International Journal of Electronics and Communication Engineering.
ISSN 0974-2166 Volume 16, Number 1 (2023), pp. 1-15
© International Research Publication House
https://dx.doi.org/10.37624/IJECE/16.1.2023.1-15

Prediction of COVID-19 Severity by Applying Machine Learning Techniques

[1]V.Shankar Telugu Hemalatha, [2]Katkoori Preethi, [2]S Sushma, [2]K Manusha

[1]Assistant Professor, [2]Student
Electronics & Communications Engineering,
G.Narayanamma Institute Of Technology &Science (For Women), Hyderabad, India
[1]email: vshankar33@gmail.com

ABSTRACT

A intimidating spread of COVID- 19(which is also known as severe acute respiratory pattern coronavirus 2 or SARS- COV- 2) led scientists to conduct tremendous sweats to reduce the epidemic goods. Artificial intelligence (AI) methods that are quick and precise are needed to support croakers in their assessments of a case's inflexibility and mortality hazard. By providing early medicine administration, pre-vaccination of rigid patients will reduce the cost to the hospital and stop instances from dying continuously. This design uses machine learning and deep learning techniques to build a vaticination model that predicts several inflexibility problems for the COVID-19 case grounded on X-ray pictures. Non-Handcrafted styles and composite handcrafted ways are applied to extract features. Principal Component Analysis(PCA) is being incorporated to choose the most vital features, and also Machine and Deep learning ways are applied. To extract characteristics, non-handmade styles and composite handcrafted methods are used. To choose the most crucial features, Principal Component Analysis (PCA) is used with machine and deep learning techniques. The cases are categorised and classified as pneumonia, severe, or normal. The PCA features enabled the Bagging, Ada Boost, KNN(K- nearest neighbors), and XGBoost classifiers to perform stylishly with 97% of accuracy, 98% of precision, and recall of 95%. This study proposes a new prophetic frame for the inflexibility and mortality threat of COVID-19 cases to help hospitals, croakers, and medical installations in their decisionmaking about which cases need to get attention first before others, and at the same time, to keep hospitals ’ coffers for high- threat precedence cases.

Keywords: Preprocessing, Feature extraction, Feature Selection, Confusion Metrix, Performance Analysis.

I. INTRODUCTION

The World Health Organization (WHO) Country Office in China initially received word of a pneumonia of unknown cause discovered in Wuhan, China, on December 31, 2019. Since then the number of cases of coronavirus are increasing along with the high death toll. In just a few days, the virus has started to spread widely from one city to the whole world. Coronaviruses are a large virus family that may cause everything from a typical cold to more serious illnesses like Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). MERS-CoV and SARS-CoV are the names of these two diseases which are spread by coronaviruses. In 2002, SARS was first seen in China, and in 2012, MERS was first seen in Saudi Arabia. In 2019, is recent virus seen in Wuhan, China is called SARS-COV2.

A critical activity that reduces the mortality rate, uses hospital resources, and benefits doctors is predicting the early-stage severity risk of any sort of illness in their decision-making. Figuring out the COVID-19 patients' severity risk is therefore a crucial task with numerous benefits, including ensuring that each patient receives the appropriate medical care in accordance with their severity, making the most effective utilisation of the hospital's amenities by providing the high-risk patient top attention and aiding doctors in making choices that will enhance the patient's care.

The three main approaches for finding COVID-19 are X-ray imaging, reverse transcription-polymerase chain reaction and computed tomography (CT). Although RT-PCR is the effective type, it is the most expensive, is not offered in all hospitals, and requires a long time to obtain results. Therefore, for early detection and treatment of this disease, many doctors rely on chest radiographic imaging, which might include CT and X-rays. Although CT is a very sensitive instrument, its findings can only be seen after a long period depending on when symptoms first appear. As a result, CT is challenging to utilize in routinely monitoring patients.

Despite the fact that chest X-ray (CXR) radiography is one of the most widely used and accessible techniques for the quick assessment of lung diseases, it is less sensitive than CT and RT-PCR. Since X-ray findings may be seen quickly and the technique is inexpensive, it can be done regularly to check on the patient's condition.

Feature extraction is a difficult task, particularly when the size of the data set is tiny. Values of numerical information or intensity of pixels that provide useful information regarding the local and/or global variations of an image's pixels are called an image's attributes. Finding a feature vector representation of the input photos and successfully extracting the most important factors pertinent to the desired application's goal constitute the procedure. Today, for feature extraction non-handcrafted (deep learning) and handcrafted feature extraction are the two main methods utilised. More known or meaningful features are manually extracted through handcrafted techniques.

The quality and effectiveness of the prediction model are heavily influenced by the selection of the most efficient image features, which is a critical stage. Filter-based, Forward Selection, Recursive Feature Elimination (RFE), Backward Elimination, Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA) are all approaches for feature selection.

As this COVID-19 is spread from person to person, Artificial intelligence-based techniques can play a pivotal role in preventing the spread of this virus. A common method to categorise conventional machine learning is the process by which a prediction-making algorithm learns to increase its accuracy. There are four basic approaches in which supervised learning is used for the proposed system. Different algorithms come under supervised learning, like Random Forest, K-Nearest Neighbors, and so on. As a result, this study created a prediction model based on a collection of chest X-ray images of public to forecast various types of patient severity risks. It could predict if the patient needs to be admitted to the intensive care unit (ICU) or even herald his death.

In this proposed work, Adaboost, XgBoost, K- Nearest Neighbors, and Bagging algorithms are used. Performance Analysis is to be done by taking some performance metrics such as Precision, Accuracy, etc. Comparison between these classifiers gives us the best model for further usage.

The proposed system has the following goals:

1. Preparing the Dataset required and classifying the data into training data and testing data.
2. Pre-process the data and extract the features from the images.
3. Reducing the features for better results by using Principal Component Analysis.
4. Classifying data using different machine learning algorithms such as Adaboost, XgBoost, Bagging, and K-Nearest Neighbor.
5. Analyzing the results based on performance metrics including Precision, Accuracy, and F-score.
6. Comparing the different algorithms by creating a performance graph and selecting one best model for further analysis.
7. Creating a Graphical User Interface to upload an Xray image and knowing the condition of that particular image using the previously selected model.

