

Ganaprathyusha Puluputhuri Muni
DATA 240 Department of Applied Data Science
San Jose State University

Neha Sharma
DATA 240 Department of Applied Data Science
San Jose State University

Nivedita Venkatachalam
DATA 240 Department of Applied Data Science
San Jose State University

***Abstract*--Pneumonia is a severe lung inflammation, especially fatal for children aged five or younger as it contributed to 14% of deaths in that age group in the year 2019. While X-ray images are crucial to detect pneumonia in patients, their addition poses challenges as well. As part of this study, the authors utilized a set of 5 863 X ray images of pediatric patients taken at the Guangzhou Women and Children's Medical Center and labelled either Normal or Pneumonia. Specifically, the authors used deep learning architectures including CNN, ResNet, DenseNet, InceptionV3 and transfer learning techniques to classify the images. Model evaluation metrics include accuracy, precision, recall, and F1 score. With the X-ray image of a patient active, we utilized SHAP Deep Explainer for model interpretation; these highlighted areas of the image which had a significant impact on the diagnostic decision made by artificial intelligence, assuring accountability for AI-powered diagnostics. As expected, ResNet50 gave the best single results but an ensemble model which used ResNet50 and other approaches did better than everyone. This paper explores the use of deep learning to automate the process of diagnosing pneumonia, enhance the ability of doctors to make decisions, and development of pediatric radiology.**

***Keywords*— CNN, ResNet, DenseNet, SHAP, Pneumonia**

I. INTRODUCTION

Pneumonia is a respiratory infection that impacts people of all age groups. It is an infection that affects humans' lungs and prevents them from breathing normally. The signs of pneumonia are Inflamed air sacs that are filled with pus. Pathogens like viruses and bacteria enter the lungs through the air we breathe and rapidly multiply in number causing inflammation and impairing the regular function of lungs and thus reducing the supply of oxygen into the blood stream.

The World Health Organization estimates that pneumonia causes around 2.5 million deaths worldwide each year. Furthermore, around 740,180 children under the age of five died from pneumonia in 2019, making nearly 14% of all deaths in this age group. These devastating numbers show how timely and accurate diagnosis is required.

Early detection is critical and essential to improve the outcomes for pneumonia patients. Timely intervention can prevent complications and reduce mortality. Chest X-ray imaging plays an important role in diagnosing pneumonia. Healthcare providers can identify these lung abnormalities with the help of these X-ray images. Even for skilled radiologists, it can be difficult to identify pneumonia from these chest X-ray pictures because of characteristics that overlap with those of other respiratory disorders and differences in imaging quality.

These problems stated above can be overcome by utilizing technologies like deep learning and artificial intelligence. A variety of computer vision, deep learning, and machine learning approaches can be applied to evaluate chest X-ray pictures with impressive accuracy. These computers will be able to spot minute irregularities and trends that a human reviewer would overlook. These systems can assist radiologists by providing faster, more reliable diagnosis. Furthermore, these solutions have a lot of potential in environments with limited resources, where access to qualified radiologists is restricted.

II. LITERATURE REVIEW AND RECENT WORK

The utilization of AI and deep learning in medical image analysis has gained significant momentum in recent years. Deep learning algorithms particularly

have shown remarkable potential in improving diagnostic accuracy and efficiency in medical imaging.

Rajpurkar et al. (2017) introduced ChexNet where a deep learning model was trained to detect pneumonia from chest X-rays. The study demonstrated that the algorithm's performance was comparable to that of expert radiologists. This marked as a breakthrough that highlighted the feasibility of using deep learning technologies to support clinical decision-making. ChexNet also emphasized the advantages of deep learning in handling and processing large volumes of medical imaging data with speed, precision and reliability.

Hwang et al. (2019) conducted a systematic review analyzing over 50 studies focused on AI applications in medical imaging. Their findings consistently highlighted the effectiveness of convolutional neural networks in medical image classification tasks like pneumonia detection. This review also emphasized the robustness of CNN models in identifying and distinguishing minute and subtle patterns in imaging data, which are often hard and not reliable to identify through manual process.

Kumar et al. (2020) explored transfer learning techniques to address the challenge of limited size of the medical datasets like in the case of disease detection through images. Adapting pre-trained models such as ResNet demonstrated how these pre trained model architectures could be used to achieve high accuracy in classifying medical images and detection of diseases even with small and domain-specific datasets. This approach reduced computational requirements and also highlighted the scalability of transfer learning in the context of healthcare.

Another crucial advancement in AI for medical diagnostics is the incorporation of explainable AI (XAI) methods. Zhang et al. (2018) emphasized the importance of visualization techniques like saliency maps and heatmaps to provide interpretability in the decision-making process of any model. Their study showed how these visualizations and tools help us highlight regions of interest in chest X-rays. This offers radiologists a recommendation on how and why the AI model is providing a particular recommendation.

Despite these promising developments, challenges remain in the implementation of AI in medical imaging. Critical challenges such dataset bias, lack of generalizability, and the requirement for strong validation across varied patient groups were highlighted by Irvin et al. (2019) in an article published in Nature Machine Intelligence. According to the study, a lot of AI models are trained on datasets that could not accurately reflect patient demographics around the world, which could result in differences in diagnostic accuracy depending on the context. This emphasizes how crucial thorough model validation and the creation of just AI systems are.

Collectively, these findings show how artificial intelligence can revolutionize the detection of pneumonia, especially in environments with limited resources and restricted access to skilled radiologists. AI has the potential to greatly lower the burden of disease by facilitating quicker and more precise diagnostics. To address current issues, such as enhancing model transparency, guaranteeing generalizability, and creating ethical frameworks for clinical deployment, more research is necessary. These initiatives will be essential to maximizing AI's potential to enhance medical results.

