IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304345 |

Applying Different Machine Learning Techniques for Prediction of COVID-19 Severity

Syed Mohammed Althaf Basha[1], Dabbara Harshitha[2], Yerragudi Anish Fathima[3], Srinivasulu Akash[4], Panga Ashok, Dr. S. Kannappan[5]

UG Student, Dept. of ECE, GATES Institute of Technology, Gooty, India[12 3 4]

Professor, Dept. of ECE, GATES Institute of Technology, Gooty, India[5]

ABSTRACT: Due to the increase in the number of patients who died as a result of the SARS-CoV-2 virus around the world, researchers are working tirelessly to find technological solutions to help doctors in their daily work. Fast and accurate Artificial Intelligence (AI) techniques are needed to assist doctors in their decisions to predict the severity and mortality risk of a patient. Early prediction of patient severity would help in saving hospital resources and decrease the continual death of patients by providing early medication actions. Currently, X-ray images are used as early symptoms in detecting COVID-19 patients. Therefore, in this research, a prediction model has been built to predict different levels of severity risks for the COVID-19 patient based on X-ray images by applying machine learning techniques. To build the proposed model, CheXNet deep pre-trained model and hybrid handcrafted techniques were applied to extract features, two different methods: Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) were integrated to select the most important features, and then, six machine learning techniques were applied. For handcrafted features, the experiments proved that merging the features that have been selected by PCA and RFE together (PCA + RFE) achieved the best results with all classifiers compared with using all features or using the features selected by PCA or RFE individually. The XGBoost classifier achieved the best performance with the merged (PCA + RFE) features, where it accomplished 97% accuracy, 98% precision, 95% recall, 96% f1-score and 100% roc-auc. Also, SVM carried out the same results with some minor differences, but overall it was a good performance where it accomplished 97% accuracy, 96% precision, 95% recall, 95% f1-score and 99% roc-auc. On the other hand, for pre-trained CheXNet features, Extra Tree and SVM classifiers with RFE achieved 99.6% for all measures.

KEYWORDS: Chest X-rays, COVID-19, deep learning, handcrafted techniques, machine learning, mortality prediction, severity prediction.

.

I. INTRODUCTION

Predicting the severity risk of any disease at an early stage is a crucial task and has many effects, like reducing the mortality rate, consuming hospital resources, and supporting doctors in their decision making. In the current critical period, during the spread of coronavirus around the world and the increasing number of patients and deaths, the number of COVID-19 patients reached nearly 230 million while the number of deaths was 4.7 million around the world till now during writing this research, according to statistics from Johns Hopkins University [1].

The United States is the head of the countries, followed by Brazil, India, France, Russia, Italy, and many other countries. The reasons behind this growth in numbers are the high prevalence of COVID-19, late diagnosis, and lack of resources in many hospitals to absorb this pandemic. Therefore, predicting the severity risk of COVID-19 patients is a critical task and has many positive outcomes, such as providing the required health care for each patient according to his severity, good consumption of hospital resources that give the highest priority to the high-risk patient, and assisting doctors in making their decisions that will lead to improvement in the patient's treatment.

Three main resources that could be used to detect COVID-19: X-ray images, computed tomography (CT) and reverse transcription-polymerase chain reaction (RT-PCR). The best type is RT-PCR, but it is very expensive, not available in all hospitals and takes a lot of time to get the results. Therefore, many doctors depend on chest radiological imaging such as X-rays and CT for early diagnosis and treatment of this disease [2]. CT is a very sensitive tool, but its results

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304345 |

can be observed after a long time according to the onset of symptoms, where normal CT takes from zero to two days to see its findings [3], so CT is difficult to use in monitoring patients periodically. Chest-X ray (CXR) radiography is less sensitive than CT and RT-PCR, but it is one of the most commonly used and accessible methods for rapid examination of lung conditions. X-ray findings are observed in a short time and it is not an expensive technique, so it can be used periodically to monitor the patient's status.

Feature extraction is a big challenge, especially when the dataset size is small. The features of an image are quantitative data values or pixel intensities that hold meaningful information about pixels of an image in terms of local and/or global variations. It is the process of locating a feature vector representation of input images, and effectively isolating the most critical variables relevant to the aim of the intended application. Currently, two main approaches are used for feature extraction in X-ray images, namely non-handcrafted (deep learning) and handcrafted feature extraction. The deep learning technique extracts local features from images as much as possible, but it is more suitable for large datasets. Most of the published research that has relied on chest X-ray images in its work so far focuses either on the diagnosis the disease itself or the distinction between COVID-19 and other types of pneumonia, as in [4]–[9]. These studies depended on the Convolutional Neural Network (CNN) techniques and different pre-trained models like ResNet, DenseNet, CheXNet, Xception, VGG, and others. In [10], the authors used X-ray images to detect specific severity scores of COVID-19 as a regression problem with a pre-trained deep learning model called DenseNet where X-ray images were scored retrospectively by experts in terms of the extent of lung involvement, which is called geographic extent score (range 0-8), as well as the degree of opacity, which is named lung opacity score (range 0-6), and mean absolute error (MAE) was calculated to evaluate the model. Unfortunately, sometimes deep learning models may suffer from over-fitting problems [11], cause high bias because they extract unknown and abstract features [12], and need high-dimensional datasets to obtain higher performance. To overcome such problems, some researchers used pre-trained transfer learning models to take advantage of the potential of deep learning techniques.