II.RELATED WORKS

M. Awad and R. Khanna.,[1] study suggested a system that categorizes data using support vector machines.

Z. Q. Lin, L. Wang, and A. Wong.,[2] developed a model using Deep Convolutional neural networks . It is hoped that COVIDx, an open-access benchmark dataset, can be used and improved upon by researchers as well as amateur data scientists to speed the development of highly accurate yet practical deep learning solutions. COVIDx was created using the COVID-Net open access benchmark dataset.Prioritizing the cases and accelerating those patients' care are their objectives.

T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya.,[3] have developed a model to offer reliable binary classification diagnostics (COVID vs. NoFindings) and multi-class classification i.e., COVID vs. NoFindings vs. Pneumonia. The DarkNet model was used in their study as a classifier for a real-time object identification system called you only look once (YOLO). 17 convolutional layers are been implemented and different filtering is done on each layer.

A. I. Khan, J. L. Shah, and M. M. Bhat [4] used chest X-ray images and proposed CoroNet, which is a Deep Convolutional Neural Network model to automatically detect COVID-19 infection. The proposed model is pre-trained on the ImageNet dataset and trained end-to-end. The collection is compiled from two publicly available datasets that contain X-ray images of COVID-19 and other types of chest pneumonia. Based on the Xception architecture is this model.

Using chest X-ray pictures, N. Habib, M. M. Hasan, M. M. Reza, and M. M. Rahman [5] suggested an ensemble methodbased pneumonia diagnosis. To extract features from provided X-ray pictures, CheXNet and VGG-19, two deep convolutional neural networks (CNNs), are trained and employed.

In their retrospective analysis, A. Bernheim, X. Mei, and colleagues[6] evaluated the chest CTs of 121 symptomatic COVID-19-infected patients from four centers in China from January 18, 2020, to February 2, 2020.They attempted to correlate the most frequent CT abnormalities with the interval between the beginning of symptoms and the initial CT scan (early, 0–2 days, 36 patients; middle, 3-5 days, 33 patients; late, 6–12 days, 25 patients). On imaging, bilateral and peripheral ground-glass, and consolidative lung opacities were the telltale signs of COVID-19 infection. It's worth noting that 20/36 (or 56% of the initial patients) had a normal CT scan.

III.THE PROPOSED MODEL

The process of classification involves several key steps. Firstly, the X-ray image is taken as input from the dataset and pre-processing should be done which resizes and Images should be normalised. Then, several feature extraction methods are used to extract the features. The selection of important features from the image is to be done by applying some feature selection techniques. Finally, Several models are built by using various machine learning classifiers. Then, a model which is more accurate is chosen, and develop a GUI where an image is uploaded to know whether they are normal or suffering from lung opacity or covid or pneumonia.

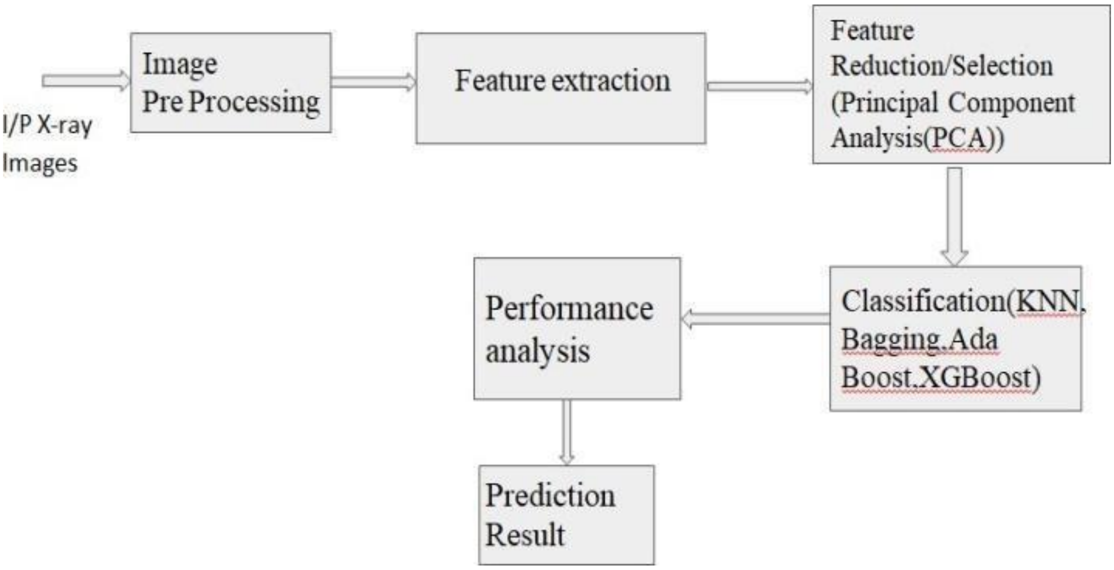

I/P X-ray Images
Image Pre Processing
Feature extraction
Feature Reduction/Selection (Principal Component Analysis(PCA))
Classification(KNN, Bagging,Ada Boost,XGBoost)
Performance analysis
Prediction Result

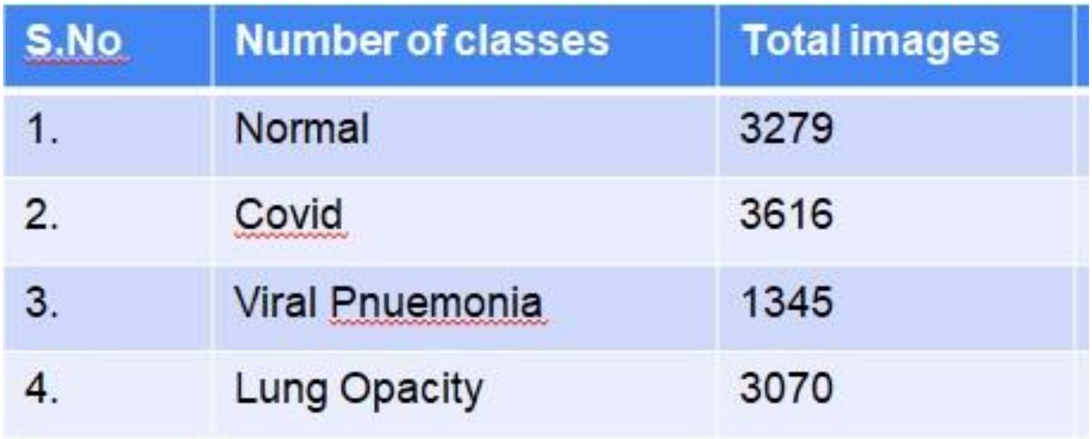
Fig. 1. Block Diagram of Proposed system.

