Results in Physics 21 (2021) 103813

Contents lists available at ScienceDirect

Results in Physics

journal homepage: www.elsevier.com/locate/rinp

Prediction modelling of COVID using machine learning methods from B-cell dataset

Nikita Jain [a], Srishti Jhunthra [a], Harshit Garg [a], Vedika Gupta [a], Senthilkumar Mohan [b], Ali Ahmadian [c,d,*], Soheil Salahshour [e], Massimiliano Ferrara [f]

[a] Department of Computer Science & Engineering, Bharati Vidyapeeth's College of Engineering, 110063 New Delhi, India
[b] School of Information Technology and Engineering, Vellore Institute of Technology, Vellore 632014, India
[c] Institute of IR 4.0, The National University of Malaysia, Bangi 43600 UKM, Selangor, Malaysia
[d] School of Mathematical Sciences, College of Science and Technology, Wenzhou-Kean University, Wenzhou, China
[e] Faculty of Engineering and Natural Sciences, Bahcesehir University, Istanbul, Turkey
[f] ICRIOS – The Invernizzi Centre for Research in Innovation, Organization, Strategy and Entrepreneurship, Bocconi University – Department of Management and Technology, Via Sarfatti, 25Milano (MI) 20136, Italy

ARTICLE INFO

Keywords:
SARS-CoV
SARS-CoV-2
Coronavirus
Support vector machine (SVM)
Naïve Bayes
K – nearest neighbors (KNN)
Logistic regression
Ensembles
Gradient boosting
Random forest
XGBoost
AdaBoost
Multilayer perceptron (MLP)
B-cells
COVD-19

ABSTRACT

Coronavirus is a pandemic that has become a concern for the whole world. This disease has stepped out to its greatest extent and is expanding day by day. Coronavirus, termed as a worldwide disease, has caused more than 8 lakh deaths worldwide. The foremost cause of the spread of coronavirus is SARS-CoV and SARS-CoV-2, which are part of the coronavirus family. Thus, predicting the patients suffering from such pandemic diseases would help to formulate the difference in inaccurate and infeasible time duration. This paper mainly focuses on the prediction of SARS-CoV and SARS-CoV-2 using the B-cells dataset. The paper also proposes different ensemble learning strategies that came out to be beneficial while making predictions. The predictions are made using various machine learning models. The numerous machine learning models, such as SVM, Naïve Bayes, K-nearest neighbors, AdaBoost, Gradient boosting, XGBoost, Random forest, ensembles, and neural networks are used in predicting and analyzing the dataset. The most accurate result was obtained using the proposed algorithm with 0.919 AUC score and 87.248% validation accuracy for predicting SARS-CoV and 0.923 AUC and 87.7934% validation accuracy for predicting SARS-CoV-2 virus.

Introduction

B-cells, where B stands for bursa of Fabricius [1,2]. This is a unique organ that is found in birds only, where B cells are mature enough. B-cells are those kinds of cells that fight against bacteria and viruses by building a Y-shaped structured protein known as antibodies. These antibodies are specific to different pathogens capable of surrounding the surface of a cell and set it for destruction marked by other immune cells. B-cells respond in vivo, producing a large amount of antigen specified antibodies by obtaining different epitope regions of the proteins [1,2]. They perform the binding operation of antibodies to other antigen proteins. Apart from B-cells, coronavirus is a pandemic that causes respiratory infection to a human. Coronavirus consists of a family of viruses that are responsible for the occurrence of this pandemic. The family includes some deadly viruses like SARS-CoV, SARS-CoV-2, MERS-CoV, and many more categorized into classes [3,4]. Moderate acute respiratory coronavirus syndrome (SARS-CoV) [5] and severe acute respiratory coronavirus syndrome (SARS-CoV-2) [6,7] have caused the majority of coronavirus cases among all other members of the family, particularly in children and elderly people. The comparison between the two coronavirus family members, SARS-CoV and SARS-CoV-2, clearly shows that SARS-CoV-2 is more dangerous than SARS-CoV [8]. These two viruses are the foremost reason for the spread of coronavirus. Thus, predicting if a patient is suffering from any of the two viruses would help form accurate patients. Therefore, considering the B-cell dataset obtained from COVID-19/SARS B-cell epitope prediction dataset [1,2].

* Corresponding author.
E-mail address: ahmadian.hosseini@gmail.com (A. Ahmadian).

https://doi.org/10.1016/j.rinp.2021.103813
Received 9 November 2020; Received in revised form 25 December 2020; Accepted 30 December 2020
Available online 17 January 2021

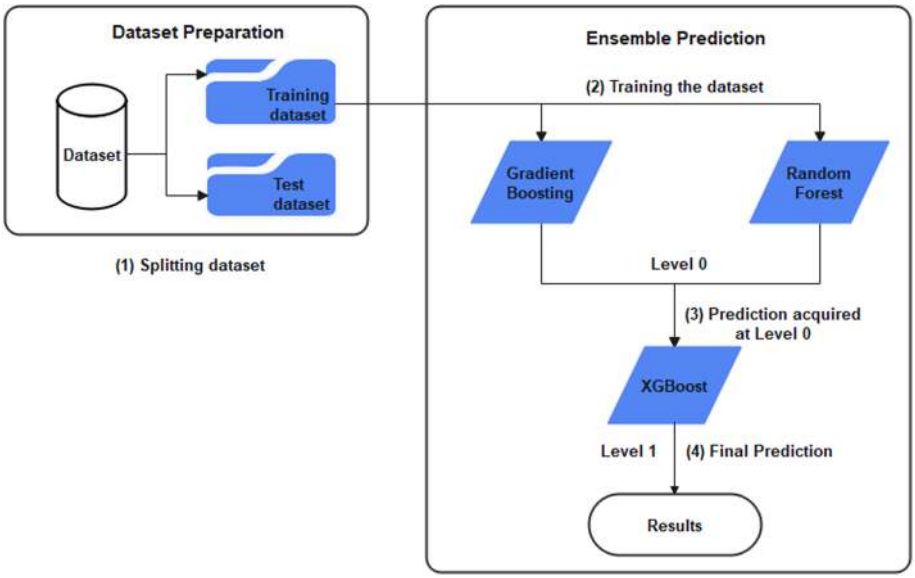

Fig. 1. Stacked ensemble.

This dataset was developed in research and observation done from the immune epitope database (IEDB) [9] and Universal Protein Resource (UniProt) [10]. The dataset contains three sub-datasets where the attributes depict that SARS-CoV and SARS-CoV-2 could be predicted using it. The B-cell dataset includes complete information about the patient's health status, which could help train the machine and predict the test cases. Thus, various machine learning models such as SVM, KNN, AdaBoost, Gradient Boosting, neural network (NN), etc., are used for the implementation. The paper first focuses on predicting SARS-CoV using a B-cell dataset. After applying all the desired machine learning models and neural networks. The B-cell dataset is combined with the SARS-CoV dataset, which predicts the SARS-CoV-2 virus in a human body. Thus, the paper entirely focuses on the prediction of SARS-CoV and SARS-CoV-2 using various machine learning models and algorithms. The paper contributes by:

1. Using the B-cells dataset containing the number of proteins and peptides in a human body and all the information related to them were used to analyze and predict SARS-CoV and SARS-CoV-2 viruses.
2. We are proposing an algorithm of a stacked ensemble formed using different baseline models and predefined ensembles.