II. MATERIALS AND METHODS

The details of the proposed framework for COVID-19 severity level prediction are presented in this section. First, the overall architecture of the framework is described, then the applied methodology for predicting the severity discussed. The proposed framework consists of four main phases, as shown in Fig. 1. Firstly, the input X-ray dataset is passed to data pre-processing to resize and normalize the images. Then, different feature extraction methods are applied to extract the features. After that, feature selection techniques are executed to select the most important features in the images, and finally, different machine learning classifiers are applied to build the models.

A.THE PROPOSED ARCHITECTURE

The proposed framework consists of four main phases, as shown in Fig. 1. Firstly, the input X-ray dataset is passed to data pre-processing to resize and normalize the images. Then, different feature extraction methods are applied to extract the features. After that, feature selection techniques are executed to select the most important features in the images, and finally, different machine learning classifiers are applied to build the models.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304345 |

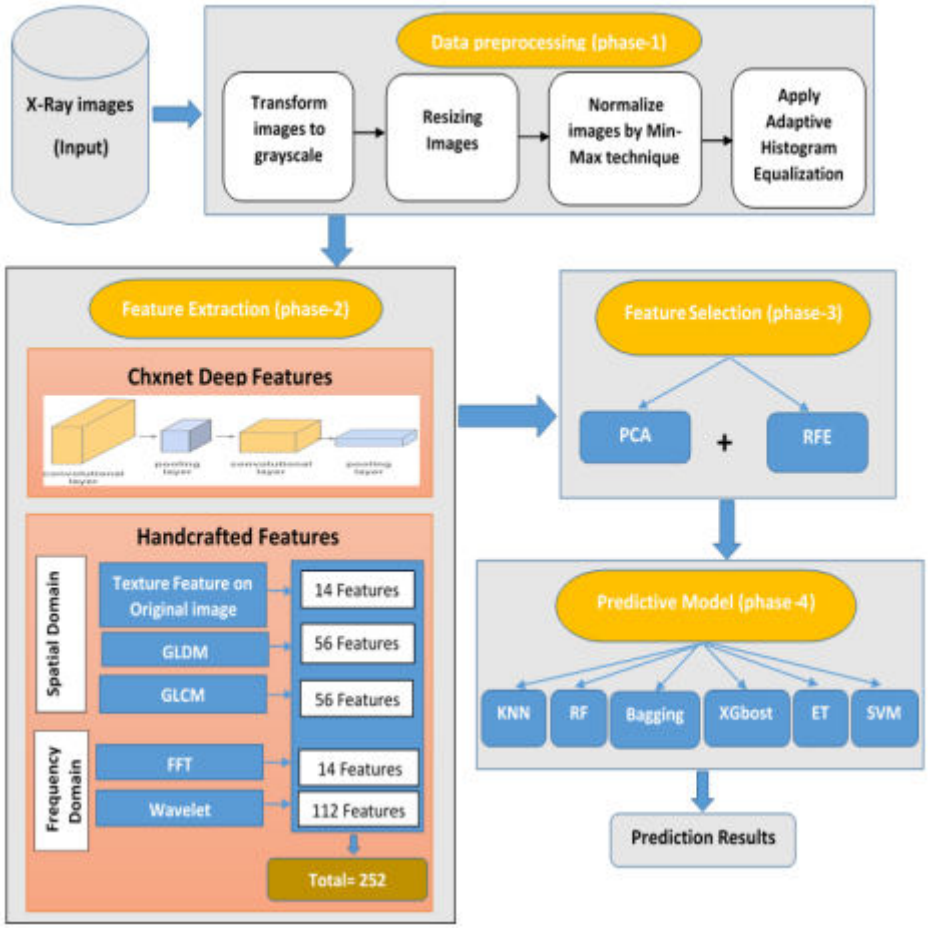


FIGURE 1.The proposed architecture

B. PROPOSED METHODOLOGY

The main phases of the proposed framework are briefly described in this section.

1) PHASE-1 (DATA PREPROCESSING)

The data pre-processing phase aims at preparing the data to be used in the prediction model. Usually, data are messy and come from different sources with different sizes and resolutions. So, this phase is crucial for cleaning up and normalizing the data to reduce the complexity and increase the accuracy of the prediction model. Different types of transformations could be executed according to the dataset like, re-sizing, rotating, shifting, and normalizing, and so on.

2) PHASE-2 (FEATURE EXTRACTION)

In this study, due to the small size of the used dataset and the motivating results of using the handcrafted techniques and pre-trained models for extracting features from medical images in other published papers [8], [9], [13], [14], [29]–[31], so the features would be extracted using two different techniques: the pre-trained CheXNet deep model and a set of handcrafted descriptors.

CheXNet Deep Features: CheXNet [32] is a 121-layer convolutional neural network based on the DenseNet architecture [33]. It trained over 100 000 frontal view chest X-ray images for 14 pneumonia diseases. The performance of CheXNet was compared with four academic radiologists who annotated a test set and is said to exceed average radiologist performance in detection.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304345 |

3) PHASE-3 (FEATURE SELECTION)

Two different methods are used for selecting the most significant features: Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). They are fast, popular in many domains, have fewer parameters, simple implementation, and require low computations.

