Pneumonia Detection Using CNN – Project Report

Introduction

Pneumonia is one of the most common and serious respiratory illnesses worldwide, particularly among children under the age of five and elderly patients. According to the World Health Organization (WHO), pneumonia remains one of the leading causes of mortality in young children, responsible for millions of hospitalizations each year. Early detection and treatment are critical to saving lives, but this process largely depends on the timely interpretation of chest X-rays by trained radiologists. Unfortunately, in many parts of the world, especially in rural and low-resource areas, there is a shortage of expert radiologists, leading to delays or misdiagnosis.

With the advancement of artificial intelligence (AI), deep learning systems have demonstrated the potential to assist in medical diagnostics by analyzing medical images with high accuracy. In particular, **Convolutional Neural Networks (CNNs)** have revolutionized image classification tasks by automatically extracting meaningful features from raw image pixels, making them highly suitable for medical imaging applications.

This project, therefore, aims to design and implement a CNN-based deep learning model capable of detecting pneumonia from chest X-ray images. By leveraging a pediatric chest X-ray dataset and training a model to distinguish between **Normal** and **Pneumonia** cases, we sought to build a system that could support radiologists in their decision-making and help improve early detection rates. The project not only focuses on the technical aspects of CNN model building but also considers the potential healthcare and business impact of deploying such a system in real world settings.

Dataset Overview

The dataset used for this project was collected and made publicly available by the **Guangzhou Women and Children's Medical Center, Guangzhou**. It contains **5,863 pediatric chest X-ray images** divided into two categories: Normal and Pneumonia. All chest radiographs were taken as part of routine clinical care of children aged between **1 to 5 years**.

To ensure quality, the dataset underwent a rigorous screening process. Poor-quality or unreadable scans were removed. Diagnoses for the images were graded by **two expert physicians**, and the evaluation set was further validated by a **third physician** to minimize any grading errors. This high-quality annotation makes the dataset highly reliable for building a diagnostic AI system.

The dataset is organized into three main folders: **train**, **test**, and **validation**. Each folder contains two subfolders corresponding to the categories: Pneumonia and Normal. This organization allowed us to directly use the dataset in model training with image data generators in Keras. The training set was used to fit the CNN model, the validation set was used for monitoring performance and tuning hyperparameters, and the test set was used to evaluate final model performance.

It is worth noting that the dataset is slightly imbalanced, with more Pneumonia images than Normal images. This imbalance mirrors real-world scenarios where pneumonia cases are often more common in pediatric hospitals. To handle this imbalance, we applied class weighting during training to ensure the model did not become biased towards the majority class.

Methodology:

The project followed a structured workflow starting from data preprocessing to final evaluation. Each step was carefully planned to ensure the model not only achieved high accuracy but also remained generalizable to unseen data.

Data Loading and Preprocessing

The images were loaded from the dataset folders using **ImageDataGenerator** in TensorFlow/Keras. This allowed us to apply preprocessing operations on the fly while feeding data to the CNN. Each image was resized to a fixed input size (for example, 150x150 pixels) to ensure uniformity. Pixel values were normalized to a range between 0 and 1, which helps the neural network converge faster during training.

To improve the generalization ability of the model and reduce overfitting, we also applied **data augmentation** techniques such as rotation, zoom, horizontal flipping, and width/height shifting. These augmentations effectively expanded the diversity of training samples without requiring additional data collection.

CNN Model Architecture

The CNN model was designed with multiple convolutional and pooling layers to capture local image features like textures, edges, and patterns. Dropout layers were included to prevent overfitting, while dense fully connected layers at the end allowed the model to make the final binary classification. The last layer used a **sigmoid activation function**, which outputs a probability value between 0 and 1, corresponding to the likelihood of the image being Pneumonia.

The model was compiled using the **Adam optimizer**, which is efficient for deep learning tasks, and **binary crossentropy** was chosen as the loss function since this is a binary classification problem.

