Journal of Theoretical and Applied Information Technology

31st December 2023. Vol.101. No 24
© 2023 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PREDICTION OF LUNG DISEASE SEVERITY BY APPLYING MACHINE AND DEEP LEARNING TECHNIQUES

SURYA PRAKASH GUTTULA [1,*], SANJEEV KUMAR GUPTA [2], LAXMI SINGH [3], K VIDYA SAGAR [4]

[1, 3] ECE Department, Rabindranath Tagore University, Mendua, Bhopal, Madhya Pradesh 464993.

[2] Dean, Rabindranath Tagore University, Mendua, Bhopal, Madhya Pradesh 464993.

[4] EIE Department, VNR Vignana Jyothi Institute of Technology, Hyderabad 500090.

suryaprakash.guttula@gmail.com

ABSTRACT

Virus infected diseases are increasing rapidly. SARS covid -19 is one emerged into human body to extinct the human life. Prediction of the rapid changes and meticulous interpretation of the type of decease is challenging. Various stages of the risk severity prediction and interpretation is challenging. This paper discussed various machine learning algorithms applied on X-radiation chest images to predict the severity of the decease. Significant features are extracted using Principal component analysis (PCA). Bagging, Ada boost, XG boost, KNN machine learning methodologies applied to achieve the reliable performance. Bagging methodology shows dominant performance over other machine learning methodologies with 98.82% precision value and 98.67% accuracy. The F1 score and recall value is significantly good with 98.755 and 98.69 respectively. Bagging methodology is much reliable to interpret X-radiation images for Sars Covid-19 infections.

Keywords: *SARs Covid-19, X-radiation images, Bagging, Ada boost, KNN, XG boost, PCA.*

1. INTRODUCTION

The work is focused on Sars covid severity detection and prediction. This paper aimed to analyze the X-radiation (X-radiation) images. The raw X-radiation image is pre-processed in four stages.

Stage1: The image is transformed into grey scale level.

Stage2: The grey scale level image is resized. For better interpretation of the severity level, without losing valued information, the image is resized.

Stage 3: Image normalization is done using min. max approach

Stage 4: Adaptive histogram equalization is applied to improve the contrast levels of the image.

The image size is again resized to 224X 224 and CheXNet model is applied to extract insight features of the image. CheXNet is a convolutional neural network with five convolution blocks. It is composing of 121-layers. Twenty thousand X-radiation images are trained for pneumonia infections. Max-pool sublayer is used to extract significant features of the image. CheXNet model generate 9216 significant features. Image features are extracted at both spatial domain and frequency domain. With Grey level Co-occurrence matrices 56 features are extracted. The difference matrices built in four directions to extract another set of 56 features. The feature extractions with frequency domain considered the rate change of pixel values in special domain. For texture feature extraction FFT is used.

Fourteen significant features are extracted with wavelet transform. The image is further processed with Principal Component Analysis (PCA) and recursive feature elimination (RFE) methods. The dataset dimensionality is minimized with PCA. This is essential to retain the original features of the image. The original data is projected into reduced space using eigen vectors of correlation and co-variance matrix. The projected data represents the variance in the dataset. The first components describe the maximum data variance and remaining components represents the variability in the rest of the dataset. The optimal features of the image dataset are selected with Recursive Feature Elimination (RFE) approach. RFE applied to extract significant influence features to enhance the reliability of the classification.

K Nearest Neighbors' (KNN), Random Forest (RF), Bagging, extreme Gradient Boosting (XGBoost), and Ada boost are considered to classify the image and to develop predictive model. KNN is used to classify new test points based on the maximum election of nearest neighbors (K) in the training dataset. With decision tree approach the variance is significantly high. To reduce the variance, Bootstrap aggregating methodology is applied. This approach accumulates the prediction of maximum weak learners (decision Tree) and build a strong learner to improve the accuracy of the output. New training sets are generated using original dataset. This dataset is labelled as bootstrap dataset. The decision tree is built with each new bootstrap dataset. The original dataset is fed with the newly built classifiers and then final decision is made. The classification results are aggregated and produce the end results with random forest methodology.

Extra tree produced an output by randomly considering subset of features. The computational speed is much faster with Extra tree than other classifiers. Extreme Gradient boosting (XG Boost) correct the errors by adding each new tree with ensemble model. This iterative process proceeds until no further modifications required. Then the result is predicated by adding the trees to minimize the loss. Supportive vector machine is used to achieve optimal decision boundary with significantly maximum margin.