Principal Component Analysis (PCA) is an orthogonal transformation that converts a group of possibly correlated features into a smaller number of interrelated features called principal components [34]. The objective of PCA is to reduce the dimensionality of the dataset while retaining most of the original variability in the data. It is done by projecting the original dataset into the reduced space of PCA using the eigenvectors of the correlation/covariance matrix.

The resulting projected data are linear combinations of the original data describing most of the variance in the data where the first component contains the largest amount of data variance. After that, each subsequent component has the remaining data variability as possible. Algorithm 1 presents the steps of PCA for reducing dimensionality

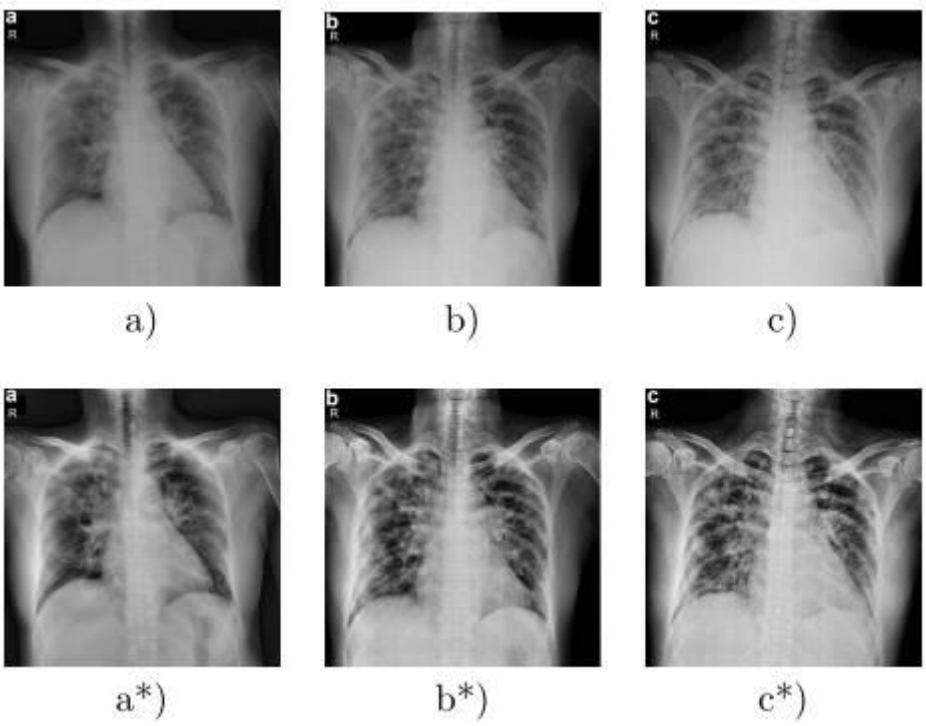


FIGURE 2. (a), (b), and (c) are examples of images before applying the preprocessing methods, while (a*), (b*), and (c*) are examples of images after applying the preprocessing

4) PHASE-4 (PREDICTION MODELS)

To build a robust predictive model, different machine learning classifiers were compared. Six classifiers: K Nearest Neighbors (KNN), Random Forest (RF), Extra Tree (ET), Bagging, extreme Gradient Boosting (XGBoost), and Support vector machine (SVM) were used. K Nearest Neighbors (KNN) is a supervised algorithm developed by Thomas Cover [36] for classification and regression problems.

It uses a feature similarity method to predict the label of a new given point, which further means that the new test point will be classified according to the majority vote of the nearest K neighbours in the training set, where K is the number of neighbours. It is characterized as simple, easy-to-implement, depending on only a single parameter (K), and effective classifiers.

In this section, all the details of the used dataset, feature selection techniques, machine learning classifiers, and evaluation metrics are described.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304345 |

III. EXPERIMENTAL SETUP

A. DATASET

A public-available dataset developed by Cohen JP [25] has been used. The dataset contains data about COVID-19 patients and other pneumonia patients. It includes information about the patient like patient-id, age, X-ray image, type of disease, survival or not, and went-ICU or not.

The cohort of the paper was built by first excluding all patients less than 18 years old; selecting only the confirmed COVID-19 patients that had a positive RT–PCR test: 40% female and 60% male patients; and then, using the variables that represent patient status, such as survival or not, as well as went-ICU to classify the patients.

The aforementioned variables are used to label the severity of patients. Actually, two main variables are used to identify the severity level, namely survival and went-ICU variables. The classification of the severity is done according to the following rules

- if survival is false, it is called high severity;
- if survival and went-ICU are true, it is termed moderate severity;
- if survival is true and went-ICU is false, it is named low severity.

B. FEATURE SELECTION METHODS

1) PRINCIPAL COMPONENT ANALYSIS (PCA)

The aforementioned PCA steps were applied to the original data. Fig. 4 shows the variance ratio of different components for handcrafted and it is noted that 24 components represent 95% of the variance of the original data, so the number of selected principal components is 24 for handcrafted features. On the other hand, the variance ratio of CheXNet components is 103 components, which represents 95% of the variance of the original features.

2) RECURSIVE FEATURE ELIMINATION (RFE)

There are two types of hyper parameters required: the number of features to be selected and the algorithm to be used as an objective function in the feature selection process. Various algorithms have been tried, like gradient boosting, logistic regression, decision tree, random forest, and perceptron with different numbers of features. Every outcome of selected features by these experiments has been tested on the used machine learning classifiers in this study. Due to the large number of results, only the best hyper-parameter selection.

