Artificial & Computational Intelligence

The classical supervised machine learning-based approaches for predicting the COVID-19 infection: an exploratory analysis on classification performance

[1]Prabhat Kumar, [2]S Suresh
Department of Computer Science, Institute of Science
[1]prabhat.kumar13@bhu.ac.in, [2]suresh.selvam@bhu.ac.in
Banaras Hindu University, Varanasi – 221 005, India

Abstract

The World Health Organization (WHO) has declared the COVID-19 as pandemic throughout the world. The COVID-19 has affected a total of 213 countries and territories around the world and more than 28 million confirmed cases, half of which have been in the Americas. Over the 169 COVID-19 vaccine candidates are under the development phase. Moreover, 26 vaccine candidates are processed towards the human trial phase. For proper and appropriate treatments, we need clinical decision support systems embedded with supervised machine learning-based approaches. In this paper, we have designed the clinical prediction system using classical machine learning approaches and experiment on laboratory finding data. Our model predicts that which patients are likely infected with COVID-19 infection. The prediction performances of our models are evaluated based on the accuracy rate, F1-score, precision, recall, and kappa. The experimental dataset has been derived from Hospital Israelita Albert Einstein at Sao Paulo, Brazil, which included the records of 600 patients from 18 laboratory findings with 10% COVID-19 disease infected patients. Our prediction models have been validated with a train-test split approach, 10-folds approach, and AUC-ROC curve score. The experimental results showed that infected patients with COVID-19 disease are identified at an accuracy of 94.16% random forest classifier. We may request the medical experts for a recommendation of our prediction model used as clinical assistance for predicting the COVID-19 infections, implemented source code including the experimental dataset available at https://github.com/prabhat-parth/Classical_machine_learning_on_COVID-19.

Keywords: Artificial intelligence, COVID-19, Coronavirus, Clinical Support System, Machine Learning.

1. Introduction

In December 2019, the national authorities in China have identified the novel coronavirus in a cluster of hospital admitted patients, infected with pneumonia due to unknown reasons. Further, the World Health Organization (WHO) has received detailed information from the National Health Commission China, regarding the unknown outbreak i.e. associated with seafood in Wuhan city on 11th and 12th January 2020. The first lab-confirmed novel coronavirus (2019 - nCoV) from Wuhan province reported to the WHO by the Ministry of Public Health, Thailand on 13 January 2020 [1]. The WHO has declared the novel coronavirus (n-CoV) as public health emergency throughout the world on 30 January 2020 [2]. The initial clinical characteristics of CoV are as fever, dry cough, tiredness, headache, loss of taste or smell, difficulty breathing or shortness of breath, chest pain or pressure, and loss of speech or movement [3]. The earlier detection of CoV infection helps to prevention from infection to other human beings and communal transmission. The common transmission way of CoV is occurred via directly communicated with the infected person through coughing or sneezing and immediately touched or used the infected surfaces or objects. However, some scientific publications and news outlets have reported that the airborne transmission of CoV. In the context of CoV, the WHO continues to carefully observe the presence of CoV in airborne but continues to recommend airborne precautions in special circumstances [4]. The people carried out the appropriate precautions to slow down the infection rate of CoV by washing hand regularly using shop or sanitizers, social distancing (3-feets) in public or crowd place, avoid unnecessary touching the organs (eyes, nose, and mouth), and self-isolation while felling minor symptoms [5].

The WHO has hired the health professionals and scientists and working in collaboration to accelerate the research process for controlling the coronavirus pandemic [6]. The vaccines are the natural way to save human life. The world is suffering from the mid-era of the COVID-19 pandemic. Over the 169 COVID-19 vaccine candidates are currently in under development phase and out of these 26 candidates processed towards the human trials phase [7]. The clinical characteristics of COVID-19 varied regularly, health practitioners are facing difficulties while treating infected patients with limited hospitality resources. The Artificial Intelligence (AI) based clinical support system became essential equipment for analyzing the data, effective patterns learning and assist in decision-making processes. Over the last two decades, AI has achieved countless milestones in the field of the health-caring system. The AI-based system works very efficiently to accomplish the various clinical tasks such as biomedical information processing, radiology, pathology, ophthalmology, dermatology, etc [8]–[10]. However, the machine learning-based approaches have already contributed for early detect and predict the human health issues such as latent diseases [11], Health Monitoring System [12], Brain Stroke [13], early-stage disease risk prediction [14], Acute Kidney Injurious prediction [15], etc.

In this paper, we used the classical supervised machine learning approaches for predicting COVID-19 infected patients. We have applied and comparatively analyze the total of twelve supervised machine learning classifiers on laboratory datasets for finding the infected patients. The prediction performance of our classifiers is evaluated using accuracy rates, F1-score, precision, recall scores, and kappa. This paper has achieved the objectives, are summarized as follows:

1. Introducing the classical machine learning-based prediction system for predicting the COVID-19 infected patients using laboratory data rather than Chest X-ray or CT-Scan Images.
2. We have comparatively analyzed the performance of machine learning algorithms mention in this paper. Further, we also analyzed the experimental results with recently published research works.
3. This research work will motivate the researchers for architecting the more affective models and including additional parameters such as genders, travel details, previous medical treatment details, etc for enhancing the prediction outcomes of COVID-19 infection.
4. Finally, we have released the source code of our implemented model on GitHub including the experimental dataset and user manual.

