COVID-19 Infection Prediction Using Efficient Machine Learning Techniques Based on Clinical Data

Bilal Abdualgalil and Sajimon Abraham
School of Computer Science, Mahatma Gandhi University, Kerala, India
Email: bsaa85@gmail.com, sajimabraham@rediffmail.com

Waleed M. Ismael
Hohai University, Chanzhou Campus, Jiangsu, China
Email: Waleed.m@hhu.edu.cn

***Abstract*—COVID-19 (coronavirus disease) has spread worldwide and has become a pandemic, which causes by the SARS-CoV2 virus. Because the number of cases increases daily, interpreting the laboratory findings takes time, resulting in limitations of findings. Because of these limitations, the need for a clinical decision-making system with predictive algorithms has arisen. By identifying diseases, predictive algorithms would be able to reduce the strain on healthcare systems. In this work, we developed clinical predictive models using machine learning techniques with the help SMOTE+ENN Hybrid technique and laboratory data to develop models that can accurately predict which patients will receive COVID-19. To evaluate our prediction models in this work, precision, F1-score, recall AUC, and Accuracy evaluation metrics are employed. From 600 patients and 10 laboratory findings, the different models are tested and validated with 10-fold cross-validation and holdout cross-validation approaches. The experimental results show that our predictive models can correctly identify patients with COVID-19 with an accuracy of 98.28%, an F1-score of 98.27%, a precision of 98.23%, a recall of 98.32%, and an AUC of 98.32% in the holdout cross-validation approach, and an accuracy of 97.42%, and F1-score of 97.82%, a precision of 97.63%, a recall of 98.05%, and an AUC of 92.66% in 10-fold cross-validation approach. The results of the experiments showed that all machine learning models in the holdout cross-validation approach outperformed the 10-fold cross-validation approach. Finally, to help medical experts with accurately prioritizing resources, predictive models based on laboratory findings have been discovered that can assist in predicting COVID-19 infection and assisting medical professionals to identify which medical resources are most valuable.**

***Index Terms*—artificial intelligence, SARS-CoV2, machine learning, COVID-19, SMOTE+ENN, Imbalanced dataset**

I. Introduction

It was found in Wuhan, China, on December 31, 2019, that the virus SARS-CoV2, which causes coronavirus diseases (COVID-19), had since spread worldwide [1]. A COVID-19 disease outbreak has been declared a pandemic by the World Health Organization (WHO), insisting on providing the required tools, mechanics, and resources to assist in identifying those who are most at risk of infection and death. A response to COVID-19 can vary greatly from one person to another. More than 80% of those infected with the virus develop mild to moderate symptoms and are well enough to return to normal activity before going to the hospital [2], [3]. Furthermost, patients begin with minor symptoms, such as a dry cough and fatigue. In addition to those symptoms already mentioned, patients may also experience paralysis and/or the inability to breathe [4]. Chronic disease patients are older, male, and at an increased risk with each decade over 50 years old [3]. People who suffer from medical issues, such as diabetes, cancer, cardiovascular disease, and chronic respiratory disorders, are the most susceptible to infection with COVID-19 disease [2]. Although no particular COVID-19 treatments exist, much ongoing clinical research is evaluating potential treatments. People can avoid infection by washing their hands, staying at home, covering their mouth and nose when coughing or sneezing, and not smoking. These measures do not treat the disease, but they can help prevent it and slow its spread.

Different laboratory studies reported different results at the beginning of the COVID-19 outbreak [5], [6].

Most cases are mild, and patients' clinical results vary significantly [7], [8]. As a result, identifying risk groups solely based on some characteristics, such as gender and age, would be difficult. It is also important to predict who is at a higher risk of mortality and a higher chance of getting ill. Clinical outcomes vary greatly between patients. Because of all of these limits, such decisions must be made by AI-assisted systems. In healthcare systems, clinical decision support using AI is being deployed [9]-[11].

In medical data such as epilepsy [12], [13], neuromuscular diseases [14], [15], heart rhythms [16], [17], etc., machine learning classifiers are very effective in interpreting such diseases. Furthermore, studies have

Manuscript received November 5, 2021; revised April 17, 2022; accepted April 18, 2022.

doi: 10.12720/jait.13.5.530-538

proven that machine learning techniques are also effective in predicting clinical findings such as biomedical studies [18], [19], viral disease [20], and cancer [21]. These techniques work well for predicting COVID-19 infection as well however, it a challenge remains in using machine learning techniques to predict COVID-19 infection based on clinical data which are mostly imbalanced and select of important findings, all of these factors affect the accuracy of models in predicting COVID-19 infection.

In this work, we applied ten machine learning models to build a system for predicting COVID-19 infection. The performance of the adopted models is measured using the Accuracy, Precision, Recall, F1 score, and AUC evaluation metrics. To the best of our knowledge, no work has utilized machine learning models, select the important features, and the SMOTE+ENN hybrid technique to predict COVID-19 infection based on laboratory findings. This was accomplished in this work.

Researchers and scientists may be encouraged by this work to validate models with a variety of laboratory data.

This paper introduces two-fold contributions, as follows:

- Developing a prediction approach for COVID-19 disease based on clinical datasets containing laboratory findings using machine learning models and the SMOTE+ENN hybrid technique,
- Ensuring that the prediction approach is accurate and effective for this new pneumonia.

This work is organized as follows. Section II explained related work. Section III describes the proposed work. Section IV provides the experimental results and discussion. Finally, Section V includes the conclusion and potential future work.

II. Related Work

Predicting clinical tasks is critical for healthcare systems. Heart failure risk [22], pneumonia mortality [23], [24], and critical care mortality risk [25]-[27] are just a few of the areas where used clinical predictive models with computer-aided. Medical experts are unable to make better decisions about clinical findings when using these systems. COVID-19 clinical predictive model was developed using recent methodological advances in this work. In the literature, there are few similar studies on COVID-19 clinical prediction.