III. DATASET OVERVIEW

The dataset being utilized contains images from the Guangzhou Women and Children's Medical Center in Guangzhou. This dataset focuses on pediatric patients between the ages of one and five. This dataset is available on Kaggle, and this dataset consists of 5,863 chest X-ray images in jpeg format. These images are categorized into two classes which are Pneumonia and Normal. Furthermore, these images are distributed into organized into train, test, and validation subsets for each category. The train set has 1341 images of x-rays which are categorized as normal and 3875 images that are classified as being affected by pneumonia. Similarly, the test has 234 normal ones and 390 affected images. Additionally, the validation set has 8 images in both the categories.

All these images were captured during routine clinical care. This dataset was subjected to stringent quality control, and every chest X-ray was taken as part of standard clinical care. To guarantee accuracy and dependability, scans that were of poor quality or illegible were eliminated. To reduce grading errors, a third expert examined the evaluation set after two skilled doctors assessed the diagnoses. These ensure that the model built on this dataset will provide reliable classifications.

IV. IMPLEMENTATION PLAN

The project implementation flowchart given below outlines the step-by-step process for building a deep learning-based pneumonia detection system using chest X-ray images.

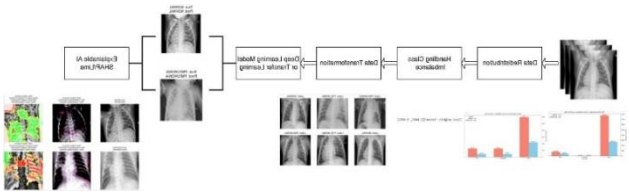

Fig.1. Architecture diagram for Data Pipeline

A methodical procedure for detecting pneumonia using chest X-ray pictures is described in the project implementation flowchart as seen in Figure 1. To guarantee balanced representation, it starts with data redistribution, dividing the dataset into training, validation, and test sets. Class weights are computed and added to the loss function in order to rectify the class imbalance and guarantee equitable representation of the minority PNEUMONIA class. After that, data transformation is performed to standardize inputs and enhance training stability. This includes scaling, grayscale-to-RGB conversion, and normalization. To properly classify the photos, a deep learning model or transfer learning technique is then used. Lastly, predictions are interpreted using explainable AI techniques like SHAP and LIME, which highlight important X-ray regions that impacted the model's judgments and guarantee transparency and reliability.

V. DATA PREPROCESSING AND TRANSFORMATION

The data distributed into train, test and validate sets in Kaggle have the following distribution in Figure 2 and we see class distribution of Normal and Pneumonia Cases in Training, Validation, and Test Sets

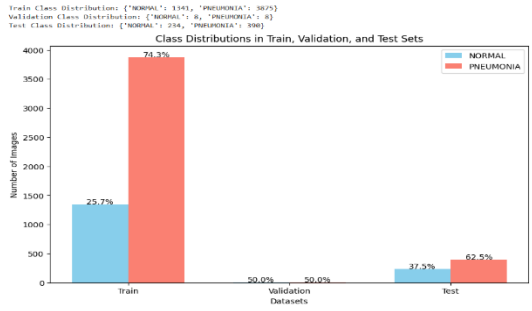

Fig.2. Uneven data split ratios

As shown in the above figure 2, the dataset contains 5,856 images, with 5,216 images (89.07%) in the training set, 16 images (0.27%) in the validation set, and 624 images (10.66%) in the test set. This distribution is highly skewed and has an underrepresented validation set. This distribution can have a negative impact on model validation. Next, we try to redistribute the data into new test, train and validation sets to overcome this issue. The image below shows the counts of test, train and validate sets post redistribution. As seen in Figure 3 we have redistributed Data Ratios and Class-Wise Distribution Summary

```
Redistributed Data Ratios:
Training Set: 4099 images (70.00% of total)
Validation Set: 878 images (14.99% of total)
Test Set: 879 images (15.01% of total)

Class-wise Distributions:
Train Class Distribution: {'NORMAL': 1108, 'PNEUMONIA': 2991}
Validation Class Distribution: {'NORMAL': 237, 'PNEUMONIA': 641}
Test Class Distribution: {'NORMAL': 238, 'PNEUMONIA': 641}
```

Fig.3. Redistributed ratios.

This relocated distribution has 4,099 images (70.0%) in the training set, 878 images (14.99%) in the validation set, and 879 images (15.01%) in the test set. This redistribution ensured that there was a good balance among test, train and validation counts. On the contrary, this distribution had the training set which consisted of 74% Pneumonia and 26% Normal images, while both the validation and test sets contained approximately 73% Pneumonia and 27% Normal images. This shows that the classes of normal and pneumonia are not balanced. This can lead to inaccurate results.

The redistributed dataset still exhibits a significant class imbalance where the normal class is much more prevalent than the pneumonia class across the training, validation, and test sets. To overcome the issues above we distribute the data using class weights. The pneumonia class predominantly dominates over the normal class. Class weights are allocated in order to rectify this imbalance and guarantee that the model could efficiently learn from both classes. The formula was used to determine the class weights. The formula of the class weights is as below figure 4. Class Weight Calculation Formula for Imbalanced Datasets

$$w_c = \frac{\text{total number of samples}}{\text{number of classes} \times \text{number of samples in class } c}$$

Fig.4. Equation used to correct imbalanced classes.