Although the results obtained from this dataset are proficient enough for making predictions but apart from the contribution, the paper could further be improved if the dataset:

1. Contained a proficient amount of 1 as a target value. The inclusion could improve accuracy even to a greater extent.
2. Suppose it included the labeled attributes of the SARS-CoV-2 virus in the dataset. The accuracy of the test set, thereby, could not be predicted. If it was labeled, then the models could have easily verified. Though, the test prediction for SARS-CoV acquired an appreciable accuracy.

The paper comprises six sections. Section "Related work" consists of related work where similar work performed by various authors are studied and summarized. Section "Material" contains a discussion concerning all models and algorithms implemented on the dataset—followed by Section "Methodology", which includes the methodology used for analyzing the dataset. Its features and their importance are discussed in this section. After the methodology section comes Section "Results and discussion", which comprises the results and discussion, all the results are adequately summarized. Section "Conclusion" contains the conclusion withdrawn after analyzing the results obtained for the prediction of SARS-CoV and SARS-CoV-2.

Related work

In the literature, many research and work have been accomplished by various authors for predicting SARS-CoV and SARS-CoV-2, causing coronavirus. [11] proposed their study on the SARS-CoV-2 vaccine by using mass spectrometry-based bioinformatics as a source of predicting identifiers. They came out with results of predicting coronavirus family epitopes using the spectrometry-based mechanism successfully. [12] also came up with their homology and bioinformatic approach to predict the target immunes that are responsive to SARSA-CoV-2. Their methodology helped in the prediction of SARS-CoV-2. The proposed approach was independently identified in various regions of the human body. Thus, their study concluded that the conserved immune regions contain many implications for designing vaccines against a variety of coronavirus. [13] studied the outbreak pandemic with great ease and thereby predicted the epidemic outbreak's impacts over the global chains. This, therefore, stimulated the analysis of SARS-CoV and SARS-CoV-2, causing coronavirus outbreak. This study demonstrated the simulation model for the epidemic analysis and gave a unique case study on supply chain risks. After looking at all the investigations and studies [14] decided to propose their study based on predicting SARS-CoV-2 infection using diagnostic samples. This study used the diagnostic samples dataset and conducted successful research on predicting whether a human being is suffering from SARS-CoV-2. Apart from all the predictions made by different [6] decided to do a preliminary study on the origin of SARS-Cov-2. According to the authors, the infection of SARS-CoV-2 has started spreading all across the world in the early march of the year 2020 and has not been growing with an exponential limit. The authors also proposed the theory behind the origin of SARS-CoV-2, causing coronavirus, and termed it as the seventh member of the coronavirus family infecting human beings. After analyzing different authors' analyses and results [15] predicted epitope surfaces of antigen proteins that would help produce the vaccine. Their study was based entirely upon the LSTM network, which gave the desired results and was considered a valuable prediction for the vaccine production scheme. However, SARS-CoV-2 spreading predictions [16] studied the virus's

structure in early April of 2020. They formulated the virus's structure in different patients like the velpatasvir, ledipasvir, and other drug candidates.

Thus, a lot of research and predictions are made by various authors using different analysis schemes and techniques for predicting SARS-CoV and SARS-CoV-2, causing coronavirus. This paper also introduces another technique for predicting the coronavirus family's viruses, i.e., prediction of SARS-CoV and SARS-CoV-2 using B-cells. Section "Material" discusses all the machine learning models and algorithms used in the paper for the prediction of coronavirus.

Material

Ensemble learning algorithm/proposed work

Ensembles refer to the algorithm that is formed by the combination of several machine learning models into a single proficient model [17]. Ensembles are used to improve the accuracy by combining weak predicting models to form a robust predicting model. The ensemble used in this study is a stacked ensemble consisting of random forest and gradient boosting at the inner layer and XGBoost on the outer layer. (ref. to Fig. 1)

Algorithm 1:. *Preprocessing function for the dataset*

1. **function** preprocessing(bcell,sars,covid)
2. bcell_sars = concatenate(bcell,sars)
3. X_Train = bcell_sars.drop(['target'])
4. Y_Train = bcell_sars['target']
5. X_Train['peptide_length'] = X_Train['end_position'] - X_Train['start_position'] + 1
6. X_Train.drop(['start_position','end_position'])
7. ex = ExtraTreesClassifier()
8. ex.fit(X_Train,Y_Train)
9. feature_importance = ex.feature_importances_
10. **function** Normalize(feature_importance)
11. **return** normalized scores
12. **end function**
13. **function** analysis(feature_importance)
14. plot graph
15. unproductive_features = ['parent_protein_id','peptide_seq','protein_seq']
16. **return**(unproductive_features)
17. **end function**
18. unproductive_features.append(['start_position','end_position'])
19. dataframes = [bcell,sars,covid,bcell_sars]
20. **for** each dataframe
21. dataframe['peptide_length'] = dataframe['end_position'] - dataframe['start_position'] + 1
22. dataframe.drop(unproductive_features)
23. **end for**
24. **return** dataframes
25. **end function**

Algorithm 2:. *Ensemble learning algorithm/Proposed model*

1. **function** proposedModel
2. x_train,x_val,y_train,y_val = SplitData(x,y,test_size = 0.2)
3. x_train_base,x_test_base,y_train_base,y_test_base = SplitData(x_train,y_train,test_size = 0.3)
4. **for** each b ∈ F **do** // F is the list of random forest and Gradient boosting models
5. b = train(x_train_base,y_train_base)
6. z = predict(x_test_base)
7. y_pred_base = append(z)
8. **end for**
9. meta_learner = XGBClassifier(estimators = 1000,max_depth = 4, learning_rate = 0.005)
10. meta_learner = train(y_pred_base,y_test_base)
11. y_pred = predict(x_val)
12. accuracy = compare(y_pred,y_val)
13. **return** y_pred,accuracy
14. **end function**

Supervised learning models

Supervised learning is a technique used to prepare a group of decision rules that can help predict a known outcome [18]. These rules are termed as models. Thus, supervised learning models are those used to indicate an existing outcome by creating a group of decision rules. There are various types of supervised learning models. The following are the models that are used in this study:

1. Support vector machine (SVM)
2. K – nearest neighbors (KNN)
3. Naïve Bayes
4. Random Forest
5. Gradient Boosting
6. Logistic Regression
7. AdaBoost
8. XGBoost

These are all the supervised learning models that were studied and implemented to train and test the machine on the dataset to predict SARS-CoV and SARS-CoV-2, respectively.

Support vector machine (SVM)

A support vector machine (SVM) is a simple but powerful supervised learning approach in predicting data [19]. Classification and regression problems can be resolved using SVM. SVM is useful in preparing the high dimensional feature spaces, which is useful while solving classification tasks. After considering the regularization parameter as 100 and probability equal to "True," the model acquired efficient results.

K – nearest neighbors (KNN)

Unlike SVM (ref. to Section "Ensemble learning algorithm/proposed work"), K – nearest neighbors are also used to solve both classification and regression problems [20]. But KNN is different from the other algorithms. KNN is a lazy learning algorithm in which no model is constructed. In this algorithm, predictions are made straight from the training dataset. In this study, we have considered n nearest neighbors as 15.

Naïve Bayes

Naïve Bayes is a classification supervised learning model which is based upon Bayes' Theorem [21]. It is a family of algorithms where every family member shares a common principle, i.e., the Bayes' Theorem. This means that during the classification process, every pair is independent of each other. Thus, the naïve bayes model is trained based on a "gaussian nb classifier."