The authors of [28] applied machine learning algorithms to predict coronavirus clinical severity. Data was gathered from Wenzhou Main Hospital and Cangnan People's Hospital in Wenzhou, China, but it is not publicly available because the information is confidential. The authors were considered eleven clinical features and six different classifiers: Logistic Regression (LR), k-Nearest Neighbour (KNN), two different Decision Trees (TD), Random Forests (RF), and support vector machines (SVM). The performance of these classifiers has been evaluated using an accuracy score only. The SVM classifier achieved the best accuracy with 80%.

According to a study [29], machine learning classifiers were applied to predict the diagnosis of COVID-19. Clinical data was provided by the Israelita Albert Einstein Hospital in Sao Paulo, Brazil, which included 18 clinical findings.

Five different classifiers used in the study included SVM, random forests, neural networks, logistic regression, and gradient boosted trees. Classifiers in the study were evaluated using AUC, sensitivity, specificity, F1-score, Brier score, positive predictive value, and negative predictive value. Both SVM and random forest classifiers achieved the best AUC values with 0.847.

Another study in [30] proposed a clinical prediction model for COVID-19 disease. Similar to [29], data were collected at the Hospital Israelita Albert Einstein in Sao Paulo, Brazil. The authors used different machine learning classifiers such as RF, NN (Neural Network), LR, SVM, and XGB (Gradient Boosting). The performance of classifiers was evaluated using determining sensitivity, specificity, and AUC scores. The XGB classifier achieved the best performance, with a 66 % AUC score.

Another study [31] applied six different deep learning classifications on patient laboratory findings like ANN, CNN, LSTM, RNN, CNNLSTM, and CNNRNN with considered 18 clinical findings. Clinical data used in the study was obtained from the Israelita Albert Einstein Hospital in Sao Paulo, Brazil. The classifiers were evaluated using the following evaluation metrics: accuracy, F1-score, precision, recall, and AUC scores, where the LSTM model achieved the best accuracy with 86.66%, F1-score with 91.89%, the precision with 86.75%, recall with 99.42%, and AUC with 62.50%.

In reference to the Related Works Section, all previous studies on COVID-19 infection prediction did not use techniques to rebalance data to obtain the best models for COVID-19 infection prediction. This is due to the fact that healthcare data are inherently imbalanced, in addition to containing unimportant features (symptoms), which are processed in the preprocessed stage. These issues are attributed to important features' selection and data imbalance, making most infection COVID-19 prediction models biased toward a majority class. Therefore, there are important research gaps in the COVID-19 infection prediction system, namely, important feature selections, and data rebalancing. By using the COVID-19 dataset, this work fills these gaps, which were accomplished through the use of machine learning techniques. Based on clinical data and findings, our proposed classifier is effective in COVID-19 infection prediction.

III. Proposed Work

A Flowchart proposed in this work is illustrated by Fig. 1 to predict COVID-19 infection. The goal was to find the most effective model for predicting COVID-19 infection. The framework utilized in this work included data description and pre-processing steps that were applied to the COVID-19 clinical data and dataset balancing stage. Moreover, machine learning techniques were used to predict COVID-19 infection and the information and parameters about the developed machine learning techniques are given. Finally, the performance of all machine learning models used in this work was evaluated using different evaluation metrics (accuracy, AUC, precision, recall, and F1-scores).

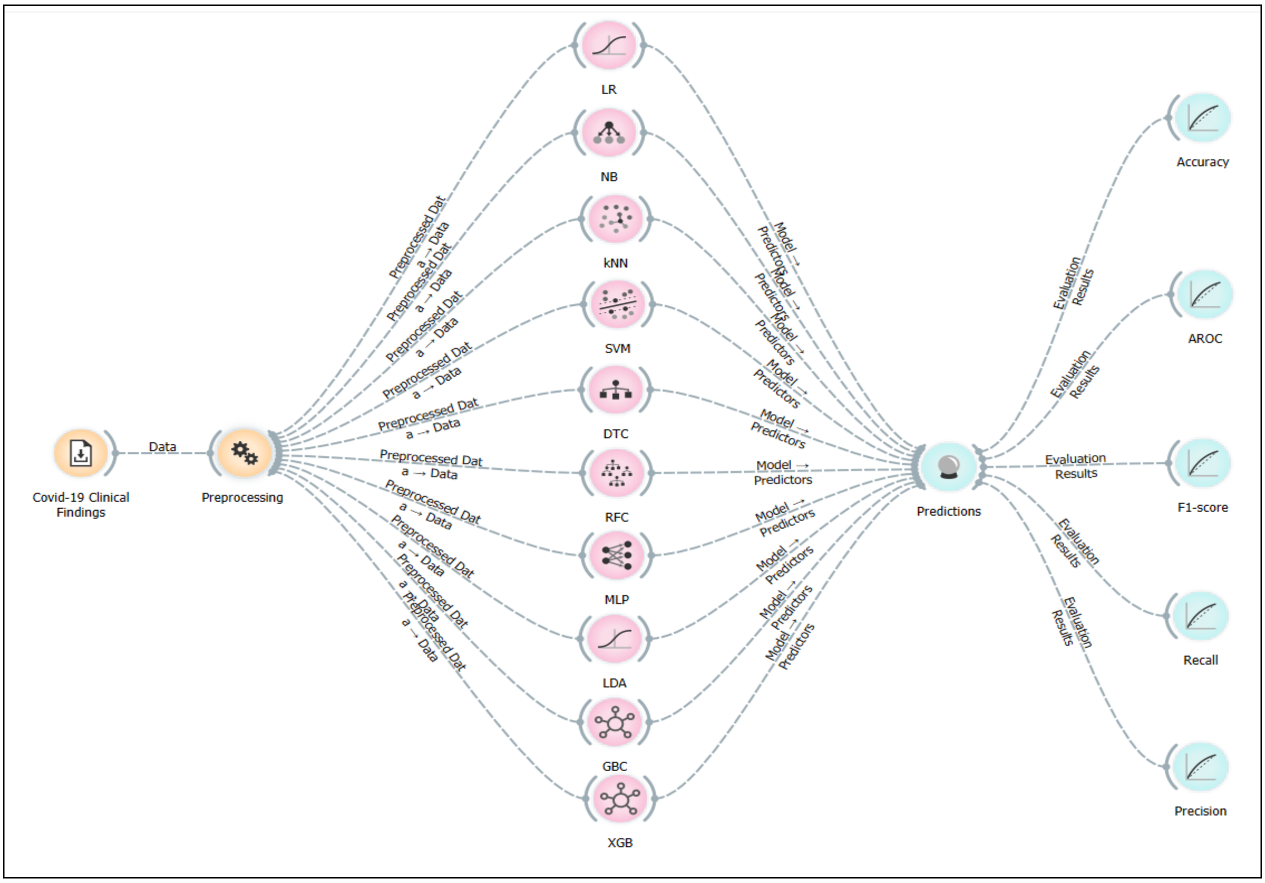
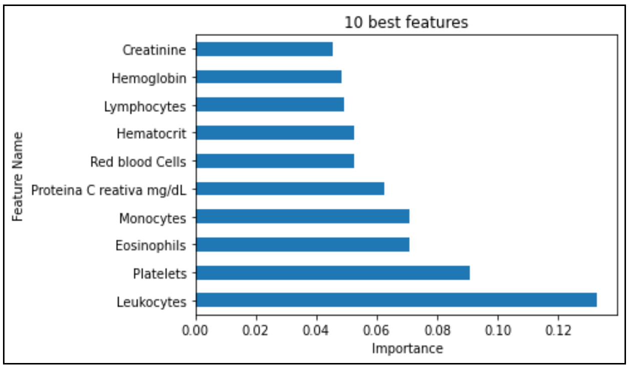