The weight for the pneumonia class was estimated to be around 0.6852, and the weight for the normal class of about 1. 8497.The calculations are as below. Furthermore, the loss function was modified to penalize inaccurate predictions on the minority class more severely to give this class a higher priority. Better generalization was made possible by these modifications, which balanced the model's attention on the two classes. The class imbalance was thus successfully reduced, enabling the model to perform better in accurately identifying and categorizing the

pneumonia class. In the end, this method improves the model's accuracy in crucial tasks involving rare classes by making it more capable of managing the dataset's diverse class distributions as seen in figure 5.

For NORMAL:

$$w_{\text{NORMAL}} = \frac{4,099}{2 \times 1,108} = \frac{4,099}{2,216} \approx 1.8497$$

For PNEUMONIA:

$$w_{\text{PNEUMONIA}} = \frac{4,099}{2 \times 2,991} = \frac{4,099}{5,982} \approx 0.6852$$

Fig.5. Class weights calculation.

The redistributed sets with class weights are distributed as below figure 6.

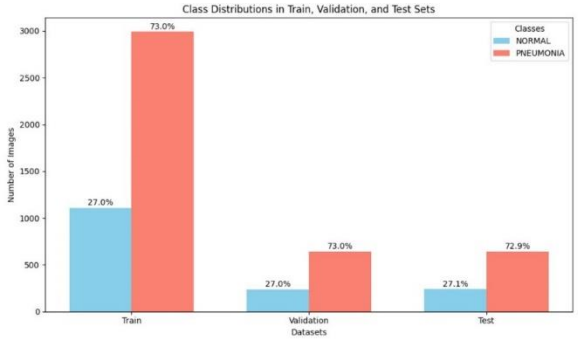

Fig.6. Class Distributions in Train, Validation and Test Sets

Next, we try to show a sample of data from both classes across all 3 sets. The samples are shown below figure 7 for the train set. The same is done for test dataset and the validation set.

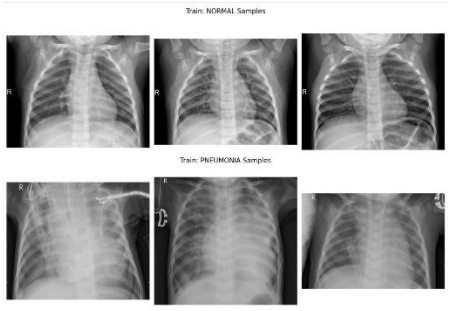

Fig.7.Sample X-Ray Images.

The data we have is now transformed. To standardize the data and guarantee compliance with the deep learning model used for pneumonia detection, the feature engineering and data transformation procedures applied to the chest X-ray dataset were essential. To resolve discrepancies in the original image sizes and facilitate effective and consistent processing throughout the training, validation, and test sets, all images were first downsized to a uniform dimension of 256x256 pixels. Additionally, to comply with the input specifications of most deep learning models, which are made to analyze RGB images, the grayscale photos were also transformed to three channels. This transformation guarantees that the model architecture can process the input efficiently even while it does not add color information. Furthermore, pixel intensity values were standardized to fall between [-1, 1], which speeds up model convergence and improves numerical stability during training. Resizing lowers the possibility of errors during training and increases computing efficiency by ensuring that the model receives input of consistent dimensions. Utilizing pre-trained models or architectures tuned for three-channel inputs is made possible by converting grayscale to RGB compatibility. Additionally, normalization lessens the effect of outlier pixel values, preventing the model from being skewed by extreme intensities and allowing it to concentrate on learning significant patterns. All of these actions improve the pneumonia detection system's overall effectiveness, generalizability, and dependability. Once transformed the image features are as shown below figure 8.

```
Shape of one batch of images: torch.Size([32, 3, 256, 256])
Number of images in the batch: 32
Image dimensions (C x H x W): 3 x 256 x 256
```

Fig.8. Images sizes after normalization.

The plot illustrates the pixel intensity distribution for the normal and pneumonia image classes in the dataset. Pixel intensity values ranging from 0 (black) to 255 (white) on the x-axis and their corresponding frequencies on the y-axis, the graphic displays the pixel intensity distribution for the normal and pneumonia picture classes in the dataset. The normal class displays fewer brighter pixels (100–150) that correspond to bones and tissues, and a strong peak in the lower intensity range (0–50) that reflects dark parts like air-filled lungs. Pneumonia pictures, on the other hand, show a wider range of intensity, suggesting greater variability brought on by brighter areas from fluid accumulation and inflammation. Because of the healthy lung structure, normal X-rays usually show homogeneous, darker regions; however, pneumonia instances show lighter, high-intensity areas because of inflammation or consolidation. As seen in Figure 9 for Pixel Intensity Distribution for Normal and Pneumonia X-Ray Images

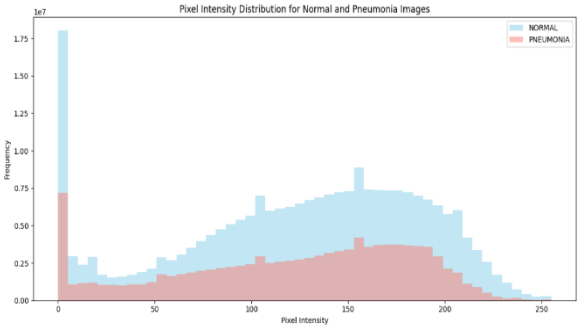

Fig.9. Pixel Intensity distribution

VI. MODELLING

Now, we will try and build the following models to classify the chest X-ray images. We will build the following models.