This paper unfolded as section 2 contains the related works regarding the prediction of COVID-19 infections. Section 3 addresses the experimental dataset, working properties, and initial configuration of the methodology used in this paper. Section 4 details the evaluation metrics and discuss the experimental results using the split and test approach, 10-folds cross-validation approach, AUC-ROC curve score, and comprehensively analyzed with recently published works. Finally, section 5 concludes our research work and future scope.

2. Related works

The machine learning-based approaches used the laboratory data as an experimental dataset for learning efficient patterns, extracting the features, and predicting results. Previously, the machine learning-based clinical systems have widely used as assisting tools for handling the various health-related issues such as diagnosis of breast cancer [16], early detection of gastric cancer [17], brain pathology identification [18], computer-aided drug discovery [19], health care facilities management [20]. The author Sarwar et al. have used the hybrid machine learning-based ensemble technique for diagnosis diabetes that assured the accuracy rate of 98.60% [21]. The broad range of medical challenges like the COVID-19 pandemic is handling through the AI-based approaches. The perfect prediction of diagnosis helps to reduce the health practitioner's effort and provide cost-effective solutions [22]. The author Wang et al. developed the deep convolutional neural network-based COVID-Net for detecting the COVID-19 infected patients using chest X-Ray (CXR) images. The COVID-Net used the open-access benchmark dataset that contained the 13,975 CXR images across 13,870 patients. The experimental results were evaluated using Positive Predictive Value (PPV) and achieved high PPV for COVID-19 cases (98.9% PPV) [22]. [23] proposed a machine learning model for predicting COVID-19 affected patients using historical data and achieved an accuracy rate of 70% to 80%. The Jiang et al obtained the data from the institutional ethics board of Wenzhou Central Hospital and Cangnan People's Hospital in Wenzhou, China, and applied the Logistic Regression, KNN (k=5), Decision Tree based on Gain Ratio & Gini Index, Random Forests, and Support Vector Machine (SVM) classifiers. Among all these classifiers, SVM has achieved a high accuracy rate of 80% [23]. The Batista et al have used the neural networks, random forests, gradient boosting trees, logistic regression, and SVM classifiers on a clinical dataset obtained from Hospital Israelita Albert Einstein in São Paulo, Brazil [24]. The SVM and random forest classifiers have achieved the achieved the experimental results as AUC = 0.847, Sensitivity = 0.677, Specificity = 0.850, F1-score = 0.724, Brier score = 0.160, PPV = 0.778, and NPV = 0.773, respectively. The machine learning classifiers including Logistic Regression (LR), Neural Network (NN), Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (XGB) was used in [25]. The XGB was obtained the best classification performance result as Area under the ROC Curve (AUC) = 0.66, area under the precision-recall curve (AUPR) = 0.21, Sensitivity = 0.75, Specificity = 0.49.

The deep learning methods Artificial Neural Networks (ANN), Convolutional Neural Network (CNN), Long Short Term Memory (LSTM), Recurrent Neural Network (RNN), CNN + LSTM, and CNN + RNN) were contributed for predicting the COVID-19 infection in [26]. The experimental results were evaluated using the split-test and 10 fold cross-validation approaches. The hybrid deep neural networks (CNN + LSTM) have achieved the high prediction score, mentioned as accuracy = 86.66%, F1-score = 91.89%, precision = 86.75%, recall = 99.42%, and AUC = 62.50%.

3. Experimental Procedure

This section outlines the necessary background details regarding the experimental dataset and methodology used in this paper.

3.1. Experimental dataset description

The experimental dataset used in this paper, obtained from Hospital Israelita Albert Einstein at Sao Paulo Brazil and accessed through [25]. The samples of SARS-CoV-2 infected patients were collected in the early month of 2020 and available on [27]. This dataset contained a sample record of 5644 patients, collected from 111 different laboratories. In the dataset, a total of 10% of patients were infected, around 6.5% required hospitalization, and 2.5% required critical caring. However, the rest of 90% of patients reported as SARS-CoV-2 negative. The gender information of patients has no mentioned in the experimental dataset. This dataset contains the total ten-columns (Patient ID, Patient age quantile, SARS-Cov-2 exam result (negative/positive), Patient admitted to the regular ward (yes/no), Patient admitted to the semi-intensive unit (yes/no), Patient admitted to intensive care unit (yes/no), Hematocrit, Hemoglobin, Platelets, and Mean platelet volume). We applied the split-test approach and randomly divided the dataset into training (80%) and testing (20%) respectively for training and validating our prediction models. Furthermore, the 10-fold cross-validation has also been applied to balancing the evaluation rates of experimental models.