Figure 1. Flowchart of the proposed work. The orange icon represents the dataset and pre-processing, both of which are laboratory findings in this work. The pink ones are machine learning models such as LR, NB, KNN, SVM, DTC, RFC, MLP, LDA, GBC, and XGB. All of these models were used to predict No findings and COVID-19. The light blue icon represents machine learning models' predictors, and next, the Accuracy, Precision, Recall, AUC, and F1-Scores were used to evaluate results. (For color interpretation, the reader is directed to the web version of this page.)

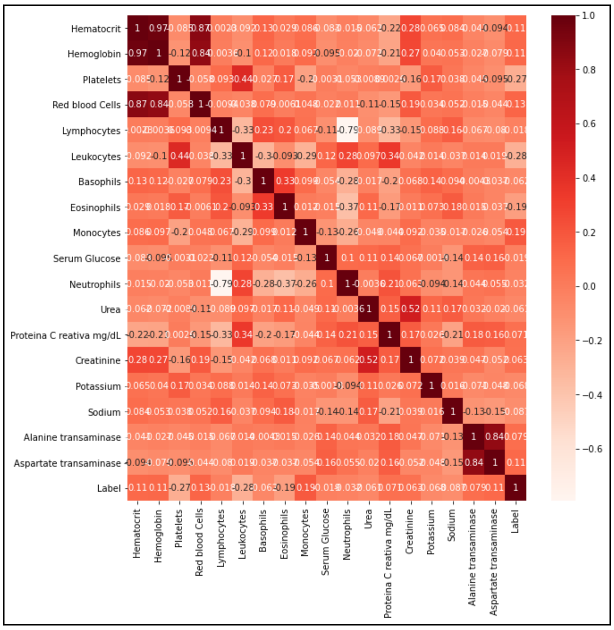
A. Data Description

The dataset, which was obtained through [31], contains the laboratory findings of patients seen at the Hospital Israelita Albert Einstein in Sao Paulo, Brazil, and is available in https://github.com/burakalakuss/COVID-19-Clinical/tree/master/Clinical%20Data. In the early months of 2020, samples from patients were collected to detect SARS-CoV2. According to [28]-[31], there is no gender information in the dataset, as well as 18 laboratory findings from 600 patients, including 520 with no findings and 80 with COVID-19.

B. Preprocessing



(a)



(b)

Figure 2. Feature importance and correlation, (a) 10 laboratory findings Important selected from a dataset, and (b) 18 laboratory findings from a dataset.

The data pre-processing module contains three steps: data standardization, feature selection, and the SMOTE+ENN Hybrid technique. We standardized all features and normalized them by the Z-Score method. To perform feature selection, we used the extra tree method in [32] to select the 10 most important features as shown in Fig. 2(a) out of 18 findings as shown in Fig. 2(b), and finally, the SMOTE+ENN hybrid technique for dataset balancing.

C. Dataset Balancing

In this work, the technique used to resample the dataset to be more balanced is SMOTE+ENN, which is a hybrid of the Synthetic Minority Oversampling Technique (SMOTE) and Edited Nearest Neighbours (ENN) technique. It was developed by [33]. SMOTE is the most popular oversampling technique and can be combined with many different undersampling techniques. SMOTE works by selecting a random sample of the minority class examples. The k nearest neighbours of that sample are chosen, and a synthetic example was created from a randomly selected point in that region.

ENN work by selecting examples for deletion. This rule involves using k=3 nearest neighbours to locate those examples in a dataset that are misclassified and deleting them.

After applying the SMOTE+ENN hybrid technique, the new balanced dataset includes 366 patients (class #0) with no findings and 507 patients (class #1) with COVID-19, as shown in Fig. 3(b) compared with the original dataset as shown in Fig. 3(a).

- Original dataset shape Counter ({0: 520, 1:80}).
- Resampled dataset shape Counter ({1: 507, 0: 366}) after applying the SMOTE+ENN hybrid technique.