Random forest

Random forest combines the tree predictors in such a way that each tree is dependent on the random vector's sampled value independently [22]. It can also be used as a classifier for training and testing purposes. Random forest divides the dataset features into a subset, which is used to build random decision trees. These trees combine built a classification tree, which is used to make predictions. The random forest also takes an estimator (n_estimator), i.e., the number of trees and max (maximum) features, which means the maximum features to be selected for the tree as hyperparameters. In this study, the n_estimator is 100, and the max features parameter is 5 to train and test our random forest model.

Gradient boosting

Gradient Boosting is another supervised learning model used as a problem solver for both regression and classification problems [23]. It predicts the ensembles of weak models and combines them to form a

strong supervised learning model. The gradient boosting model has a parameter known as the estimator, which means the number of boosting stages required by the model. Having the value of n_estimator as 100, the model observed to predict results more accurately.

Logistic regression

Logistic regression is a form of binary regression [24]. A statistical model uses a logistic function approach to build an independent binary variable that plays an essential role in prediction making. For training and predicting SARS-CoV and SARS-CoV-2 viruses, the regularization parameter was taken as 100 while training the logistic regression model.

AdaBoost

AdaBoost stands for Adaptive Boosting. AdaBoost was the first successful boosting algorithm designed for binary classification problems [25]. This algorithm can be used for various types of problems like regression, classification and can also be used as an ensemble. This algorithm uses an iterative approach, i.e., it keeps on correcting the errors from weak classifiers and finally making a robust combined classifier. In this study, AdaBoost uses a decision tree classifier with the tree's maximum depth as 2, where the maximum depth of the tree is the hyperparameter.

Xgboost

XGBoost is an optimized solution of the gradient boosting model. It is highly efficient, portable, and flexible [26]. It uses the gradient boosting framework and provides a parallel tree boosting technique, which can solve a variety of problems with high accuracy and in a short period. Some of the hyperparameters of XGBoost are the boosting stage, the tree's maximum depth, and the learning rate. In this study, the n_estimators of the XGBoost model is taken as 1000, the maximum depth of the tree as 4 with a learning rate of 0.005 to train and test the model.

Neural network algorithms

A neural network is a part of learning algorithms where different layer formation techniques are used to predict and analyze the data. Multilayer perceptron (MLP) [27] is a class of artificial neural networks consisting of numerous features and fully connected hidden layers. In this study, MLP consisted of 3 Hidden layers with one dropout layer followed by each hidden layer. Each hidden layer consisted of 64, 32, and 16 neurons, respectively. After the series of all the hidden and dropout layers, the resulted output at the output layer used sigmoid as its activation function

Further, Section "Methodology" contains the methodology used for training and testing the dataset for both SARS-CoV and SARS-CoV-2, causing coronavirus infection to human beings.

Methodology

To predict the viruses: SARS-CoV and SARS-CoV-2, the COVID-19/SARS B-cell epitope prediction dataset was used to get accurate results. Thus, the dataset was first needed to be transformed and then implemented. Below Section "Dataset description" contains the overview of how the dataset was transformed and then divided to predict SARS-CoV and SARS-CoV-2 viruses, respectively.

Dataset description

The COVID-19/SARS B-cell epitope prediction dataset consists of three sub-datasets. These sub-datasets were used sequentially to make the respective predictions, i.e., SARS-CoV and SARS-CoV-2 virus. The three sub-datasets are:

1. B-cell dataset:

Table 1
Attributes.

S. no.	Attributes	Description	Data type
1.	parent_protein_id	Unique parent protein ID	Categorical
2.	protein_seq	Parent protein sequence	Categorical
3.	start_position	Start position of the peptide	Numerical
4.	end_position	End position of the peptide	Numerical
5.	peptide_seq	Peptide sequence	Categorical
6.	chou_fasman	Peptide feature, Beta turn	Numerical
7.	emini	Relative surface accessibility	Numerical
8.	kolaskar_tongaonkar	Antigenicity	Numerical
9.	parker	Hydrophobicity	Numerical
10.	isoelectric_point	Protein feature	Numerical
11.	aromacity	Protein feature	Numerical
12.	hydrophobicity	Protein feature	Numerical
13.	stability	Protein feature	Numerical

The B-cell dataset was mainly used for training models. It consisted of the patient's data to be trained in the machine learning models. This dataset consists of 14,387 rows for all combinations of 14,362 peptides and 757 proteins. Compared with SARS-Cov, the prediction results helped to calculate the accuracy and precision learning models applied models' accuracy and precision dataset. After analyzing the SARS-CoV results, a combination of the B-cell dataset and the SARS-CoV dataset was obtained. Thus, the combined dataset supported predicting SARS-CoV-2, which is the primary coronavirus reason.

2. SARS-CoV dataset:

The SARS-CoV dataset was also used as the training dataset while predicting SARS-CoV-2, majorly spreading worldwide. The obtained is a labeled dataset that consists of 520 rows in total. This dataset promoted to predict the models' accuracy and precision applied to the B-cell dataset, further aided in predicting the SARS-CoV virus prediction.

3. COVID dataset:

The COVID dataset was the target dataset. This dataset was used in the testing of SARS-CoV-2. This dataset was not much of use, but it enhanced the results by testing the model using this data, which gave the accuracy and AUC scores of the models applied. (ref. to Section "Results and discussion").

Further, in Section "Data analysis", the datasets were analyzed, and the required attributes were selected using feature engineering.

Data analysis

All the three datasets obtained from the COVID-19/SARS B-cell epitope prediction dataset has 13 attributes. These attributes contain information about the patient for predicting the SARS-CoV and SARS-CoV-2 virus.

Table 1 shows the description of all the attributes of the dataset. All three datasets, i.e., the B-cell dataset, SARS-CoV dataset, and COVID dataset, comprise the same attributes mentioned in Table 1. Thus, all the attributes were studied and analyzed. Section "Feature selection" introduces the feature engineering concept applied to the dataset to select the most proficient attributes.

Feature selection

Though the given dataset was perfectly balanced, selecting the desired features that carry a high weightage would help analyze more accurate and precise results. Thus, feature selection is a methodology used to obtain the importance of different attributes [28]. Depending on the significance of attributes, their weightage increases, which indulge in attaining accurate results. In this paper, all three datasets contain 13

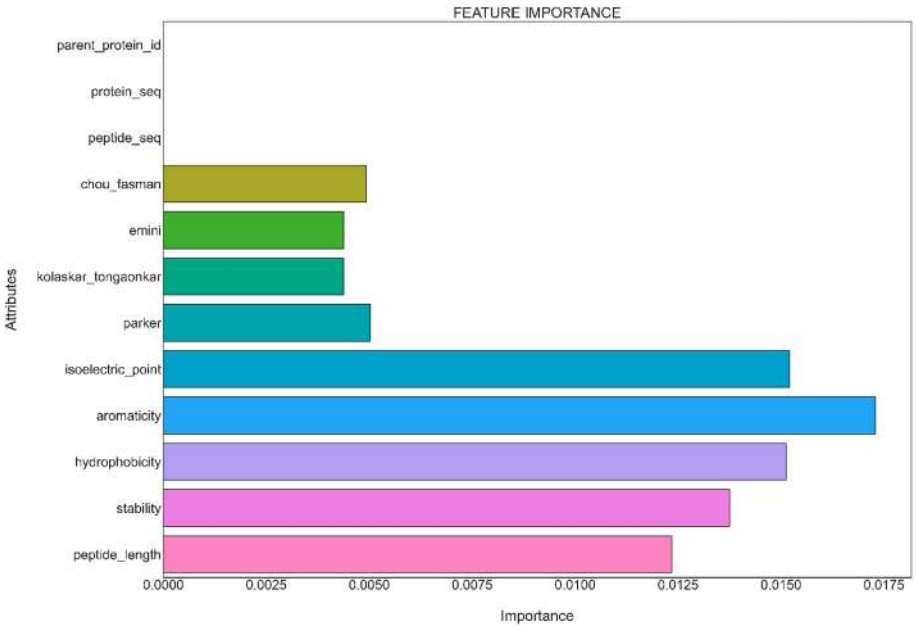

Fig. 2. Feature importance.

similar attributes (ref. to Table 1). Two attributes labeled start_position and end_position contained the starting and ending position of the peptides. These two attributes showcase the same functionality, i.e., why we formulated a combined attribute: peptide_length. Eq. (1) shows the formula of peptide_length attribute, which is formed after combining start_position and end_position.

$$peptide\_length = start\_position - end\_position \tag{1}$$

After applying feature importance on all the 13 attributes. The feature importance plot was created in which three attributes, i.e., parent_protien_id, protein_seq, and peptide_seq, showed null importance, as shown in Fig. 1.