3.2. Methodology

This section served as a reasonable framework for classical supervised machine learning approaches to build the clinical predictive models for predicting the COVID-19 infection. We have used the logistic regression, K-Neighbors classifier, support vector classifier, decision tree classifier, random forest classifier, AdaBoost classifier, GaussianNB, linear discriminant analysis, quadratic discriminant analysis model as experimental methodology. The logistic regression classification approach is mostly used for binary classification problems and predicting the probability of a categorical dependent variable. This model is frequently used for risk predicting in chronic diseases [28], Trauma and Injury Severity Score (TRISS) [29], diabetes [30], heart disease identification [31], and breast cancer [32]. The K-Nearest Neighbors is a supervised machine learning algorithm and classifies using the similarity measure score based on hamming or standardized distance function. This method is widely used in voice disorder identification [33], Brain tumor classification [34], etc. The SVM extended its application areas in regression and outlier detection including the classification task. The best classification result achieved using a hyperplane that has the largest distance to the nearest training data point of any class. This popular method is widely accepted in skin disease detection [35], heart disease diagnosis [36], etc. The decision tree uses the concept of the divide-and-conquer approach and classifies the instances based on homogeneous properties [37]. The research areas of decision tree demonstrated in various health care applications such as Hepatocellular carcinoma (HCC) in clinical data [38], Opioid Use Disorder (OUD) understanding [39], etc. The random forest is applied in both classification and regression techniques. Usually, the set of decision trees are drawn from a randomly selected subset of the training set and then prediction result obtained by merging these trees. The application areas of random forest applied in various health caring applications such as identification of human vital functions related to the respiratory system [40], prognosis prediction [41], etc. The AdaBoost or Adaptive Boosting classifier follows an iterative approach that learns using incorrectly classified instances and fits the additional copies of the classifier for generating strong classifiers. The relevant application area of AdaBoost is the prediction of lung cancer [42], pinus diseased recognition [43], etc. The Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and use continuous data. This method is applied to solve various problems such as diabetes prediction [44], a prediction model for the detection of cardiac arrest [45], etc. The Linear Discriminant Analysis (LDA) classifier uses the prediction value, estimated by the probability new inputs set that belong to each class. The highest probability value designates the output and produces the prediction result. This classifier is mostly preferred in modeling the effects on human health [46], detection of epileptic seizures using EEG signals [47], etc. The quadratic discriminant analysis is used as both classifier and dimensionality reduction technique. This approach is a variation of the LDA classification technique that also allows for non-linear separation of data. This method is applied in various application areas such as epileptic seizure detection [48], pre-diagnosis of Huntington's disease [49], etc.

3.3. Experimental Models Configuration

In this section, we are addressing the configuration details of the machine learning methods used in this paper. The Scikit-learn is a machine learning library, used in the Python programming language and contained the various classifications, regression, and clustering algorithms. We have used the sklearn package for implementing the logistic regression, K-neighbors classifier, support vector classifier, decision tree classifier, random forest classifier, AdaBoost classifier, GaussianNB, linear discriminant analysis, and quadratic discriminant analysis. These methods are publically accessible with full documentation and can be imported from the sklearn library [50]. The initial values of parameters for each classifier and reference section contain the user guideline URL, mentioned in table 1.

4. Experimental Results and Discussion

In this section, we have discussed the evaluation metrics, experimental results for predicting the COVID-19 infection, and comparative analysis of the result with recently published research work. The experimental results are evaluated using the split-test approach, 10-folds cross-validation, and ROC-AUC score.

4.1. Evaluation Metrics

To evaluate the classification performance of models, we use accuracy (A), precision (P), recall (R), and F1-score (F1). For a binary classification problem, the confusion matrix holds the entries of True Positive (TP), False Positive (FP), True Negatives (TN), and False Negatives (FN). The diagonal entries in confusion metrics hold the correct prediction TP and incorrect prediction denoted by TN. The classifier estimates the wrong prediction, referred to as FP and FN. The accuracy evaluates the ratio between the number of correct predictions and the total number of input samples. The classification model was observed as correct when the number of predicted samples equals to the total number of samples. For the multiclass classification problem, the number of classes is initiated by the value of K. The precision equation measures the number of correct positive results divided by the number of positive results predicted by the classifiers. The recall value measures the number of correct positive results divided by the number of all relevant samples. The F1-score is primarily used to test the model's accuracy and the outcome score varies between 0 and 1 value. The high precision value and low recall value achieves a great accuracy rate but avoid the large number of samples that are difficult to classify.

Table 1 Configuration details of machine learning classifiers.

Classifier	Scikit-learn method	Parameters	References
Logistic Regression	sklearn.linear_model.LogisticRegression	C=1.0, max_iter=100, penalty='l2', solver='lbfgs', tol=0.0001	[51]
K-Neighbors	sklearn.neighbors.KNeighborsClassifier	leaf_size=30, metric='minkowski', n_neighbors=3, p=2	[52]
Support Vector	sklearn.svm.SVC	C=0.025, cache_size=200, degree=3, kernel='rbf', max_iter=-1	[53]
Decision Tree	sklearn.tree.DecisionTreeClassifier	criterion='gini', min_samples_leaf=1, presort='deprecated', splitter='best'	[54]
Random Forest	sklearn.ensemble.RandomForestClassifier	n_estimators=100, min_samples_split=2, min_samples_leaf=1	[55]
AdaBoost	sklearn.ensemble.AdaBoostClassifier	algorithm='SAMME.R', learning_rate=1.0, n_estimators=50	[56]
GaussianNB	sklearn.naive_bayes.GaussianNB	var_smoothing=1e-09	[57]
Linear Discriminant Analysis	sklearn.discriminant_analysis.LinearDiscriminantAnalysis	solver='svd', tol=0.0001	[58]
Quadratic Discriminant Analysis	sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis	reg_param=0.0, tol=0.0001	[59]

Table 2 illustrates the equation to measure the classification accuracy, precision, recall, and F1-score, extracted from the confusion matrix. The ROC-AUC (Receiver Operating Characteristics - Area under the Curve) is frequently used to evaluate the classification and prediction model's performance. This examines the model's ability while distinguishing between positive and negative classes. The higher AUC score indicates the better model at the prediction of patients with infected or not infected. The ROC-curve is plotted with False Positive Rate (FPR) on X-axis and True Positive Rate (TPR) on Y-axis (figure 1).