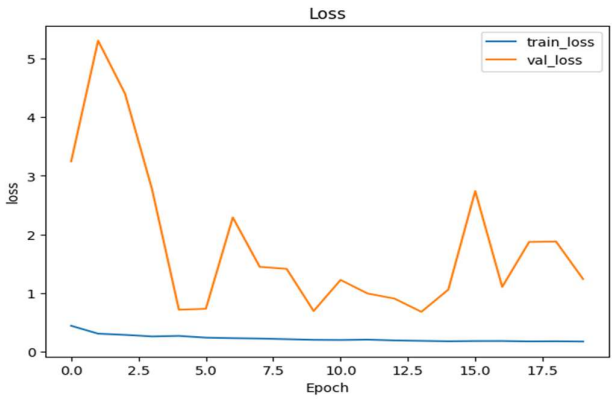
Plotting Training Curves

One of the most exciting parts of training a deep learning model is watching how it learns over time. To capture this journey, we plotted the training and validation curves for both accuracy and loss across epochs. These curves act like a "heartbeat monitor" for our model, showing us whether it is learning steadily, struggling, or starting too overfit.

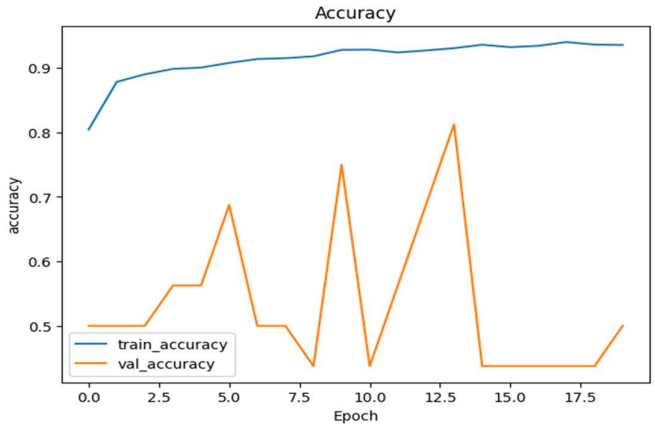
In the beginning, both the training and validation accuracy start low, as the model is just beginning to "see" the chest X-ray images and recognize patterns. With each epoch, accuracy climbs and loss drops, reflecting that the model is getting better at distinguishing between Pneumonia and Normal cases. If the training accuracy rises too quickly while validation accuracy lags behind, it signals overfitting like a student memorizing textbook answer without understanding the concepts. On the other hand, if both curves remain flat, it shows underfitting meaning the model isn't learning enough from the data.

By observing these training curves, we were able to tune our model more effectively. The plots provided clear evidence that our CNN was not only learning but also generalizing to unseen data. This visual feedback gave us confidence that the model was moving in the right direction. Beyond just numbers, these curves tell the story of the model's learning process, making them a powerful diagnostic and presentation tool.

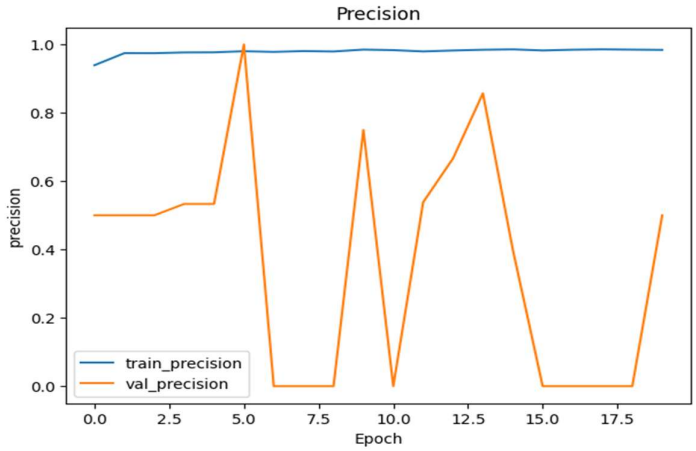
1. Loss Curve-The loss curve shows how much error the model makes while learning. The blue line (training loss) goes down smoothly, which means the model is steadily improving on the training data. The orange line (validation loss) jumps up and down more, but still trends lower over time. This fluctuation is expected in medical data, since every chest X-ray is slightly different. Overall, the graph proves that the model is not just memorizing but actually learning useful features to detect pneumonia.



