STA 141C Final Project

# Pneumonia Prediction by X-Ray Imaging

Sung Woo Bak, Jonathan Casas-Ramirez, Adam Hetherwick, Jose Arreola Muñoz

Instructor: Dr. Nina Dörnemann

03/18/2024

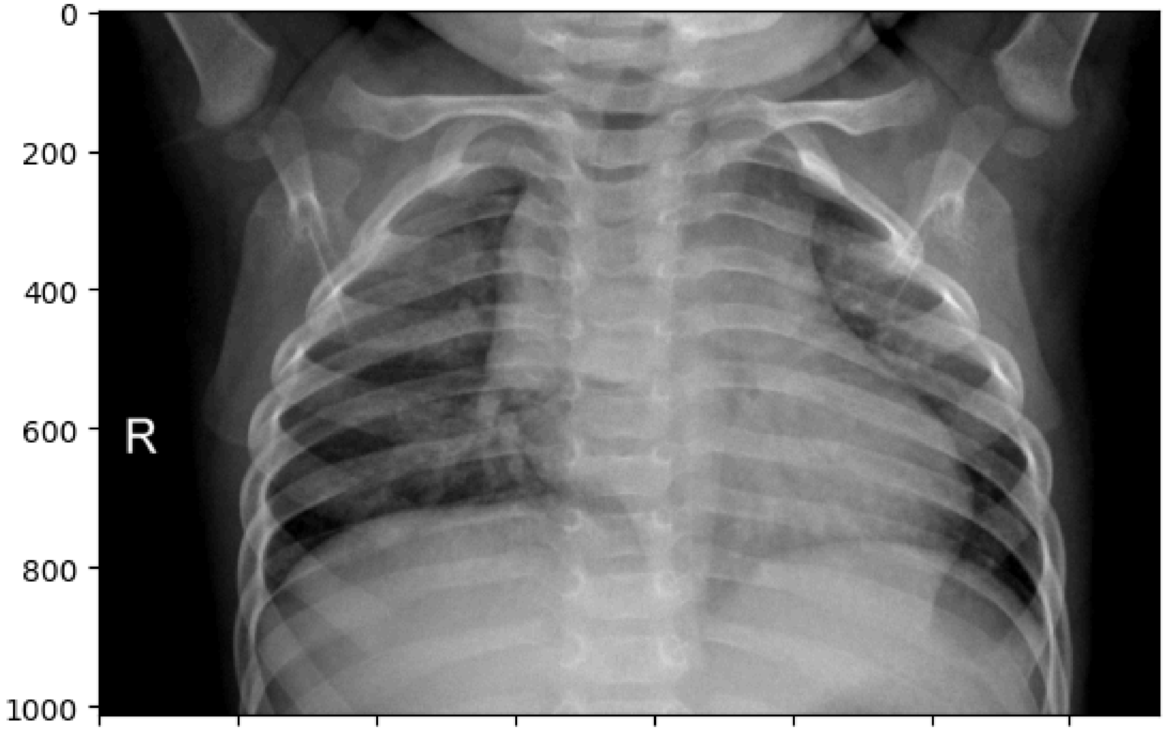


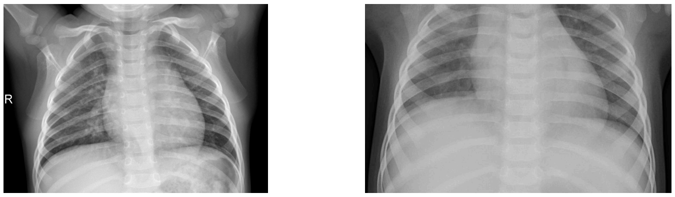
Figure 1: Chest X-ray (CoronaHack Chest X-Ray Database)

