COVID Detection using X-ray Images

Yuvraj Soni, B21EE089
Department of Electrical Engineering
Indian Institute of Technology Jodhpur
Rajasthan, India
soni.19@iitj.ac.in

Shreejan Kumar, B21EE088
Department of Electrical Engineering
Indian Institute of Technology Jodhpur
Rajasthan, India
kumar.301@iitj.ac.in

Vasubhya Diwan, B21AI044
Department of Computer Science
Indian Institute of Technology Jodhpur
Rajasthan, India
diwan.1@iitj.ac.in

Abstract

The COVID-19 pandemic has posed a significant threat to global public health. Machine learning can aid in detecting COVID-19 cases using chest X-ray images. This is a classification problem where a machine learning model is trained to classify an input chest X-ray image as COVID-19 positive or negative. In this study, we explore various machine learning techniques to detect COVID-19 cases from chest X-ray images. We used a publicly available dataset of chest X-ray images and compared the performance of different machine learning algorithms such as logistic regression, random forest, and convolutional neural networks. We evaluated the performance of the models using metrics such as accuracy, precision, recall, and F1 score. Our results show that the convolutional neural network achieved the highest overall performance, indicating its effectiveness in detecting COVID-19 cases from chest X-ray images. These findings demonstrate the potential of machine learning in COVID-19 detection, highlighting the importance of accurate and timely diagnosis to control the spread of the disease.

I. INTRODUCTION

The COVID-19 pandemic has presented a significant challenge to global public health. One of the areas where machine learning can play a crucial role is in the early detection of COVID-19 cases. Chest X-ray images can provide valuable diagnostic information for COVID-19 detection. In this project, we explore the use of machine learning techniques for COVID-19 detection using chest X-ray images. The task is to classify an input chest X-ray image as either COVID-19 positive or negative, which is a simple binary classification problem. However, the challenge lies in developing accurate models, especially in the absence of a balanced dataset. In this study, we aim to develop a machine learning model that can accurately detect COVID-19 cases from chest X-ray images. We evaluate various machine learning algorithms, including convolutional neural networks, and compare their performance using metrics such as accuracy, precision, recall, and F1 score. The results demonstrate the effectiveness of machine learning in COVID-19 detection and emphasize the importance of developing accurate models for timely and effective diagnosis.

