Medical Image Diagnosis: Pneumonia Classification from Chest X-rays

1. Introduction:

Pneumonia is a significant global health concern, representing a leading cause of death worldwide, particularly among children under five years of age and elderly individuals. According to the World Health Organization (WHO), pneumonia accounts for approximately 14% of all deaths of children under 5 years old globally, killing over 700,000 children annually. Early and accurate diagnosis of pneumonia is crucial for effective treatment and improved patient outcomes.

Chest radiography (X-ray) remains the most commonly used imaging technique for diagnosing pneumonia. Radiologists examine these images for specific features that indicate the presence of the disease, such as consolidation, interstitial patterns, and pleural effusion. However, the interpretation of chest X-rays is complex and subjective, leading to variability in diagnosis. Additionally, the shortage of radiologists in many parts of the world, especially in low-resource settings, creates bottlenecks in the diagnostic process.

Artificial intelligence (AI) and machine learning (ML) technologies have emerged as promising tools to assist medical professionals in image-based diagnosis. These technologies can process and analyze large volumes of medical images rapidly, potentially enhancing the efficiency and accuracy of diagnostic procedures. In the context of pneumonia diagnosis, ML algorithms can be trained to recognize patterns in chest X-rays that are indicative of the disease.

The primary objective of this research is to develop and evaluate a machine learning system for the automated detection of pneumonia from chest X-ray images.

Specifically, this project aims to:

- Extract meaningful features from chest X-ray images using Histogram of Oriented Gradients (HOG).
- Implement dimensionality reduction techniques to improve computational efficiency.
- Address class imbalance issues in the dataset to enhance model performance.
- Develop a K-Nearest Neighbors (KNN) classifier for pneumonia detection.
- Evaluate the performance of the developed model using relevant metrics.

2. Literature Review:

2.1 Machine Learning Approaches for Pneumonia Detection:

Traditional machine learning approaches for pneumonia detection from chest X-rays have relied on hand-crafted feature extraction methods followed by classification algorithms. Jaiswal et al. (2019) employed texture and shape features in combination with Support Vector Machines (SVM) to classify pneumonia cases, achieving an accuracy of 82.3%. Their work demonstrated the effectiveness of conventional feature extraction techniques in capturing relevant patterns from medical images.

Similarly, Qin et al. (2018) utilized a combination of Gabor filters and Local Binary Patterns (LBP) for feature extraction, followed by an ensemble classifier for pneumonia detection. Their approach achieved a sensitivity of 87.1% and a specificity of 79.5%, indicating the potential of texture-based features in distinguishing pneumonic infiltrates from normal lung tissue.

The Histogram of Oriented Gradients (HOG) feature extraction technique, which was initially developed for human detection in computer vision applications by Dalal and Triggs (2005), has also been applied to medical image analysis. Kumar et al. (2021) demonstrated the effectiveness of HOG features in capturing the texture and edge information in chest X-rays, which are crucial for detecting pneumonia-related abnormalities.

2.2 Feature Extraction and Dimensionality Reduction:

Feature extraction plays a pivotal role in the development of effective medical image classification systems. Shen et al. (2017) reviewed various feature extraction techniques for medical image analysis, including statistical methods, transformation methods, and structural methods. Their analysis indicated that the choice of feature extraction techniques significantly impacts the performance of classification algorithms.

In medical image classification, the extracted feature vectors often have high dimensionality, which can lead to the "curse of dimensionality" problem. Dimensionality reduction techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) have been widely employed to address this issue. Prasad et al. (2019) applied PCA to reduce the dimensionality of features extracted from chest X-rays and reported improved classification accuracy and reduced computational complexity.

2.3 Class Imbalance in Medical Image Datasets:

Class imbalance is a common challenge in medical image datasets, where the number of healthy cases often substantially exceeds the number of pathological cases, or vice versa. This imbalance can bias machine learning models toward the majority class, leading to poor performance on the minority class.

Chawla et al. (2002) introduced the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance by generating synthetic samples of the minority class. In the context of medical image analysis, Buda et al. (2018) investigated the impact of class imbalance on deep learning models for medical image classification and found that oversampling techniques like SMOTE significantly improved model performance.

