Pneumonia Detection

Abdul Moiz Abid (CS-21140)

Syed Ismaeel Hamza(CS-21147)

July 2024

Abstract

This report presents a machine learning approach to detecting pneumonia from chest X-ray images. The study involves data preprocessing, exploratory data analysis, feature engineering, model development, and evaluation. Results show that our model achieves a high accuracy in distinguishing between normal and pneumonia-affected images.

1 Introduction

Pneumonia is a serious respiratory condition that can be life-threatening if not diagnosed and treated promptly. Early and accurate detection of pneumonia can significantly improve patient outcomes and reduce the burden on healthcare systems. Traditional methods of pneumonia diagnosis, which rely on clinical examination and radiographic analysis by medical professionals, can be time-consuming and subject to human error.

In recent years, machine learning techniques have shown great potential in automating medical image analysis, offering the promise of faster and more accurate diagnoses. This study aims to develop an automated system for pneumonia detection using chest X-ray images and various machine learning algorithms. Specifically, we will explore the performance of three popular classification models: Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN).

The choice of these models is motivated by their widespread use and proven effectiveness in image classification tasks. Logistic Regression, a linear model, is known for its simplicity and interpretability. SVM, a powerful and versatile classifier, is capable of handling both linear and non-linear data through the use of kernel functions. KNN, a non-parametric method, is easy to implement and often performs well with small to medium-sized datasets.

By comparing the performance of these models, we aim to identify the most effective approach for pneumonia detection in chest X-ray images. The performance of each model will be evaluated based on metrics such as accuracy, precision, recall, F1 score etc. Additionally, we will discuss the advantages and limitations of each method in the context of medical image analysis.

This study not only contributes to the ongoing research in medical image analysis but also provides insights into the practical application of machine learning techniques for pneumonia detection. The development of an accurate and reliable automated system has the potential to assist healthcare professionals in making timely and informed decisions, ultimately improving patient care and outcomes.

2 Data Preprocessing

The dataset consists of chest X-ray images categorized into normal and pneumonia classes. Effective preprocessing is crucial to enhance the quality of the images and ensure that the machine learning models can learn from the data effectively. The preprocessing steps include resizing images to a uniform dimension, converting them to grayscale, and normalizing pixel values. These steps help in standardizing the input data and reducing computational complexity.

2.1 Image Resizing

All images were resized to 64x64 pixels to ensure uniformity in input dimensions for the model. This size was chosen to balance the trade-off between preserving important features and reducing the computational load. Resizing the images allows the model to process data more efficiently and ensures that each input image has the same scale, which is essential for consistent model training.

2.2 Grayscale Conversion

Images were converted to grayscale to reduce computational complexity and focus on essential features. Color information is often not necessary for detecting pneumonia in X-ray images, and converting to grayscale reduces the data size from three channels (RGB) to one. This simplification helps in highlighting the structural details of the lungs, which are critical for identifying signs of pneumonia.

2.3 Normalization

Pixel values were normalized to a range of [0, 1] to standardize the input data for the machine learning model. Normalization helps in mitigating the effect of varying lighting conditions in the X-ray images and ensures that the pixel values are on a consistent scale. This step is essential for improving the convergence rate of the learning algorithms and enhancing their performance. Normalization is achieved by dividing each pixel value by the maximum possible value (255 for 8-bit images), resulting in values between 0 and 1.

3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in understanding the dataset's characteristics and uncovering patterns that can inform the model development process. In this study, we perform EDA on the chest X-ray images to gain insights into the distribution of the data, visualize sample images, and analyze various statistical properties of the images.

3.1 Image Distribution

We begin by examining the distribution of the images across the two classes: normal and pneumonia. This helps us understand the balance of the dataset and identify any potential class imbalance issues, which is important for ensuring that the model does not become biased towards the more prevalent class.

3.2 Sample Image Visualization

To get an initial visual understanding of the data, we display a sample of images from both the normal and pneumonia classes. Observing sample images allows us to identify any visible differences between the two classes and gain intuition about the features that might be useful for classification.

3.3 Class Distribution

The distribution of images across the two classes is visualized using a count plot. This plot provides a clear picture of the number of images in each class, helping us to identify if any class has a significantly higher number of images. Understanding class distribution is essential for assessing whether any rebalancing techniques are needed during model training.

3.4 Image Statistics

We calculate and analyze various statistical properties of the images, such as dimensions and mean pixel values. This helps in understanding the general characteristics of the images in each class.

3.4.1 Dimensions

We explore the dimensions of the images in both classes to check for consistency and any notable differences. Consistent image dimensions are crucial for feeding the images into machine learning models, which often require uniform input sizes.

3.4.2 Mean Pixel Values

The mean pixel value distribution is analyzed to understand the brightness levels of the images in each class. Differences in mean pixel values can indicate variations in image acquisition settings or inherent differences between normal and pneumonia-affected lungs.

3.5 Aspect Ratios

We calculate the aspect ratios of the images to identify any patterns or differences between the two classes. Aspect ratio analysis helps in understanding the shape and orientation of the images, which can be relevant for feature extraction and model performance.

3.6 Pixel Intensity Histogram

We plot the intensity histograms of the images to analyze the distribution of pixel intensities. This provides insights into the contrast and brightness characteristics of the images in each class. Understanding pixel intensity distributions helps in preprocessing steps like normalization and contrast adjustment.

3.7 Brightness and Contrast

The brightness and contrast of the images are key factors that can affect the performance of image classification models. We analyze the distributions of these metrics for both classes to understand their impact.

3.7.1 Brightness Distribution

We visualize the distribution of brightness values across the images in each class. Brightness levels can influence how well the model distinguishes between different classes, making it important to ensure that there is no significant bias in brightness.