3.1 Dataset

Researchers from Qatar University, Doha, Qatar, and the University of Dhaka, Bangladesh, along with their associates from Pakistan and Malaysia, have developed a database of chest X-ray pictures for COVID-19-positive cases as well as normal and viral pneumonitis images.They worked with medical professionals.

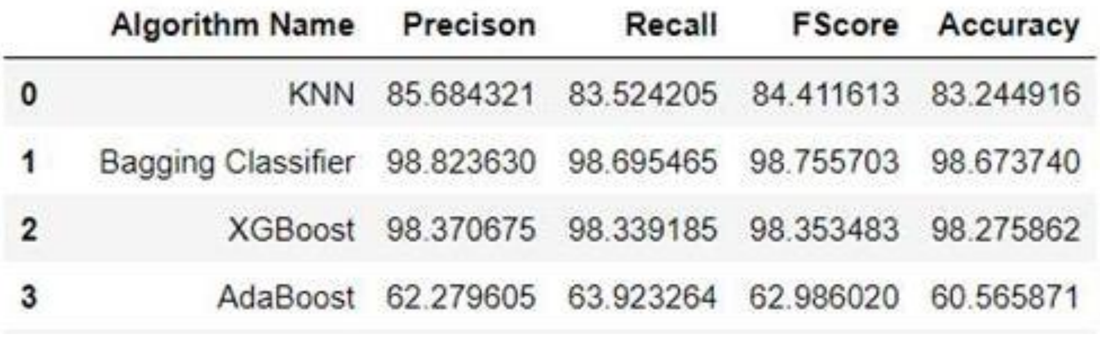
They published 219 COVID-19, 1341 regular, and 1345 viral pneumonia chest X-ray (CXR) images in the initial publication. The COVID-19 class was expanded to 1200 CXR pictures in the initial upgrade. In the 2nd update, they have increased the database to 3616 COVID-19 positive cases along with 10,192 Normal, 6012 Cold, and 1345 Viral Pneumonia images.

TABLE 1: Dataset details



S.No	Number of classes	Total images
1.	Normal	3279
2.	Covid	3616
3.	Viral Pnuemonia	1345
4.	Lung Opacity	3070

TABLE 2: Comparison Table



	Algorithm Name	Precison	Recall	FScore	Accuracy
0	KNN	85.684321	83.524205	84.411613	83.244916
1	Bagging Classifier	98.823630	98.695465	98.755703	98.673740
2	XGBoost	98.370675	98.339185	98.353483	98.275862
3	AdaBoost	62.279605	63.923264	62.986020	60.565871

The dataset is split into 80/20 train/test sets and categorised by patient-id to guarantee that each patient's Xray images are distributed to only one (train/test) set. The model has been tested on 2262 photos after being trained on 9048 images.

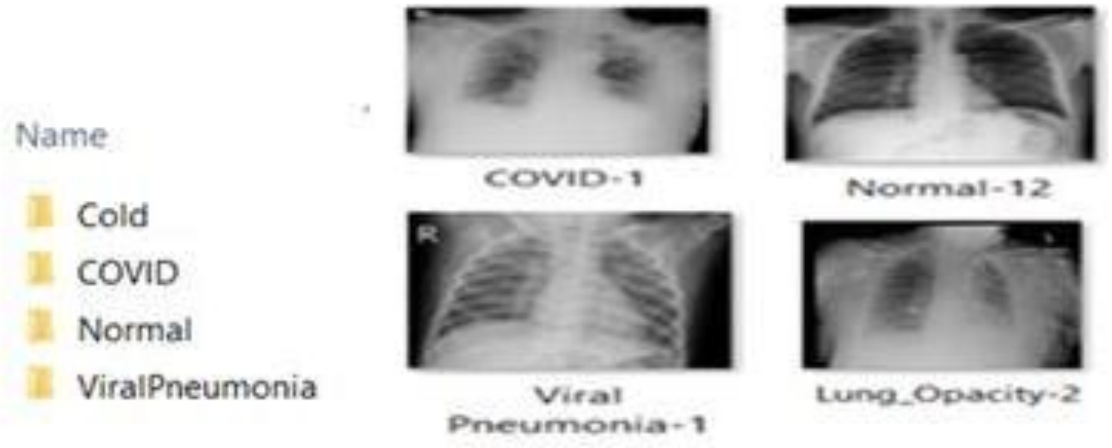


Figure 2. Input Images

3.2 Pre-processing

Preparing the data for the prediction model is the aim of the data pre-processing phase. Data frequently lack organisation and originate in a variety of sizes and resolutions from various sources. simplify and normalize the data, minimize complexity, and improve model accuracy, this phase is essential. According to the dataset, many forms of modifications, such as resizing, rotating, shifting, and normalizing, could be carried out. The four preparation procedures are applied one after another to the dataset. Convert the image to grayscale first to improve the presentation of diagnostically important information and to improve the informational value of radiography. Second, resize the image to 512 X 512 dimensions for handcrafted features to ensure that the image preserves the most useful information regarding the patient's level of severity If not, due to the image being less than 512 X 512, information about the entire image may lose its original allocated class. In improving the medical image, adaptive histogram equalization (AHE) is then applied after normalizing the image Using the min-max approach, reconfigure the image pixels between zero to one The preprocessed and normalized image is shown in fig .3.