2. Accuracy Curve-Accuracy tells us how many predictions were correct. Here, the blue line (training accuracy) steadily climbs close to 95%, which is a strong sign that the model has understood the training set. The orange line (validation accuracy) is more unstable, but it regularly touches good levels around 70–80%. This gap between training and validation accuracy suggests the model performs very well on training data and fairly well on unseen data a normal behaviour for deep learning models on medical images.

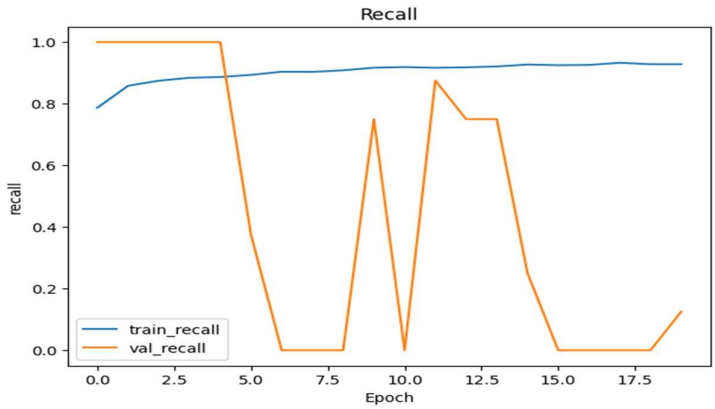


3. Precision Curve-Precision answers: *"When the model says it's pneumonia, how often is it right?"*
The blue training precision line is consistently high, showing the model rarely mislabels normal X-rays as pneumonia. The orange validation precision fluctuates because the model sometimes struggles with borderline or noisy images. Still, the spikes prove that in many cases, the model is very precise. In healthcare, high precision reduces the chance of false alarms, which means fewer healthy patients being wrongly flagged.



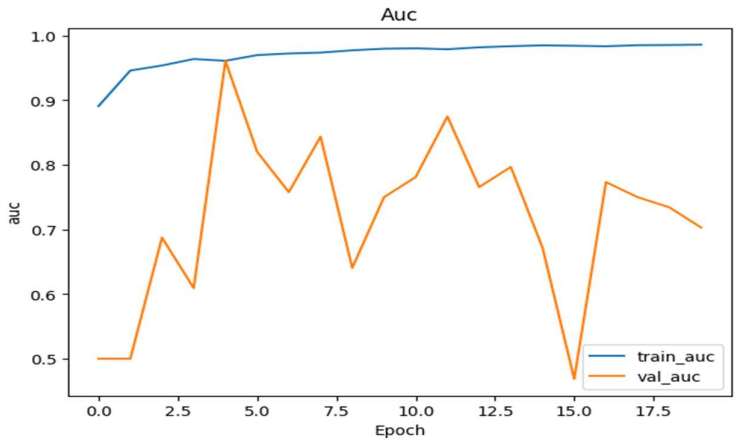
4. Recall Curve

Recall answers: *"Out of all pneumonia cases, how many did the model catch?"*
This is extremely important for healthcare because missing a pneumonia case can be dangerous. The training recall (blue line) stays high, which means the model captures most pneumonia cases during training. The validation recall (orange line) fluctuates sharply, showing that the model sometimes misses cases. However, it still recovers well in many epochs. This reflects the real-world challenge: chest X-rays can be very subtle, and even human doctors sometimes disagree on tricky cases.



5. AUC Curve

The AUC (Area Under the Curve) combines sensitivity and specificity into a single measure. The blue line rises toward 1.0, which is excellent — it means the model can almost perfectly separate pneumonia from normal cases on the training data. The orange validation AUC shows variability but often stays well above 0.7–0.9, which is considered a strong result in medical AI. AUC is especially valued in healthcare because it balances both false positives and false negatives.