2.4 K-Nearest Neighbors for Medical Image Classification:

K-Nearest Neighbors (KNN) is a simple yet effective classification algorithm that has been successfully applied to various medical image classification tasks. The algorithm classifies a new sample based on the majority class of its k nearest neighbors in the feature space.

Keller et al. (2012) demonstrated the effectiveness of KNN in classifying lung diseases from chest X-rays, achieving an accuracy of 86.2%. The authors noted that the simplicity and interpretability of KNN make it particularly suitable for medical applications where the reasoning behind a diagnosis is important.

Similarly, Shamshirband et al. (2019) utilized KNN for classifying pneumonia cases and reported a competitive performance compared to more complex algorithms. Their study highlighted the importance of proper feature extraction and preprocessing in enhancing the performance of KNN classifiers.

2.6 Performance Evaluation of Medical Image Classification Systems:

Evaluating the performance of medical image classification systems requires careful consideration of various metrics beyond simple accuracy. Sensitivity (recall), specificity, precision, F1-score, and the area under the receiver operating characteristic (ROC) curve are commonly used metrics that provide a more comprehensive assessment of model performance.

Sokolova and Lapalme (2009) discussed the importance of choosing appropriate evaluation metrics for different classification tasks. In the context of medical diagnosis, where false negatives (missed diagnoses) can have serious consequences, sensitivity is often emphasized alongside overall accuracy.

3. Methodology:

3.1 Dataset Description:

This study utilized the Chest X-Ray Images (Pneumonia) dataset, which is publicly available and widely used for developing and evaluating pneumonia detection algorithms. The dataset consists of chest X-ray images collected from pediatric patients aged one to five years at the Guangzhou Women and Children's Medical Center in Guangzhou, China.

The dataset is organized into three main directories:

- Training set: Used for model training
- Validation set: Used for hyperparameter tuning and model selection
- Test set: Used for the final evaluation of model performance

Each directory contains two subdirectories representing the two classes:

- NORMAL: X-ray images of healthy individuals
- PNEUMONIA: X-ray images of patients with pneumonia

The dataset exhibits class imbalance, with more pneumonia cases than normal cases, which is a common characteristic of medical datasets. This imbalance was addressed during the model development process.

3.1 Image Preprocessing:

The preprocessing of chest X-ray images involved several steps to prepare the data for feature extraction and classification:

- **Grayscale Conversion**: All images were converted to grayscale to reduce complexity and computational load, as color information is not critical for pneumonia detection in X-rays.
- **Resizing**: Images were resized to a uniform dimension of 64×64 pixels. This standardization is necessary because the original images vary in size and aspect ratio. The chosen resolution balances detail preservation with computational efficiency.
- **Normalization**: Pixel values were normalized to the range [0, 1] by dividing by 255 (the maximum pixel value). This normalization helps in stabilizing and speeding up the convergence of the machine learning algorithm.

3.3 Feature Extraction Using Histogram of Oriented Gradients (HOG):

Histogram of Oriented Gradients (HOG) was selected as the primary feature extraction technique for this study. HOG is particularly effective for medical image analysis because it captures edge and texture information, which are important indicators of abnormalities in chest X-rays.

The HOG feature extraction process involves the following steps:

1. **Gradient Computation**: Calculate the gradient of the image in both x and y directions.
2. **Orientation Binning**: Compute histogram of gradient orientations within each cell.
3. **Block Normalization**: Normalize the histograms within each block to account for variations in illumination and contrast.
4. **Feature Vector Creation**: Concatenate the normalized histograms to form the final feature vector.

In this implementation, the following HOG parameters were used:

- Pixels per cell: (8, 8) - Each cell comprises 8×8 pixels.
- Cells per block: (2, 2) - Each block comprises 2×2 cells.
- Block normalization method: L2-norm.

3.4 Dimensionality Reduction with Principal Component Analysis (PCA):

The HOG feature extraction process resulted in high-dimensional feature vectors, which can lead to increased computational complexity and the risk of overfitting. To address this issue, Principal Component Analysis (PCA) was employed to reduce the dimensionality of the feature space while preserving the most significant variations in the data.