Fig. 2 shows a bar graph representation of feature importance using the concept of feature engineering. Feature engineering is the process of selecting a variety of features from a raw dataset using data mining techniques. Selecting features helps to acquire more accurate and optimized results. Thus, it improves the overall performance of machine learning models. The graphs clearly show that parent_protein_id, protein_seq, and peptide_seq hold zero importance than all the other attributes. Thus, it clearly shows that all these attributes are redundant and can be excluded. Therefore, after removing these three attributes mentioned above. The final dataset comprises nine attributes that will significantly predict SARS-CoV and SARS-CoV2, respectively.

After removing all the three attributes, individual attribute analysis helped check whether the dataset accommodates the given parameter. For this graph for each attribute was plotted against the dataset provided. The results obtained, as shown in Fig. 3, concluded that the given dataset is normally distributed. For each attribute taken into consideration, there are no left out points in the dataset. This shows that the final dataset, which contains all the nine attributes, is applicable and should be considered for the training and testing of different machine learning models.

Thus, Fig. 3 clearly shows that the dataset is not properly maintained, and all the essential features do not contain any left out point in the dataset. After obtaining the final dataset, we obtained the training and test dataset. Various machine learning models and algorithms such as SVM, KNN, Naïve Bayes, random forest, AdaBoost, gradient boosting, XGBoost, MLP-NN, and the stacked ensemble (ref. to Section "Material") was applied to the dataset. The dataset was not trained and applied on different machine learning models and algorithms mentioned above after splitting.

Fig. 4 shows the overall description of the complete methodology used in this paper. The figure indicates the dataset extraction, feature selection, and the different combinations of the dataset made to predict SARS-CoV and SARS-CoV-2, respectively. The Figure also shows all the models and stacked ensemble used to predict the viruses causing coronavirus.

Various machine learning models have been applied to the dataset after acquiring the adequate dataset for predicting SARS-CoV and SARS-CoV-2 causing coronavirus disease (ref. Section 3). Section "Results and discussion" contains the results and discussion after applying all the models and ensembles on the final dataset.

Results and discussion

The B-cell dataset was mainly used to predict the SARS-CoV virus belonging to the coronavirus family,. The B-cell dataset was split into validation and test dataset. After splitting, various machine learning models were applied to it. The dataset's testing results provided the confusion matrix and ROC curve of all the machine learning models and algorithms. The proposed ensemble's training time varied between 2 and 4 s. The confusion matrix is a tabular form representation describing the machine learning models' performance on a set of test data [29]. It is also known as the error matrix used to calculate the accuracy of the model applied such as used in the following models [30–33]. It represents the truth and the false rate of the models. At the same time, the ROC curve refers to the receiver operating characteristic curve. It is the graphical representation of the false positive rate versus the true positive rate. It is used to analyze the AUC score, i.e., the area under the curve, which measures all possible classification thresholds' overall performance. The AUC value depicts the accuracy by subjecting more true values, resulting in increased accuracy of the model, as shown in the Eq. (2).

$$AccuracyScore = \frac{CorrectPredictions}{AllPredictions} = \frac{TP + TN}{TP + FN} \tag{2}$$

TP and TN refer to the true positive value and true negative value, whereas FN refers to the false negative value. All these values could be

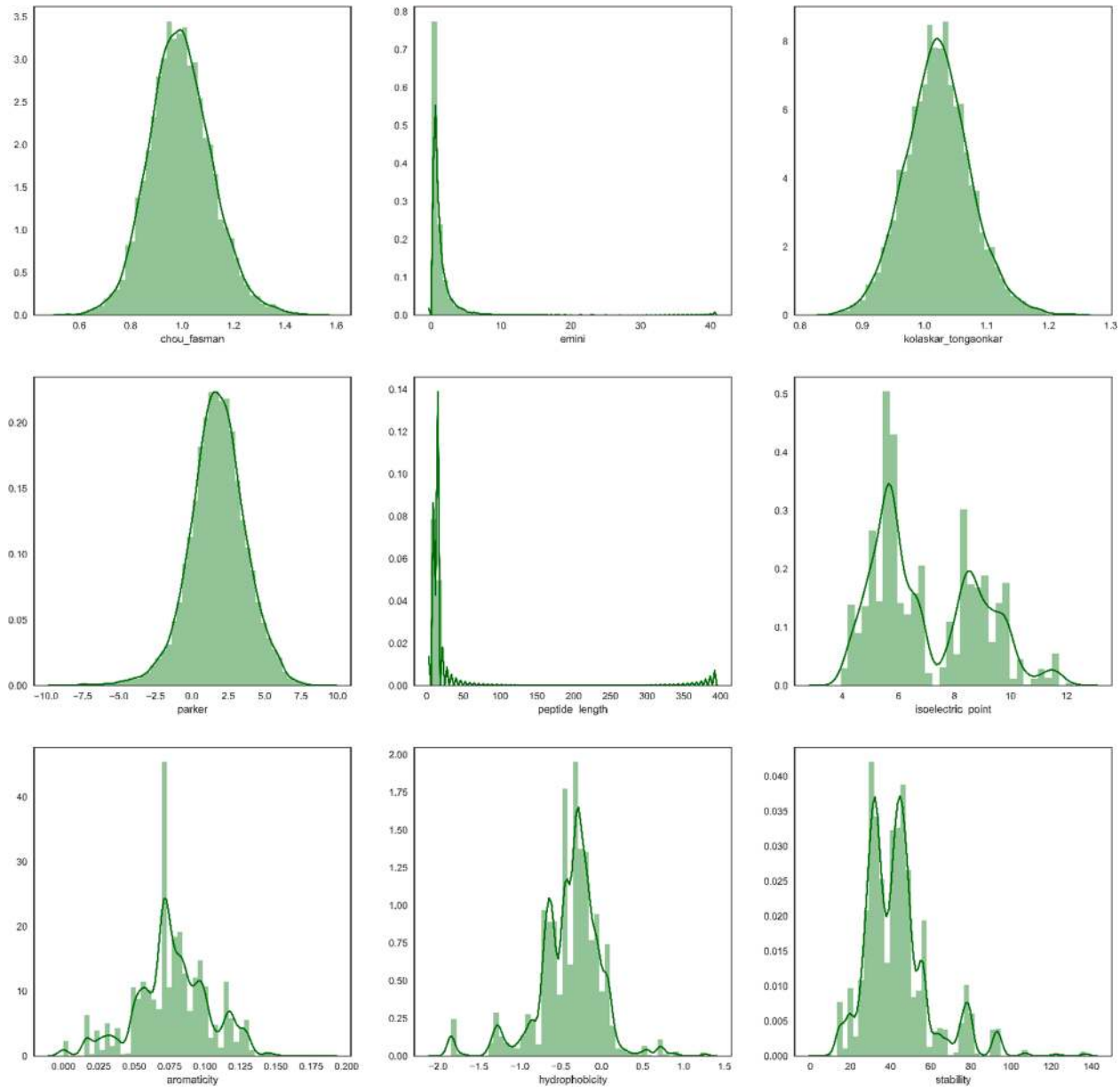

Fig. 3. Data distribution plots of selected attributes.

easily obtained by the confusion matrix, as shown in Fig. 5.