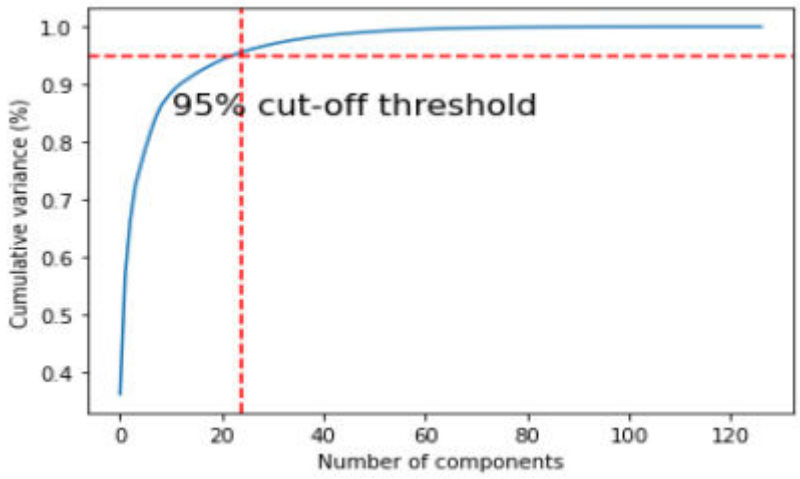


FIGURE 4. The number of PCA components for the handcrafted features.

values are mentioned here, where the best estimator algorithm is perceptron and the optimal number of features is 28 for handcrafted and 100 for CheXNet deep features that performed well with the classifiers.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304345 |

C. MACHINE LEARNING CLASSIFIERS

For the hyper-parameters of the used classifiers, the grid search algorithm and a try and error approach were applied to find out the best values for them. Table. 1 reports the hyper-parameter values of the used classifiers.

TABLE 1. Hyper-parameters of each classifier.

Classifier	Parameters
KNN	n_neighbors=10.
RF	n_estimators=100; max_depth=6; max_features=0.2.
XGboost	n_estimators=100; learning_rate=0.3; objective=multi:softprob; subsample=0.5.
Bagging	n_estimators=100; base_estimator=DecisionTree; max_features=0.2; max_samples=0.9.
ET	n_estimators=100; max_features=0.2; min_samples_split=15.
SVM	kernel=rbf; C=3; gamma=0.1.

D. EVALUATION METRICS

In order to evaluate the performance of the models, various scores were measured to ensure the results, like accuracy, macro-avg of sensitivity/recall, precision, f1 scores, and AUROC. The variable C refers to the number of classes and i 2 C indicates to a specific class. TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative, respectively.

• The precision score measures the ratio of correctly predicted positive observations to the total predicted positive observations by (4).

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}. \qquad (4)$$

• Recall/Sensitivity calculates the ratio of correctly predicted positive observations to all observations in a true class using (5).

$$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{TN}_i}. \qquad (5)$$

• F1-score is the weighted average between Precision and Recall calculations and is computed by (6). It can be documented that this score is more useful than accuracy because it takes into account false positive and false negative measurements.

$$\text{F1-Score}_i = \frac{2(\text{Recall}_i \times \text{Precision}_i)}{\text{Recall}_i + \text{Precision}_i}. \qquad (6)$$

• Accuracy simply calculates the ratio of correctly predicted observations to the total observations, as in (7).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \qquad (7)$$

• To evaluate the general performance of different classifiers, the Macro-avg score has been chosen and computed as (8) for averaging calculation by classes.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