# Introduction

Chest X-rays are regurlaly used to diagnose and detect various diseases and viruses, including pneumonia, tuberculosis, and Covid-19. Visual interpretations of these images may be misleading and lead to false diagnoses due to human error. A shocking 38.8% of patients with pneumonia were misdiagnosed upon initialization of treatment, most of them being incorrectly diagnosed with upper respiratory tract infections (Ang 2020). Incorrect diagnostics can lead to prolonged treatment for an absent disease and further development of pneumonia. It is also harmful to patients who were diagnosed as healthy when pneumonia was present. This project aims to develop an X-ray image classification model to recognize pneumonia at a higher rate than visual interpretations. This report will outline the process of handling image data into formats suitable for analysis, and discuss the models used for classification. The report will then evaluate the performance of the models using a large test subset of labeled chest X-ray images. Our project uses the Covid-19/pneumonia and CT dataset provided by Joseph Paul Cohen available both on GitHub and Kaggle.

# Exploratory Data Analysis

The dataset consists of metadata containing the image names, labels, and chest X-ray images of normal and pneumonia. We started by loading the 5,910 X-ray images, 73% of which are labeled ‘pneumonia’ and 26% of which are ‘Normal’. This 73:26 split of class posed issues for analysis, and handling this indifference is addressed later in the report. This dataset was pre-split into train and test sets with a proportion of 90% and 10%, respectively. The test set contains 390 pneumonia and 234 normal X-ray images.



Figures 2 & 3: Normal (Left) & Pneumonia (Right)

The images were stored in matrices with dimension width (pixel) x height (pixel) x 3 (RGB). Each entry represents the RGB value (from 0 to 255) of the corresponding pixel. This forms the predictive basis of our analysis. We planned to use these values as our predictors for the classification in this project. Each image is represented as a matrix and not a vector, which may cause problems in training the model. Pre-processing of these matrices is discussed in the next section.

# Methodology

The classification in this project was performed with logistic regression, linear discriminative analysis, quadratic discriminative analysis, and convolutional neural network. To train the candidate classification models, the images had to be processed into a proper format for feature extraction and accurate classification.

## Image Preprocessing

The RGB format of the X-rays was unnecessary, therefore, the image data was converted to grayscale, which lowers the dimension. These matrices were flattened and converted into vector format for proper analysis of features and predictors. Each image also had a unique size, so unifying the size and focusing on the chest (center of the image) where the important features are was crucial for analysis. The image processing was challenging due to the high dimensionality. The images were resized to 16 x 16 and 128 x 128, and the models' accuracy is tested to obtain the performance based on the resolution.