2. METHODOLOGY

The X-radiation image is pre-processed in four stages. Stage1: The image is transformed into grey scale level. Stage2: The grey scale level image is resized. Stage 3: Image normalization is done using min. max approach 4: Adaptive histogram equalization is applied to improve the contrast levels of the image. The image is further processed with Principal Component Analysis (PCA) and recursive feature elimination (RFE) methods. The dataset dimensionality is minimized with PCA. The optimal features of the image dataset are selected with Recursive Feature Elimination (RFE) approach. RFE applied to extract significant influence features to enhance the reliability of the classification.

K Nearest Neighbours (KNN), Bagging, extreme Gradient Boosting (XG Boost), and Adaboost are considered to classify the image and to develop predictive model. KNN is used to classify new test points based on the maximum election of nearest neighbours (K) in the training dataset. With decision tree approach the variance is significantly high. To reduce the variance, Bootstrap aggregating methodology is applied. This approach accumulates the prediction of maximum weak learners (decision Tree) and build a strong learner to improve the accuracy of the output. New training sets (bootstrap dataset) are generated using original dataset. The decision tree is built with each new bootstrap dataset. The original dataset is fed with the newly built classifiers and then final decision is made. The classification results are aggregated and produce the end results with random forest methodology. Extra tree produced an output by randomly considering subset of features. Extreme Gradient boosting (XG Boost) correct the errors by adding each new tree with ensemble model. This iterative process proceeds until no further modifications required. Then the result is predicated by adding the trees to minimize the loss. Total process showed in flow chart format in Fig 2.1 below.

3. RESULTS AND DISCUSSION

The datasets taken from open source developed by Cohen JP. Survival and went-ICU are two significant variables considered. The severity classification is done based on the intensity of the Xray image which represents the Sars covid -19. High severity represents the patient survival is false. Severity moderate level represents the patient is safe and intensive care unit required is true. Low severity represents the patient survival is true and intensive care unit requirement is false. 80 percent of the images are considered for training and 20 percent of the data sets are considered for testing. The data sets are grouped with patient Id for meticulous analysis. The variance ratio for twenty-four components shows 95% from actual data. So, the principal components elected twenty-four features are handcrafted features. Similarly, the variance ratio is 102, so the number of components is 102. The H parameters are tuned for improving the performance of various methodologies. For KNN, n-neighbours are considered as ten. For RF estimator is hundred, depth is six, For XGboost N_estimator is hundred, learning rate is 0.3, sub sample is 0.4. for Bagging classifier n_estimator is hundred, base estimator is same as decision tree, Max. Estimator is 0.25, Max. samples are 0.9.

The performance evaluation is done for ADABOOST, Bagging Classifier, XGBOOST, KNN and CNN deep learning methodology to predict COVID severity using Accuracy, Precision, ecall, Confusion Matrix and FSCORE. The performance evaluation is stated below in Figure 3.1.

The input X-radiation datasets are tagged with different labels like Normal, COVID, COLD and Pneumonia. The accuracy of a methodology is enhanced by applying PCA to elect significant features from the image. Jupiter notebook is used to run different methodologies.

Table 3.1. Quality metrics for various algorithms

Algorithm name	**Precision**	**Recall**	**F1 Score**	**Accuracy**
KNN	85.684321	83.524205	84.411613	83.244916
Bagging	98.823630	98.695465	98.755703	98.673740
XGBoost	98.370675	98.339185	98.353483	98.275862
AdaBoost	62.279605	63.923264	62.986020	60.565871
CNN	96.113944	95.152189	95.583942	95.269673

The confusion matrix for various methodologies represented in fig.3.1. The true_ labels are shown on Y-Axis and Predicted_ labels are shown on X-Axis. distinctive colour boxes in diagonal represents true prediction count. Blue colour boxes represent false prediction count. KNN methodology exhibits 84% accuracy Bagging Classifier, and we got its accuracy as 98.58%. XGBOOST produced accuracy is 98.23%. AdaBoost produced accuracy is 60%. with CNN achieved 95.26% accuracy. The prediction count with bagging classifier is superior to other classifiers. The incorrect prediction count is 1.41 percentage. The incorrect prediction count with XGBoost is 1.7 percentage only. Bagging classifier and XG boost shows reliable methodologies to predict COVID-19 severity detection.

The performance metrics of all the algorithms are represented in Fig. 4.3 The X-Axis represents various methodologies and each colour bar represents metrics (accuracy, precision, F score and Recall) of the testing images. Bagging classifier exhibits superior performance over other methodologies. XG Boost also competing with bagging classifier to analyze the X-radiation images.