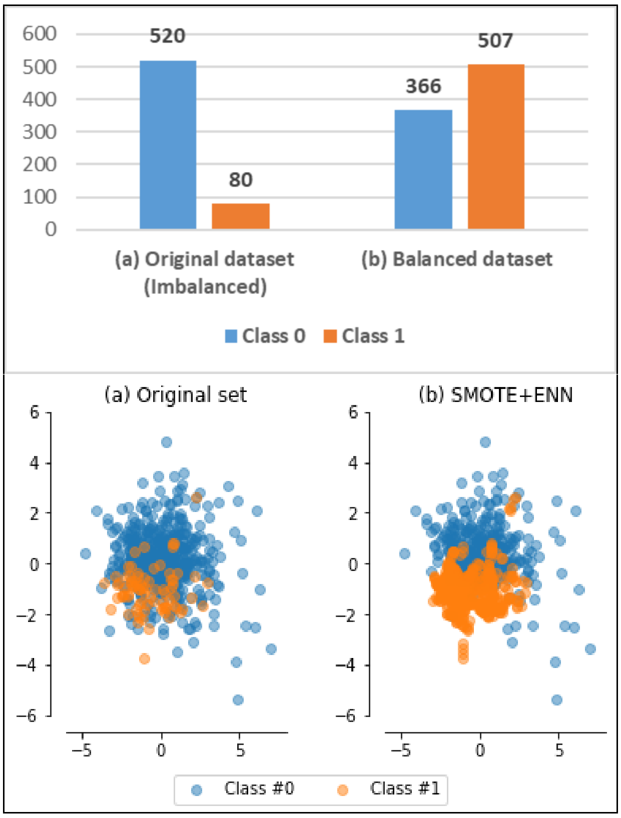


Figure 3. Dataset (a) original dataset, and (b) dataset after SMOTE+ENN hybrid technique.

We also implemented SMOTE and its other extensions, such as SMOTE + Tomek and adaptive synthetic sampling (ADASYN). The best results, however, were obtained by combining SMOTE with an ENN modification.

D. Machine Learning Techniques

Machine learning techniques are algorithms based on Artificial Intelligence (AI) to predict future outcomes from historical data. Deep Learning (DL) and Machine Learning (ML) algorithms are subsets of Artificial Intelligence (AI). It is a field that deals with computer algorithms learning and developing on their own. Deep learning and machine learning have some differences. Machine learning is about computers learning to think and act with less human intervention, and it works best with small data sets. Deep learning, on the other hand, is about computers learning to think using structures modelled after the human brain, and it works best with large data sets. In general, machine learning is used for mage Recognition, Speech Recognition, Traffic prediction, Medical Diagnosis [34], etc.

In this work, to determine COVID-19 infection using laboratory findings, we developed and evaluated clinical prediction models. Ten machine learning classification models were trained to evaluate the work: Logistic Regression (LR), Naive Bayes (NB), k-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees Classifier (DTC), Random Forest Classifier (RFC), Multilayer Perceptron (MLP), Linear Discriminant Analysis (LDA), Gradient Boosting Classifier (GBC), and Extreme Gradient Boosting (XGB).

LR model is used to predict a dependent data variable (y) by investigating the relationship between one or more existing independent variables (X). NB model is a classification method based on Bayes' Theorem and predictor independence. KNN It is also known as a "lazy learner" due to its lengthy and limited training period. The training set is used to evaluate a new instance. The distance between the new instance and the training instances is measured, and the result is calculated based on the new instance's proximity to the training instances.

SVM employs classification algorithms to solve two-group classification problems. They can categorize new text after feeding an SVM model a set of labelled training data for each category. DTC This algorithm creates a tree from the input dataset based on conditions. The tree is refined and made top-down. Conditions are used to build the branches. For example, if the dataset meets the condition, it is refined on the left branch. RFC is a decision tree-based classification algorithm. In uncorrelated forests, the algorithm builds each tree randomly to promote accurate decision-making. MLP is a neural network supplement. It has three layers: input, output, and hidden. The input layer receives the processed signal. LDA is a generalization of Fisher's linear discriminant, a method used in statistics and other fields to find a linear combination of features that characterizes or separates two or more classes of objects or events. For each weak learning model, GBC combines them to create a strong predictive model. Decision trees are commonly used in gradient boosting. And XGB is a gradient boosting decision-tree-based ensemble Machine Learning algorithm. Unstructured data prediction problems (images, text, etc.). Wide range of uses like user-defined prediction and regression problems.

To set the parameters for each Machine learning model, we used a trial and error method. The parameters of each machine learning classifier as shown in Table I. We calculated the accuracy, f1-score, precision, recall, and area under the roc curve (AUC) of each developed predictive approach to evaluate its predictive performance. We used a 10-fold cross-validation and an 80% – 20% Holdout cross-validation approach to validate the data.

TABLE I. PARAMETERS SETTINGS OF EACH MACHINE LEARNING CLASSIFIER

No.	Model	Hyper-parameters settings setting
1	LR	penalty='l2', solver='sag', C=1.0, random_state=33.
2	NB	priors=None, var_smoothing=1e-09.
3	KNN	n_neighbors= 10,weights ='uniform', algorithm='auto'.
4	SVC	kernel= 'rbf', max_iter=100,C=2.0, gamma=1.
5	DTC	criterion='entropy',max_depth=3,random_state=33.
6	RF	criterion = 'gini',n_estimators=25,max_depth=5,random_state=33.
7	MPL	activation='relu',solver='adam',learning_rate='constant', early_stopping= True,alpha=0.0001 ,hidden_layer_sizes=(100, 4),random_state=33.
8	LDA	n_components=1,solver='svd',tol=0.00001.
9	GBC	reg_param=0.1,tol=0.0001.
10	XGB	learning_rate =0.1, n_estimators=1000, max_depth=7, in_child_weight=1, gamma=0.1, subsample=0.8,colsample_bytree=0.8, objective= 'binary:logitraw', nthread=4, scale_pos_weight=1,seed=27.

E. Evaluation Metrics

To comprehend the utilized machine learning models in this work and their potentials for classification. F1 scores, recall, precision, AUC, and accuracy are used to evaluate all models and which are mathematically represented in equations (1)-(6).

$$Accuracy=\frac{(TP+TN)}{(TP+FP+TN+FN)}\times 100 \tag{1}$$

$$\mathrm{Re}call=\frac{TP}{(TP+FN)}\times 100 \tag{2}$$

$$\mathrm{Pr}ecision=\frac{TP}{(TP+FP)}\times 100 \tag{3}$$

$$F1\text{-}score=2\times\left(\frac{Precision{*}Recall}{Precision+Recall}\right)\times 100 \tag{4}$$

$$True\ Positive\ Rate\ (\,TPR)=\frac{TP}{TP+FN}\times 100 \tag{5}$$

$$False\ Positive\ Rate\,(\,FPR\,)=\frac{FP}{TN+FP}\times 100 \tag{6}$$

where *TP* is True Positive, *TN* is True Negative, *FP* is False Positive, and *FN* is False Negative.