- Convolutional Neural Network
- Inception Net
- ResNet
- ChexNet
- Ensemble Models

These models were chosen for this project for pneumonia classification based on their aptitude for processing medical images and their capacity to identify complex patterns in X-ray data. Convolutional Neural Networks (CNNs) were chosen because of their ability to automatically learn spatial feature hierarchies, which makes them very useful for identifying distinct patterns, such as those linked to pneumonia. To capture intricate patterns and improve the model's generalization across a variety of medical images, advanced CNN architectures were further utilized. To learn fine-grained information in X-ray pictures and prevent vanishing gradient problems, ResNet50 was added with its deep architecture and residual connections. To take advantage of its superior performance on medical imaging, CheXNet, a pre-trained architecture designed especially for pneumonia identification, was refined for this task. InceptionNet's special capacity to employ numerous kernel sizes per layer and capture features at different scales to detect minute changes in the images could help with image-based pneumonia classification. To improve classification resilience and lower the risk of overfitting, an ensemble model was finally put into place to incorporate predictions from various architectures. This ensured a more accurate and dependable pneumonia detection method.

Models Justification

The Convolutional Neural Network (CNN) was selected as the baseline model due to its proven ability to automatically learn spatial hierarchies of features in images, making it particularly effective for medical image classification tasks like detecting pneumonia in X-rays. By emphasizing localized patterns like edges and textures—which are essential for detecting lung abnormalities—its architecture streamlines feature extraction. Larger kernels, high-resolution inputs, and regularization strategies like dropout and weight decay were among the advanced advancements made to the CNN that improved its capacity to capture finer details while avoiding overfitting.

CheXNet and ResNet50 improved the categorization skills much more. With its residual connections, ResNet50's deep design successfully solved the vanishing gradient issue in deep networks, allowing it to pick up on minute details like fluid accumulation or inflammation in chest X-rays. The NIH ChestX-ray14 dataset was used to pre-train CheXNet, which is based on DenseNet-121 and was specifically tailored for chest X-ray classification. Through fine-tuning, it is especially well-suited for pneumonia identification in smaller datasets because of its lightweight design, which guarantees computational economy while keeping high accuracy.
Inception Net and an ensemble model rounded out the approach to improve robustness and generalization. Data augmentation approaches improved InceptionNet's capacity to generalize across the dataset, while its modular design enabled it to capture information at various scales. Lastly, to leverage the advantages of each architecture, the ensemble model aggregated predictions from several architectures, including ResNet50 and Enhanced CNN, using a weighted voting technique. This approach offered a solid answer for pneumonia categorization and increased reliability, especially in difficult situations.

Convolutional Neural Network

Convolutional Neural Network (CNN) model is implemented for pneumonia detection problem. This model processes 256x256 RGB X-ray images which ensures standardizing the inputs for efficient training. The network consists of three convolutional layers which are Conv1 with 32 filters, Conv2 with 64 filters and Conv3 with 128 filters and all 3 layers are with 3x3 kernels, stride 1, and padding 1 to extract progressive features. ReLU activation is used to ensure effective gradient flow and non-linearity. A

MaxPooling with 2x2 kernels is used to reduce spatial dimensions, enhancing efficiency and mitigating overfitting. Fully connected layers FC1 with 512 neurons and FC2 with 2 neurons aggregate the features for binary classification into "Pneumonia" or "Normal." The overall result of the CNN model is shown below figure 10:

	precision	recall	f1-score	support
NORMAL	0.93	0.94	0.94	238
PNEUMONIA	0.98	0.98	0.98	641
accuracy			0.97	879
macro avg	0.96	0.96	0.96	879
weighted avg	0.97	0.97	0.97	879

Fig.10. Classification report CNN

The model achieved high precision (0.93 for Normal, 0.98 for Pneumonia), recall (0.94 for Normal, 0.98 for Pneumonia), and F1-scores (0.94 for Normal, 0.98 for Pneumonia), demonstrating its robustness and reliability in detecting pneumonia from chest X-rays. These results underscore CNN's capability to distinguish subtle patterns in medical imaging, providing a reliable diagnostic aid.

Advanved CNN

The advanced CNN model incorporates some enhancements to address and capture the complex patterns in pneumonia detection. Some essential improvements include experimenting with various activation functions like ReLU and Leaky ReLU for optimal feature extraction. It also has a dropout of 0.3 and weight decay to reduce overfitting. Batch normalization accelerates training and improves generalization. The AdamW optimizer ensures efficient weight updates. High-resolution images and larger kernels (7x7) capture nuanced X-ray features effectively, and gradient clipping manages high-contrast regions and subtle differences between normal and pneumonia cases. The performance of the model is as seen in figure 11.

	precision	recall	f1-score	support
NORMAL	0.51	0.71	0.59	238
PNEUMONIA	0.87	0.75	0.80	641
accuracy			0.73	879
macro avg	0.69	0.73	0.70	879
weighted avg	0.77	0.73	0.75	879

Fig.11. Classification report CNN Advanced.

The model attempts to address the difficulties of differentiating minute and subtle changes in medical imaging with primary focus on efficiency and robustness. Additionally, by minimizing overfitting, regularization approaches ensured improved performance on unseen data. The model produced moderate results with an accuracy of 0.69 and a macro-average F1-score of 0.70 despite these sophisticated methodologies. These results demonstrate the possibility of improving performance and generalization through additional tuning or the use of ensemble approaches.

Inception Net

For effective feature extraction and picture categorization, InceptionNet—more especially, the InceptionV3 architecture—is a deep convolutional neural network. To efficiently capture multi-scale features, the architecture makes use of cutting-edge "Inception modules," which carry out concurrent convolutions with different kernel sizes (1x1, 3x3, and 5x5) and pooling operations. This design is especially well-suited for activities like medical imaging because it strikes a balance between high accuracy and computational economy. Aside from its feature extraction capabilities, InceptionV3 also has an auxiliary classifier that acts as a secondary branch during training, supplying extra learning signals and assisting in the fight against vanishing gradients. The model was modified in this implementation to classify chest X-ray pictures into two categories: normal and pneumonia. the assignment. To support two output classes, a fully connected layer was used in place of the pre-trained model's last layer, and task-specific auxiliary logits were modified.