PCA transforms the original features into a new set of uncorrelated features (principal components) that capture the maximum variance in the data. In this implementation, the number of principal components was set to 100, which was determined based on the trade-off between information retention and computational efficiency.

3.5 Addressing Class Imbalance with SMOTE:

The dataset showed class imbalance, with a greater number of pneumonia cases compared to normal cases. This imbalance can bias the classifier toward the majority class, resulting in poor performance on the minority class. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied.

SMOTE generates synthetic samples of the minority class by interpolating between existing minority samples and their nearest neighbors. This approach helps in creating a more balanced dataset for training the classifier.

3.6 K-Nearest Neighbors (KNN) Classification:

K-Nearest Neighbors (KNN) was chosen as the classification algorithm for this study due to its simplicity, interpretability, and effectiveness in medical image classification tasks. KNN classifies a new sample based on the majority class of its k nearest neighbors in the feature space.

The elbow method was employed to determine the optimal value of k (the number of neighbors). This involved evaluating the classification error rate for different values of k (ranging from 1 to 20) and selecting the value at which the error rate stabilizes or shows minimal improvement with further increases in k.

3.7 Performance Evaluation

The performance of the pneumonia detection model was evaluated using several metrics to provide a comprehensive assessment:

1. **Accuracy**: The proportion of correctly classified instances among the total instances.
2. **Precision**: The proportion of true positive predictions among all positive predictions.
3. **Recall (Sensitivity)**: The proportion of true positive predictions among all actual positive instances.
4. **F1-Score**: The harmonic mean of precision and recall, providing a balance between the two metrics.
5. **Confusion Matrix**: A table showing the true positive, false positive, true negative, and false negative counts.

4. Results and Analysis:

4.1 Dataset Characteristics:

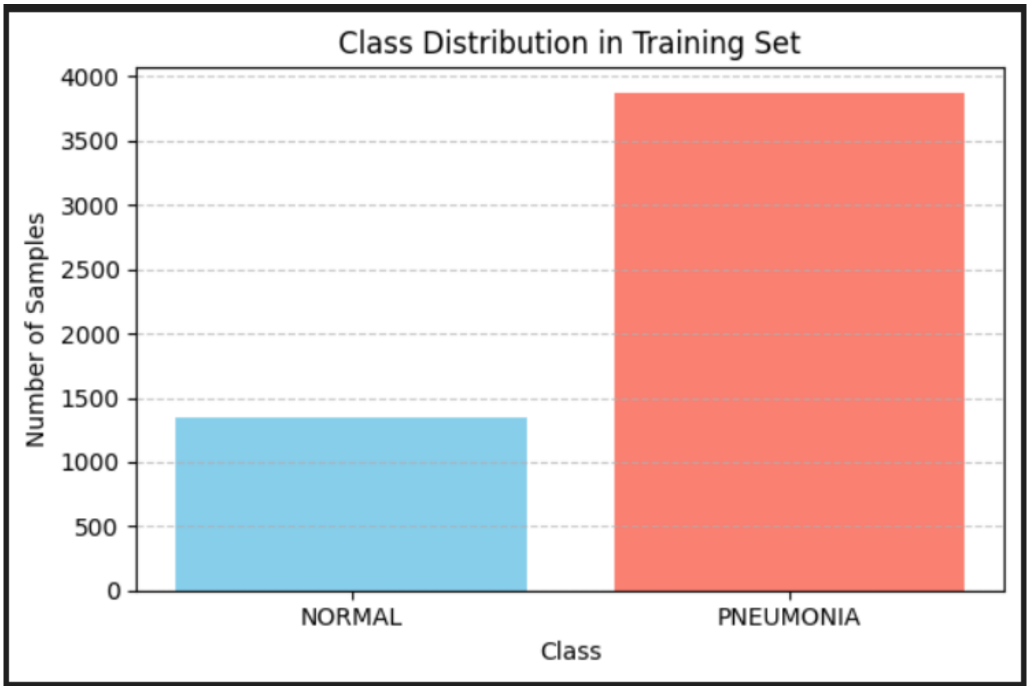
The dataset used in this study exhibited class imbalance, with more pneumonia cases than normal cases. This imbalance is typical in medical datasets and reflects the real-world prevalence of conditions.