Thus, the error approximation could also be analyzed and calculated for all the applied machine learning models and algorithms. The error computation is performed in three forms, which are known as a mean average error (MAE), root means square error (RMSE) and mean square error (MSE). Mean average error (MAE) highlights the average error of the predicted value versus actual value using the formula discussed in the Eq. (3).

$$MAE = \frac{1}{n}\sum |y - \hat{y}| \tag{3}$$

Root mean square error refers to the error estimation techniques of comparatively large- scale error computation. It takes the square root of the MSE value, i.e., the mean square value, which could be calculated by taking the square difference of the predicted and the actual value as shown in the Eq. (4).

$$MSE = \frac{1}{n}\sum (y - \hat{y})^2 \tag{4}$$

Thus, after calculating the MSE value, a straightforward calculation gives RMSE by taking the square root of the MSE value (ref. to Eq. (5)). RMSE is highly preferred over MSE and MAE as it gives more accuracy and a precise error of the models applied.

$$RMSE = \sqrt{\frac{1}{n}\sum (y - \hat{y})^2} \tag{5}$$

These are the factors that justify the predictions made using different machine learning models. It explains the accuracy and precision of the models, which showcase the overall performance of the model. The following are the results obtained after applying all different machine learning models and algorithms used for predicting SARS-CoV and SARS-CoV-2 virus. The results are arranged in a sequence where the validation and test results for predicting the SARS-CoV virus causing coronavirus are showcased first. Further, the validation output is showcased, including all the confusion matrix, ROC curves, concluding with a complete table representing all the validation scores and error approximation.

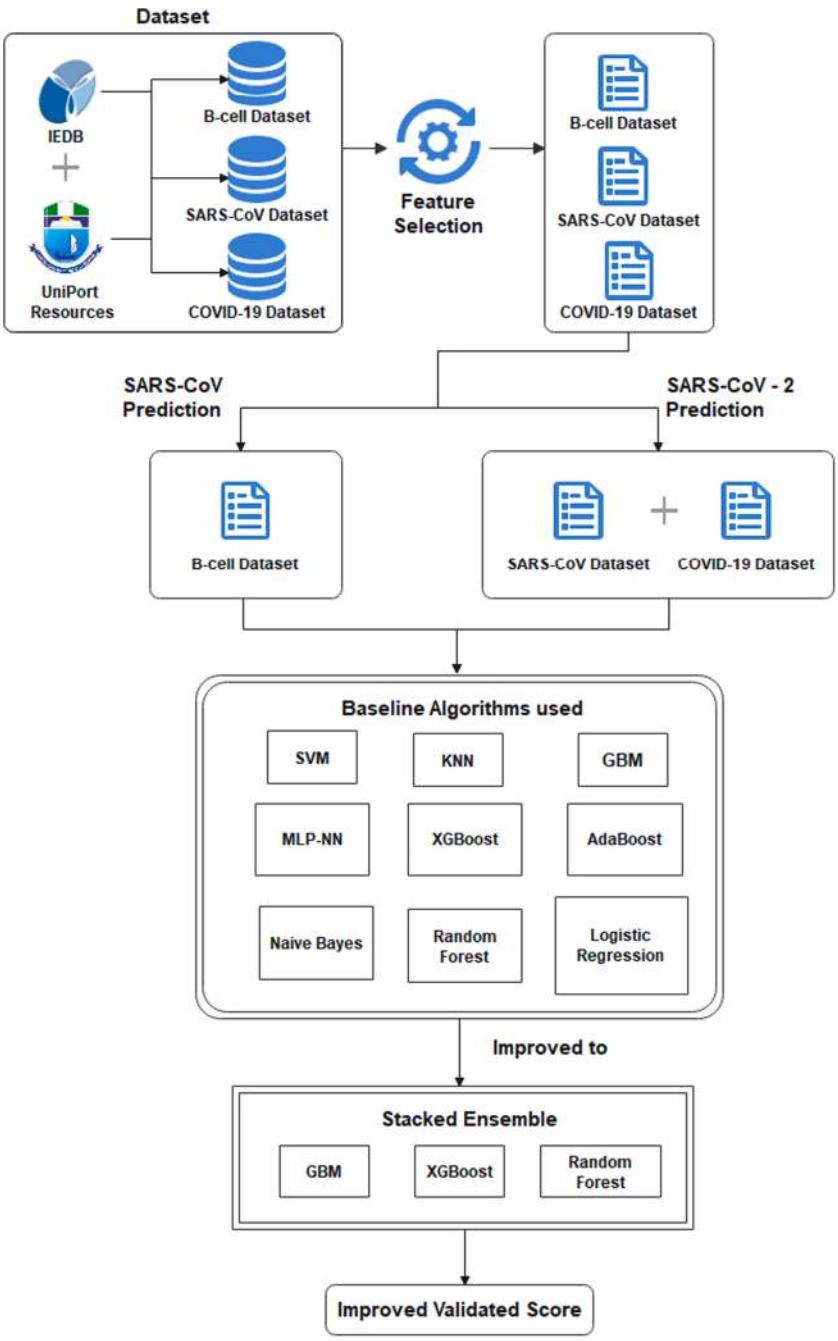

Fig. 4. Overall description.

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Fig. 5. Structure of a confusion matrix.

The first section of the results contains a complete and detailed description of the outputs obtained after applying the B-cell dataset on different machine learning models. This results section shows all the prediction outcomes of the SARS-CoV virus.

Thus, the following are the validation results obtained after applying different machine learning models and algorithms, which includes models like SVM, KNN, Naïve Bayes, Random forest, Gradient boosting, XGBoost, Logistic regression, AdaBoost, MLP-NN, and a stacked ensemble (ref. to Section "Material"):

Figs. 6 and 7 represents the confusion matrix and ROC curve obtained after predicting SARS-CoV using the B-cell dataset applied to the SVM machine learning model. The diagonal components represent the true positive and the true negative values in the prediction (ref. to Fig. 5). The ROC curve shows that the prediction is satisfactory and

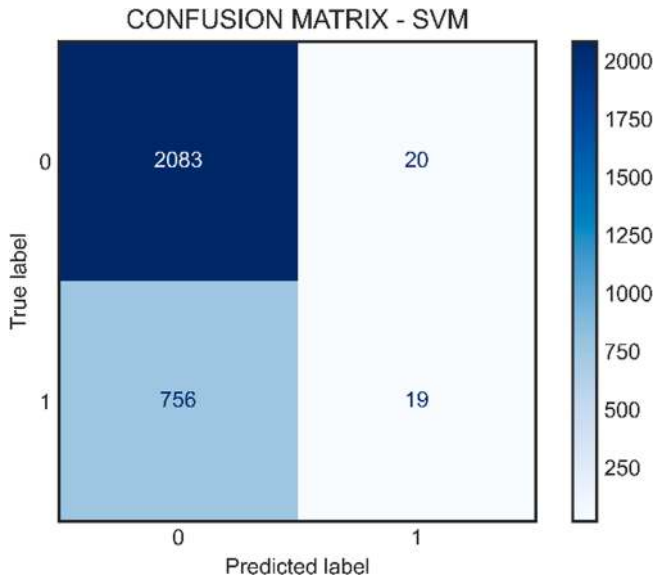

Fig. 6. Confusion matrix of SVM.