The model handled resized photos (299x299) during training to satisfy the input specifications of InceptionV3. With a learning rate of 0.001, the Adam optimizer was utilized for effective convergence, and the loss function was Cross-Entropy Loss. The primary and auxiliary outputs were both handled by

the training pipeline, guaranteeing optimal weight updates throughout each epoch. The model's state was preserved for later testing and implementation, and it demonstrated good accuracy and generalization.

The InceptionV3-based implementation demonstrated exceptional performance across several evaluation metrics. The overall classification accuracy was 97%, reflecting its reliability in distinguishing between Normal and Pneumonia cases. Class-specific metrics further highlighted its robustness: for the Normal class, the model achieved a precision of 96%, recall of 92%, and an F1-score of 94%. For the Pneumonia class, precision reached 97%, recall was 98%, and the F1-score was an impressive 98%. underscoring the model's ability to detect this critical condition accurately. The confusion matrix revealed high true positive and true negative rates, with minimal misclassifications.

Additional evaluations included the ROC Curve, which demonstrated an AUC of 0.98, confirming the model's excellent ability to distinguish between the two classes across varying thresholds. Detailed metrics, including macro and weighted averages, indicated balanced performance, even in the presence of class imbalances. The model's design and results make it an ideal candidate for clinical use, as it delivers both accuracy and interpretability, which are critical for medical decision-making as seen in figure 12.

	precision	recall	f1-score	support
NORMAL	0.96	0.92	0.94	238
PNEUMONIA	0.97	0.99	0.98	641
accuracy			0.97	879
macro avg	0.97	0.96	0.96	879
weighted avg	0.97	0.97	0.97	879

Fig.12. Classification report InceptionNet.

ROC Curve for Normal vs. Pneumonia Classification (AUC = 0.99) as seen in figure 13.

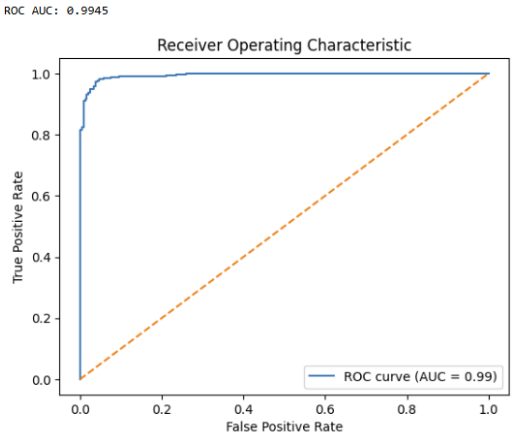

Fig.13. ROC-AUC curve for InceptionNet.

ResNet50

The ResNet50 model utilizes a pre-trained 50-layer architecture with residual connections which effectively addresses the vanishing gradient problem and enables efficient training for classification even with a smaller dataset. Custom top layers, including Global Average Pooling, a dense ReLU layer, and a final sigmoid layer, are tailored to extract task-specific features, ensuring precise binary classification of chest X-rays. The fine-tuning process focuses on adapting the last few layers to X-ray features while freezing earlier layers to preserve general representations, making the model highly adaptable to medical imaging tasks. Optimization techniques, including the Adam optimizer, binary cross-entropy loss, and dynamic learning rate adjustments with early stopping, further enhance training efficiency and performance.

The model achieved a precision value of 0.96 for normal and 0.97 for pneumonia, recall values of 0.91 for normal and 0.99 for pneumonia. F1-scores of 0.94 and 0.98, respectively, the model's evaluation metrics demonstrate its strong performance. Its excellent accuracy in detecting real positives and reducing false negatives is demonstrated by these measures. The model has remarkable reliability with an overall accuracy of 0.97, backed by weighted and macro averages of comparable variables. ResNet50's capacity to use transfer learning, bespoke architecture, and fine-tuning to produce accurate and dependable results in pneumonia classification tasks is demonstrated by this performance as seen in figure 14.

	precision	recall	f1-score	support
NORMAL	0.96	0.91	0.94	238
PNEUMONIA	0.97	0.99	0.98	641
accuracy			0.97	879
macro avg	0.97	0.95	0.96	879
weighted avg	0.97	0.97	0.97	879

Fig.14. Classification report ResNet50.

ChexNet

CheXNet is also a convolutional neural network which is particularly designed for chest X ray classification and pneumonia detection. This network is mainly based on DenseNet-121 architecture. This model supports a network pretrained on imagenet and finetuned for medical imaging tasks. DenseNet-121, a convolutional neural network, uses dense connections, allowing each layer to draw input from all previous layers. This design improves gradient flow and encourages feature reuse, making it well-suited for deep learning applications.

Images are processed by CheXNet in batches usually in torch format. Size([32, 3, 256, 256]), which denotes 32 photos with three color channels and 256x256 pixel sizes each batch. The X-ray pictures are resized as part of the model's process and then run through DenseNet-121's layers, which are organized in dense blocks and divided by transition layers. These transition layers use pooling operations and 1x1 convolutions to minimize the size of feature maps. Furthermore, traditional dense layers are replaced by Global Average Pooling (GAP), which increases spatial invariance and decreases overfitting. The final output layer creates probabilities for binary or multi-class classification using a sigmoid activation function.