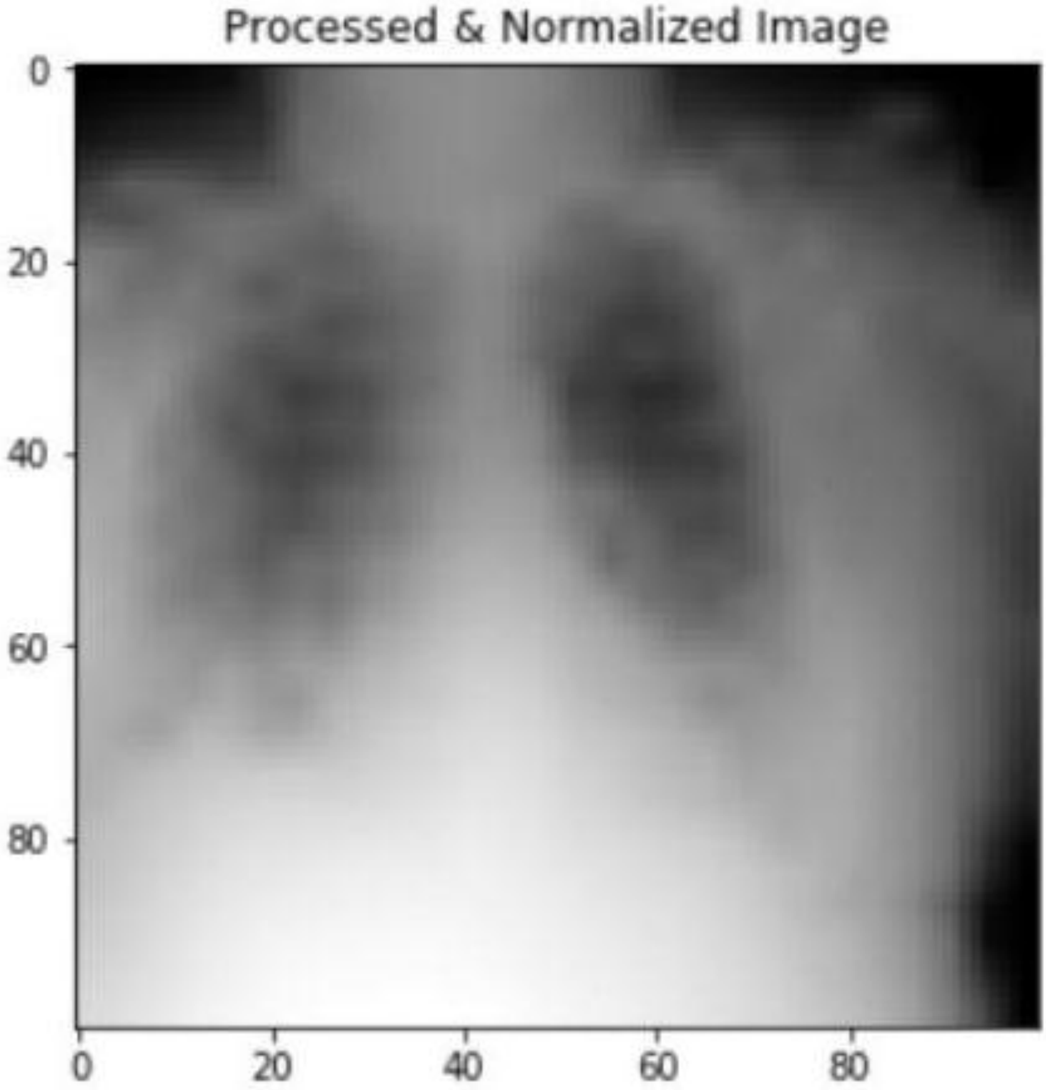


Figure 3. Pre-processed Image

3.3 Feature Extraction

The spatial and frequency domains are used to obtain the characteristics. The spatial domain is used to deal with the image's starting matrix. To extract features in this area, the source image's texture features, the Gray-Level Co-Occurrence Matrix, and the Grey Level Difference Matrix were all used.

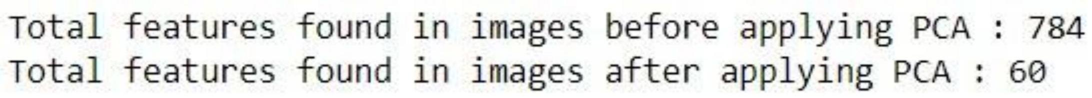
Characteristics of the source image's texture:

14 features—Area, Mean, Standard Deviation, Skewedness, Kurtosis, Energy, Entropy, Maximum, Minimum, Median, Range operation, Root Mean Square (RMS), and consistency(Uniformity)—are calculated from the image's initial matrix.Transform Discrete Wavelet.

Gray-Level Co-Occurrence Matrix : By multiplying each of the 14 texture features by their cartesian products in the four directions, 56 features are created using this model's fourdirectional co-occurrence matrices (e.g., 0, 45, 90, and 135) 3)Grey Level Difference Matrix (GLDM): The GLDM created difference matrices in four directions using the same techniques as GLCM before extracting an additional 56 features.
In a spatial matrix domain, the frequency domain focuses on how quickly pixel values change. The Fast Fourier Transform (FFT) and the Wavelet transform were employed to extract features. Finally, the picture features are created by combining all the texture characteristics that were extracted using the aforementioned approaches.