Table 2 equations for evaluating the classification performance

Evaluation Metric	Equation
Accuracy (A)	$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}$
Precision (Pk)	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$
Recall (Rk)	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
F1-score (F1k)	$2 \times (\frac{\mathrm{P_k} \times \mathrm{R_k}}{\mathrm{P_k}+\mathrm{R_k}})$

The FPR and TPR score are calculating using the expression (1) and (2), respectively. The idea behind the calculation of the AUC score (exists between 0 and 1) is the measurement of separability. The AUC score exists near the 1, which means has good separation capability. For the multi-class problem, we can plot N number of AUC ROC curves for multiple classes.

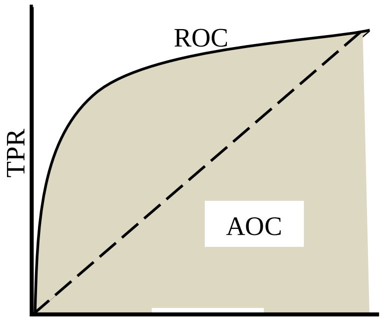

Fig. 1. AUC - ROC Curve

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \text{ ...... (1), TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \text{ ...... (2)}$$

Cohen’s kappa exhibits a more realistic view of the model’s performance when datasets are imbalanced. This is calculated with the following formula.

$$\text{Kappa (K)} = \frac{p_0 - p_e}{1 - p_e}$$

Where p_0 denotes the overall accuracy of the model and p_e measure of the agreement between the model predictions and the actual class values.

4.2. Split-Test Approach

We randomly split the clinical dataset of 80% for training and 20% for testing the model. Table 3 shows the experimental results of classical supervised machine learning models based on the split-test approach. Table 3 has shown the classification performance of logistic regression, K-Neighbors classifier, support vector classifier, decision tree classifier, random forest classifier, AdaBoost classifier, GaussianNB, linear discriminant analysis, quadratic discriminant analysis model, have reached at least 80.83% and above.

Table 3. Experimental result of machine learning classifiers using the split-test approach

Machine Learning Classification Models					
Classifiers	**Accuracy (%)**	**Precision**	**Recall**	**F1 score**	**Kappa**
Logistic Regression	0.9083	0.4545	0.5000	0.4761	0.4260
K-Neighbors	0.9166	0.5000	0.6000	0.5454	0.5000
Support Vector	0.9166	-	-	-	0.5000
Decision Tree	0.8083	0.2903	0.9000	0.4390	0.3581
Random Forest	**0.9416**	**0.6363**	**0.7000**	**0.6666**	**0.6347**
AdaBoost	0.8833	0.3750	0.6000	0.4615	0.4000
GaussianNB	0.9000	0.4375	0.7000	0.5384	0.4857
Linear Discriminant Analysis	0.9083	0.4444	0.4000	0.4210	0.3714
Quadratic Discriminant Analysis	0.9166	0.5000	0.7000	0.5833	0.5384