Cross-Entropy Loss, a common loss function for classification problems, is used in the training phase. Class weights are added to the loss function to rectify the dataset's class imbalance and guarantee that the minority class is sufficiently represented during training. The Adam Optimizer, which has a learning rate of 1e-4, is used to carry out the optimization. Adam works well for deep learning problems because of its adaptive learning capabilities, which enable more seamless convergence and effective parameter updates.

The assessment of the CheXNet model shows that it performs exceptionally well in dividing chest X-ray pictures into two groups: pneumonia and normal. The model's exceptional ability to differentiate between the two groups is demonstrated by the Receiver Operating Characteristic (ROC) Curve, which shows an AUC of 1.00. The model's accuracy in detecting true positive cases while reducing false positives is demonstrated by its high AUC. With an overall accuracy of 97%, the classification report further demonstrates its dependability. The model has good generalization for detecting healthy instances, with 92% precision, 97% recall, and 95% F1-score for the Normal class. With a precision of 99%, recall of 97%, and F1-score of 98% for pneumonia as seen in figure 15 and figure 16, the results are even more remarkable and guarantee precise identification of this serious illness.

	precision	recall	f1-score	support
NORMAL	0.92	0.97	0.95	238
PNEUMONIA	0.99	0.97	0.98	641
accuracy			0.97	879
macro avg	0.96	0.97	0.96	879
weighted avg	0.97	0.97	0.97	879

Fig.15. Classification report CheXNet.

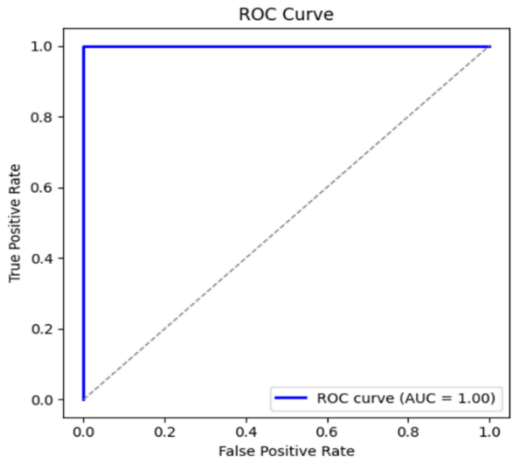

Fig.16. ROC-AUC for CheXNet.

Ensemble Models

Ensembling is a technique where we combine two or more models to make more accurate and robust predictions. This approach leverages the strength of individual models while reducing overfitting and biases. There are multiple approaches for ensembling of deep learning models. The majority voting approach is one where each model in the ensemble predicts a class and the final prediction is the class with the most votes selected. Another way is average probabilities where the average probability scores from all the selected models for each class is considered and the highest average score is selected. Weighted averaging is another way which is like average probabilities wherein the better performing model has more influence on the final predictions. The

other sophisticated approaches to ensembling are stacking and bagging. Each ensemble strategy has unique strengths and is suited for different scenarios. Majority voting and averages probabilities are simple and effective for combining diverse models. Weighted averaging and stacking offer greater flexibility and precision but require tuning or sometimes training while bagging method reduces overfitting in individual models.

For our project, we considered two models for ensemble modeling - CNN and ResNet50. We used a weighted voting ensembling approach wherein the weightage was automatically calculated based on accuracy of the models on the validation data. Based on this approach, a slightly higher weightage of 51% was given to the ResNet model while 49% weightage was with the CNN model. The weights are fairly close since the performance of either standalone model wasn't too different on the validation data. We are combining predictions based on the optimal weights that reflect the relative importance or performance of CNN and ResNet50 models and finally producing more accurate and robust predictions than when they are done individually. Each model in the ensemble techniques is evaluated to ensure consistent behavior. Prediction probabilities i.e., Pneumonia or not are generated by passing the models through a SoftMax function. This ensemble method can be adapted to various domains, such as medical diagnosis tasks where model diversity is an asset and also, this can act as a second opinion for the seeking medical advice. We can observe the evaluation metric of ROC curve with AUC below for the individual models along with the ensemble - the performance is quite across all 3 models as seen in figure 17 and figure 18.

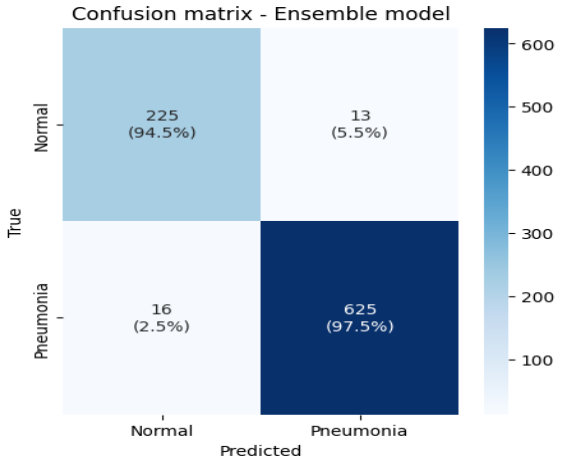

Fig.17. Confusion matrix Ensemble Method.

	precision	recall	f1-score	support
0	0.93	0.95	0.94	238
1	0.98	0.98	0.98	641
accuracy			0.97	879
macro avg	0.96	0.96	0.96	879
weighted avg	0.97	0.97	0.97	879

Fig.18. Classification report Ensemble Model.

Comparison of ROC Curves for Pneumonia Detection: Enhanced CNN, ResNet, and Ensemble Models as seen in figure 19.