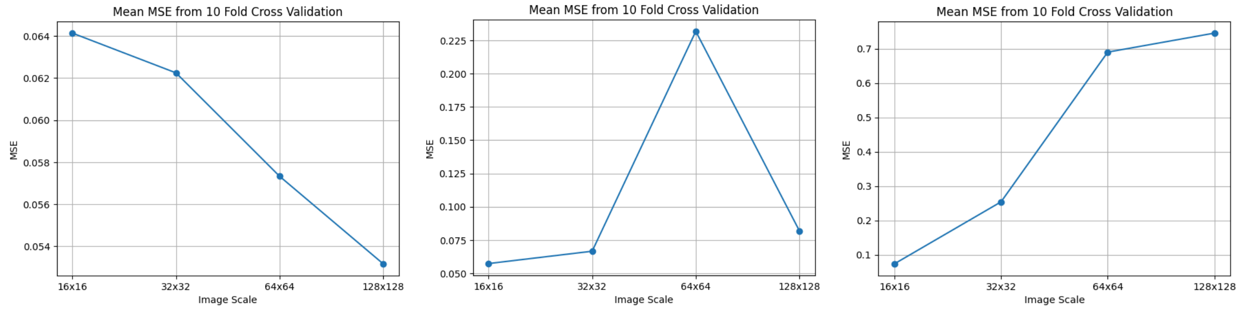
## Classification Models

The candidate models (logistic regression, LDA, QDA) are trained with preprocessed image data by different image sizes or dimensions. Then, the accuracy scores and the means of mean squared errors are computed using k-fold cross-validation (k=5, 10) on the train set. The classification rates on the test set are also obtained. This procedure is conducted to find the performance of the models and the result based on images with different dimensions. The images with lower dimensions can be chosen for efficiency if the results on low-dimension images are reliable. To perform accurate predictions on image data, a convolution neural network model was utilized as well. Convoluted neural networks are one of the strongest and most popular methods used in image classification. In this project, two crucial layers in our model are convolution and pooling. The convolutional layer is equipped with a specified number and size of kernels, allowing it to process values of sub-regions (local connectivity). This enhances the model's efficiency. The pooling layer, which follows the convolutional layer, selectively retains features (eg. MaxPooling) from each region, thereby reducing the input size. This contributes to the model's robustness. By utilizing these features, the model is constructed. Furthermore, dense layers are added after the above two layers and selected activation functions are assigned to each layer. In this project, multiple activation functions (eg. ReLU and tanh) were tested and different layers were added and removed to find the model with the highest accuracy. The final model includes two convolution layers, two max-pooling layers, and four dense layers, and Leaky ReLU and sigmoid are used as the activation functions. To avoid overfitting and obtain efficiency, multiple dropout layers (randomly switching off the nodes) are also included with a rate of 0.5. In executing the model, ADAM is used as an optimizer, and binary cross-entropy is used to calculate the loss. Due to the imbalanced proportion of the classes, class weights were added to the model when it was compiled to prevent bias. In selecting the final model, two key measures were considered: loss and accuracy. In this project, priority was given to accuracy, while also taking loss into account. After training the models for a total of 30 epochs (early stopping was implemented), the model with the highest accuracy was chosen, and the corresponding loss was also taken into consideration to finalize the selection. To find the performance of the model, the classification rate on the test set and the confusion matrix were obtained.

# Results

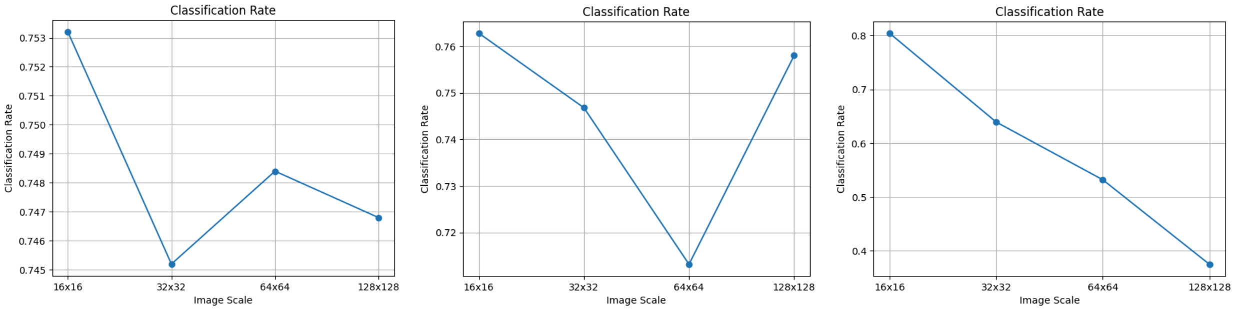
## Logistic Regression, LDA, & QDA

The performance of our models was determined by visualizing the classification rates and mean squared error on k-fold cross-validation for different image resolutions or dimensions.



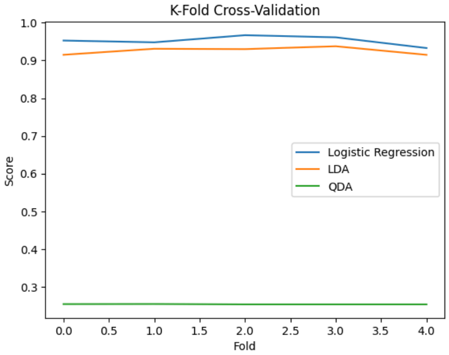
Figures 4, 5 & 6: Mean MSE from K-fold Cross-Validation (K=10)

The plots above represent the mean MSE obtained from k-fold cross-validation for Logistic Regression, LDA, and QDA at various image resolutions. As observed, Logistic Regression and LDA exhibit relatively low mean MSE, whereas QDA shows higher mean MSE. Additionally, it cannot be concluded with certainty in LDA and QDA that a higher dimension in image leads to a decrease in mean MSE.