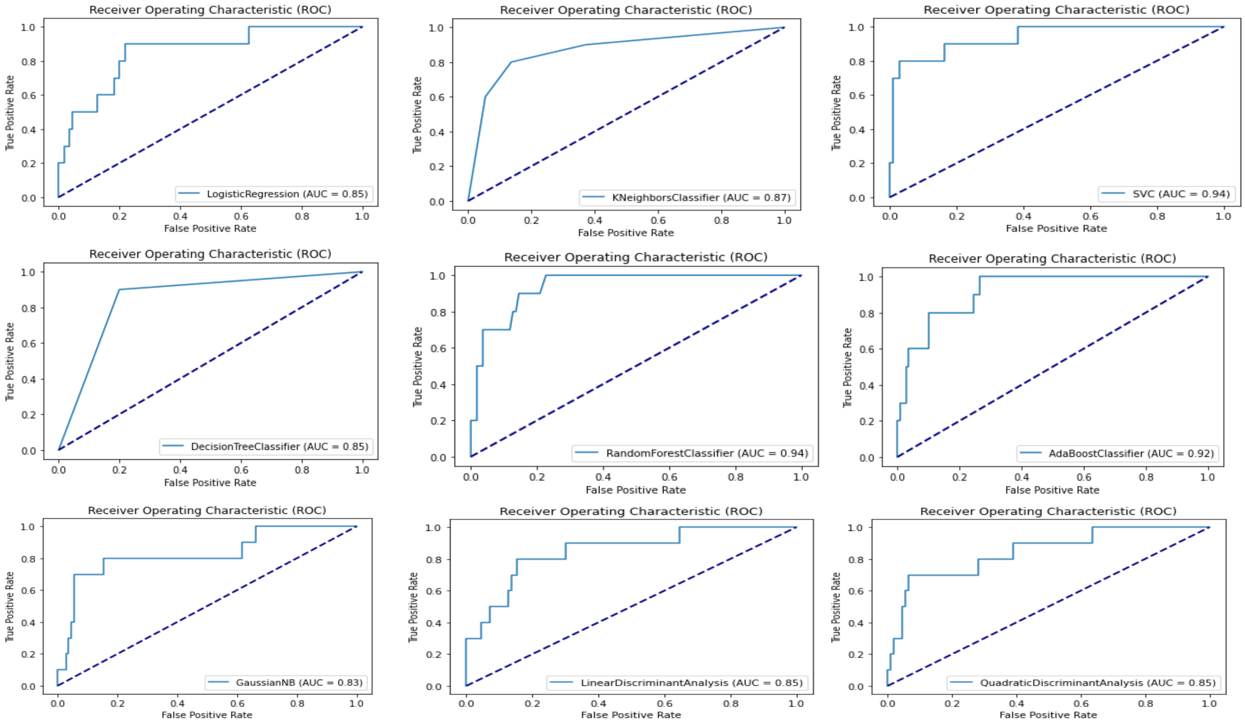

Fig. 2. Classification performance of classical machine learning approaches using the AUC score.

In terms of predictive performance, we observed that the overall best-identified model was random forest by the accuracy of 94.16%, the precision score of 63.63%, recall of 70.00%, f1-score of 66.66%, and kappa of 63.47%, respectively for predicting COVID-19 disease. This is the respective result achieved by random forest; we have generated the total 100 estimators and then choose the best classification tree. The decision tree achieved the worst classification performance on the clinical dataset with only an 80.83% accuracy rate.

4.3. 10 folds Cross-Validation Approach

In addition to the split-test approach, we tested the performance of classifiers based on 10 fold cross-validation approach. The experimental dataset is randomly partitioned into 10 equal sub-datasets in the 10 folds cross-validation technique. Out of these sub-datasets, the nine sub-datasets are retained as training the model, and the rest one sub-dataset assigned for validating the model. The cross-validation technique repeats the process 10 - times and each of the ten sub-samples is used exactly once as the validation dataset. The final result of 10-folds can be produced by aggregating the average results of each folding set. Table 4 shows the experimental results of machine learning models using 10 folds cross-validation approach.

In cases of the relatively small data sample, the 10 fold cross-validation approach is frequently used to measure the accurate classification performance of classifiers especially in health studies [24]. As can be seen in Table 4, the accuracy results of all classifier models have reached at least 81.87% and above.

Table 4. Experimental result of machine learning classifiers using a cross-validation approach

Machine Learning Classification Models	
Classifiers	**Accuracy (%)**
Logistic Regression	**0.8708**
K-Neighbors	0.8625
Support Vector	0.8541
Decision Tree	0.8187
Random Forest	0.8604
AdaBoost	0.8312
GaussianNB	0.8270
Linear Discriminant Analysis	0.8625
Quadratic Discriminant Analysis	0.8520

Table 5: Comparison of our experimental results with recently published research works.

Research Works	**Dataset Source**	**Techniques**	**Classification Methods**	**Accuracy**	**AUC**	**F1 – Score**
[23]	Wenzhou Central Hospital and Cangnan People's Hospital in Wenzhu, China	Machine Learning	Support Vector Machine	0.80	-	-
[24]	Hospital Israelita Albert Einstein at Sao Paulo, Brazil	Machine Learning	Support Vector Machine, Random Forest	-	0.87	0.72
[26]	Hospital Israelita Albert Einstein at Sao Paulo, Brazil	Machine Learning	eXtreme Gradient Boosting	-	0.66	-
[26]	Hospital Israelita Albert Einstein at Sao Paulo, Brazil	Deep Learning	CNNLSTM	92.30%	0.90	0.93

Our Work	**Hospital Israelita Albert Einstein at Sao Paulo, Brazil**	**Machine Learning**	**Random Forest**	**94.16%**	**0.94**	**0.66**

The clinical predictive performance of machine learning classification models with 10 fold cross-validation approach, the accuracy of 87.08% for the best-performing algorithm; this was a logistic regression model. The K-Neighbors and LDA classifiers followed the best classification performance with an accuracy rate of 86.25%.

4.4. Performance measurement using Area under the ROC Curve (AUC) score

The AUC score plays a vital role in the medical science research field. This can meaningfully interpret earlier disease identification and prediction from healthy subjects [60], [61]. The classification performance of models can be analyzed for predicting the best classes through the AUC score [62]. The AUC score of 0.5 means no discrimination, between 0.6 and 0.8 acceptable, excellent performance considered between 0.8 and 0.9, and more than 0.9 considered as outstanding. According to the AUC scores, all machine learning classifier models accepted as excellent performance can be used for clinical prediction of COVID-19. The clinical predictive performance of all machine learning classifier algorithms was better in comparison with the split-test strategy with an AUC of 0.94 by support vector and random forest classifiers (figure 2).

4.5. Comparative analysis of experimental results with recently published research works

As we observed in table 5, the authors of [23], [24], and [26] were used the machine learning technique i.e. support vector, random forest classifiers, and eXtreme Gradient Boosting respectively. However, [26] mentioned that the deep learning-based approach CNNLSTM has achieved the best classification performance. Apart from the classification performance of previous research works, we have used the classical supervised machine learning approaches that obtained the best experimental results as the accuracy of 94.16%, AUC score of 0.94, and F1-score of 0.66.