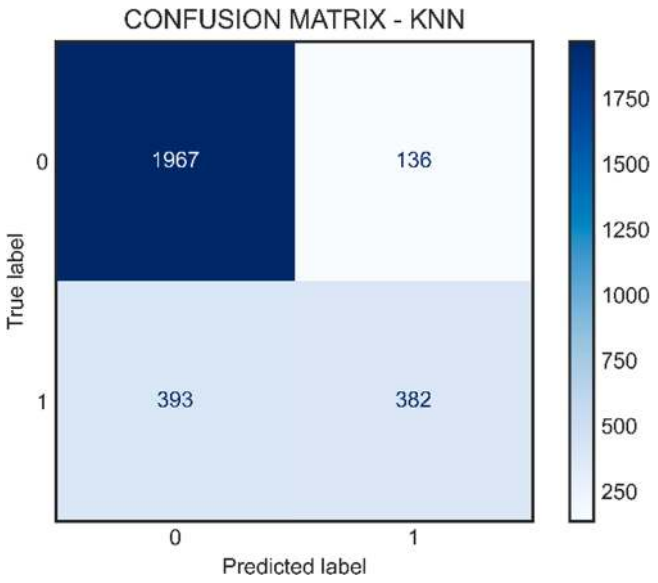

Fig. 8. Confusion matrix of KNN.

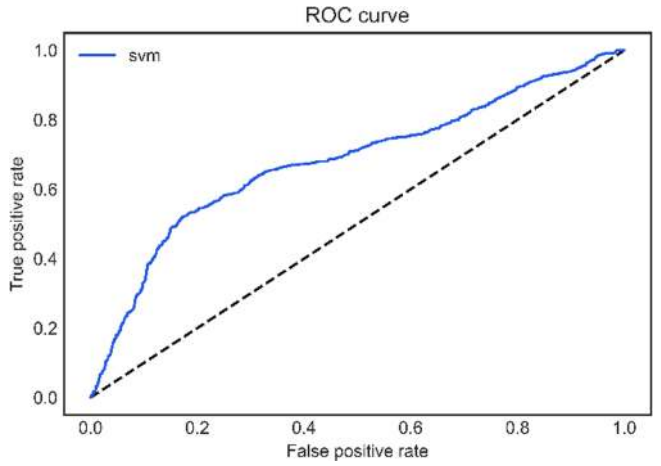

Fig. 7. ROC curve of SVM.

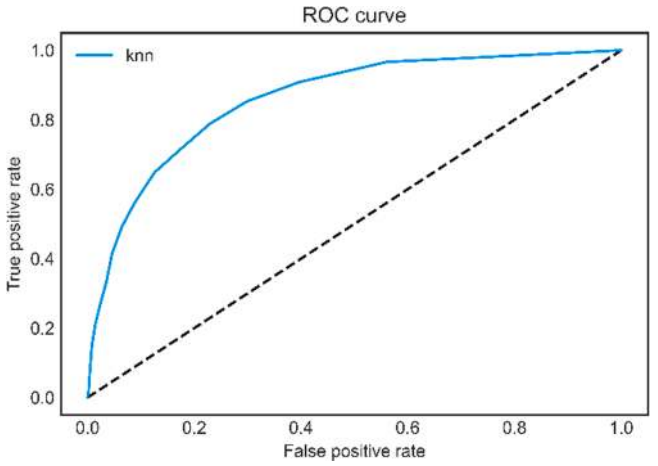

Fig. 9. ROC curve of KNN.

Table 2
SARS-CoV validation results.

Model	Validation AUC	Validation accuracy (%)	MSE	RMSE	MAE
SVM	0.682	73.4190	0.2658	0.5156	0.2658
KNN	0.858	81.6192	0.1838	0.4287	0.1838
Naïve-Bayes	0.652	72.9673	0.2703	0.5199	0.2703
Random Forest	0.909	85.5803	0.2365	0.4863	0.2365
GBM	0.868	81.7582	0.1824	0.4271	0.1824
Logistic	0.652	72.6546	0.2735	0.5229	0.2735
AdaBoost	0.869	82.6963	0.1730	0.4159	0.1730
XGBoost	0.871	81.3065	0.1869	0.4323	0.1869
Ensemble	0.919	87.2481	0.1442	0.3797	0.1442
MLP - NN	0.809	77.4843	0.2252	0.4745	0.2252

could attain an AUC score greater than 0.5 (ref. to Table 2).

Figs. 8 and 9 showcase the confusion matrix and ROC curve after applying the KNN model (ref. to Section "K – nearest neighbors (KNN)"). The ROC curve shows that the predicted output is efficient and has acquired an accurate result by obtaining an AUC score of 0.858. The ROC curve shows accurate results when the curve moves to the left top corner.

Figs. 10 and 11 shows the output obtained after applying the Naïve Bayes. Though the confusion matrix predicted a proficient number of true positive values, it also had an equivalent number of false-negative values that affected the ROC curve. The ROC curve thus did not represent many accurate results as expected. But the AUC is above 0.6, from which we can conclude that the results are just satisfactory.

Figs. 12 and 13 are the results obtained using the random forest model. These figures represent the ROC and confusion matrix obtained after predicting the SARS-CoV virus using random forest. The ROC curve has an AUC score of 0.919, which is the best score obtained among all the machine learning models applied in this paper. The ROC curve clearly shows that it can predict the most accurate results as it is entirely tilted towards the top left corner, which means that the model applied has outperformed among all the others.

After applying the random forest model, which is an inbuilt Ensemble, there are various other models as well, which are used in making predictions and analyzing more accurate results for the SARS-CoV virus causing coronavirus disease. Thus, Figs. 14 and 15 shows

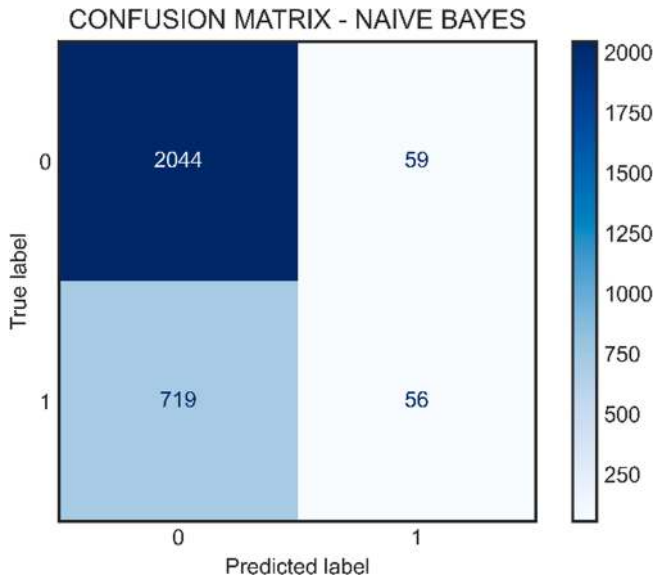

Fig. 10. Confusion matrix of Naïve Bayes.