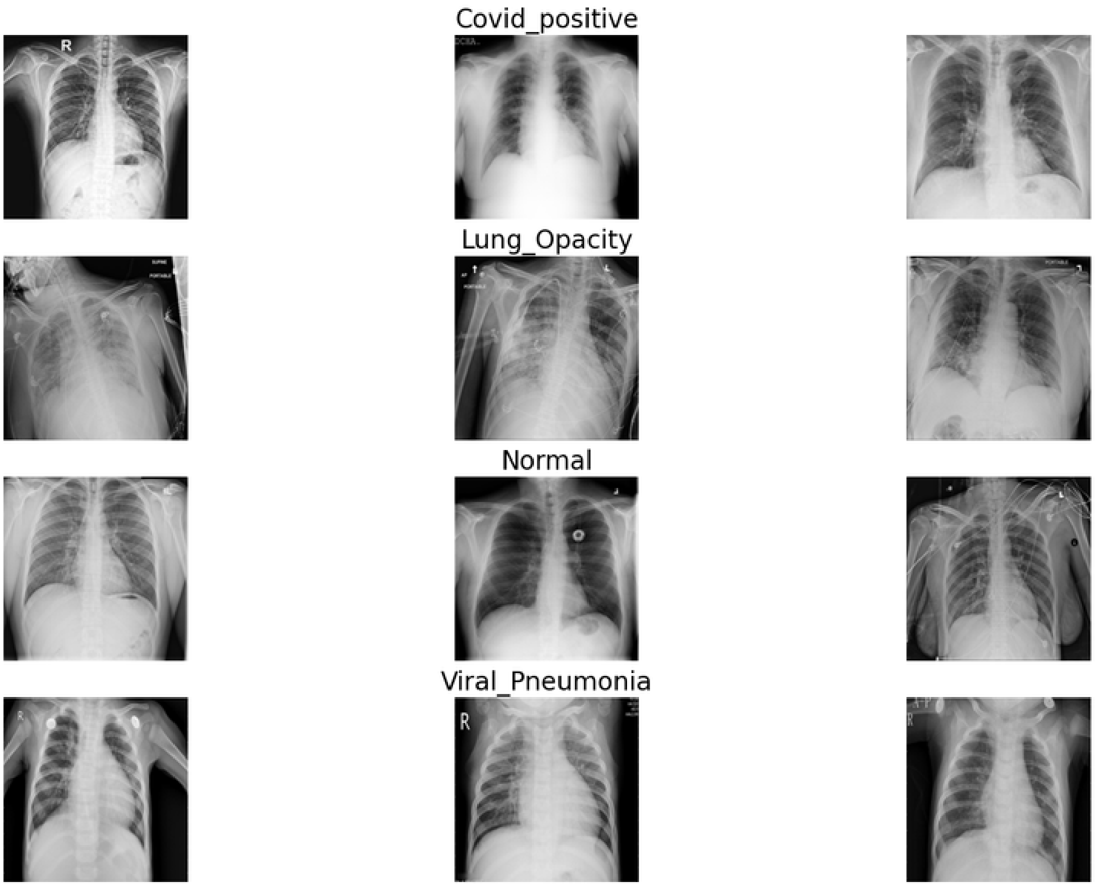


Fig. 1: Sample Images from all classes

Figure 1 displays a selection of sample images from the various classes present in the dataset. Furthermore, Figure 2.1 illustrates the distribution of available data within each class. The dataset was then encoded into two classes, 1 and 0, with class 0 consisting of the Normal, Lung Opacity, and Viral Pneumonia classes, while class 1 contained the Covid-positive class. The distribution of 0 and 1 class is shown in Fig. 2.2.

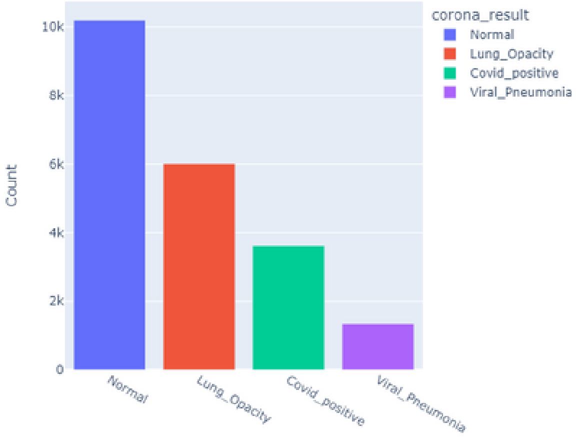


Fig. 2.1: Distribution of Classes before Encoding

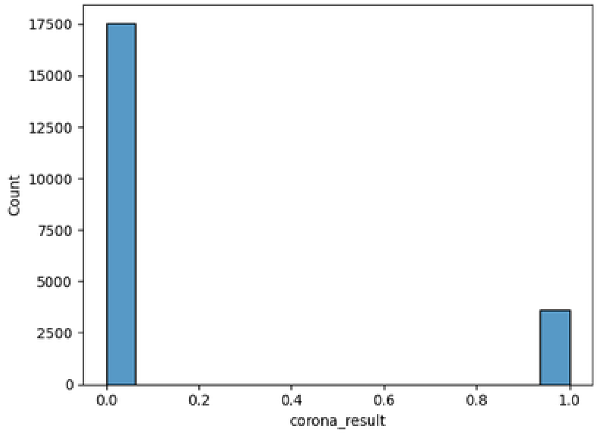


Fig. 2.2: Distribution of both Classes after Encoding

II. OUR APPROACH

We have two approaches for preparing the data for training in our problem, which involves classifying X-Ray images into COVID and non-COVID classes. The first approach involves flattening the pixel values of each image into a standard 1-D dataset, which simplifies the data and allows for simple refining techniques to be applied. However, this approach results in a loss of valuable information regarding the relative positioning of the pixels.

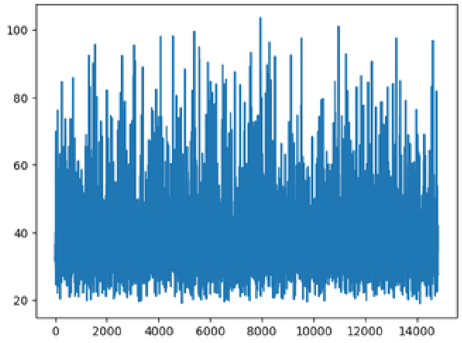
The second approach involves maintaining a 2-D dataset that includes the pixel values of each pixel in each image. This approach preserves the spatial information of the images but results in a more raw dataset that may limit the performance of the model.

We will evaluate the performance of our models based on two metrics: accuracy and recall. Accuracy measures how well the model is able to classify X-Ray images into the COVID and non-COVID classes, while recall measures the ability of the model to minimize false negative predictions, which is important in a healthcare-related problem.