5. Conclusion and Future Works

This study was extremely dedicated to designing and developing the clinical support system for predicting the COVID-19 infection. We have carried out the classical supervised machine learning approaches (logistic regression, K-Neighbors classifier, support vector classifier, decision tree classifier, random forest classifier, AdaBoost classifier, GaussianNB, linear discriminant analysis, quadratic discriminant analysis) to accomplish the clinical prediction task. The experimental data preprocessed using standardization and then fed to the experimental models. Further, the classification results were measured based on accuracy rate, F1-score, precision, recall, and kappa. To validate our model, we have used the split-test approach, 10-folds cross-validation approach, and AUC-ROC curve score. In the split-test approach, the best result achieved using random forest with an accuracy of 94.16%, a precision score of 63.63%, a recall of 70.00%, f1-score of 66.66%, and kappa of 63.47%, respectively for predicting COVID-19 disease. However, the logistic regression model has achieved an accuracy of 87.08% for the best-performing algorithm in 10 fold cross-validation approach. Our research work also carried out a major limitation with a small and imbalancing experimental dataset. The performance of our prediction model can be enhanced by increasing the size of the dataset either combining the data from different laboratories or data augmentation techniques. The future work can be extended from our paper, by adding additional parameters such as genders, travel details, previous medical treatment details, etc for enhancing the prediction rate. Based on our experimental results, we conclude that the clinical system should explore the use of artificial intelligence for prioritizing the models as clinical support systems while reducing the personalizing infection risks.

References

[1] B. Zhou, J. She, Y. Wang, and X. Ma, "Novel Coronavirus (2019-nCoV) SITUATION REPORT - 1 21 JANUARY 2020," no. JANUARY, pp. 1–14, 2020.

[2] "IHR Emergency Committee on Novel Coronavirus (2019-nCoV)." [Online]. Available: https://www.who.int/dg/speeches/detail/who-director-general-s-statement-on-ihr-emergency-committee-on-novel-coronavirus-(2019-ncov). [Accessed: 11-Aug-2020].

[3] W. Guan *et al.*, "Clinical characteristics of coronavirus disease 2019 in China," *N. Engl. J. Med.*, vol. 382, no. 18, pp. 1708–1720, 2020, doi: 10.1056/NEJMoa2002032.

[4] "Modes of transmission of virus causing COVID-19: implications for IPC precaution recommendations." [Online]. Available: https://www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations. [Accessed: 13-Aug-2020].

[5] "Advice for the public." [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public. [Accessed: 13-Aug-2020].

[6] "Potent antibodies found in people recovered from COVID-19 | National Institutes of Health (NIH)." [Online]. Available: https://www.nih.gov/news-events/nih-research-matters/potent-antibodies-found-people-recovered-covid-19. [Accessed: 13-Aug-2020].

[7] "COVID-19 vaccine." [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines. [Accessed: 26-Sep-2020].

[8] S. Kulkarni, N. Seneviratne, M. S. Baig, and A. H. A. Khan, "Artificial Intelligence in Medicine: Where Are We Now?," *Acad. Radiol.*, vol. 27, no. 1, pp. 62–70, 2020, doi: 10.1016/j.acra.2019.10.001.

[9] J. P. Rowe and J. C. Lester, "Artificial Intelligence for Personalized Preventive Adolescent Healthcare," *J. Adolesc. Heal.*, vol. 67, no. 2, pp. S52–S58, 2020, doi: 10.1016/j.jadohealth.2020.02.021.

[10] G. Rong, A. Mendez, E. Bou Assi, B. Zhao, and M. Sawan, "Artificial Intelligence in Healthcare: Review and Prediction Case Studies," *Engineering*, vol. 6, no. 3, pp. 291–301, 2020, doi: 10.1016/j.eng.2019.08.015.

[11] Y. Wang *et al.*, "Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records," *J. Biomed. Inform.*, vol. 102, no. May 2019, p. 103364, 2020, doi: 10.1016/j.jbi.2019.103364.

[12] K. Gnana Sheela and A. Rose Varghese, "Machine Learning based Health Monitoring System," *Mater. Today Proc.*, vol. 24, pp. 1788–1794, 2019, doi: 10.1016/j.matpr.2020.03.603.

[13] M. S. Sirsat, E. Fermé, and J. Câmara, "Machine Learning for Brain Stroke: A Review," *J. Stroke Cerebrovasc. Dis.*, vol. 29, no. 00, 2020, doi: 10.1016/j.jstrokecerebrovasdis.2020.105162.

[14] M. A. Hossain, R. Ferdousi, and M. Alhamid, "Knowledge-driven machine learning based framework for early-stage disease risk prediction in edge environment," *J. Parallel Distrib. Comput.*, vol. 146, pp. 25–34, 2020, doi: 10.1016/j.jpdc.2020.07.003.

[15] X. Yang *et al.*, "Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study," *Lancet Respir. Med.*, vol. 8, no. 5, pp. 475–481, 2020, doi: 10.1016/S2213-2600(20)30079-5.

[16] M. M. Rahman, Y. Ghasemi, E. Suley, Y. Zhou, S. Wang, and J. Rogers, "Machine Learning Based Computer Aided Diagnosis of Breast Cancer Utilizing Anthropometric and Clinical Features," *Irbm*, 2020, doi: 10.1016/j.irbm.2020.05.005.