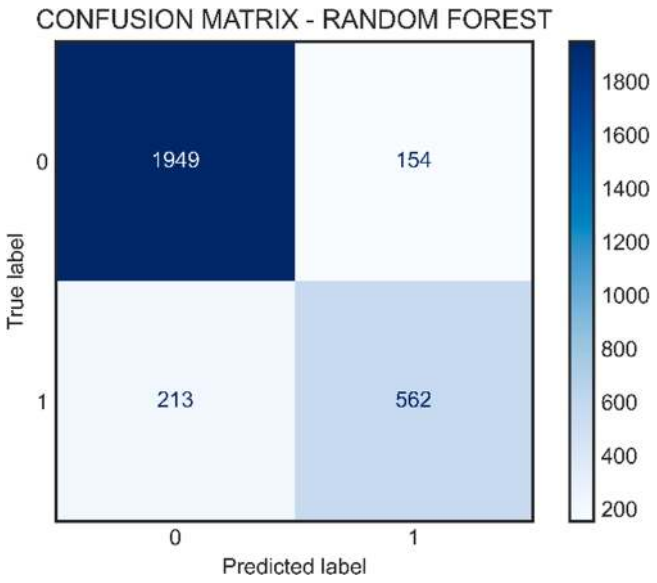

Fig. 12. Confusion matrix of Random forest.

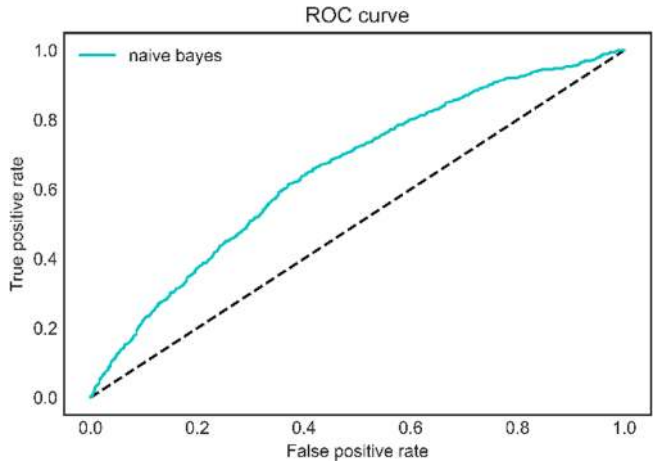

Fig. 11. ROC curve of Naïve Bayes.

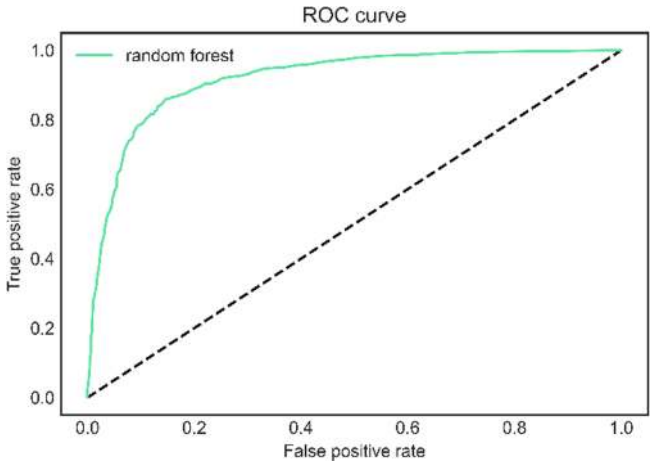

Fig. 13. ROC curve of Random forest.

another inbuilt ensemble, i.e., gradient boosting, which also indicates some précised results after applying it on the dataset. It contains many false-negative values, but it has been easily overcome by a large amount of true positive values. The ROC curve significantly shows better results with an AUC score of 0.868.

Figs. 16 and 17 shows the results obtained from logistic regression. This model is a baseline model that gave a bit low accuracy and AUC score. But, it could be used on a small dataset if required to make predictions on specific content. The ROC curve shown in the figure is satisfactory and could only be used on a small dataset to get more accurate and precise results.

AdaBoost is another inbuilt ensemble that is applied for making predictions for the SARS-CoV virus. Figs. 18 and 19 shows the confusion matrix and ROC curve obtained after applying the AdaBoost model on the B-cell dataset. The ROC curve and the confusion matrix illustrate a better model. the increasing true positive value in the confusion matrix shows the ROC curve's increased sensitivity and a higher cut off value. The observed cut-off value is the point in the graph where the ROC curve has the maximum "sensitivity + specificity − 1" value.

Figs. 20 and 21 represents the confusion matrix and the ROC curve on applying the XGBoost model. The results obtained from AdaBoost and XGBoost are similar. Both the ROC curve shows the same cut off values. Thus, a better predicting model can also be acquired.

After applying all the machine learning models, a correlation matrix of all the models was made, comparing which combination of the model specifies the best results after plotting the matrix, as shown in Fig. 22. Gradient boosting and Random forest are the two machine learning models that have the best results. Thus, taking gradient boosting and random forest at the inner layer, the two models' combination was trained and fitted on XGBoost on the outer layer. Therefore, an ensemble (ref. to Section "Ensemble learning algorithm/proposed work") proposed was used to predict more accurate and efficient results for predicting the SARS-CoV virus.

Fig. 22 shows the correlation matrix used for the ensemble selection. This represents that the combination of gradient boosting with random forest and gradient boosting with XGBoost acquires the best results, i.e., 0.92 and 0.99, respectively. The matrix also shows that XGBoost and random forest also combine to give an accurate result of 0.93, concluding that all three models could be combined to form a stacked ensemble.

Figs. 23 and 24 represents the confusion matrix and ROC curve obtained by combining gradient boosting, random forest, and XGBoost to

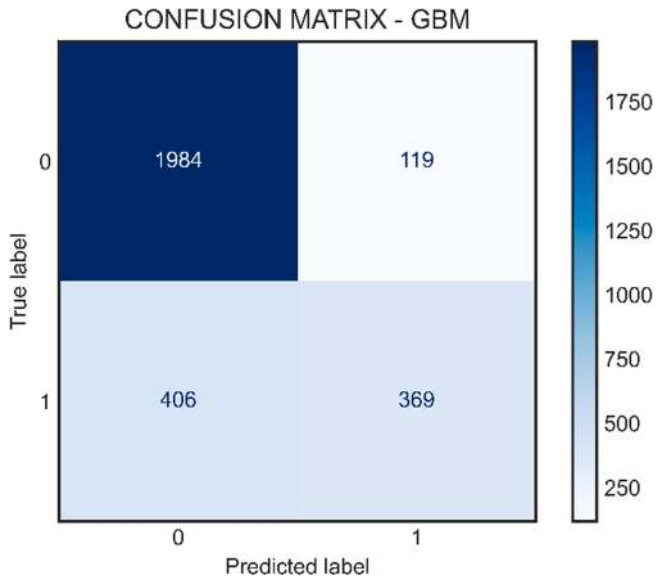

Fig. 14. Confusion matrix of Gradient boosting.

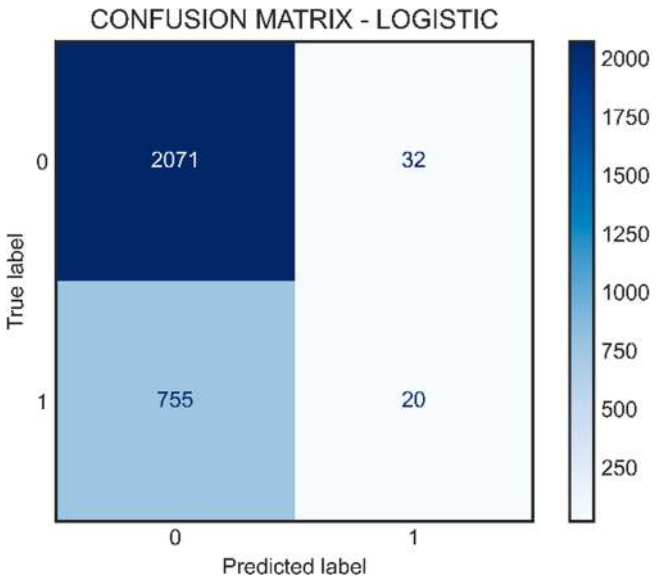

Fig. 16. Confusion matrix of Logistic regression.