IV. RESULTS AND DISCUSSION

For the prediction of COVID-19 infection, 600 patients and 10 laboratory importance findings were considered. The laboratory findings of all of the patients are samples. Ten different applications were developed and used as classifiers in machine learning models. After that, predictions were made, and the machine learning models performance was evaluated. In a 10-fold cross-validation approach, Table II shows the evaluation results for all machine learning applications models.

We observed that the XGB model had the best predictive performance for predicting COVID-19 disease, with an AUC score of 92.66%. Sample collection takes time and requires complex procedures, making it difficult to predict COVID-19 disease from laboratory findings. However, Clinical prediction results using the XGB model had a highly respected accuracy of 97.42%, an f1-score of 97.82%, a precision of 97.63%, and a recall of 98.05%. This is not a surprising outcome. The XGB is so good because XGB is a library for creating fast and high-performance gradient boosting tree models. So that XGB achieves the best results on a range of different complex machine learning problems. All machine learning models were evaluated with a 10-fold cross-validation approach are shown in Fig. 4.

In addition, we also used the Holdout cross-validation approach to test the performance of algorithms using 80% training and 20% testing sets. Due to the use of artificial intelligence applications in health studies, the k-fold cross-validation approach is usually used for this purpose, especially when the data is imbalanced and small in size [29]. However, it provides results that are less clearly defined in clinical applications. All algorithms outperformed the 10-fold cross-validation approach in terms of clinical predictive performance, with the best-performing algorithm being the XGB model, which had an AUC of 98.32%, an accuracy of 98.28%, and an f1-score of 98.27%, precision of 98.23%, and recall of 98.32%. Table III shows the results of all machine learning models evaluated using the Holdout cross-validation approach.

As given by equation (1), accuracy is calculated by dividing the number of correct predicted observations by the number of total predictions observations. The accuracy of all machine learning models was at least 85.14% or higher. As shown in Table III, the XGB model achieved the highest evaluation performance of 98.28%. The XGB model was also considered to be the best. XGB is still an excellent choice for a wide range of real-world machine learning problems. It is well-known for outperforming all other machine learning algorithms. XGB is an excellent algorithm that was initially chosen for structured data. This work demonstrated its determination in terms of speed and performance.

All of the precision, recall, F1-score, and values were more than 85.14%. As shown in equation (3), precision can be defined as the ratio of correctly predicted positive observations to total predicted positive observations. A perfect precision in Information Retrieval (IR) studies should be 100. The best-obtained precision score in this work with the XGB model was 98.23%. The recall is defined as the ratio of correctly predicted positive observations to all observations, as shown in equation (2).

The recall score required 1 for the perfect classification process. The XGB model achieved the best recall value of 98.32%. F1-score combines the recall and precision scores, as given by equation (4). This criterion considers both False Positives (FP) and False Negatives (FN). A high F1-score indicates that the classifier has few False Positives (FP) and few False Negatives (FN). In this case, the classifier identifies true threats and is unaffected by false alarms. When the value of an F1-score is 100, it is considered perfect. Like any other evaluation metric, The F1-score achieved the best F1-score of 98.27% with the XGB model.

In classification analysis, the AUC score is used to determine which model best predicts classes. AUC of 50% indicates no discrimination, 60-80% acceptable, 80-90% excellent, and over 90% outstanding [34]. AUC score of the MLP model is acceptable because it ranges from 60% to 80%. The AUC score of the NB and DTC models is excellent, ranging from 80% to 90%.

The remaining AUC scores were outstanding because all of the results were greater than 90%. All machine learning models can be used to predict COVID-19 clinically based on their AUC values. Since recall shows the percentage of actual positives detected, True Positive Rates (TPR) are essential in critical medical and clinical studies [35]. The recall is calculated by dividing the number of correctly diagnosed COVID-19 patients by the total number of COVID-19 diseased patients. Therefore, recall is an important criterion in this work. Additionally, AUC scores are important in medical research because they help classify diseases based on data collected from healthy participants [36], [37].

One of the characteristics of the work is accuracy, which indicates the extent to which the parameters of the sample agree with the population characteristics [38]. The researcher can demonstrate that the research is generalizable, dependable, and valid by testing the correctness of the models [39]. As a result, just these three evaluation measures were considered in this work. To compare the results with [28]-[31], the remaining ones were calculated. AUC values for all machine learning models using the Holdout cross-validation approach are shown in Fig. 5.

Table IV compares the results of this work with those of other studies in [28]-[31]. In their research, the authors used machine learning and deep learning techniques. The best classification in these studies was obtained with SVM, XGB, and CNNLSTM classifiers, as shown in Table IV. However, we did not use deep learning in this work. We have developed ten different machine learning models with more balanced data and more important features. These models outperformed deep learning classifiers in terms of accuracy and AUC. It proved that machine learning algorithms could be more powerful than deep learning algorithms when the data is small in size and balanced.

TABLE II. ALL MACHINE LEARNING MODELS WERE EVALUATED WITH A 10-FOLD CROSS-VALIDATION APPROACH

Model	**Accuracy**	**F1-Score**	**Precision**	**Recall**	**AUC**
LR	91.68	93.1	91.46	94.91	92.89
NB	86.38	88.09	90.77	85.91	82.66
KNN	94.98	95.84	94.44	97.34	94.24
SVC	97.27	97.67	98.09	97.32	92
DTC	85.95	88.56	85.49	92.23	87.86
RFC	96.42	96.95	96.88	97.09	93.19
MLP	90.97	92.48	90.93	94.18	62.03
LDA	90.39	92.27	88.17	96.85	89.67
GBC	97.13	97.62	96.27	99.03	87.66
XGB	**97.42**	**97.82**	**97.63**	**98.05**	**92.66**