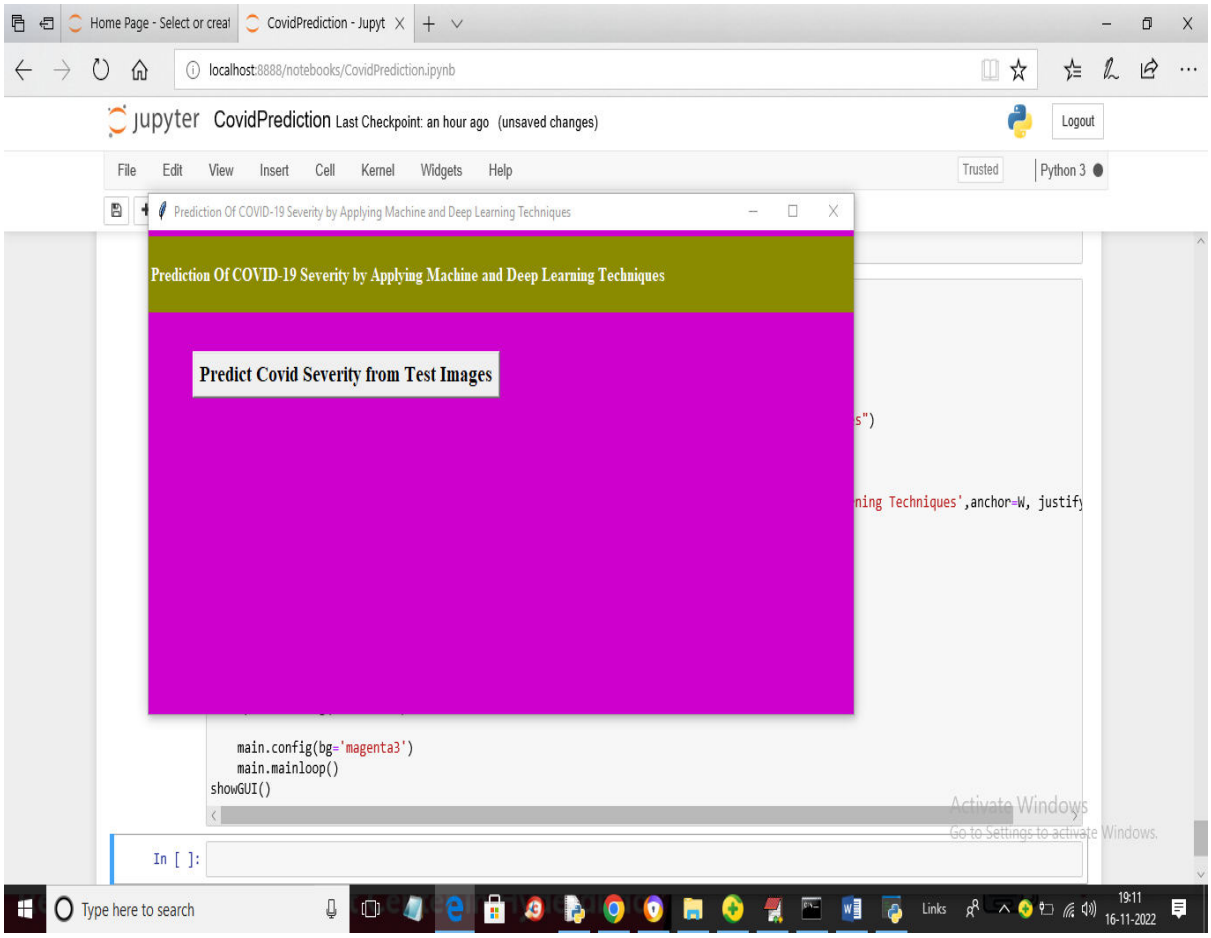
| DOI:10.15680/IJIRSET.2024.1304345 |

$$\text{Macro-avg(measure)} = \frac{\sum_i \text{measure}_i}{C}. \qquad (8)$$

• The Area Under the Receiver Operating Characteristics (AUROC) is computed. It is an important performance measurement for classification problems. It computes the capability of the model to distinguish between the different classes, which mean that the model with a higher AUC value can predict the class sample as its actual value.

IV. RESULT

In above screen in tabular format we can see performance of each algorithms and now run last block to get below screen where you can upload test image and predict severity



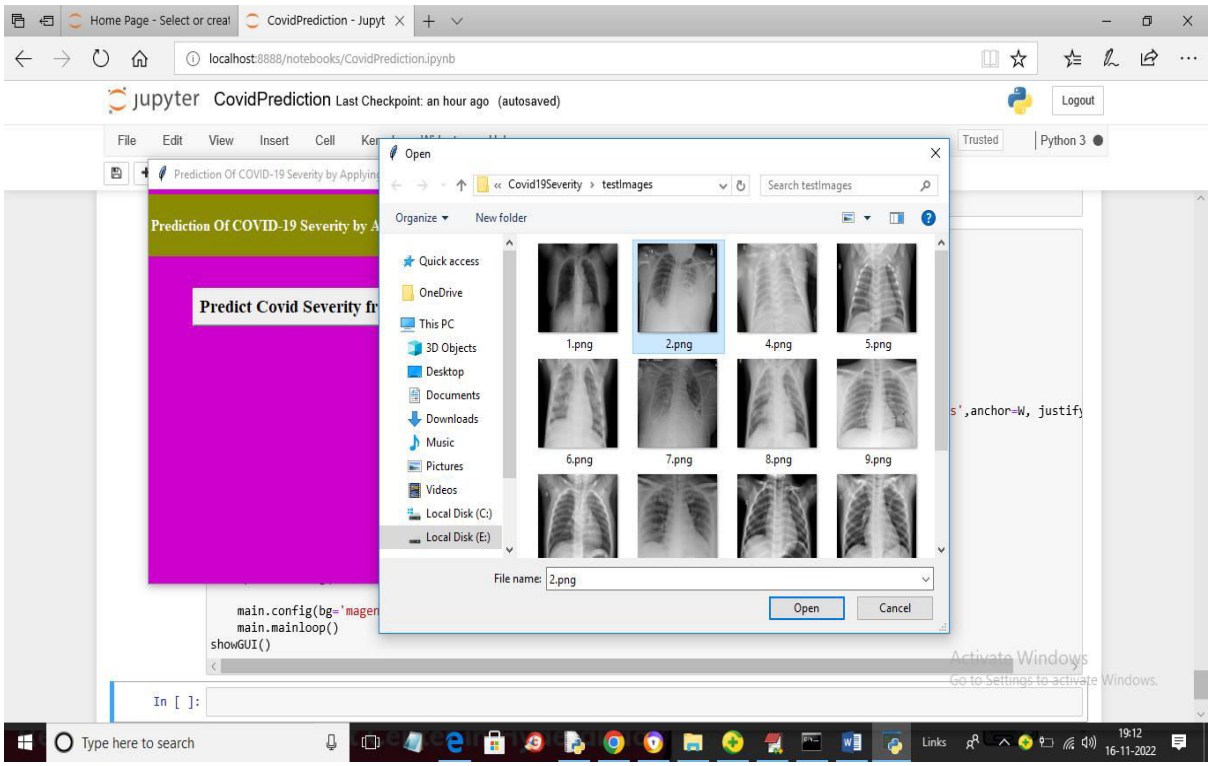
In above screen click on 'Predict Covid Severity' button to upload test image like below screen