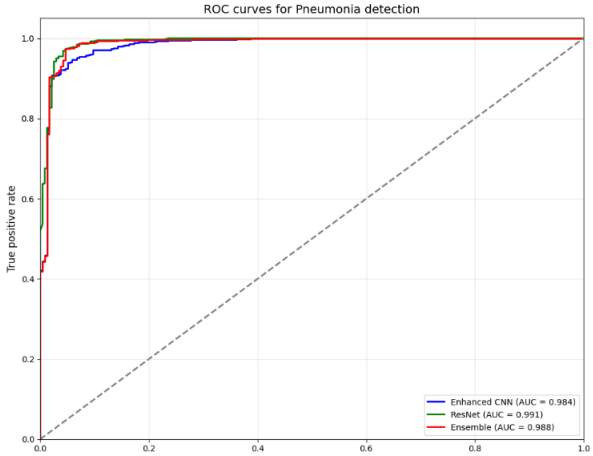

Fig.19. ROC-AUC for Ensemble Model.

Model Comparison:

Here we evaluated several deep learning models for pneumonia detection using chest X-ray images which includes CNN, ResNet50, CheXNet, InceptionNet and an ensemble model combining CNN and ResNet50. Each of the models were assessed using metrics such as accuracy, precision, recall, F1-score and AUC. Among the models ResNet50 showed the most robust performance achieving an accuracy of 97% and an AUC of 0.991 there by highlighting its ability to extract fine-grained features using the residual connections. CheXNet used a DenseNet-121 architecture and achieved a comparable accuracy with slightly higher recall for pneumonia detection thus making it ideal for reducing false negatives. InceptionNet stood out with highest AUC (0.9945) thus exhibiting its effectiveness in capturing multi-scale features through its modular design. The ensemble model, incorporating predictions from ResNet50 and CNN with weighted voting thus offered improved robustness and balanced performance across metrics, achieving an AUC of 0.988. This approach showed the value of combining complementary model architectures for enhanced reliability. Overall, ensemble models offered a more reliable and broadly

applicable solution for clinical applications even though single models such as ResNet50 and InceptionNet has shown exceptional diagnostic capabilities as seen in figure 20.

Model	Precision	Recall	F1-Score	Accuracy
CNN	N: 0.93, P: 0.98	N: 0.94, P: 0.98	N: 0.94, P: 0.98	0.97
CNN Advanced	N: 0.51, P: 0.87	N: 0.71, P: 0.75	N: 0.59, P: 0.80	0.73
ResNet50	N: 0.96, P: 0.97	N: 0.91, P: 0.99	N: 0.94, P: 0.98	0.97
CheXNet	N: 0.92, P: 0.99	N: 0.97, P: 0.97	N: 0.95, P: 0.98	0.97
InceptionNet	N: 0.96, P: 0.97	N: 0.92, P: 0.99	N: 0.94, P: 0.98	0.97
Ensemble Model	N: 0.93, P: 0.98	N: 0.95, P: 0.98	N: 0.94, P: 0.98	0.97

Fig.20. Model Comparison.

VII. EXPLAINABLE AI

SHAP (SHapley Additive exPlanations)

For our project, we have used SHAP (SHapley Additive exPlanations) which is a quantitative measure of how features contribute to the models' final predictions. It calculates the contribution of each feature for understanding the influence of input images. SHAP values are visualized using a color-coded scheme wherein blue color indicates that features are predicting negative class in our project i.e. Normal and red color indicates that features are predicting towards the positive class i.e. Pneumonia in this case. This indicates the granular understanding of how specific regions of input, such as image affects the final predictions. In our project, each pixel can be thought of voting for or against a specific class and the sum of these leads to the final predictions. For instance, in our project, SHAP can identify regions that most influence a diagnosis providing an interpretable breakdown of the prediction process. The main advantage of this approach of explainable AI is it gives a perspective of the feature importance and offers insights into the contribution of each feature or a region of the image. The drawback is that it is computationally expensive as it has high- dimensional inputs in the form of images. Interpreting the images along with SHAP requires some technical expertise.

LIME (Local Interpretable Model–agnostic Explanations)

This is another approach to explain what a model considers important in terms of specific regions in the image which leads to classifying whether the image is Pneumonia or Normal. LIME generates local explanations for a specific prediction by approximating the model's behavior around that instance. LIME works by dividing the input image into interpretable components, such as super pixels in our project. It selectively modifies or turns off parts of the input to observe how predictions change which highlights the region's most relevant to prediction. There is a digital highlighter to identify the critical areas that led to the model's prediction. In the chest X-ray, LIME highlights lung regions that strongly influence the diagnosis of pneumonia. The main advantage of this approach is it is an easy, visual and intuitive way to understand which parts of the image are important for generating predictions. However, it is relatively less useful for understanding overall behavior of the model and has lower detail when compared to SHAP. In the below figure 21, we can observe the SHAP and LIME results that are done after the ensembling of the models are done.

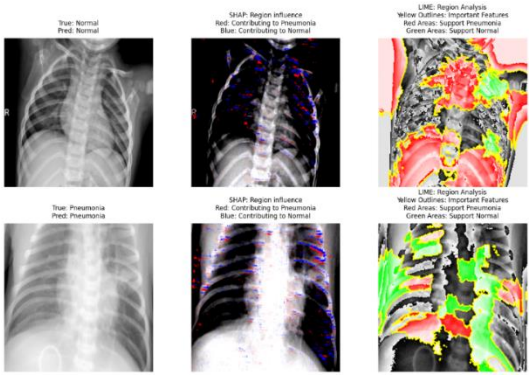

Fig.21. Explainable AI on Chest X-ray's.