After evaluating the two best models obtained from the above approaches, we will select the model with the best performance based on the defined metrics as our final solution.

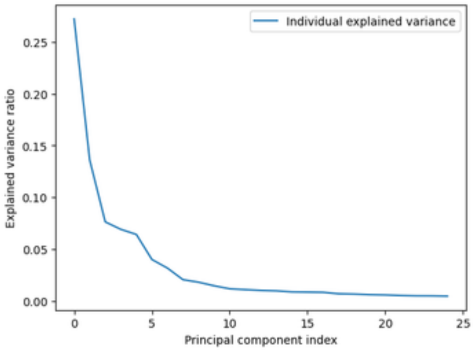
Data Manipulation

Anomaly Detection - Anomaly detection can be useful as with as much uncertainty as there is in the given problem where even humans are very much prone to misclassification, removing anomalous samples can result in enhancing the generalizability of our models as well as enhancing the performance as the quality of training dataset is increased as well. We use distance based anomaly detection which is implemented using KNN algorithm where we apply a threshold beyond which we consider the data as anomalous.



(Average distance v/s Datapoint)

PCA-PCA is a technique used for reducing the dimensionality of the dataset. For the given dataset, we pixelated the images into 50X50 which meant 2500 attributes which would result in tremendous amounts of time invested in training the models. To get around this, we apply PCA so that we could get a dataset which encapsulates as much variance as we can whilst keeping the dimensions small.



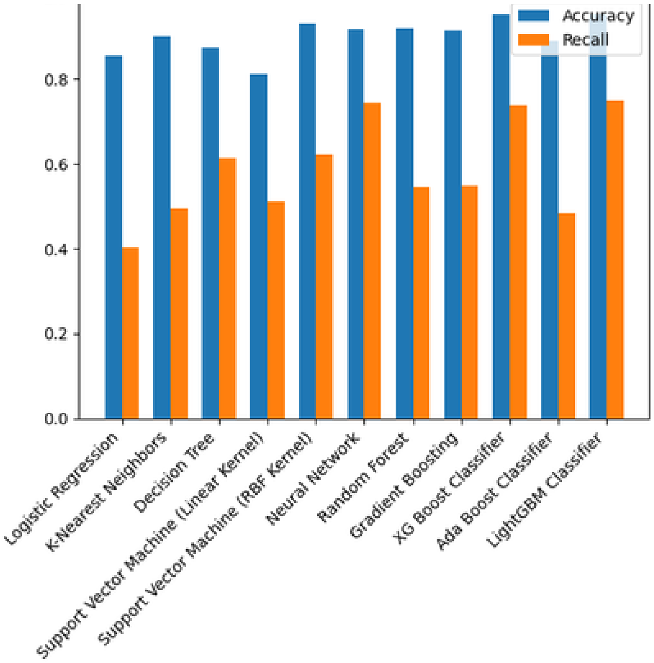
Model Selection

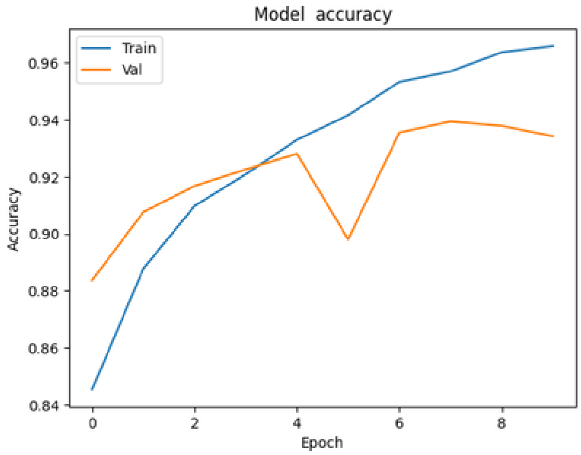
In order to select the best machine learning model for our COVID-19 detection task, we experimented with a range of different algorithms, including logistic regression, k-nearest neighbors, decision tree, support vector machines (both linear and with radial basis function kernels), neural networks, random forests, gradient boosting, AdaBoost, and LightGBM. We trained and tested each of these models on our prepared dataset, and evaluated their performance based on two key metrics: accuracy and recall.

After conducting these initial experiments, we found that the XGBoost algorithm achieved the best overall performance, with a strong balance between accuracy and recall. However, we also recognized that there was still room for improvement in terms of the model's ability to minimize false negative predictions (i.e. its recall score). So we tried tuning the hyperparameters of the model for better results but we did not see any significant change in the results even after tuning the hyperparameters.