TABLE III. ALL MACHINE LEARNING APPLICATION MODELS WERE EVALUATED USING THE HOLDOUT CROSS-VALIDATION APPROACH

Model	**Accuracy**	**F1-Score**	**Precision**	**Recall**	**AUC**
LR	93.71	93.61	94.21	93.32	93.32
NB	88	87.94	87.87	88.05	88.05
KNN	97.14	97.11	97.3	96.97	96.97
SVM	96.57	96.55	96.48	96.64	96.64
DTC	85.14	84.83	82.85	84.53	84.53
RFC	97.14	97.11	97.3	96.97	96.97
MLP	93.71	93.59	94.48	93.22	93.22
LDA	92	91.8	93.18	91.34	91.34
GBC	97.71	97.69	97.81	97.59	97.59
XGB	**98.28**	**98.27**	**98.23**	**98.32**	**98.32**

TABLE IV. COMPARISON OF EVALUATION RESULTS WITH PREVIOUS STUDIES

Work	**Dataset Location**	**No. Of features**	**AI Technique**	**Classifier**	**Accuracy**	**AUC**	**F1-Score**
[28]	Wenzhou Central Hospital and Cangnan People's Hospital in Wenzhu, China	11	ML	SVM	80.00%	-	-
[29]	Hospital Israelita Albert Einstein at Sao Paulo, Brazil	18	ML	SVM, RF	-	0.87	0.72
[30]	Hospital Israelita Albert Einstein at Sao Paulo, Brazil	18	ML	XGB	-	0.66	-
[31]	Hospital Israelita Albert Einstein at Sao Paulo, Brazil	18	DL	CNNLSTM	92.30%	0.6250	0.9189
This work	Hospital Israelita Albert Einstein at Sao Paulo, Brazil	**10**	**ML with SMOTE+ENN**	**XGB**	**98.28 %**	**98.32%**	**98.27%**

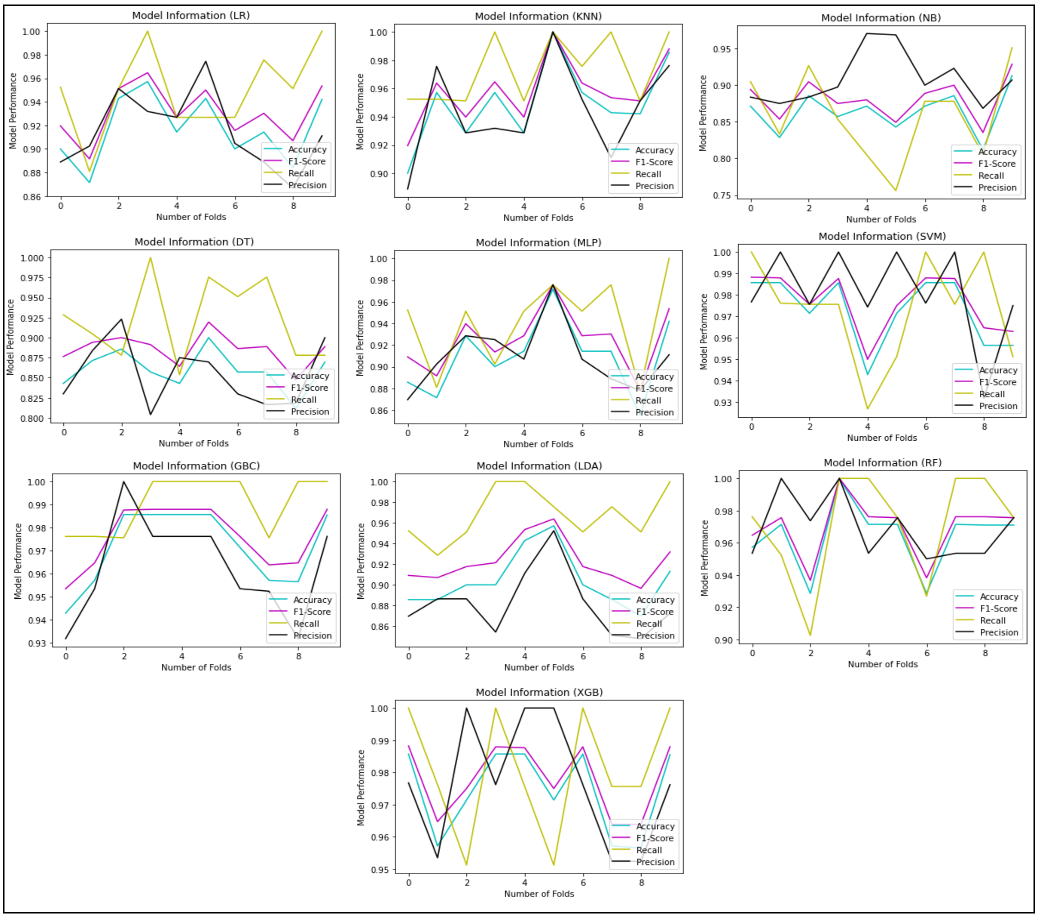


Figure 4. All machine learning models were evaluated with a 10-fold cross-validation approach.