3.4 Feature Selection

An orthogonal transformation called principal component analysis (PCA) separates a bigger collection of potentially related data into a more manageable number of interrelated features called principle components. PCA's goal is to reduce the dimensionality of the dataset while retaining the majority of the original data variability. To project the source dataset into the PCA's smaller domain, the correlation/covariance matrix's eigenvectors are used.The resulting projected data are the highest proportion of data variance is present in the first component of linear mixtures of the initial information. and accounting for the majority of the data volatility. The remaining components each have the lowest possible level of data fluctuation.



```
Total features found in images before applying PCA : 784
Total features found in images after applying PCA : 60
```

Figure 4. Features selected

3.5 Evaluation metrics

While accuracy is a popular performance indicator for classification issues, machine learning models are also frequently evaluated using a number of other metrics. Seven widely used metrics are listed below:

Accuracy: The ratio of reliable forecasts among all the assumptions in the model's projections.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: The ratio of correct predictions that were favourable among all positive predictions made by the model.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Recall: The percentage of estimations that were correct out of all real instances that were positive.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

F1 score: A calculation that combines precision and recall and takes both measurements into account.

$$\mathrm{F1} = 2 \times \frac{precision \times recall}{precision + recall}$$

Confusion matrix: A model's predictions on a data set are used to construct confusion matrices. By examining a confusion matrix, you can better grasp the advantages and disadvantages of your model as well as compare it to two others to determine which is more appropriate for your application. Traditionally, a model's predictions on a held-out test set are used to create a confusion matrix. It displays the total number of true positives, true negatives, false positives, and false negatives.

IV. CLASSIFICATION

A comparison of various machine learning classifiers was done in order to create a strong predictive model. K Nearest Neighbors (KNN), Bagging, XgBoost and Adaboost were the four classifiers employed. Thomas Cover created the supervised algorithm K Nearest Neighbors (KNN) for classification and regression issues.

4.1 Adaboost Algorithm

AdaBoost's ensemble technique trains and launches trees one following the other. AdaBoost employs boosting, which involves connecting a sequence of weak classifiers in order to ensure each weak classifier attempts to correct the categorization of data that was mistakenly classified by the weakest classifier preceding it. Enhancing accomplishes this by connecting weak classifiers to create a powerful classifier. Because decision trees used in boosting approaches have a propensity to be shallow models that do not overfit but might be biassed, they are commonly referred to as "stump." A specialised tree receives instruction to focus solely on the flaws of one additional a tree.

4.2 XgBoost Algorithm

The supervised classifier Extreme Gradient Boosting (XGBoost) is ensemble-based. Gradient tree boosting is used in scalable machine learning. Each new tree is constructed and added to the ensemble framework in order to correct the defects brought by the previous ones, up until no more improvements can be achieved. Combining the trees yields a prediction of the eventual result.

One tree is created by XGBoost at a time in a forward motion. Gradient boosting is a technique that minimises loss when adding extra trees by using the gradient descent procedure. This classifier's key advantage is that it utilizes a more regularised the formalization of the model to reduce excessive fitting and improve performance.

4.3 K-Nearest Neighbor Algorithm

According to the KNN algorithm, similar objects are close by. In other words, things that are related to one another are found nearby. For the algorithm known as KNN to work, this assumption must be sufficient. KNN encapsulates the concept of similarity (also known as distance, proximity, or closeness) using an equation we might have learnt as children—calculating the distance that exists across nodes on a graph.

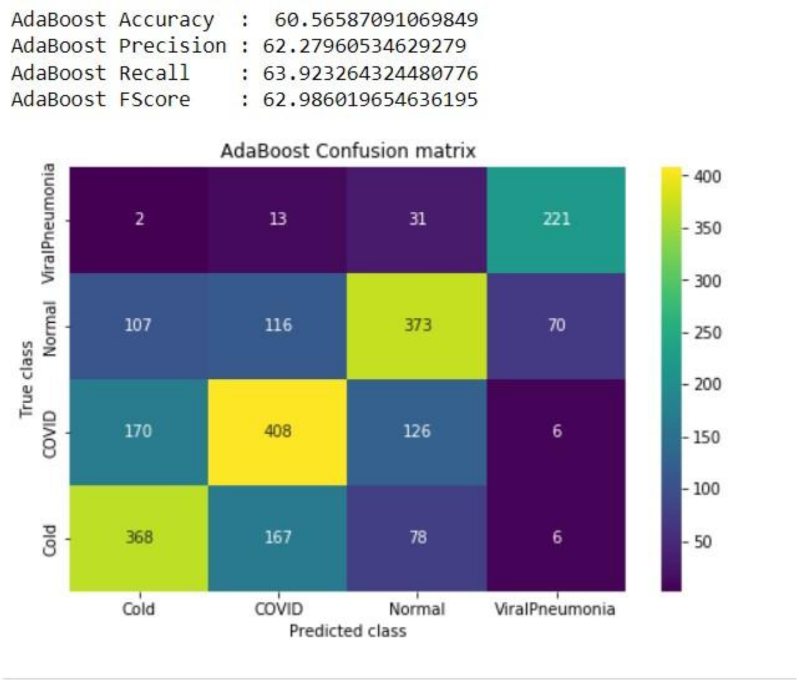
Other approaches to distance estimation exist, and one approach may be preferable depending on the situation. The direct distance, commonly referred to as the Euclidean distance, is the most popular and well-known alternative, nonetheless.

4.4 Bagging Algorithm

Bagging, also known as Bootstrap aggregating, is a method of ensemble learning that contributes in improving both the effectiveness and accuracy of the machine learning algorithms. It reduces a prediction model's variance and is used to address bias-variance trade-offs. Bagging, particularly decision tree approaches, is used to reduce overfitting of data in both classification and regression models.

Bootstrapping is a method for predicting an estimated parameter by selecting data samples at random from a population. Model predictions are aggregated to create a final prediction that takes into account all potential outcomes. The aggregate might be done based on the overall number of results or the likelihood that forecasts made using bootstrapping.