To address this issue, we decided to explore the use of convolutional neural networks (CNNs), which have shown promise in image classification tasks. We experimented with several different architectures and hyperparameter settings for our CNN model, training and evaluating each variant on our dataset to find the best-performing option.

Ultimately, we found that our CNN model outperformed all of the other algorithms we tested, achieving a significantly higher recall score than any of the previous models. As a result, we selected the CNN model as our final approach for detecting COVID-19 cases from chest X-ray images.





III. Results

XGBOOST- Xgboost is a highly versatile algorithm applicable to a variety of problems, including image classification. Xgboost can effectively deal with absent data and anomalies, making it a trustworthy algorithm for image classification tasks.

LIGHTGBM-LightGBM is efficient in the handling of high-dimensional data. This means that LightGBM can be used to classify images without requiring extensive computational resources. In addition, LightGBM can be used to model intricate relationships between features and the target variable, making it useful for tasks where traditional feature engineering methods may not be adequate.

CNN- CNNs consist of multiple layers, including convolutional layers that identify local patterns in the input image and pooling layers that downsample the output of the convolutional layers. By stacking these layers together, CNNs can learn increasingly complex patterns and relationships within the input image.

CNNs are more suitable for image classification problems due to their capacity to automatically learn hierarchical representations of features and their competence in dealing with high-dimensional image data. This makes **CNNs** more appropriate for use in picture classification tasks. Tree-based models, on the other hand, such as **LightGBM** and **XGBoost**, are more suited for structured data, although they may need manual feature engineering.

Classifier	Accuracy	Recall
Logistic Regression	84.71%	35.08%
KNN	90.08%	52.21%
Decision Tree	87.03%	64.09%
SVM (Linear Kernel)	82.81%	47.24%
SVM (RBF Kernel)	91.43%	56.08%
Neural Network	91.57%	67.13%
Random Forest	91.43%	53.31%
Gradient Boosting	90.45%	51.38%
XG Boost Classifier	95.04%	76.80%
XG Boost tuned	95.31%	76%
Ada Boost	88.37%	48.34%
LightGBM	94.39%	72.10%
CNN	93%	89%

IV. Contributions

Shreejan Kumar: In this project, I made significant contributions to the development of an accurate model for COVID detection from chest X-ray images. Specifically, I employed feature selection techniques to identify the most relevant features from the dataset, but it led to the loss of data when applied so we did not use it in the final model. Additionally, I implemented hyperparameter tuning techniques to optimize the performance of the model by finding the best combination of hyperparameters that maximized the model's performance metrics. Through this process, I improved the model's accuracy and reduced its false negative rate, resulting in a more reliable model for COVID detection. I also helped in applying the CNN model and tuned its hyperparameters for a better overall result. Overall, my contributions to this project demonstrate the importance of applying advanced techniques such as feature selection and hyperparameter tuning to improve the accuracy of machine learning models for COVID detection, which can help to prevent the spread of the virus and save lives.

Vasubhya Diwan: I made a significant contribution by applying anomaly detection and PCA techniques. These techniques were used to identify outliers in the dataset and reduce the dimensionality of the feature space, respectively. Anomaly detection was employed to identify abnormal images that could potentially lead to misclassification of the COVID-19 cases. By removing such outliers, we were able to improve the overall accuracy of the classification model.PCA was used to reduce the dimensionality of the feature space while retaining the most significant information from the input data. This allowed us to reduce the computational complexity of the model and improve its performance. Moreover, it enabled us to visualize the high-dimensional feature space in a lower-dimensional space, making it easier to interpret the results, and also contributed to applying CNN.

Yuvraj Soni: In this project, I have made significant contributions to the detection of COVID-19 using chest X-ray images. Firstly, I performed an extensive exploratory data analysis (EDA) of the dataset to gain insights into the distribution of the data and identify any potential issues or biases. The EDA allowed me to visualize the distribution of the data, identify any class imbalances, and understand the relationships between different variables in the dataset. This information was crucial in informing the pre-processing steps required for the model development. Secondly, I applied a state-of-the-art deep learning model, Convolutional Neural Network (CNN), to classify the chest X-ray images as COVID-19 positive or negative. The CNN model was trained on the pre-processed dataset and achieved high accuracy in classifying the images. The use of a CNN model allowed for the automatic extraction of relevant features from the images and the learning of hierarchical representations of the features, which contributed to the model's high performance.