Figures 7, 8 & 9: Classification Rate on Test Set

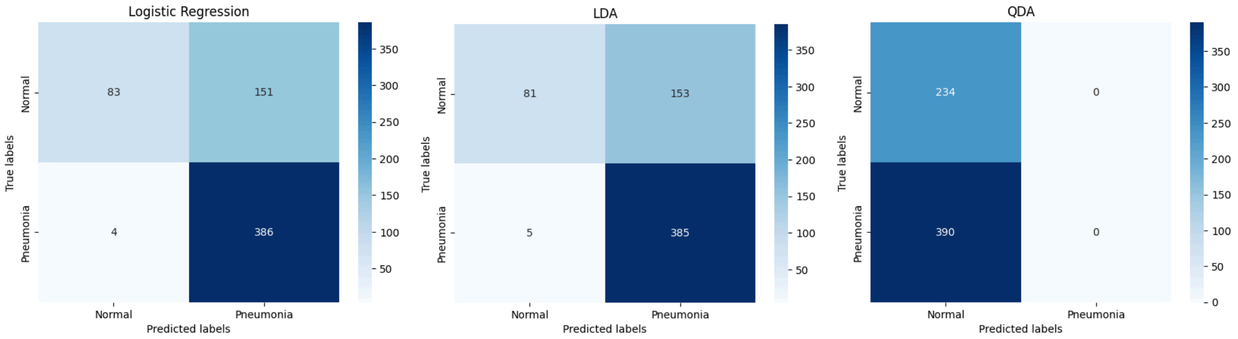
The plots above display the accuracy rate for each model on the test set. Logistic Regression and LDA exhibit classification rates of over 70%, whereas QDA shows a decrease as the image dimensions increase. This suggests that the boundary for binary classification in this project is linear. However, achieving the highest accuracy of only 70% is disappointing. To evaluate the performance of these models, the k-fold cross-validation scores, classification rates, and confusion matrices were computed using 256 by 256 image data.



| Classification Rate | |
|---|---|
| Logistic | 0.7516 |
| LDA | 0.7468 |
| QDA | 0.375 |

Figures 10 and 11: Scores from K-fold Cross-Validation (K=5) and Classification Rate

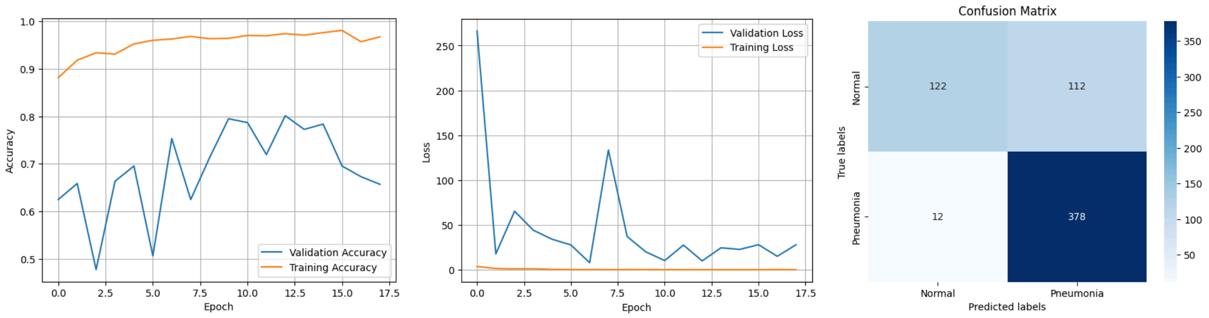
The plot above represents the scores obtained from k-fold classification (K = 5) for each model, while the table shows the classification results for the test set. Logistic regression and LDA exhibit high scores above 0.9, whereas QDA shows considerably lower scores less than 0.3. Additionally, when examining the classification rates for the test set, logistic regression and LDA demonstrate rates of over 70%, whereas QDA shows a rate in the 30% range.