3.7.2 Contrast Distribution

Similarly, the contrast distribution is examined to understand the variability in pixel intensity within each image. Higher contrast generally helps in better feature extraction, making it easier for the model to learn discriminative features.

3.8 Interpretation of Results

The findings from the EDA provide valuable insights into the dataset. Key observations include the distribution of image dimensions, mean pixel values, aspect ratios, and the overall brightness and contrast levels. These insights will guide the subsequent preprocessing and model development steps, ensuring that the models are well-suited to the characteristics of the data.

4 Feature Engineering

To enhance the performance of the machine learning models, feature selection and normalization techniques were employed. Initially, the SelectKBest method with the ANOVA F-value (f_classif) was utilized to select the 500 most relevant features from the dataset. This step is crucial in reducing dimensionality, removing redundant features, and focusing on the most informative aspects of the data, thereby improving model efficiency and accuracy. Following feature selection, standard scaling was applied to the selected features using StandardScaler. This normalization process involves transforming the features such that they have a mean of zero and a standard deviation of one, ensuring that the features contribute equally to the model and enhancing the convergence rate during training. The combined approach of feature selection and scaling prepares the dataset for optimal performance with machine learning algorithms.

5 Models

This section describes the various models employed in the study for pneumonia detection using chest X-ray images. We compare the performance of logistic regression, support vector machine (SVM), and k-nearest neighbors (KNN) models, including both custom implementations and sklearn-based models.

5.1 Logistic Regression

Custom Logistic Regression

- **True Positives (TP):** 69
- **True Negatives (TN):** 30
- **False Positives (FP):** 24
- **False Negatives (FN):** 5

Interpretation: The custom logistic regression model correctly identified 69 instances of pneumonia and 30 normal instances. However, it also misclassified 24 normal cases as pneumonia and 5 pneumonia cases as normal. This indicates that while the model has a good sensitivity (low FN), its specificity is relatively lower (higher FP).

Sklearn Logistic Regression

- **True Positives (TP):** 68
- **True Negatives (TN):** 40
- **False Positives (FP):** 14
- **False Negatives (FN):** 6

Interpretation: The sklearn logistic regression model correctly identified 68 pneumonia cases and 40 normal cases. It misclassified 14 normal cases as pneumonia and 6 pneumonia cases as normal. This model shows a better balance between sensitivity and specificity compared to the custom implementation, with fewer false positives and a similar number of false negatives.

5.2 K-Nearest Neighbors (KNN)

Scratch KNN

- **True Positives (TP):** 68
- **True Negatives (TN):** 35
- **False Positives (FP):** 19
- **False Negatives (FN):** 6

Interpretation: The Scratch KNN model correctly identified 68 pneumonia cases and 35 normal cases. It misclassified 19 normal cases as pneumonia and 6 pneumonia cases as normal. The performance of this model lies between the custom logistic regression and sklearn logistic regression, with a slightly better specificity than the custom logistic regression but not as good as the sklearn logistic regression.

Sklearn KNN

- **True Positives (TP):** 68
- **True Negatives (TN):** 35
- **False Positives (FP):** 19
- **False Negatives (FN):** 6

Interpretation: The sklearn KNN model shows a similar pattern to the scratch KNN model with a slight improvement. Compared to the custom and sklearn SVM models, the KNN models have higher false positives but similar true positives. Both SVM models tend to have slightly better performance metrics, especially in terms of lower false negatives and slightly better precision and recall.

5.3 Support Vector Machine (SVM)

Custom SVM

- **True Positives (TP):** 68
- **True Negatives (TN):** 39
- **False Positives (FP):** 15
- **False Negatives (FN):** 6

Interpretation: The custom SVM model has a relatively high number of true positives and true negatives, indicating good performance. The false positives and false negatives are reasonably low, though the custom SVM seems to misclassify a few instances.

Sklearn SVM

- **True Positives (TP):** 70
- **True Negatives (TN):** 39
- **False Positives (FP):** 15
- **False Negatives (FN):** 4

Interpretation: The sklearn SVM model performs similarly to the custom SVM, but it has slightly fewer false negatives and more true positives, which suggests that the sklearn SVM might be marginally more accurate.

5.4 Overall Comparison

Sensitivity: All three models have high sensitivity, with sklearn logistic regression and Scratch KNN having similar performance. **Specificity:** Sklearn logistic regression shows the highest specificity, followed by Scratch KNN, and then the custom logistic regression.

In summary, the sklearn logistic regression model appears to perform the best overall, with a good balance between high sensitivity and high specificity. The custom logistic regression has higher sensitivity but at the cost of higher false positives. Scratch KNN performs moderately well but does not surpass the sklearn logistic regression in either metric.

Model	TP	TN	FP	FN	Accuracy
Custom Logistic Regression	69	30	24	5	77.34%
Sklearn Logistic Regression	68	40	14	6	84.38%
Scratch KNN	68	35	19	6	80.47%
Sklearn KNN	68	35	19	6	80.47%
Custom SVM	68	39	15	6	83.59%
Sklearn SVM	70	39	15	4	85.16%

Table 1: Confusion Matrix and Accuracies Summary for Different Models

6 Conclusion

In this project, we developed and evaluated multiple models for pneumonia detection using chest X-ray images, including logistic regression, SVM, and KNN, with both custom and sklearn implementations. The sklearn logistic regression model demonstrated the best overall performance with a balanced sensitivity and specificity. Our custom models, while effective, exhibited higher false positive rates compared to their sklearn counterparts. The SVM models, both custom and sklearn, showed competitive results with slightly lower false negatives. This comparative analysis highlights the importance of selecting appropriate models and implementation methods for medical image classification tasks to ensure high accuracy and reliability.