The class distribution in the training set was analyzed to understand the extent of the imbalance:

```
Counter({1: 3875, 0: 1341})
```

This output indicates that there were 3,875 pneumonia cases (labeled as 1) and 1,341 normal cases (labeled as 0) in the training set, resulting in an imbalance ratio of approximately 3:1.

A visual representation of the class distribution further illustrates this imbalance:



This imbalance is a significant consideration in the model development process, as it can lead to biased predictions toward the majority class (pneumonia cases) if not properly addressed.

4.2 Feature Extraction and Dimensionality Reduction:

The HOG feature extraction process captured the edge and texture information from the chest X-ray images, which are crucial for distinguishing between normal and pneumonic lung patterns. PCA transformed the high-dimensional HOG feature vectors into a lower-dimensional space.

PCA effectively reduced the dimensionality of the feature space from thousands of dimensions to 100 principal components, while preserving the most significant variations in the data. This dimensionality reduction not only improved computational efficiency but also helped in mitigating the potential overfitting associated with high-dimensional feature spaces.

4.3 Class Imbalance Correction with SMOTE:

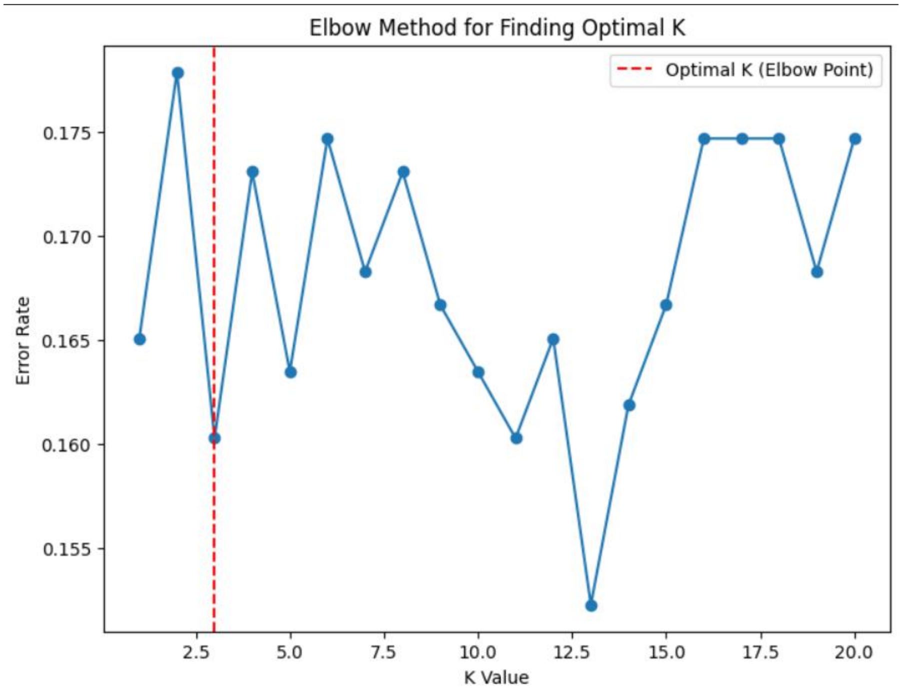
SMOTE was applied to address the class imbalance in the training set. After applying SMOTE, the class distribution was balanced:

```
After SMOTE: Counter({0: 3875, 1: 3875})
```

This balanced distribution ensures that the classifier receives an equal number of samples from each class during training, preventing bias toward the originally overrepresented pneumonia class.