[17] Y. Horiuchi *et al.*, "Performance of a computer-aided diagnosis system in diagnosing early gastric cancer using magnifying endoscopy videos with narrow-band imaging (with videos)," *Gastrointest. Endosc.*, 2020, doi: 10.1016/j.gie.2020.04.079.

[18] A. Gudigar, U. Raghavendra, A. Hegde, M. Kalyani, E. J. Ciaccio, and U. Rajendra Acharya, "Brain pathology identification using computer aided diagnostic tool: A systematic review," *Comput. Methods Programs Biomed.*, vol. 187, p. 105205, 2020, doi: 10.1016/j.cmpb.2019.105205.

[19] I. E. Ebhohimen, L. Edemhanria, S. Awojide, O. H. Onyijen, and G. Anywar, *Advances in computer-aided drug discovery*. Elsevier Inc., 2020.

[20] E. Iadanza and A. Luschi, *Computer-aided facilities management in health care*, Second Edi. Elsevier Inc., 2019.

[21] A. Sarwar, M. Ali, J. Manhas, and V. Sharma, "Diagnosis of diabetes type-II using hybrid machine learning based ensemble model," *Int. J. Inf. Technol.*, vol. 12, no. 2, pp. 419–428, 2020, doi: 10.1007/s41870-018-0270-5.

[22] J. Bullock, A. Luccioni, K. H. Pham, C. S. N. Lam, and M. Luengo-Oroz, "Mapping the Landscape of Artificial Intelligence Applications against COVID-19," pp. 1–32, 2020.

[23] X. Jiang *et al.*, "Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity," *Comput. Mater. Contin.*, vol. 63, no. 1, pp. 537–551, 2020, doi: 10.32604/cmc.2020.010691.

[24] A. F. de M. Batista, J. L. Miraglia, T. H. R. Donato, and A. D. P. Chiavegatto Filho, "COVID-19 diagnosis prediction in emergency care patients: a machine learning approach," 2020, doi: 10.1101/2020.04.04.20052092.

[25] P. Schwab, A. D. Schütte, B. Dietz, and S. Bauer, "predCOVID-19: A Systematic Study of Clinical Predictive Models for Coronavirus Disease 2019," pp. 1–8, 2020.

[26] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection," *Chaos, Solitons and Fractals*, vol. 140, p. 110120, 2020, doi: 10.1016/j.chaos.2020.110120.

[27] "Diagnosis of COVID-19 and its clinical spectrum | Kaggle." [Online]. Available: https://www.kaggle.com/einsteindata4u/covid19. [Accessed: 18-Aug-2020].

[28] S. Nusinovici *et al.*, "Logistic regression was as good as machine learning for predicting major chronic diseases," *J. Clin. Epidemiol.*, vol. 122, pp. 56–69, 2020, doi: 10.1016/j.jclinepi.2020.03.002.

[29] P. J. Schluter, "The Trauma and Injury Severity Score (TRISS) revised," *Injury*, vol. 42, no. 1, pp. 90–96, 2011, doi: 10.1016/j.injury.2010.08.040.

[30] E. B. Aboagye-Mensah *et al.*, "The association of ideal cardiovascular health with self-reported health, diabetes, and adiposity in African American males," *Prev. Med. Reports*, vol. 19, no. December 2019, p. 101151, 2020, doi: 10.1016/j.pmedr.2020.101151.

[31] H. Ahmed, E. M. G. Younis, A. Hendawi, and A. A. Ali, "Heart disease identification from patients' social posts, machine learning solution on Spark," *Futur. Gener. Comput. Syst.*, vol. 111, pp. 714–722, 2019, doi: 10.1016/j.future.2019.09.056.

[32] F. Morais-Rodrigues *et al.*, "Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression," *Gene*, vol. 726, no. November 2019, 2020, doi: 10.1016/j.gene.2019.144168.

[33] L. Chen, C. Wang, J. Chen, Z. Xiang, and X. Hu, "Voice Disorder Identification by using Hilbert-Huang Transform (HHT) and K Nearest Neighbor (KNN)," *J. Voice*, 2020, doi: 10.1016/j.jvoice.2020.03.009.

[34] K. Kaplan, Y. Kaya, M. Kuncan, and H. M. Ertunç, "Brain tumor classification using modified local binary patterns (LBP) feature extraction methods," *Med. Hypotheses*, vol. 139, no. February, 2020, doi: 10.1016/j.mehy.2020.109696.

[35] V. R. Balaji, S. T. Suganthi, R. Rajadevi, V. Krishna Kumar, B. Saravana Balaji, and S. Pandiyan, "Skin disease detection and segmentation using dynamic graph cut algorithm and classification through Naive Bayes classifier," *Meas. J. Int. Meas. Confed.*, vol. 163, p. 107922, 2020, doi: 10.1016/j.measurement.2020.107922.

[36] S. M. S. Shah, F. A. Shah, S. A. Hussain, and S. Batool, "Support Vector Machines-based Heart Disease Diagnosis using Feature Subset, Wrapping Selection and Extraction Methods," *Comput. Electr. Eng.*, vol. 84, p. 106628, 2020, doi: 10.1016/j.compeleceng.2020.106628.