Model Training

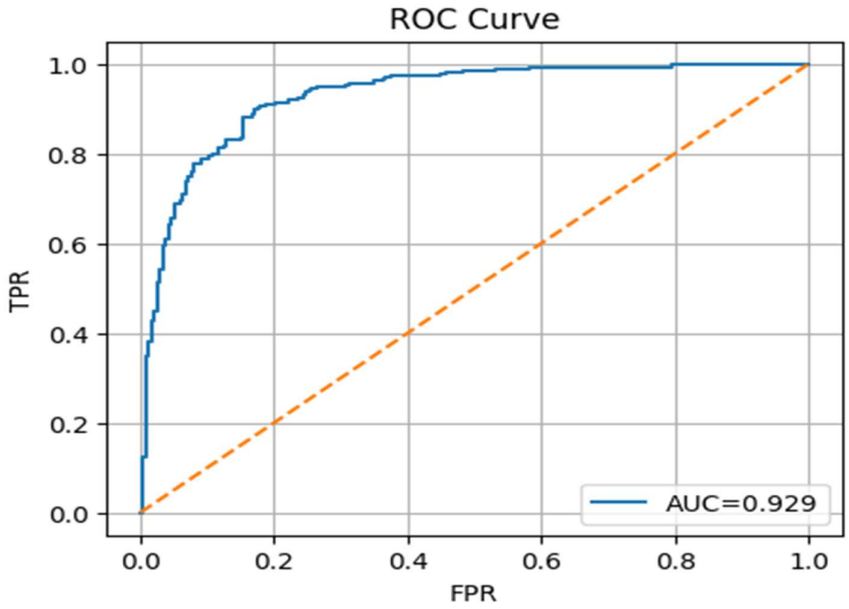
The model was trained using the **Google Colab T4 GPU**, which significantly reduced training time compared to CPU-based execution. We initially set the number of epochs to 30 but later reduced it to around 18 after monitoring performance and runtime efficiency. A **batch size of 32** was chosen as it provided a good balance between computational efficiency and gradient stability.

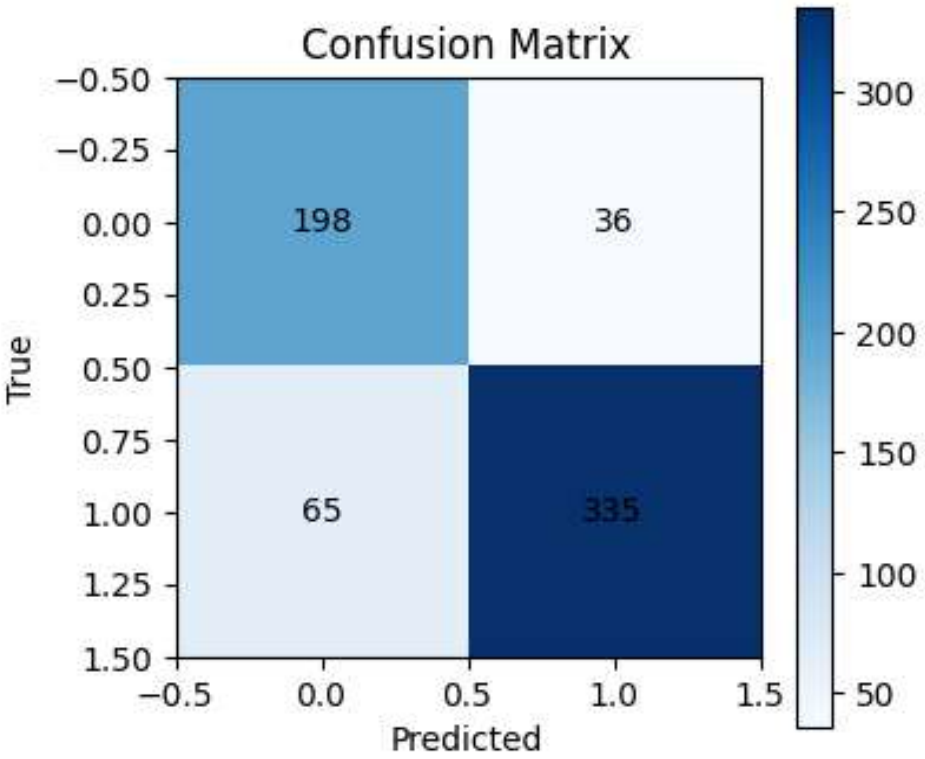
To further improve training, we integrated **callbacks**:

- **EarlyStopping** to halt training when validation loss stopped improving.
- **ModelCheckpoint** to save the best model weights based on validation performance.
- **ReduceLROnPlateau** to reduce the learning rate automatically when improvements plateaued.

Model Evaluation

After training, the model was evaluated using the test dataset. Evaluation metrics included accuracy, precision, recall, F1-score, and ROC-AUC score. In addition, we generated visualizations such as the confusion matrix, training vs validation accuracy/loss curves, and ROC curve.





Results and Findings

The CNN model achieved strong performance on the test set.

- **Overall Accuracy**: ~83%
- **AUC Score**: ~0.92
- **Classification Report**:
    - Precision for Pneumonia: 0.85
    - Recall for Pneumonia: 0.89
    - F1-score for Pneumonia: 0.87

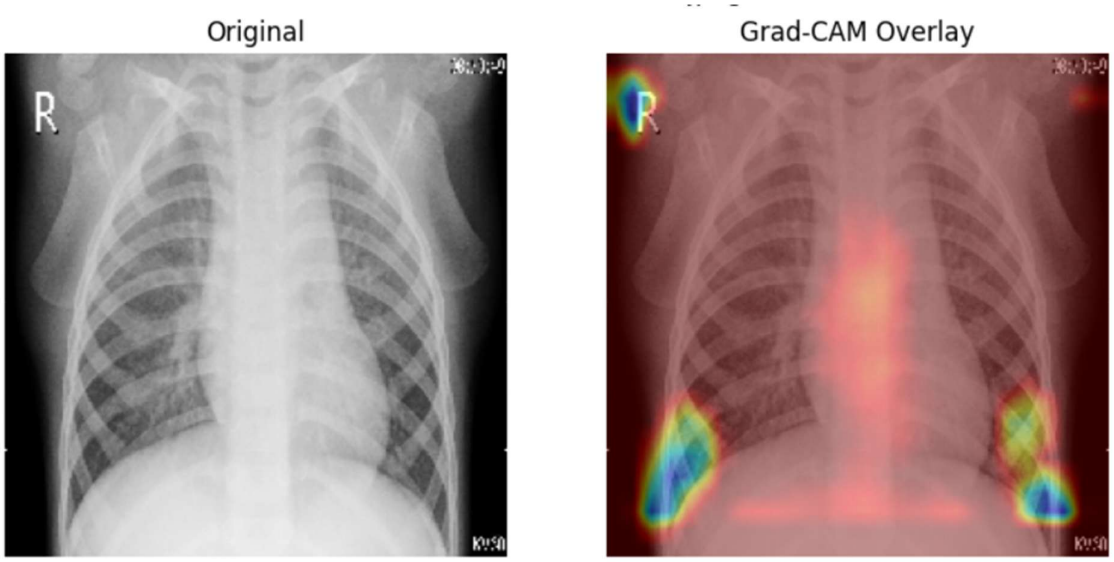
```
Classification Report:

              precision    recall  f1-score   support

      NORMAL       0.80      0.73      0.76       234
   PNEUMONIA       0.85      0.89      0.87       400

    accuracy                           0.83       634
   macro avg       0.82      0.81      0.82       634
weighted avg       0.83      0.83      0.83       634
```

These results demonstrate that the model is particularly effective at detecting pneumonia cases, as indicated by the high recall. In healthcare applications, a high recall is crucial because missing a pneumonia case could have severe consequences. Although precision for Normal cases was slightly lower, the trade-off favors patient safety, which is a desirable outcome in medical diagnostics.