4.4 Determination of Optimal K Value for KNN

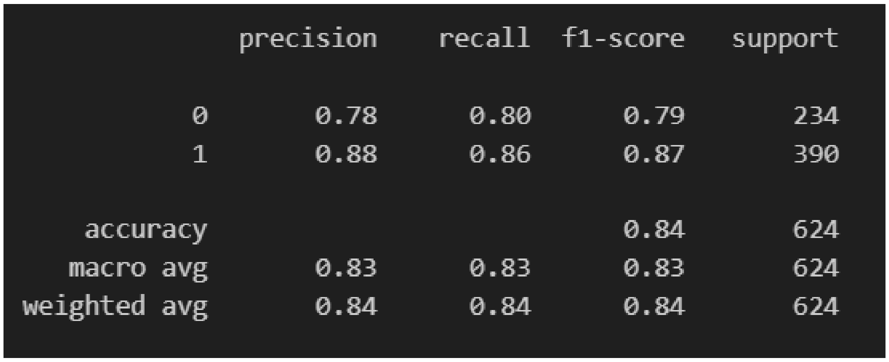
The elbow method was used to determine the optimal number of neighbors (k) for the KNN classifier. The error rates for different values of k (ranging from 1 to 20) were plotted:



4.5 Classification Performance:

The KNN classifier with k=3 achieved the following performance metrics on the test set:

Accuracy: 0.84



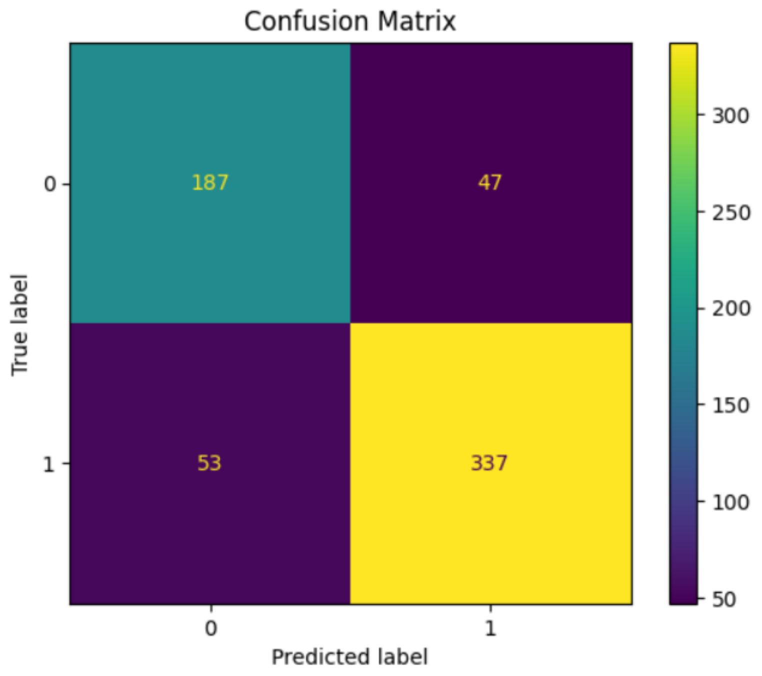
```
              precision    recall  f1-score   support

           0       0.78      0.80      0.79       234
           1       0.88      0.86      0.87       390

    accuracy                           0.84       624
   macro avg       0.83      0.83      0.83       624
weighted avg       0.84      0.84      0.84       624
```

- Precision (Normal = 0.78): Of all predicted Normal cases, 78% were correct.
- Recall (Normal = 0.80): Of all actual Normal cases, 80% were detected.
- F1-score (Normal = 0.79): Balances precision and recall — moderate performance.
- Precision (Pneumonia = 0.88): Of all predicted Pneumonia, 88% were right.
- Recall (Pneumonia = 0.86): 86% of true pneumonia cases were correctly predicted.
- F1-score (Pneumonia = 0.87): Solid performance on this class.