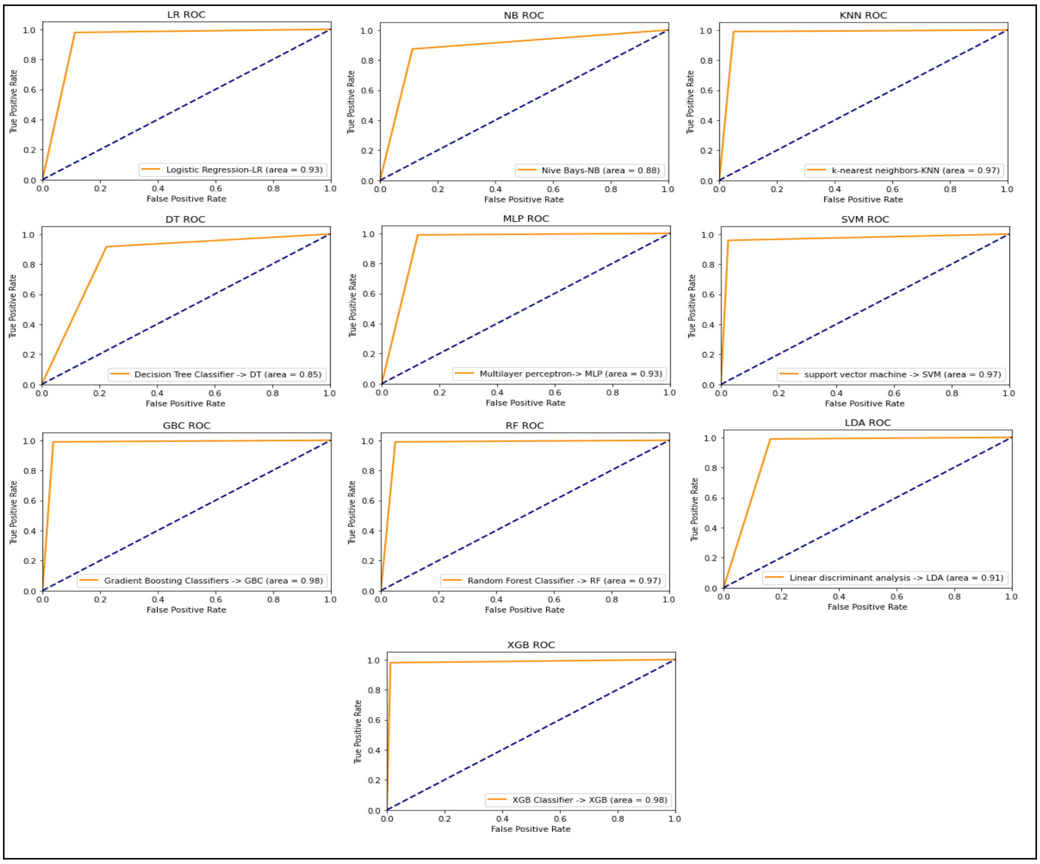


Figure 5. The AUC values were calculated for all machine learning models with the Holdout cross-validation approach.

V. Conclusion and Future Work

In this work, machine learning techniques based on laboratory data were used to predict the COVID-19 epidemic. Ten different machine learning techniques were used to analyze various laboratory data. In the initial stage of the work, the data were standardized. The selection features technique was applied to select the important features in a pre-processing stage. After that, the SMOTE+ENN hybrid technique resampled the dataset to be more balanced and then used as inputs for machine learning models. We applied cross-validation approaches such as 10-fold and Holdout cross-validation to split the dataset into training and testing sets.

After that, classification models were built, and their performance was evaluated using accuracy, F1-score, precision, recall, and AUC scores.

The XGB machine learning model produced the most meaningful results in the 10-fold cross-validation approach, with an accuracy of 97.42%, an F1-score of 97.82%, a precision of 97.63%, a recall of 98.05%, and an AUC score of 92.66%. Although this validation approach is widely used, it was not satisfactory in this work compared to the holdout cross-validation approach. In the Holdout cross-validation approach, the XGB model achieved the best accuracy, F1-score, precision, recall, and AUC values: 98.28%, 98.27%, 98.23%, 98.32%, and 98.32%, respectively. All of the machine learning models used in this work have an accuracy of at least 85.14%. Precision and recall values can be inferred in the same way.

The results of the experiments showed that all machine learning models in the holdout cross-validation approach outperformed the 10-fold cross-validation approach.

Our modeling techniques show the relevance of early COVID-19 detection and treatment. Finally, and through the results of our experiments, we found evidence to recommend that machine learning application models may be used to predict COVID-19 infection based on laboratory findings. Based on the results of our work, we recommend that healthcare systems be used to investigate the use of individual prediction models to improve the prioritization of healthcare resources and inform patient care.

In future studies, Artificial Intelligence (AI) techniques and an increase in the volume of data can be used to give early detection and treatment of COVID-19 disease.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Bilal Abdualgalil conducted the work; Sajimon Abraham analyzed the data; Waleed M. Ismael wrote the paper; all authors had approved the final version.

References

[1] World Health Organization. (2020). Report of the WHO-China joint mission on coronavirus disease (COVID-19). [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf

[2] World Health Organization. (2020). Health topics, coronavirus. [Online]. Available: https://www.who.int/health-topics/coronavirus#tab=tab_3

[3] National Institute of Infection Diseases. (2020). Field briefing: diamond princess COVID-19 cases. [Online]. Available: https://www.niid.go.jp/niid/en/2019-ncov-e/9407-covid-dp-fe-01.html

[4] C. D. Rio and P. N. Malani, "Novel coronavirus: Important information for clinicians," *J. Am. Med. Assoc.*, vol. 323, no. 11, 2020.

[5] H. Li, C. Li, and H. G. Liu, "Clinical characteristics of novel coronavirus cases in tertiary hospitals in Hubei Province," *Chin. Med. J.*, vol. 133, no. 9, pp. 1025-1031, 2020.

[6] C. Huang, *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *Lancet*, vol. 395, no. 10223, pp. 497-506, 2020.

[7] K. Q. Karm, *et al.*, "A well infant with coronavirus diseases 2019 (COVID-19) with high viral load," *Clin. Infect. Dis.*, vol. 10, 2020.

[8] Y. Bai, *et al.*, "Presumed asymptomatic carrier transmission of COVID-19," *J. Am. Med. Assoc.*, vol. 323, no. 14, pp. 1406-1407, 2020.

[9] F. Jiang, *et al.*, "Artificial intelligence in healthcare: Past, present and future," *Stroke Vasc. Neurol.*, vol. 2, no. 4, 2017.

[10] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthcare J.*, vol. 6, no. 2, pp. 92-98, 2019.

[11] S. Reddy, J. Fox, and M. P. Purohit, "Artificial intelligence-enabled healthcare delivery," *J. R. Soc. Med.*, vol. 112, no. 1, pp. 22-28, 2019.

[12] T. B. Alakus and I. Turkoglu, "Detection of pre-epileptic seizure by using wavelet packet decomposition and artificial neural networks," in *Proc. 10th International Conference on Electrical and Electronic Engineering*, 2017, pp. 511-515.

[13] N. Memarian, S. Kim, S. Dewar, J. Engel, and R. J. Staba, "Multimodal data and machine learning for surgery outcome prediction in complicated cases of mesial temporal lobe epilepsy," *Comput. Biol. Med.*, vol. 64, no. 1, pp. 67-78, 2015.