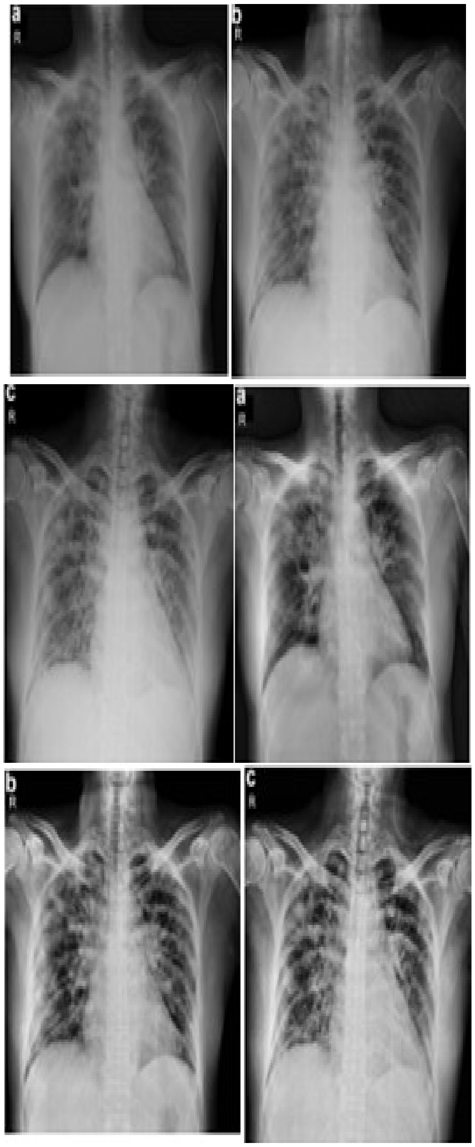
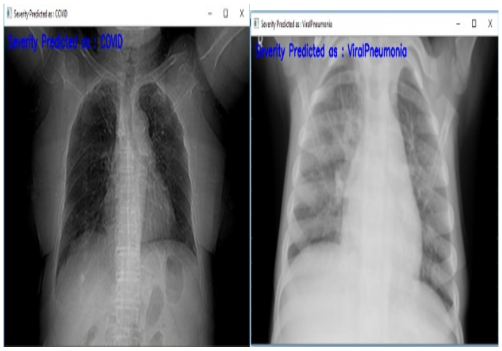


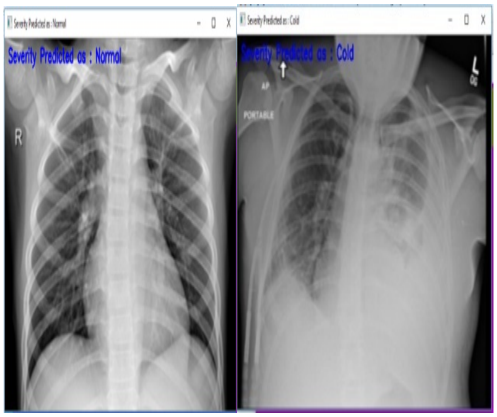
Fig 3.1. (a), (b), (c) are original Images and (d), (e), (f) are preprocessed images



Severity Predicted as : COVID

Severity Predicted as : ViralPneumonia

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

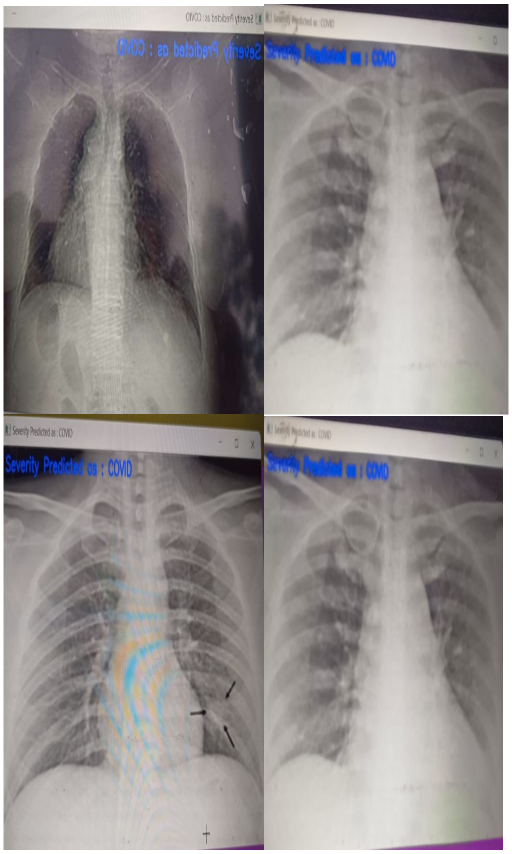


Severity Predicted as : Normal

Severity Predicted as : Cold

Fig 3.2 Predicted images set 1

Pre-processed images are represented in Fig.3.2. d, e, f represents the quality images with good resolution.
Fig3.2. (a) represents covid severity is very high.
Fig.3.2. (b) represents viral pneumonia.
Fig.3.2. (c) represents normal image.
Fig.3.2. (d) represents lungs infected due to cold.