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

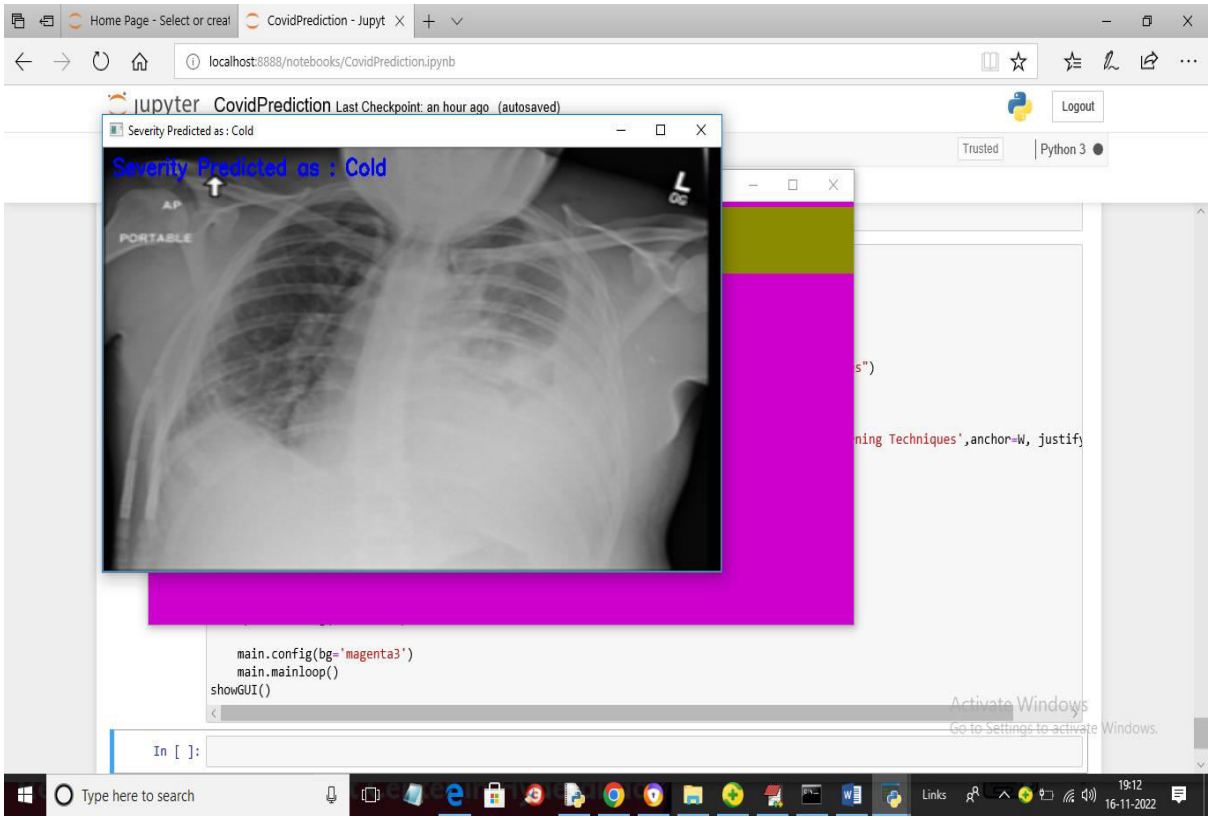
| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304345 |



In above screen selecting and uploading 2.png and then click on 'Open' button to get below prediction