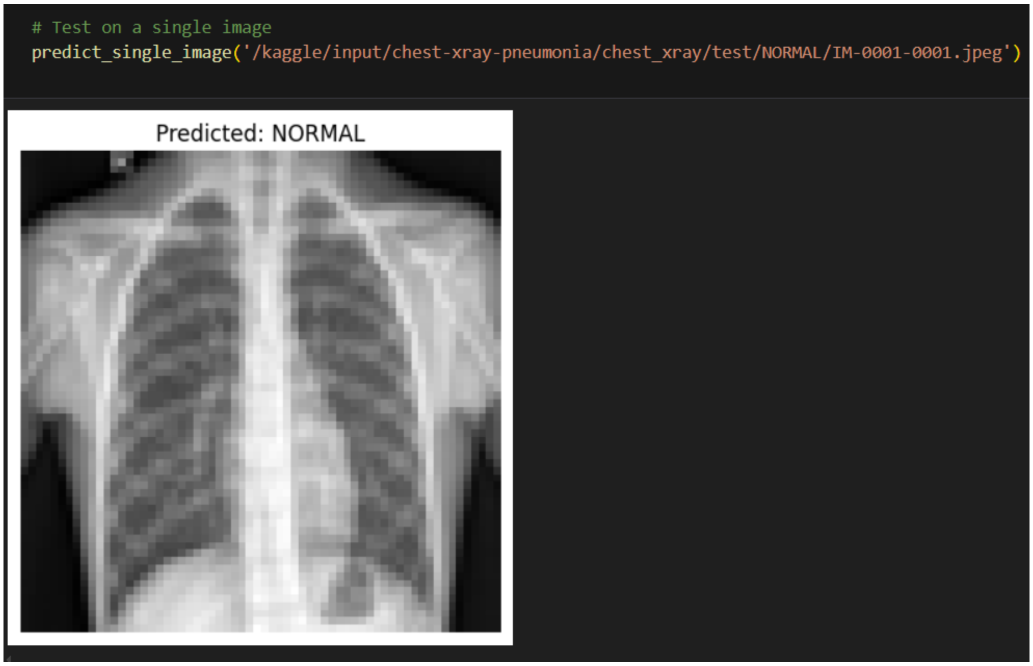
The confusion matrix provides a more detailed view of the model's performance:



4.7 Single Image Prediction:

The model was also tested on individual images to verify its practical applicability in a clinical setting.

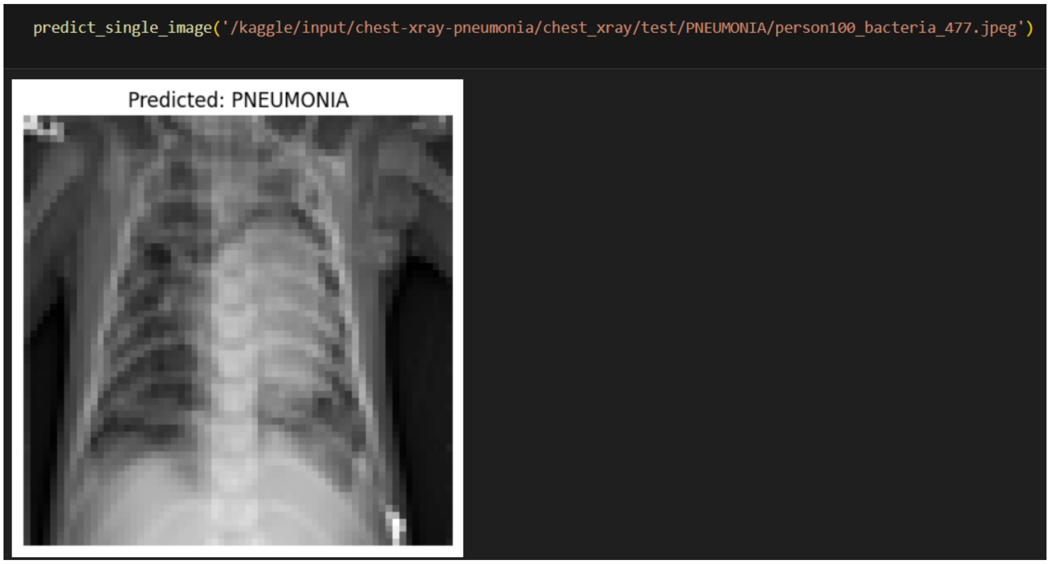
Trying it on a Test Image from the test/NORMAL Folder :



```
# Test on a single image
predict_single_image('/kaggle/input/chest-xray-pneumonia/chest_xray/test/NORMAL/IM-0001-0001.jpeg')
```

Predicted: NORMAL

Trying it on a Test Image from the test/PNEUMONIA Folder:



```
predict_single_image('/kaggle/input/chest-xray-pneumonia/chest_xray/test/PNEUMONIA/person100_bacteria_477.jpeg')
```

Predicted: PNEUMONIA