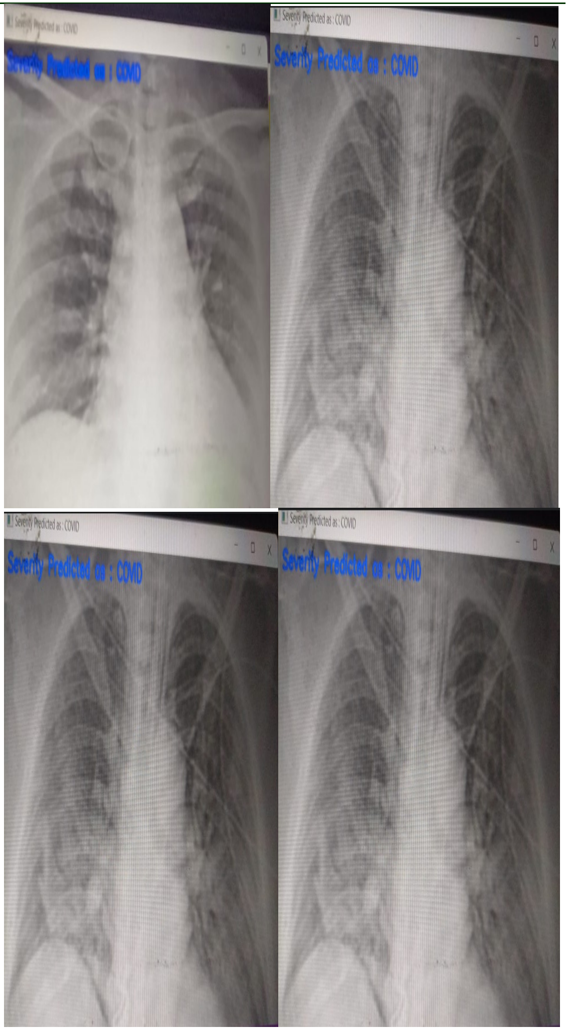


Severity Predicted as : COVID

Severity Predicted as : COVID

Severity Predicted as : COVID

Severity Predicted as : COVID



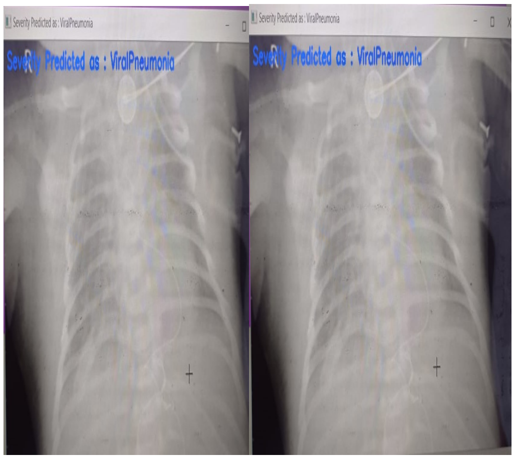
Severity Predicted as : COVID

Severity Predicted as : COVID

Severity Predicted as : COVID

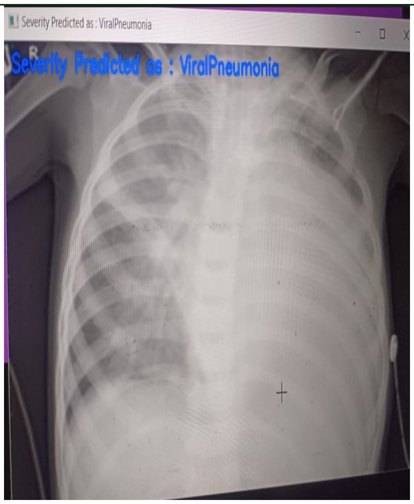
Severity Predicted as : COVID

Fig 3.3 Predicted image Set 2 Covid Detected X ray Images



Severity Predicted as : ViralPneumonia

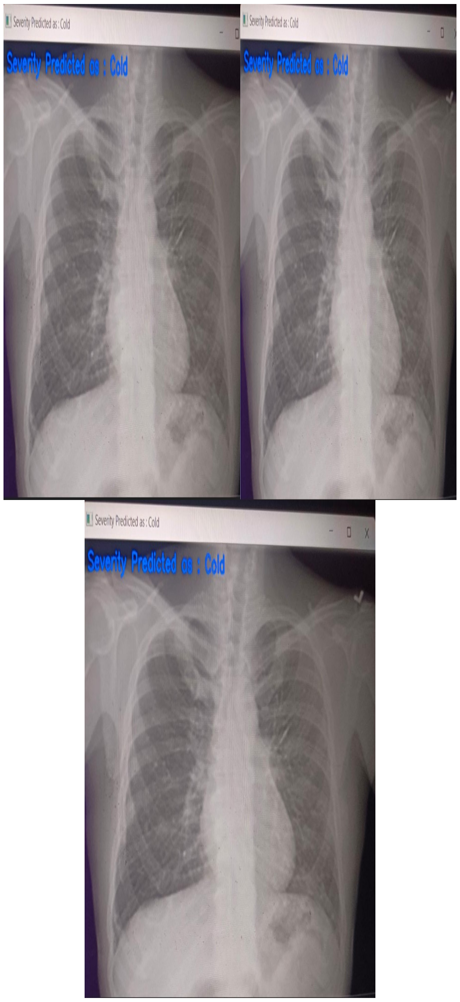
Severity Predicted as : ViralPneumonia



Severity Predicted as : ViralPneumonia

Severity Predicted as : ViralPneumonia

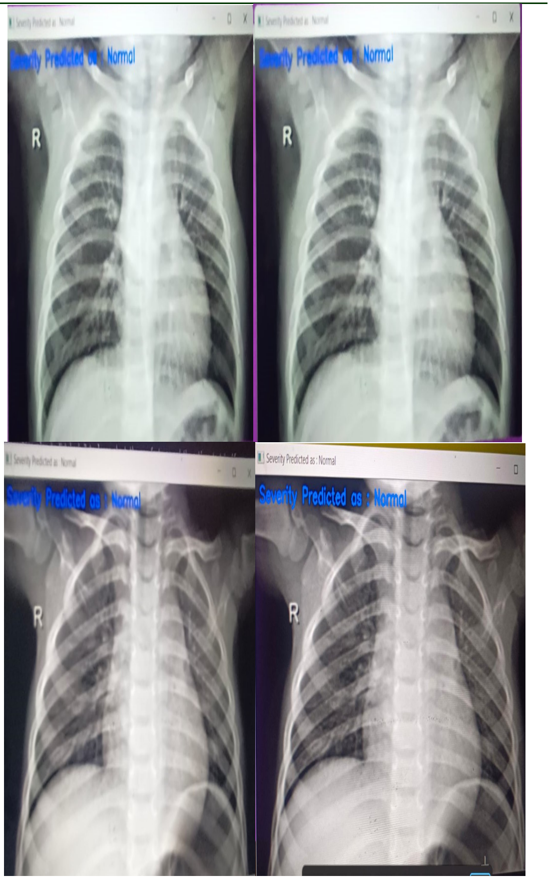
Fig 3.4 Predicted Image set 3 Viral Pneumonia detected X ray Images



Severity Predicted as : Cold

Severity Predicted as : Cold

Fig 3.5 Predicted Image set 4 cold detected X ray Images



Severity Predicted as : Normal

Severity Predicted as : Normal

Fig 3.6 Predicted Image set 5 Normal detected X ray Images

Posteroanterior chest X-Rays (PA) fig.3.4. are considered SARS COVID-19 detection with chest X-RAY findings from the above images shows ground glass opacities and ground glass opacities with morphology. Patchy multifocal distribution is also traced. It had been observed the multifocal distribution in the lower field is bi lateral and peripheral only. Rounded bilateral peripheral opacities are traced in the lower portion of PA chest X-ray images. Both the lungs are extensively involved with peripheral opacities. Ground glass opacities are traced at central and lower lobe. Both the lobes are involved, and pleural effusion is found. These findings suspect severe covid-19 infection. Pneumonia PA Chest X-ray images Fig.3.5. shows lobar consolidation and masses involvement of left and right-side lungs. Pleural effusion found at lower region of the lungs including cavitation at lower region. These findings may be suspected covid-19 with pneumonia. Fig 3.6. are considered cold detection with chest X-RAY findings shows masses involvement of left and right-side lungs. Fig 3.6 shows normal images which means no involvement of any of the above mentioned disorders.