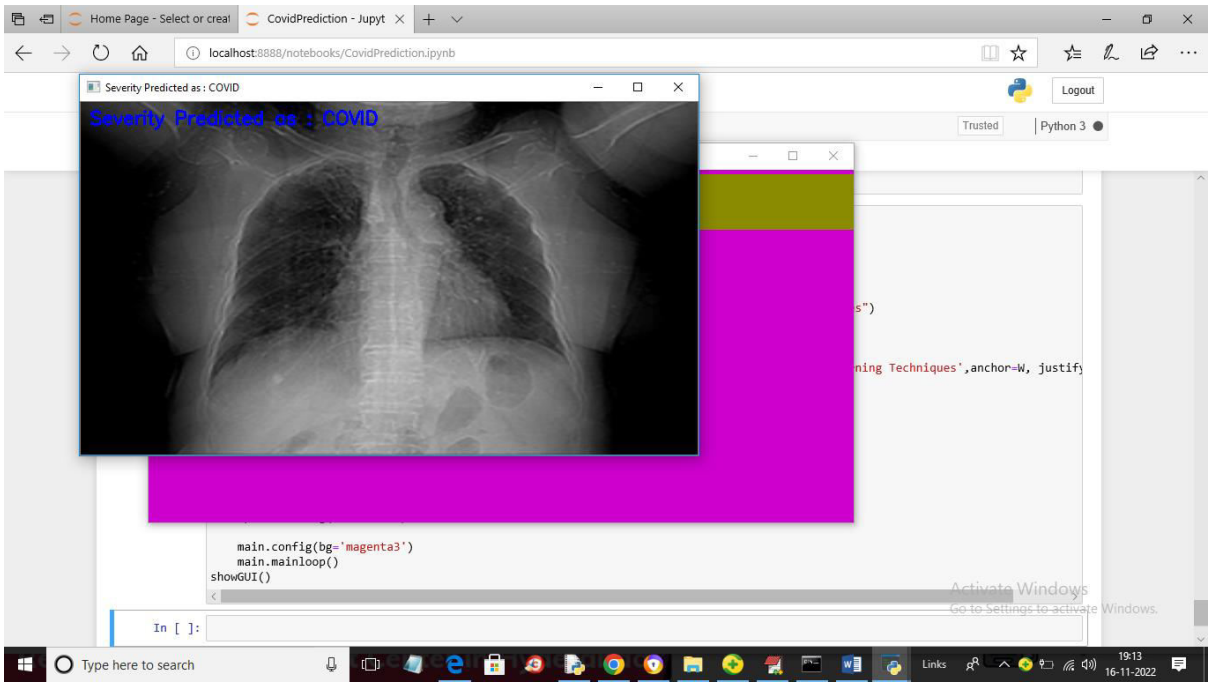
In above screen severity predicted as COLD and similarly you can upload and test other images and below are the other images severity

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

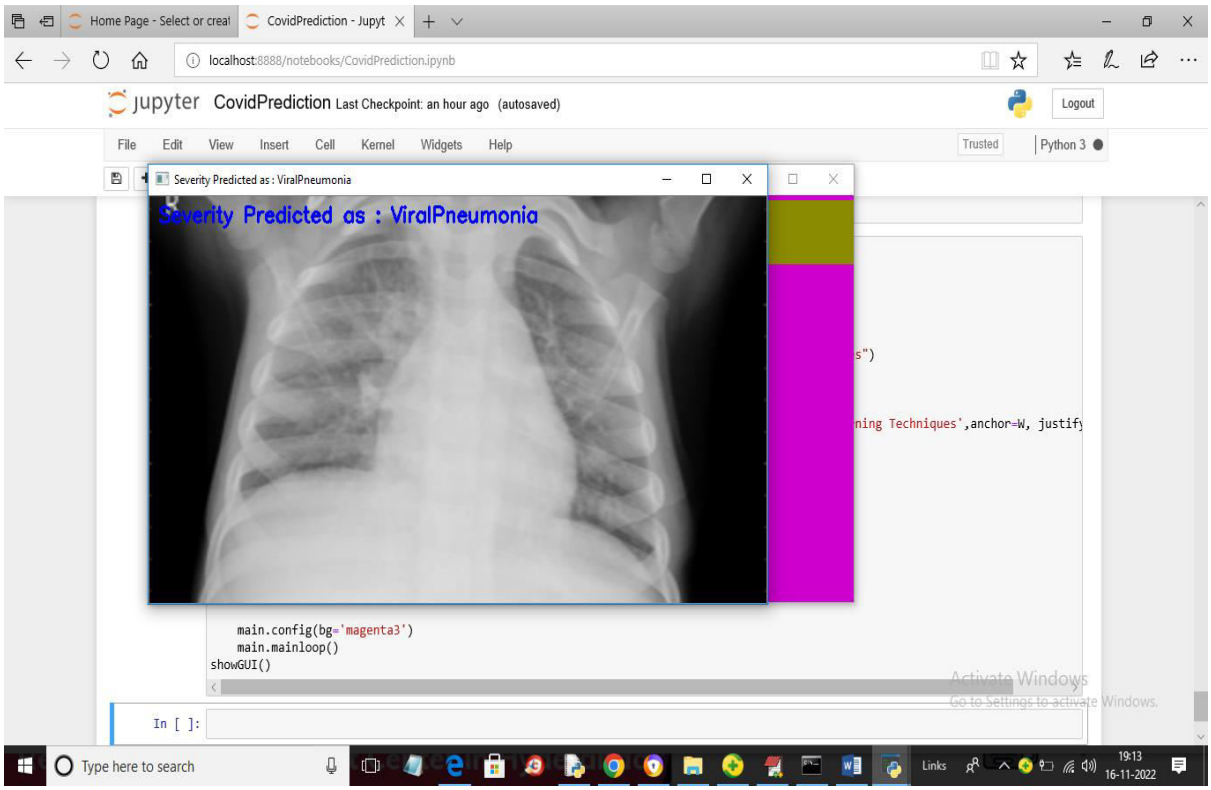
| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304345 |