This visualization of the original X-ray alongside the overlaid SHAP and LIME plots is useful in understanding how deep learning models make decisions when classifying chest X-rays for pneumonia. 1 case of Normal and Pneumonia each are considered to explain the nuances of either explainable AI approaches. As explained earlier, SHAP uses blue/red coloring to show regions influencing the model's decision with red areas indicating features contributing to Pneumonia diagnosis while blue areas indicate features supporting normal classification. In the Pneumonia case (bottom), there's more red highlighting in areas that likely suggest infection while in the normal case, more blue highlighting in clear lung regions. This is an excellent way of augmenting a medical professional's decision of diagnosis by looking at the X-rays.

LIME plots have yellow outlines marking important regions with red areas supporting pneumonia diagnosis and green areas supporting normal classification. The plots show how different regions contribute to the model's final decision. There are some overlapping areas between the SHAP and LIME plots as well - these can be considered as the most important regions.

VIII. APPLICATIONS

Such an approach can act as a decision support tool to help radiologists by highlighting suspicious areas, acting as an AI powered "second opinion" with explanations. This can potentially reduce the chance of missed diagnoses or incorrect diagnosis. This approach can also play a crucial role in medical training and education. Such a tool can help in training new radiologists by showing what areas of the X-ray to look for before diagnosing the patient. This also provides visual feedback on diagnostic reasoning and overall helps build trust in AI systems by making decisions transparent. In the future, in emergency situations like another pandemic, this approach could also be used for emergency diagnosis to provide quick initial assessment with explanations. The doctor will provide human oversight while leveraging AI capabilities, but this can help in quickly filtering out healthy cases and focusing on patients that require urgent care. Doctors and radiologists can participate in the feedback loop and help improve model performance by identifying areas where the model needs improvement which facilitates collaboration between AI developers and medical professionals and helps in AI adoption by the medical industry. Overall, this approach shows how explainable AI can make deep learning more practical and trustworthy in medical applications, where understanding the reasoning behind decisions is crucial for patient care.

IX. CONCLUSION

In this project we try to highlight the potential of deep learning-based methods like CNN, ResNet50, CheXNet and ensemble models for automating pneumonia detection using chest X-ray images. By leveraging techniques such as transfer learning and explainable AI tools like SHAP and LIME where in the proposed solutions achieved high accuracy and interpretability. Out of all models we tried, ResNet50 emerged to be the most effective single model, while ensemble model further improved the performance by showcasing the advantages of combining models for robust predictions. Explainable AI methods enhanced transparency there by providing insights into decision-making process.

This project also underscores the importance of addressing challenges such as class imbalance and dataset quality to ensure reliable model performance in real-world applications. In addition, integration of explainable AI strengthens the useability of these models in clinical settings thus helping radiologists and reducing errors. For future research we can focus on expanding datasets, improving generalizability across diverse populations and refining model interpretability to improve the adoption of AI in medical diagnostics.

REFERENCES

[1] D. J. Alapat, M. V. Menon, and S. Ashok, “A Review on Detection of Pneumonia in Chest X-ray Images Using Neural Networks,” Journal of Biomedical Physics and Engineering, vol. 12, no. 6, Dec. 2022, doi: https://doi.org/10.31661/jbpe.v0i0.2202-1461.

[2] M. F. Hashmi, S. Katiyar, A. G. Keskar, N. D. Bokde, and Z. W. Geem, “Efficient Pneumonia Detection in Chest Xray Images Using Deep Transfer Learning,” Diagnostics, vol. 10, no. 6, p. 417, Jun. 2020, doi: https://doi.org/10.3390/diagnostics10060417.

[3] R. Kundu, R. Das, Z. W. Geem, G.-T. Han, and R. Sarkar, “Pneumonia detection in chest X-ray images using an ensemble of deep learning models,” PLoS ONE, vol. 16, no. 9, p. e0256630, Sep. 2021, doi: https://doi.org/10.1371/journal.pone.0256630.

[4] F. Shu and J. Shu, “An eight-camera fall detection system using human fall pattern recognition via machine learning by a low-cost android box,” Scientific Reports, vol. 11, no. 1, Jan. 2021, doi: https://doi.org/10.1038/s41598-021-81115-9.

[5] L. Wang and A. Wong, “COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images,” arXiv:2003.09871 [cs, eess], Apr. 2020, Available: https://arxiv.org/abs/2003.09871

[6] A. A. Reshi et al., “An Efficient CNN Model for COVID-19 Disease Detection Based on X-Ray Image Classification,” Complexity, vol. 2021, pp. 1–12, May 2021, doi: https://doi.org/10.1155/2021/6621607.

[7] Y. Oh, S. Park, and J. C. Ye, “Deep Learning COVID-19 Features on CXR using Limited Training Data Sets,” IEEE Transactions on Medical Imaging, pp. 1–1, 2020, doi: https://doi.org/10.1109/TMI.2020.2993291.

[8] N. M. Elshennawy and D. M. Ibrahim, “Deep-Pneumonia Framework Using Deep Learning Models Based on Chest X-Ray Images,” Diagnostics, vol. 10, no. 9, p. 649, Aug. 2020, doi: https://doi.org/10.3390/diagnostics10090649.

[9] A. Kareem, H. Liu, and P. Sant, “Review on Pneumonia Image Detection: A Machine Learning Approach,” Human-Centric Intelligent Systems, vol. 2, no. 1–2, pp. 31–43, May 2022, doi: https://doi.org/10.1007/s44230-022-00002-2.

[10] “Explain PyTorch MobileNetV2 using the Partition explainer — SHAP latest documentation,” Readthedocs.io, 2018. https://shap.readthedocs.io/en/latest/example_notebooks/image_examples/image_classification/Explain%20MobilenetV2%20using%20the%20Partition%20explainer%20%28PyTorch%29.html