Figures 12, 13 & 14: confusion matrices for Logistic, LDA & QDA

The above figures represent the confusion matrices for the test set predictions of each model. In the case of QDA, all predictions were classified as "Normal." As seen earlier, QDA is not a suitable model for this project. One noteworthy point is that when the actual class is "Normal," most models, except for QDA, predicted a substantial portion as "Pneumonia." This could be due to the imbalance in the data. To address this, one approach is to create a subset of the training set and train the model on it. This subset should ideally balance the classes to mitigate the impact of class imbalance on the model's performance.

## Convolution Neural Network



Figures 15, 16, & 17: Convolutional Neural Network Accuracy Rate, Loss, Confusion Matrix

The plots above display the accuracy and loss obtained during the training of a CNN model, along with the confusion matrix on the test set. The yellow line represents the accuracy rates and loss for the train set and the blue line represents the test set. At epoch 13, the model achieved the highest accuracy rate of 0.8013 and a relatively low loss of 9.906. Subsequently, the model trained at epoch 13 was used to make predictions on the test set, as shown in the confusion matrix above. In this matrix, consistent errors are observed compared to the previous models. However, the relatively low error rate and the highest classification rate compared to other models suggest that selecting this model is justified.

## Discussion and Outlook

A separate subset with equal class proportions was extracted from the training set, and the model was trained on it. From this process, a mean score of 0.976 was obtained through k-fold cross-validation (k = 5), with the highest accuracy rate reaching 0.8285 and a loss of 4.3174. Although relatively high accuracy rates were observed across different epochs, significant differences were not evident. This issue can be handled with modification of the layers, learning rate, selection of different activation functions, etc to find the optimal model for X-ray image classification. However, due to the size of the image data used in this project and the limitation of equipment, other methods could not be further applied. Other similar studies also had the highest accuracy rates of around 80 percent, indicating the results of this project are near maximized.

There exists an alternate approach to improve the performance of the models in this project, which is to first apply the Gray Level Co-occurrence Matrix to extract important features from each image. GLCM analyzes the co-occurrence of pixel intensities and their directions to form features such as roughness, smoothness, contrast, etc. From those chosen features, they can be used as input for LDA and QDA model comparison. The main downside of this approach is that prior knowledge about how specific diseases look is important to determine the features. This can be useful for future predictive projects, and to work around models with dimension limitations.

## Conclusion

Convolution neural networks had exemplary performance due to the increased flexibility and higher dimension compatibility. After vectorizing the chest X-ray images into a consistent format, Logistic Regression, LDA, QDA, and CNN models were trained to predict the presence of pneumonia in patients. The unbalanced split in pneumonia to normal could cause the models to be biased toward the dominant group but the result using the subset of the train set with equal proportion failed to provide a difference. Modifying the model or different methods for image processing may improve the performance. Throughout this project, we have learned that the high dimension of our input vector may be non-compatible with Logistic Regression, LDA, and QDA. The implementation of CNN to work around these limitations was crucial, as the predictive rate was relatively higher. Modeling with CNN provided the highest classification rate of over 80 percent, 0.8013, or 0.8285 in this project.

# References

**Covid-19/Pneumonia Dataset** (5933 Images)

Joseph Paul Cohen

- https://github.com/ieee8023/covid-chestxray-dataset

**In-text citations**

1. Ang, CS. “Misdiagnosis of Community-Acquired Pneumonia in Patients Admitted to Respiratory Wards, Penang General Hospital.” *The Medical Journal of Malaysia*, U.S. National Library of Medicine, 1 July 2020, pubmed.ncbi.nlm.nih.gov/32723999/.
2. Cohen, Joseph Paul. “Covid-Chestxray-Dataset.” 1 Oct. 2020, Accessed 3 Mar. 2024.
3. Rahman, Tawsifur. “Tuberculosis (TB) Chest X-Ray Database.” 1 Aug. 2020, Accessed 3 Mar. 2024.