In above screen COVID detected



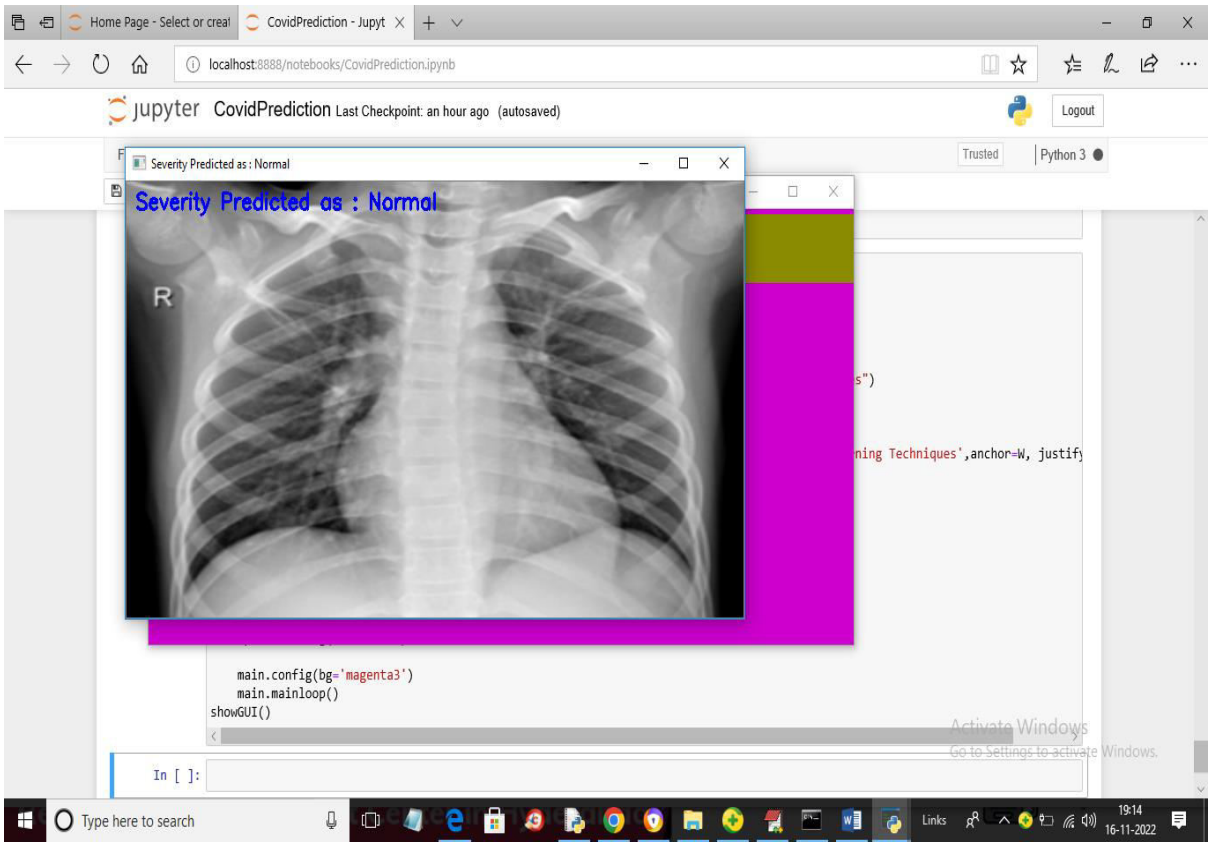
In above screen pneumonia detected

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304345 |



In above screen Normal is predicted

V. CONCLUSION

This study proposes a new predictive framework for the severity and mortality risk of COVID-19 patients to help doctors, hospitals, and medical facilities in their decision making about which patients need to get attention first before others, and at the same time, to keep hospitals' resources for high-risk priority patients. The proposed model is based on a public X-ray image dataset for confirmed patients with COVID-19 disease. The dataset is classified into three severity classes: high, moderate and low severity labels. The high severity class means that a patient may die, while the moderate severity class refers that a patient will need to enter the ICU, whereas the low severity indicates that a patient will not need to enter the ICU. Pre-trained deep CheXNet and hybrid handcrafted techniques were applied to extract the features from X-ray images, then two feature selection techniques were merged together: PCA and RFE, and many predictive models were built based on machine learning algorithms like KNN, Random Forest (RF), XGboosting, Bagging, Extra Tree, and SVM to compare and ensure the results. Many experiments were executed and the results revealed that, for handcrafted features, merging the selected features by PCA and RFE (PCA C RFE) achieved the best results with allclassifiers where the number of used features is 52, which is nearly 25% of the original number of extracted features (252). Also, XGboost and SVM surpassed other classifiers with an accuracy of 97% and 100% roc-AUC with (PCA C RFE) selected features. On the other hand, for CheXNet deep features, RFE has achieved promising results with all measures by all classifiers and 99.6% for all measures with SVM and Extra tree classifiers where the number of selected features is 100 from originally 9216 features.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304345 |

REFERENCES

[1] COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). [Online]. Available: https://coronavirus.jhu.edu

[2] Z. Y. Zu, M. D. Jiang, P. P. Xu, W. Chen, Q. Q. Ni, G. M. Lu, and L. J. Zhang, ''Coronavirus disease 2019 (COVID-19): A perspective from China,'' Radiology, vol. 296, no. 2, pp. E15–E25, Aug. 2020.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] M. Rahimzadeh and A. Attar, ''A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of xception and ResNet50 V2,'' Informat. Med. Unlocked, vol. 19, Jan. 2020, Art. no. 100360, doi: 10.1016/j.imu.2020.100360.

[8] 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.

[9] P. R. A. S. Bassi and R. Attux, ''A deep convolutional neural network for COVID-19 detection using chest X-rays,'' Res. Biomed. Eng., vol. 2, pp. 1–10, Apr. 2021.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details