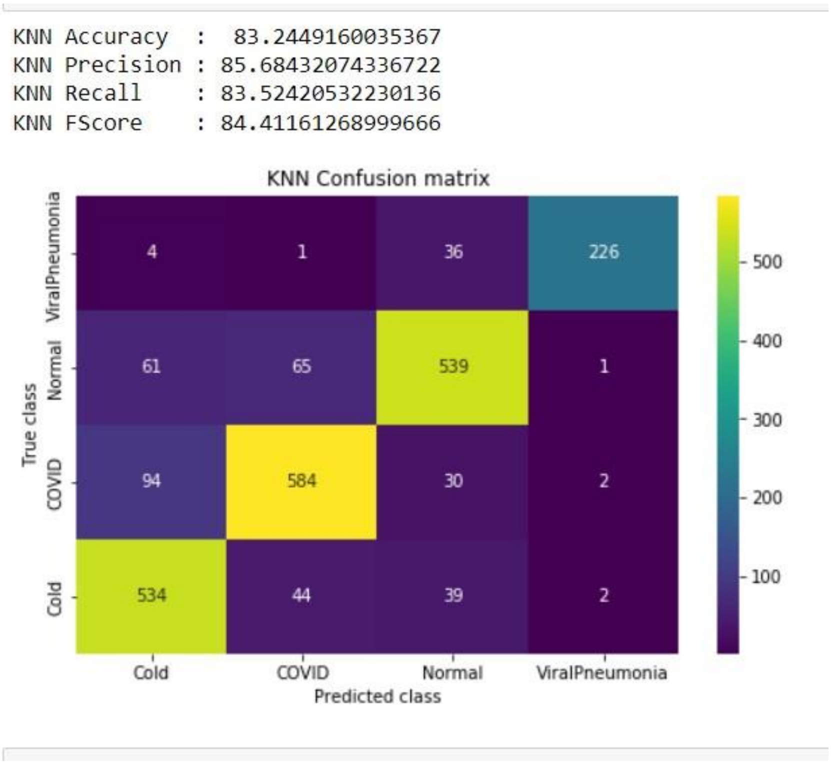
V. RESULTS



```
AdaBoost Accuracy  :  60.56587091069849
AdaBoost Precision : 62.27960534629279
AdaBoost Recall    : 63.923264324480776
AdaBoost FScore    : 62.986019654636195
```

	Cold	COVID	Normal	ViralPneumonia
ViralPneumonia	2	13	31	221
Normal	107	116	373	70
COVID	170	408	126	6
Cold	368	167	78	6

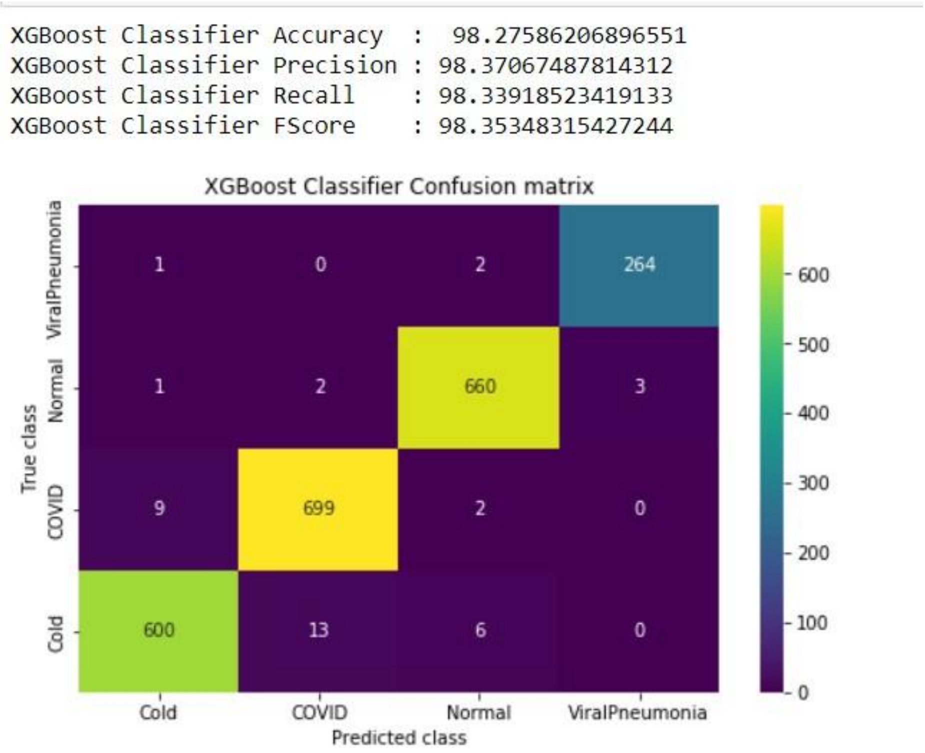
Figure 5. Confusion matrix for Adaboost algorithm



```
KNN Accuracy  :  83.2449160035367
KNN Precision : 85.68432074336722
KNN Recall    : 83.52420532230136
KNN FScore    : 84.41161268999666
```

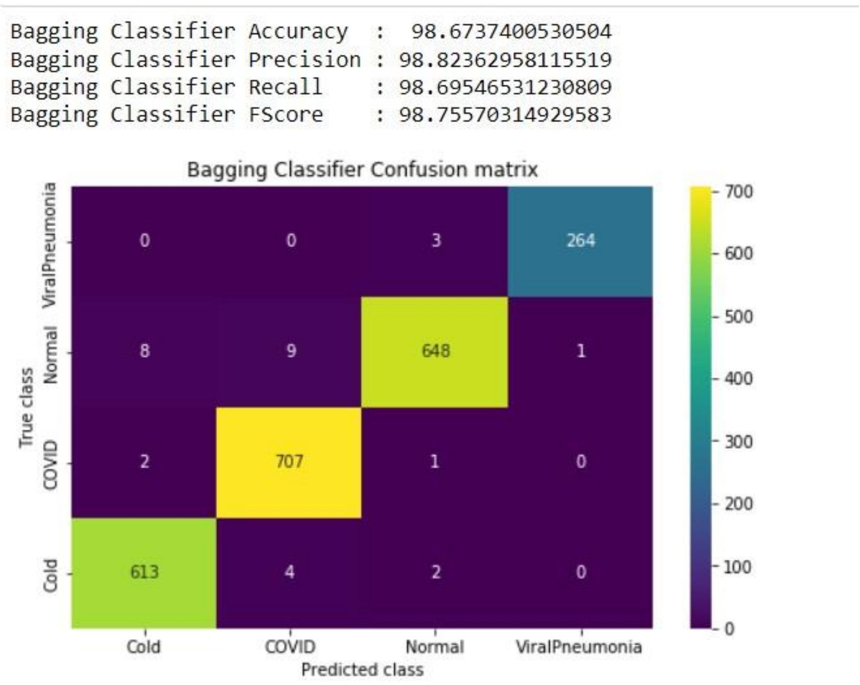
	Cold	COVID	Normal	ViralPneumonia
ViralPneumonia	4	1	36	226
Normal	61	65	539	1
COVID	94	584	30	2
Cold	534	44	39	2