[14] J. Yousefi and A. Hamilton-Wright, "Characterizing EMG data using machine-learning tools," *Comput. Biol. Med.*, vol. 51, pp. 1-13, 2014.

[15] P. A. Karthick, D. M. Ghosh, and S. Ramakrishnan, "Surface electromyography based muscle fatigue detection using high-resolution time-frequency methods and machine learning algorithms," *Comput. Methods Programs Biomed.*, vol. 154, pp. 45-56, 2018.

[16] M. Alfaras, M. C. Soriano, and S. Ortin, "A fast machine learning model for ECG-based heartbeat classification and arrhythmia detection," *Front. Phys.*, vol. 7, 2019.

[17] C. A. Ledezma, X. Zhou, B. Rodriguez, P. J. Tan, and V. Diaz-Zuccarini, "A modeling and machine learning approach to ECG feature engineering for the detection of ischemia using pseudo-ECG," *PLoS ONE*, vol. 14, no. 8, 2019.

[18] K. Munir, H. Elahi, A. Ayub, F. Frezza, and A. Rizzi, "Cancer diagnosis using deep learning: A bibliographic review," *Cancers (Basel),* vol. 11, no. 9, p. E1235, 2019.

[19] V. Andriasyan, *et al.*, "Deep learning of virus infections reveals mechanics of lytic cells," bioRxiv, 2019.

[20] A. W. Senior, *et al.*, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, pp. 706-710, 2020.

[21] G. Bosco and M. A. Gangi, "Deep learning architectures for DNA sequence classification," *Lect. Notes Comput. Sci.*, pp. 162-171, 2017.

[22] J. Wu, J. Roy, and W. F. Stewart, "Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches," *Med. Care*, vol. 48, no. 6, pp. 106-113, 2010.

[23] G. F. Cooper, *et al.*, "An evaluation of machine-learning methods for predicting pneumonia mortality," *Artif. Intell. Med.*, vol. 9, no. 2, pp. 107-138, 1997.

[24] C. Wu, R. Rosenfeld, and G. Clermont, "Using data-driven rules to predict mortality in severe community acquired pneumonia," *Plos ONE*, vol. 9, no. 4, p. e89053, 2014.

[25] G. Clermont, D. C. Angus, S. M. DiRusso, M. Griffin, and W. T. Linde-Zwirble, "Predicting hospital mortality for patients in the

intensive care unit: A comparison of artificial neural networks with logistic regression models," *Crit. Care Med.*, vol. 29, no. 2, pp. 291-296, 2001.

[26] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits, "Unfolding physiological state: Mortality modelling in intensive care units," in *Proc. the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 75-84.

[27] A. E. W. Johnson, T. J. Pollard, and R. G. Mark, "Reproducibility in critical care: A mortality prediction case study," *Proc. Mach. Learn. Res.*, vol. 68, pp. 361-376, 2017.

[28] X. Jiang, *et al.*, "Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity," *Compu. Mater. Continua.*, vol. 63, no. 1, pp. 537-551, 2020.

[29] A. F. Batista, J. L. Miraglia, T. H. R. Donato, and A. D. P. C. Filho, "COVID-19 diagnosis prediction in emergency care patients: A machine learning approach," medRxiv, 2020.

[30] P. Schwab, A. D. Schütte, B. Dietz, and S. Bauer, "predCOVID-19: A systematic study of clinical predictive models for coronavirus disease 2019," arXiv:2005.08302, 2020.

[31] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection," *Chaos, Solitons & Fractals*, vol. 140, p. 110120, 2020.

[32] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3-42, 2006.

[33] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20-29, 2004.

[34] F. Xing, *et al.*, "Deep learning in microscopy image analysis: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4550-4568, 2017.

[35] J. N. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment," *J. Thoracic. Oncol.*, vol. 5, no. 9, pp. 1315-1316, 2010.

[36] A. Avati, K. Jung, S. Harman, L. Downing, A. Ng, and N. H. Shah, "Improving palliative care with deep learning," *BMC Med. Inform. Decis. Mak.*, vol. 18, no. 4, 2018.

[37] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) curve analysis for medical diagnostic test evaluation," *Caspian. J. Intern. Med.*, vol. 4, no. 2, pp. 627-635, 2013.

[38] A. N. Kamarudin, T. Cox, and R. Kolamunnaage-Dona, "Time-dependent ROC curve analysis in medical research: Current methods and applications," *BMC Med. Res. Methodol.*, vol. 17, no. 53, 2017.

[39] L. Wynants, *et al.*, "Prediction model for diagnosis and prognosis of covid-19: Systematic review and critical appraisal," *BMJ*, vol. 369, 2020.

Bilal Abdualgalil is a PhD research scholar in AI, Mahatma Gandhi University, Kerala, India. He received his Master degree in Master of Computer Application (MCA) from JNT University, Hyderabad, India in 2018, And he got his B.Sc. degree from Thamar University, Yemen in 2009. His research interests include Artificial Intelligence (Deep learning and Machine learning) in healthcare domain.

Dr. Sajimon Abraham is Professor in Computer Applications in School of Management and Business Studies, Mahatma Gandhi University Kottayam Kerala. He received Ph.D. in Computer Science from Mahatma Gandhi University in 2015 in the area of Spatio-Temporal Data Mining. He is additionally holding the position of Director of University Centre for International Co-operation and worked in Royal University of Bhutan for 3 years as Computer Science Faculty Member and Data Base Architect under Ministry of External Affairs, Govt of India. He has published more than 90 research articles in various journals and his research area includes Data Analytics, Spatio-Temporal Data Mining, E-learning and applications of AI in business domain.

Waleed M. Ismael is an assistant professor in Information and Communication Engineering, majoring in IoT Engineering. He received his B.Sc. in Computer Science from Thamar University, Yemen, in 2006. In 2009, he received a postgraduate diploma in Geoinformatics from ITC institute, Hollande. He completed his M.Sc. in Geoinformatics, Osmania University, India. His research interests include WSN reliability, Data fusion, Geoinformatics, and deep learning.