4. CONCLUSION

The features extracted from X-radiation images using PCA and RFE are further processed with Random Forest, Bagging, XG boosting and KNN methods. The selected '52'features are extracted with PCA and RFE are merged to achieve better results. The merged features are processed with bagging classifier shows improved results over other methods. The precision is 98.82% and accuracy is 98.69% and specifically F1 Score is 98.755. Ada boost methodology shows inferior results. The Ada boost methodology shows metrics such as accuracy of 38.10 %, precision is 36.54%, F1 Score is 35.76% and recall value is 34.77% less values over Bagging Classifier. XG boost methodology is much compete with bagging classifier. The metrics of XG boost precision id 0.453%, accuracy is 0.398%, F1 Score is 0.402% and Recall value is 0.356 lesser values over Bagging classifier.

For meticulous estimation of the results Bagging classifier is much superior over other methodologies specifically for Sars Covid infected X-radiation images. Computed tomography (CT) imaging is expensive over X-radiation imaging. Processing PA chest X-ray images and interpreting the severity of the decease is significantly beneficial to minimize medical diagnosing expenses.

REFERENCES

[1] 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 (2): Aug. 2020, pp. E15–E25.

[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-radiation images,'' Sci. Rep., vol. 10(1): Nov. 2020, pp 12-21.

[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-radiation images,'' Comput.Biol.Med., Vol.121: Jun. 2020.

[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-radiation images'' Comput. Methods Programs Biomed, vol. 196: Nov. 2020.

[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(6): Oct. 2020, pp. 1–9.

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

[7] J. P. Cohen, L. Dao, K. Roth, P. Morrison, Y. Bengio, A. F. Abbasi, B. Shen, H. K. Mahsa, M. Ghassemi, H. Li, and T. Duong, ''Predicting COVID-19 pneumonia severity on chest X-radiation with deep learning,'' Cureus, vol. 12: Jul. 2020

[8] D. Al-Karawi, N. Polus, S. Al-Zaidi, and S. Jassim, ''Artificial intelligence-based chest X-radiation test of COVID-19 patients,'' Int. J. Comput. Inf. Eng., vol. 14(10): 2020, pp. 353–359.

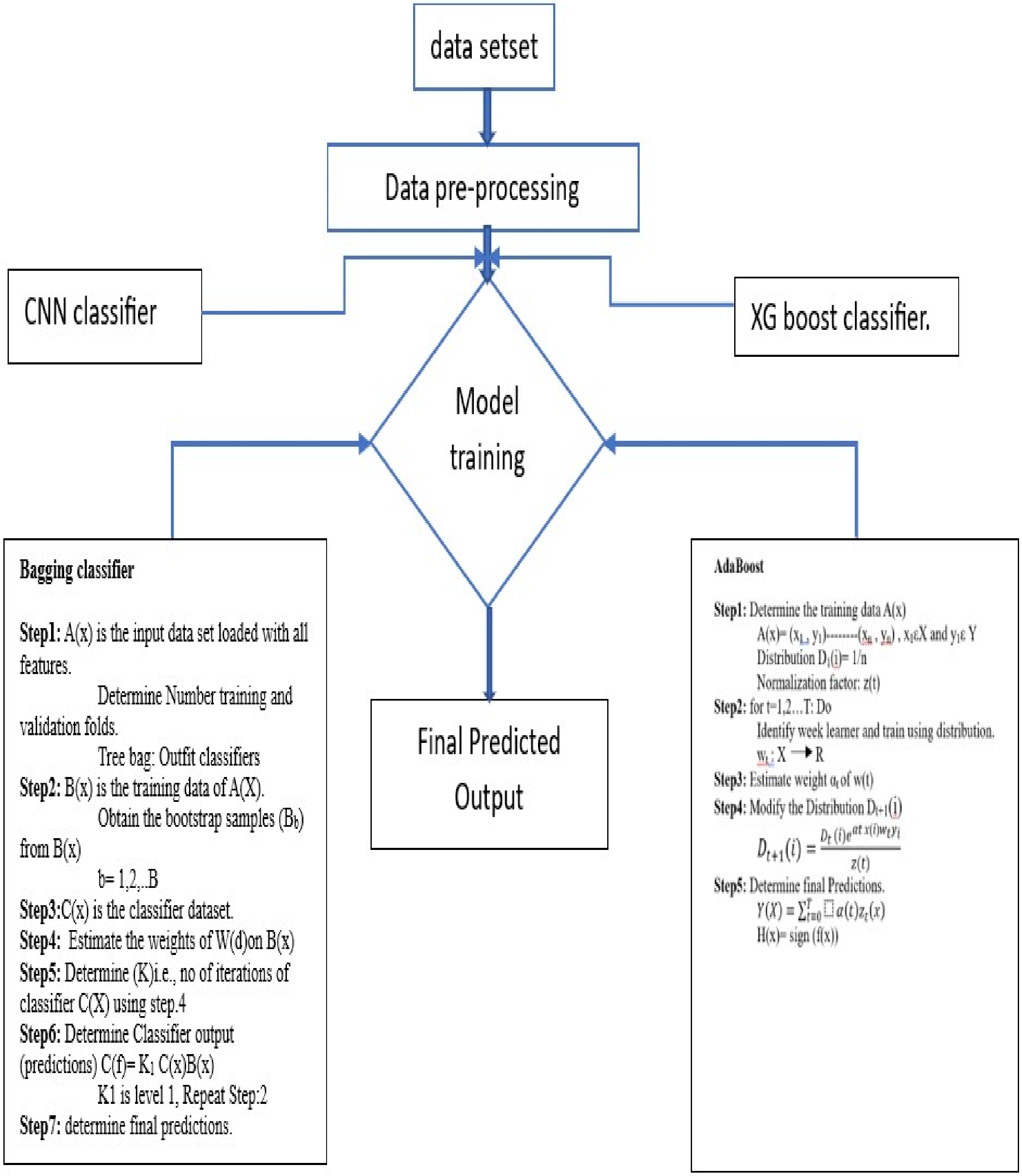
[9] M. A. Elaziz, K. M. Hosny, A. Salah, M. M. Darwish, S. Lu, and A. T. Sahlol, ''New machine learning method for image-based diagnosis of COVID-19'' PLoS ONE, vol. 15(6) Jun. 2020.