Figure 6. Confusion matrix for KNN algorithm



```
XGBoost Classifier Accuracy  :  98.27586206896551
XGBoost Classifier Precision : 98.37067487814312
XGBoost Classifier Recall    : 98.33918523419133
XGBoost Classifier FScore    : 98.35348315427244
```

Figure 7. Confusion Matrix for XgBoost Algorithm



```
Bagging Classifier Accuracy  :  98.6737400530504
Bagging Classifier Precision : 98.82362958115519
Bagging Classifier Recall    : 98.69546531230809
Bagging Classifier FScore    : 98.75570314929583
```

Figure 8. Confusion Matrix for Bagging Algorithm

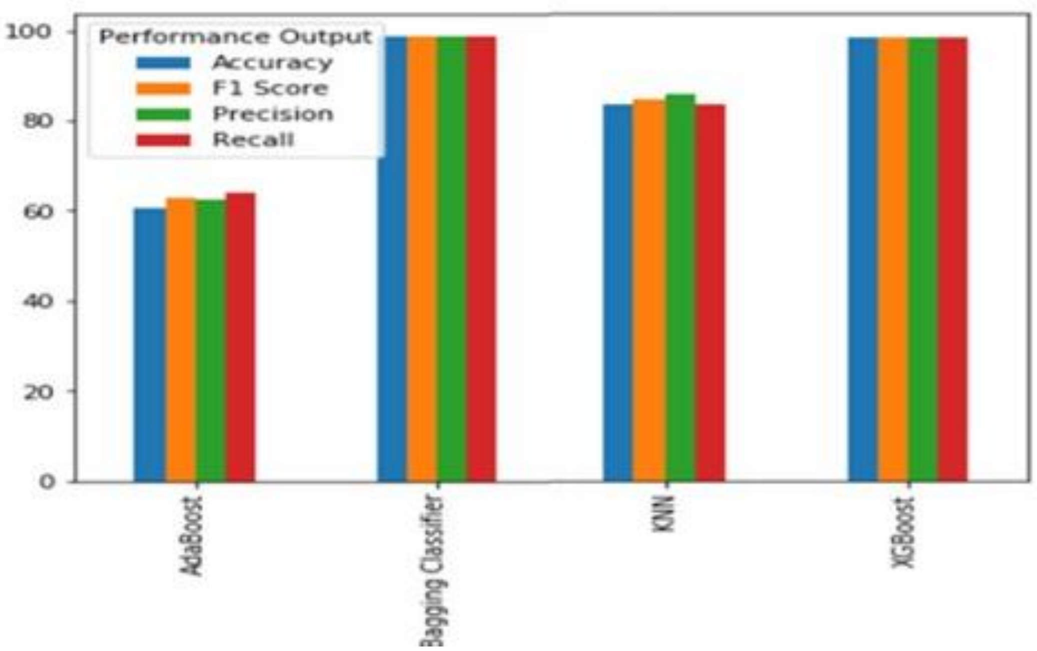


Figure 9. Performance Graph

Prediction Of COVID-19 Severity by Applying Machine and Deep Learning Techniques
Prediction Of COVID-19 Severity by Applying Machine and Deep Learning Techniques
Predict Covid Severity from Test Images

