PRML Course Project: COVID Detection Using X-Ray Images

Soham Deshmukh (B21EE067)

Priyanshi Shah (B21EE062)

Bikas Singh (B21EE015)

Abstract – *The global pandemic of COVID- 19 affects global health care and lifestyle, and its early detection is crucial for controlling the spread of cases and mortality. Actual diagnostic leader is the Reverse transcription Polymerase chain reaction (RT-PCR), but result times and costs are high, necessitating the development of additional rapid and inexpensive diagnostic instruments. This paper describes our experience with COVID-19 detection using chest X-ray (CXR) images. On CXR images, we attempt to implement unsupervised, supervised, and deep learning methods for binary image classification (COVID or Non-COVID). Data preprocessing pipeline included reduction techniques - PCA and LDA. We analysed models such as Random Forest classifier, Decision Tree classifier, GaussianBayes, and XGBoost in supervised learning. For the Deep Learning approach, we created a custom pipeline that utilises segmentation and CNN models. The best models obtained a COVID-19 detection accuracy of approximately 96%.*

I. Introduction

Classification is a learning technique in which the predicted values (categorical variables) belong to a specific number of classes. It is used for detecting fraudulent emails, recognising images, determining the presence or absence of a disease, and recognising digits. The output variable can belong to either two cases (binary classification) or multiple classes (multi-class classification).

II. Overview of the pipeline
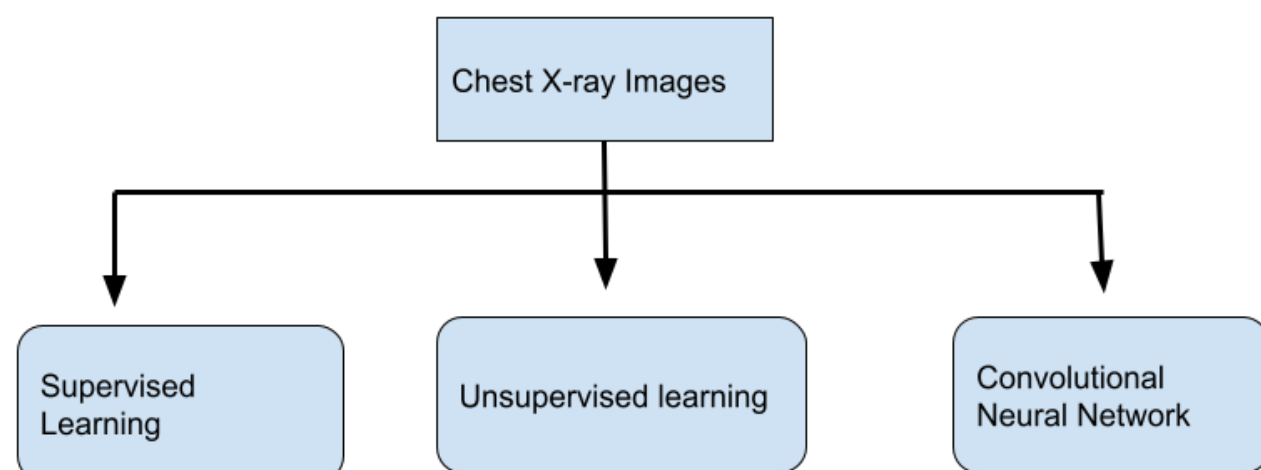
III.   Importing the Dataset

The dataset comprises of chest X-ray scans for COVID, non-COVID (including Viral Pneumonia) cases. The data has been derived from [Github](#), [SIRM](#), [Github-CXNET](#) and [Eurorad](#).

Retrieval of data from these sources is performed using various parsing techniques (including wget and git cloning of repositories). The image data is then converted into numpy arrays using PIL.Image library and all the images are resized for uniformity in the standard (250*250) shape.

To create the dataset, all the images are combined in an array.

IV.   Data Preprocessing and Analysis

The array is then converted to a Pandas dataframe. Class labelling (for supervised purposes) is done on the basis of the category of the dataset the image belongs to - COVID or Non-COVID. Each datapoint in this dataframe has 62500 attributes - the number of pixel values per image.

Considering the complexity of the data and the computational cost involved, the number of attributes in the dataset must be reduced. For this, dimensionality reduction techniques such as PCA and LDA are used.

The advantages of PCA are :

- Reducing the time and storage space required.
- Removing multi-collinearity which improves the interpretation of the parameters of the machine learning model. Considering the complexity of the data and the computational cost involved, the number of principal components is set as 20.

The advantages of LDA are :

- Extracting the most relevant features that best separate the classes in the data, resulting in improved classification performance.
- Increasing the separation between the points from distinct classes

Considering the computational cost involved, relevance of the dataset and the overall impact of the transformation, the number of attributes are reduced to 20 using these reduction techniques.

V.   Supervised Learning Techniques

Gaussian Naive Bayes: It is based on Bayes' theorem. It is a probabilistic method that assumes independent Gaussian characteristics. It is "naive" because it thinks features are independent, which is not always true. The algorithm chooses the class with the highest probability based on input features. Gaussian Naive Bayes is fast and easy for high-dimensional data.

Decision Tree Classifier: It does classification and regression using supervised learning. The decision tree starts at the root node and displays a feature with branches representing its possible values. Using feature value, the algorithm chooses a path at each node. The process continues until a leaf node makes the ultimate decision. Interpretability and ability to handle category and numerical data make the decision tree classifier popular.

Random Forest Classifier: It is a supervised learning algorithm that uses ensemble learning method for classification. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. In case of Random Forest, these algorithms would be Decision Trees.

XGBoost: It is a popular machine learning algorithm for classification and regression tasks. Gradient boosting creates a strong learner from numerous weak learners. Iteratively adding trees that minimise error optimises the loss function using decision trees as weak learners. Industry uses XGBoost for fraud detection and customer churn prediction due to its speed and accuracy. Regularisation and early halting reduce overfitting and improve generalisation.

VI.   Unsupervised Learning Techniques

K Means: It organises data points into k clusters based on their similarity or distance. The algorithm initialises k centroids, assigns data points to the closest centroid, updates centroids, and repeats until convergence. It assumes spherical clusters, identical size, and similar densities and converges to local optima.

VII.   Deep Learning Techniques

VIII.   Comparative Study of Models

The comparison of these models is done by means of performance metrics such as accuracy score, precision and recall. As for the clustering models (KMeans), we shall use the acccuracy and Silhouette score as a metric.