[10] A. Zargari Khuzani, M. Heidari, and S. A. Shariati, ''COVID-classifier: An automated machine learning model to assist in the diagnosis of COVID- 19 infection in chest X-radiation images,'' Sci. Rep., vol. 11[1]: May 2021.

[11] J. Rasheed, A. A. Hameed, C. Djeddi, A. Jamil, and F. Al-Turjman, ''A machine learning-based framework for diagnosis of COVID-19 from chest X-radiation images,'' Interdiscipl. Sci., Comput. Life Sci., vol. 13[1]: Jan. 2021, pp. 103–117.

[12] R. Mostafiz, M. S. Uddin, N.-A. Alam, M. M. Reza, and M. M. Rahman, ''COVID-19 detection in chest X-radiation through random forest classifier using a hybridization of deep CNN and DWT optimized features,'' J. King Saud Univ.-Comput. Inf. Sci., Dec. 2020.

[13] A. Akgundogdu, ''Detection of pneumonia in chest X-radiation images by using 2D discrete wavelet feature extraction with random forest,'' Int. J. Imag. Syst. Technol., vol. 31[1], Oct. 2020, pp. 82–93.



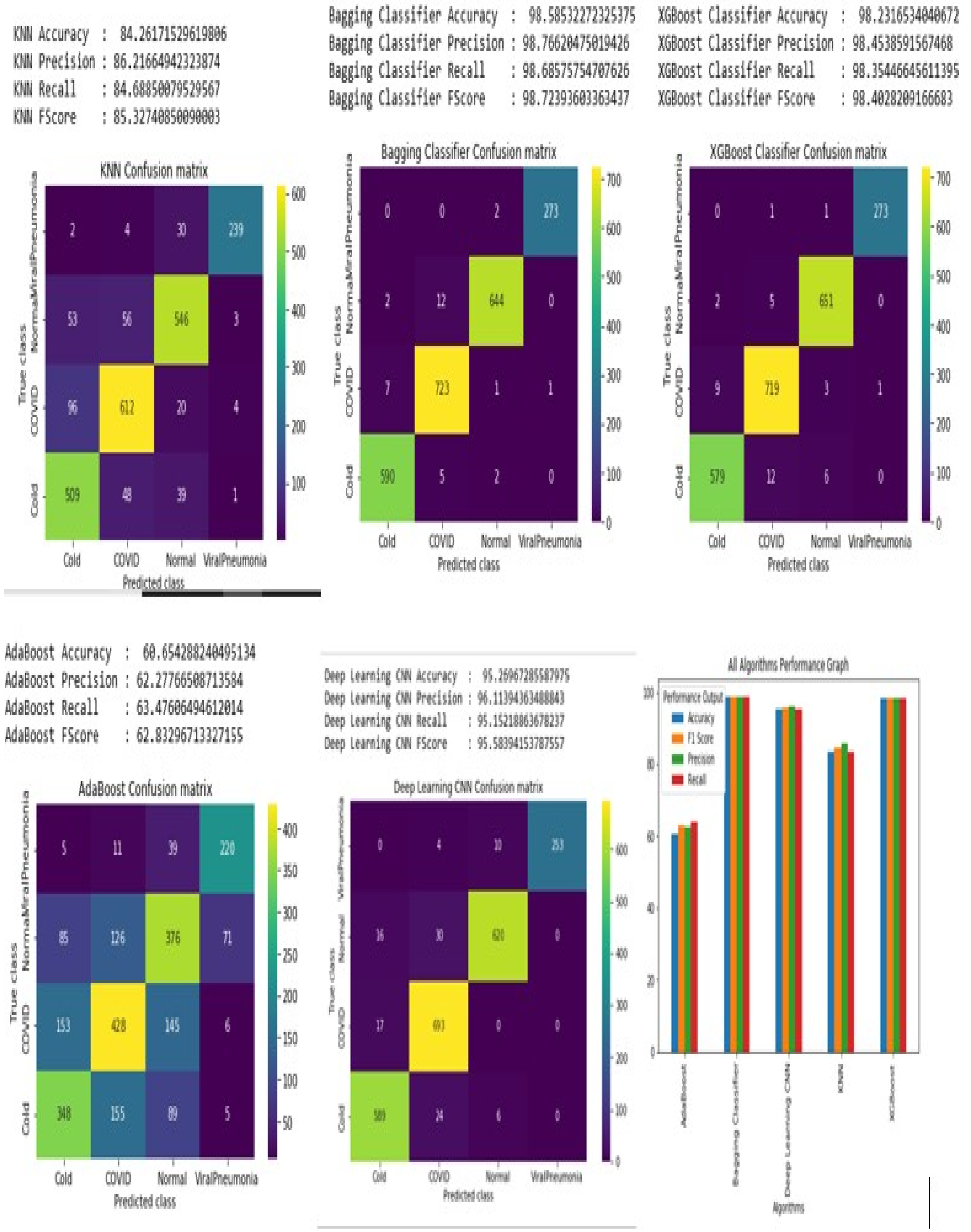
```
Bagging classifier

Step1: A(x) is the input data set loaded with all features.
        Determine Number training and validation folds.
        Tree bag: Outfit classifiers
Step2: B(x) is the training data of A(X).
        Obtain the bootstrap samples ($B_b$) from B(x)
        b= 1,2,..B
Step3:C(x) is the classifier dataset.
Step4:  Estimate the weights of W(d)on B(x)
Step5: Determine (K)i.e., no of iterations of classifier C(X) using step.4
Step6: Determine Classifier output (predictions) C(f)= $K_1$ C(x)B(x)
        K1 is level 1, Repeat Step:2
Step7: determine final predictions.
```

```
AdaBoost