Figure 10. Graphical user Interface for uploading an image

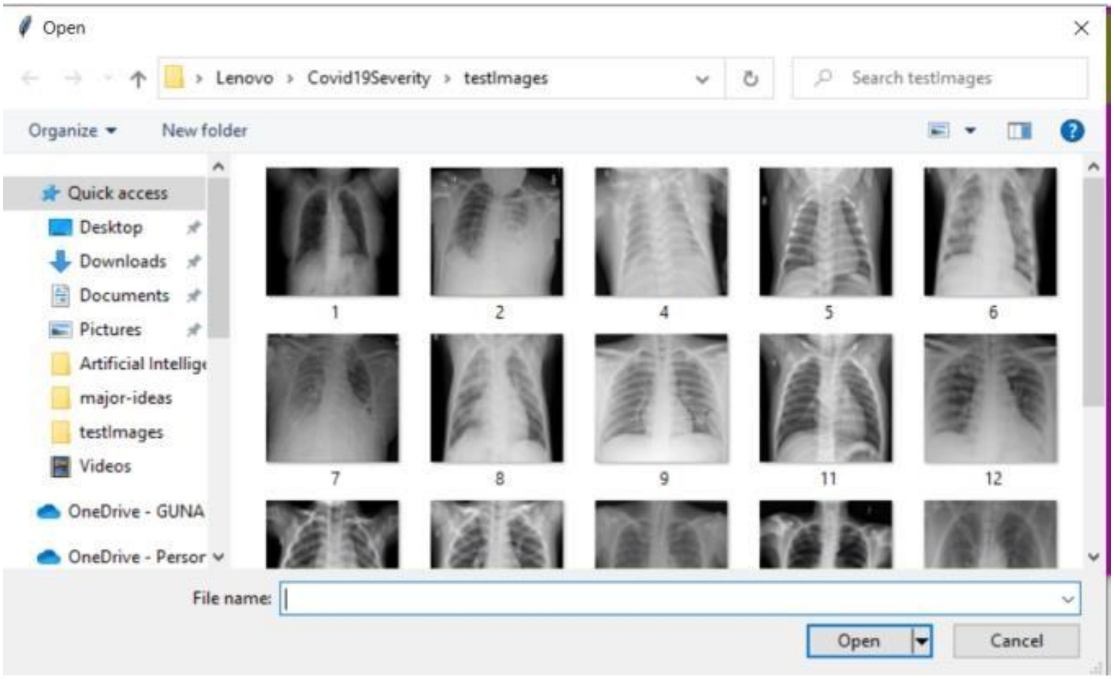
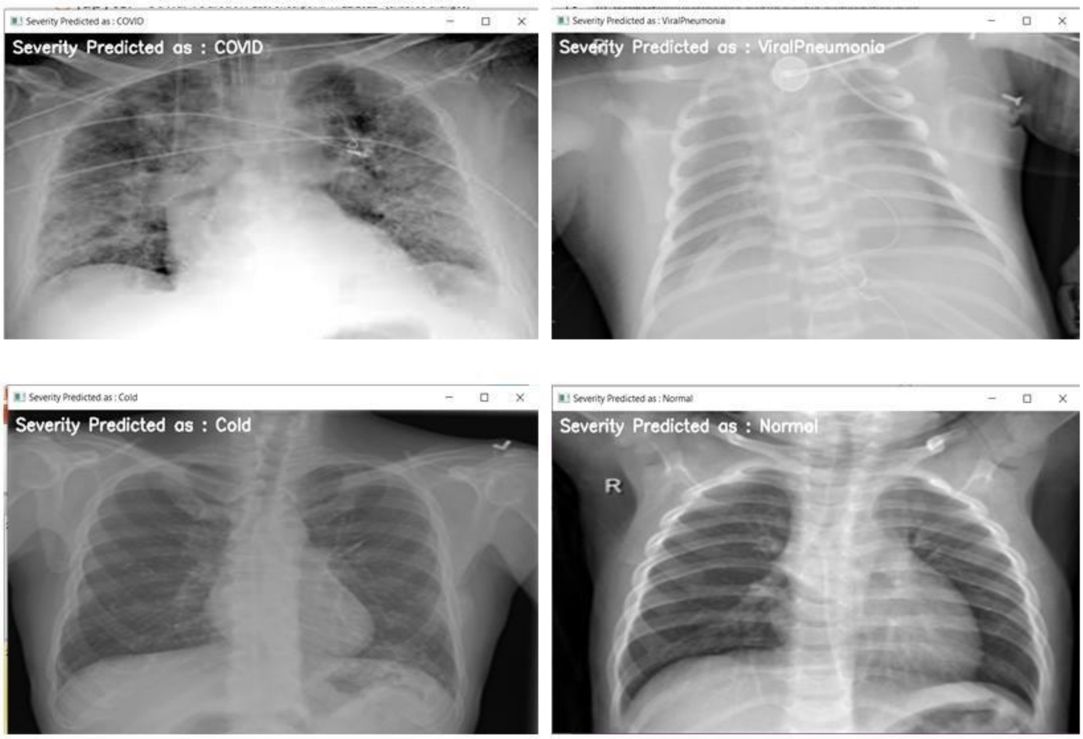


Figure 11: Selecting an image


Severity Predicted as : COVID

Severity Predicted as : ViralPneumonia

Severity Predicted as : Cold

Severity Predicted as : Normal

Figure 12. Results shown from GUI

VI. CONCLUSION

In order to help doctors, hospitals, and other healthcare providers choose which patients need immediate attention before others while also protecting hospital resources for highrisk priority patients, the objective of this study is to develop a cutting-edge prediction paradigm for COVID-19 severity and mortality risk. The proposed model is based on a set of confirmed COVID-19 illness patients' X-ray pictures that are openly accessible.

Pneumonia, colds, covids, and normal comprise the four severity classifications for the dataset. Hybrid handmade techniques were applied to X-ray images in order to extract the features. PCA was used to choose the features, and numerous machine learning prediction models, such as KNN, XGboosting, Bagging, and Adaboost, were developed to compare and guarantee the findings.

Numerous tests were carried out, and the findings showed for handcrafted features, features selected generated the best outcomes with all classifiers. Approximately 7.65% of the features that were initially obtained were used, or 60 features. (784). The bagging technique fared better than other classifiers with a 98% accuracy when using PCA-selected features.

REFERENCES

[1] M. Awad and R. Khanna, “Support Vector Machines for Classification,” Efficient Learning Machines, Berkeley, CA, Apress, pp. 39-66, 2015, doi: 10.1007/978-1-4302-5990-9_3.

[2] L. Wang, Z. Q. Lin, and A. Wong, ‘‘COVID-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images,’’ Sci. Rep., vol. 10, no. 1, Nov. 2020, Art. no. 19549.

[3] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya, ‘‘Automated detection of COVID-19 cases using deep neural networks with X-ray images,’’ Comput. Biol. Med., vol. 121, Jun. 2020, Art. no. 103792, doi: 10.1016/j.compbiomed.2020.103792.

[4] A.I Khan,J.L Shah, and M.M Bhat,’’CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest X-ray images,’’ Comput. Methods Programs Biomed., vol. 196, Nov. 2020, Art. no. 105581, doi: 10.1016/j.cmpb.2020.105581.

[5] N. Habib, M. M. Hasan, M. M. Reza, and M. M. Rahman, ‘‘Ensemble of CheXNet and VGG-19 feature extractor with random forest classifier for pediatric pneumonia detection,’’ Social Netw. Comput. Sci., vol. 1, no. 6, pp. 1–9, Oct. 2020.

[6] A. Bernheim, X. Mei, M. Huang, Y. Yang, Z. A. Fayad, N. Zhang, K. Diao, B. Lin, X. Zhu, K. Li, S. Li, H. Shan, A. Jacobi, and M. Chung, ‘‘Chest CT findings in coronavirus disease-19 (COVID-19): Relationship to duration of infection,’’ Radiology, vol. 295, no. 3, Jun. 2020, Art. no. 200463, doi: 10.1148/radiol.2020200463.