The ROC curve indicated a strong separation between classes, further confirming the robustness of the model. Training and validation curves also suggested that the model generalized well without significant overfitting, thanks to the use of dropout layers and data augmentation.

Model Interpretability with Grad-CAM

While deep learning models, particularly Convolutional Neural Networks (CNNs), achieve high accuracy in image classification tasks, they are often considered "black-box" models because their decision-making process is not easily interpretable. In the context of healthcare, interpretability is crucial for building trust among clinicians and ensuring that model predictions align with meaningful clinical features. To address this, we integrated Gradient-weighted Class Activation Mapping (Grad-CAM) into our pneumonia detection pipeline. Grad-CAM generates heatmaps that highlight the most important regions of the chest X-ray images that influenced the model's decision. For example, if the CNN predicts pneumonia, the Grad-CAM visualization shows which lung areas were most critical for that prediction. This not only helps in validating whether the model is focusing on medically relevant regions (such as lung opacities) rather than irrelevant artifacts, but also makes the system more reliable for real-world clinical use. Including Grad-CAM adds a vital interpretability layer to the project, bridging the gap between high-performing AI models and human decision-making in healthcare.



Discussion

The project demonstrated the potential of CNN-based models in analyzing medical images and assisting in disease diagnosis. By focusing on pneumonia detection in pediatric chest X-rays, we addressed a critical healthcare problem. The relatively high accuracy and strong recall achieved by the model indicate its usefulness as a decision-support tool for radiologists.

However, certain limitations must be acknowledged. The dataset, while high quality, was restricted to pediatric patients aged 1 to 5. Therefore, the model may not generalize well to adult populations or to X-rays obtained with different imaging protocols. Additionally, the dataset had more pneumonia cases than normal cases, which may have influenced model performance despite the use of class weighting. Future improvements could involve using larger and more diverse datasets and experimenting with advanced architectures like ResNet or DenseNet for improved performance.

Business and Healthcare Impact

The deployment of AI models such as this one can bring significant benefits to both healthcare systems and business stakeholders.

From a healthcare perspective, the model can serve as an **assistive diagnostic tool**, helping radiologists quickly identify pneumonia cases and prioritize urgent patients. In rural or resource-constrained regions where radiologists are scarce, such a system could provide preliminary diagnoses, allowing for faster treatment and reducing mortality rates.

From a business standpoint, hospitals and diagnostic centers can integrate AI models into their **Picture Archiving and Communication Systems (PACS)** to streamline workflows and reduce turnaround times. Pharmaceutical companies and health-tech startups can also build mobile or cloud-based applications around such AI systems to provide diagnostic support at scale. Ultimately, AI-powered diagnostic tools can lead to **cost savings**, improved efficiency, and better patient outcomes.

Future Work

Future directions for this project include:

- **Transfer Learning**: Using pretrained models like ResNet or VGG to improve accuracy and reduce training time.
- **Larger Dataset**: Expanding the dataset to include adult X-rays and diverse imaging sources.
- **Real-Time Application**: Deploying the model as a web or mobile app using frameworks like Streamlit or Flask.
- **Explainable AI**: Adding techniques such as Grad-CAM to visualize which parts of the X-ray the model focused on when making predictions, improving transparency and trust.

Conclusion

This project successfully demonstrated the application of deep learning in detecting pneumonia from chest X-ray images. The CNN model achieved strong performance with an accuracy of 83% and an AUC of 0.91, showing particular strength in identifying pneumonia cases.

While challenges such as dataset imbalance and population generalization remain, the project highlights how AI can augment radiology practices, especially in resource-limited settings. By combining technical innovation with medical expertise, AI-driven tools like this one can play a vital role in **improving healthcare accessibility and outcomes worldwide**.

- Sample Predictions with X-ray images