[37] R. Panigrahi and S. Borah, "Classification and Analysis of Facebook Metrics Dataset Using Supervised Classifiers," in *Social Network Analytics*, Elsevier, 2019, pp. 1–19.

[38] P. Radha and R. Divya, "An Efficient Detection of HCC-recurrence in Clinical Data Processing using Boosted Decision Tree Classifier," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 193–204, 2020, doi: 10.1016/j.procs.2020.03.196.

[39] A. S. Wadekar, "Understanding Opioid Use Disorder (OUD) using tree-based classifiers," *Drug Alcohol Depend.*, vol. 208, no. November 2019, p. 107839, 2020, doi: 10.1016/j.drugalcdep.2020.107839.

[40] K. Proniewska, A. Pregowska, and K. P. Malinowski, "Identification of Human Vital Functions Directly Relevant to the Respiratory System Based on the Cardiac and Acoustic Parameters and Random Forest," *Irbm*, vol. 1, pp. 5–10, 2020, doi: 10.1016/j.irbm.2020.02.006.

[41] J. Li *et al.*, "A multicenter random forest model for effective prognosis prediction in collaborative clinical research network," *Artif. Intell. Med.*, vol. 103, no. September 2019, p. 101814, 2020, doi: 10.1016/j.artmed.2020.101814.

[42] C. Tan, H. Chen, and C. Xia, "Early prediction of lung cancer based on the combination of trace element analysis in urine and an Adaboost algorithm," *J. Pharm. Biomed. Anal.*, vol. 49, no. 3, pp. 746–752, 2009, doi: 10.1016/j.jpba.2008.12.010.

[43] G. Hu, C. Yin, M. Wan, Y. Zhang, and Y. Fang, "Recognition of diseased Pinus trees in UAV images using deep learning and AdaBoost classifier," *Biosyst. Eng.*, vol. 194, pp. 138–151, 2020, doi: 10.1016/j.biosystemseng.2020.03.021.

[44] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," *Procedia Comput. Sci.*, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.

[45] S. Layeghian Javan, M. M. Sepehri, M. Layeghian Javan, and T. Khatibi, "An intelligent warning model for early prediction of cardiac arrest in sepsis patients," *Comput. Methods Programs Biomed.*, vol. 178, pp. 47–58, 2019, doi: 10.1016/j.cmpb.2019.06.010.

[46] A. P. Worth and M. T. D. Cronin, "The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects," *J. Mol. Struct. THEOCHEM*, vol. 622, no. 1–2, pp. 97–111, 2003, doi: 10.1016/S0166-1280(02)00622-X.

[47] L. Chanel, D. Nkengfack, D. Tchiotsop, R. Atangana, and D. Wolf, "EEG signals analysis for epileptic seizures detection using polynomial transforms , linear discriminant analysis and support vector machines," vol. 62, no. August, 2020, doi: 10.1016/j.bspc.2020.102141.

[48] M. F. Bari and S. Anowarul Fattah, "Epileptic seizure detection in EEG signals using normalized IMFs in CEEMDAN domain and quadratic discriminant classifier," *Biomed. Signal Process. Control*, vol. 58, p. 101833, 2020, doi: 10.1016/j.bspc.2019.101833.

[49] N. Georgiou-Karistianis *et al.*, "Automated differentiation of pre-diagnosis Huntington's disease from healthy control individuals based on quadratic discriminant analysis of the basal ganglia: The IMAGE-HD study," *Neurobiol. Dis.*, vol. 51, pp. 82–92, 2013, doi: 10.1016/j.nbd.2012.10.001.

[50] "Supervised learning — scikit-learn 0.23.2 documentation." [Online]. Available: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning. [Accessed: 31-Aug-2020].

[51] "sklearn.linear_model.LogisticRegression — scikit-learn 0.23.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. [Accessed: 31-Aug-2020].

[52] "sklearn.neighbors.KNeighborsClassifier — scikit-learn 0.23.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html. [Accessed: 31-Aug-2020].

[53] "sklearn.svm.SVC — scikit-learn 0.23.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html. [Accessed: 31-Aug-2020].

[54] "sklearn.tree.DecisionTreeClassifier — scikit-learn 0.23.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html. [Accessed: 31-Aug-2020].

[55] "sklearn.ensemble.RandomForestClassifier — scikit-learn 0.23.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Accessed: 31-Aug-2020].

[56] "sklearn.ensemble.AdaBoostClassifier — scikit-learn 0.23.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html. [Accessed: 31-Aug-2020].

[57] "sklearn.naive_bayes.GaussianNB — scikit-learn 0.23.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html. [Accessed: 31-Aug-2020].

[58] "sklearn.discriminant_analysis.LinearDiscriminantAnalysis — scikit-learn 0.23.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html. [Accessed: 31-Aug-2020].

[59] "sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis — scikit-learn 0.23.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis.html. [Accessed: 31-Aug-2020].

[60] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," 2013.

[61] A. N. Kamarudin, T. Cox, and R. Kolamunnage-Dona, "Time-dependent ROC curve analysis in medical research: Current methods and applications," *BMC Med. Res. Methodol.*, vol. 17, no. 1, pp. 1–19, 2017, doi: 10.1186/s12874-017-0332-6.

[62] J. N. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment," *J. Thorac. Oncol.*, vol. 5, no. 9, pp. 1315–1316, 2010, doi: 10.1097/JTO.0b013e3181ec173d.