Step1: Determine the training data A(x)
        A(x)= ($x_{1,}$ $y_1$)--------($x_n$ , $y_n$) , $x_1$εX and $y_1$ε Y
        Distribution $D_1(i)$= 1/n
        Normalization factor: z(t)
Step2: for t=1,2…T: Do
        Identify week learner and train using distribution.
        $w_t$ : X → R
Step3: Estimate weight $\alpha_t$ of w(t)
Step4: Modify the Distribution $D_{t+1}(i)$
        $D_{t+1}(i) = \frac{D_t(i)e^{\alpha t\, x(i) w_t y_i}}{z(t)}$
Step5: Determine final Predictions.
        $Y(X) = \sum_{t=0}^{T} \alpha(t) z_t(x)$
        H(x)= sign (f(x))
```

Fig 2.1 Flow Chart For Severity Classification Of Covid Using X- Radiation Images



```
KNN Accuracy  :  84.26171529619806
KNN Precision : 86.21664942323874
KNN Recall    : 84.68850079529567
KNN FScore    : 85.32740850090003
```

```
Bagging Classifier Accuracy  :  98.58532272325375
Bagging Classifier Precision : 98.76620475019426
Bagging Classifier Recall    : 98.68575754707626
Bagging Classifier FScore    : 98.72393603363437
```

```
XGBoost Classifier Accuracy  :  98.2316534040672
XGBoost Classifier Precision : 98.4538591567468
XGBoost Classifier Recall    : 98.35446645611395
XGBoost Classifier FScore    : 98.4028209166683
```

```
AdaBoost Accuracy  :  60.654288240495134
AdaBoost Precision : 62.27766508713584
AdaBoost Recall    : 63.47606494612014
AdaBoost FScore    : 62.83296713327155
```

```
Deep Learning CNN Accuracy  :  95.26967285587975
Deep Learning CNN Precision : 96.11394363488843
Deep Learning CNN Recall    : 95.15218863678237
Deep Learning CNN FScore    : 95.58394153787557
```

Fig 3.1 Confusion Matrix For Various Algorithms