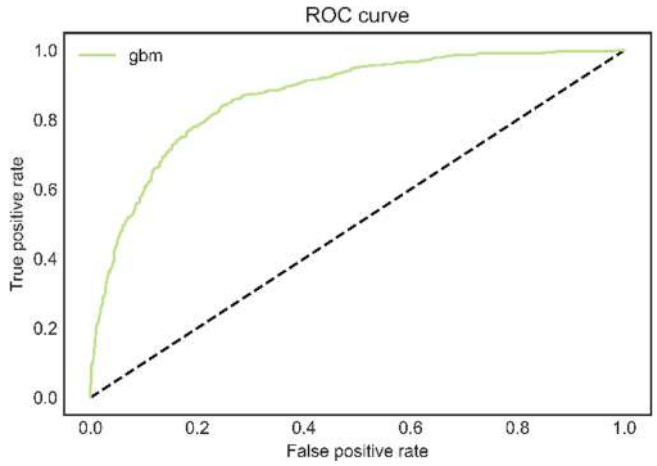

Fig. 15. ROC curve of Gradient boosting.

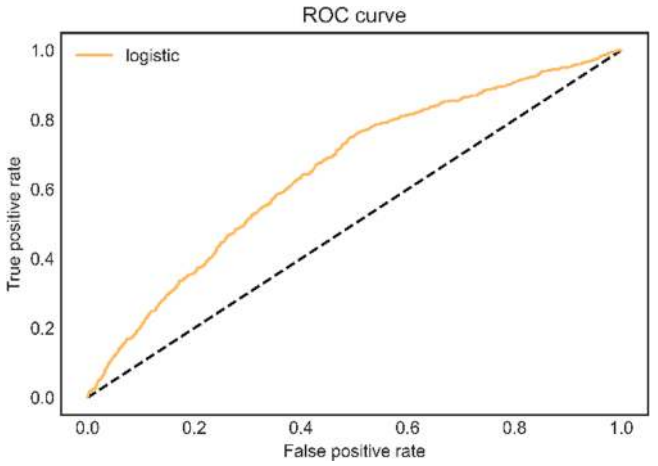

Fig. 17. ROC curve of Logistic regression.

form an ensemble. The confusion matrix and ROC curve show good results. It could be considered as a perfect model for making predictions for SARS-CoV causing coronavirus disease.

After applying all the machine learning models, a neural network, i.e., MLP-NN, was also used to make predictions based on the SARS-CoV virus's B-cell dataset (ref. to Section "Neural network algorithms"). After applying the neural network on the dataset Figs. 25 and 26 represents the confusion matrix and ROC obtained. The AUC score and accuracy obtained are exceptionally proficient and can be used efficiently for prediction purposes.

Fig. 27 represents the combined ROC curve of all the models applied for predicting the SARS-CoV virus. The combined ROC curve shows that random forest acquires the best output, followed by the ensemble outputs. All the ROC curves could be compared using Fig. 27.

On applying all the models and algorithms on the B-cell dataset, Table 2 represents the final results of calculating the validation accuracy, AUC score, MSE, MAE, and RMSE. The tabulated results of all the machine learning and algorithms applied to the B-cell dataset for predicting the SARS-CoV virus, causing coronavirus disease. The Table shows proposed ensemble is the algorithm that has performed the best out of all the models and ensembles applied in this paper. It has scored a validation accuracy of 87.2481% and a validation AUC of 0.919, which means that the model is efficient. The accuracy obtained is also good enough to make for making predictions on the given dataset. Thus, the Table also shows the error approximation is optimized in the case of a random forest. Thus, the overall results show that random forest has outperformed all the ensembles and machine learning models followed by the ensemble results as shown in Table 2.

After analyzing and discussing the validation for predicting the SARS-CoV virus, the B-cell dataset and SARS-CoV dataset (ref. to Section "Dataset description") were combined to predict the SARS-CoV-2 virus causing coronavirus disease. The processed dataset was evaluated with different machine learning models that helped predict the validation of the models.

The following are the results obtained after applying various machine learning models and algorithms:

Figs. 28 and 29 shows the confusion matrix and the ROC curve of SVM. The ROC curve shows a validation accuracy of 0.652, which is just a satisfactory result. The confusion matrix shows that the number of true positive value is greater than the false negative value. This represents

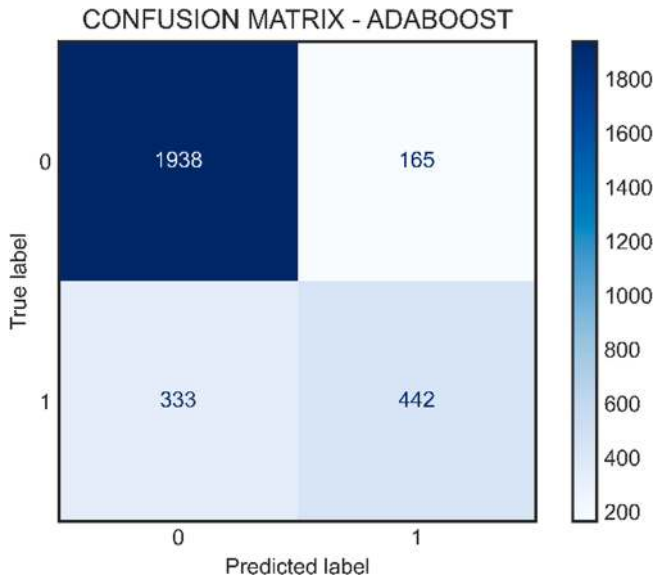

Fig. 18. Confusion matrix of AdaBoost.

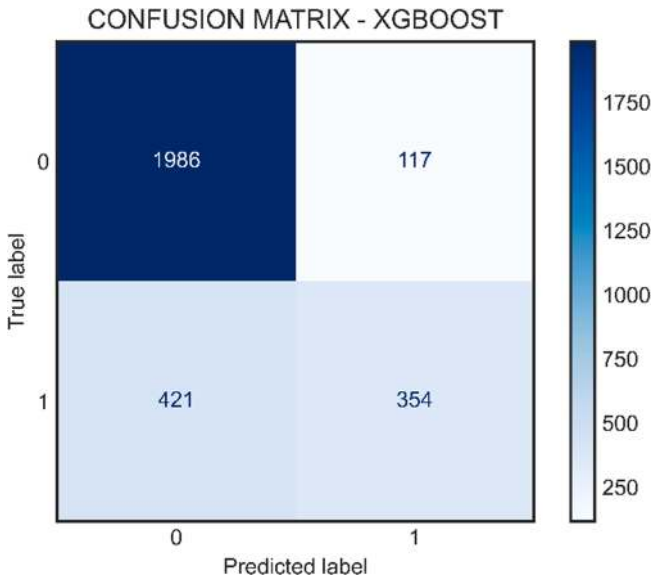

Fig. 20. Confusion matrix of XGBoost.

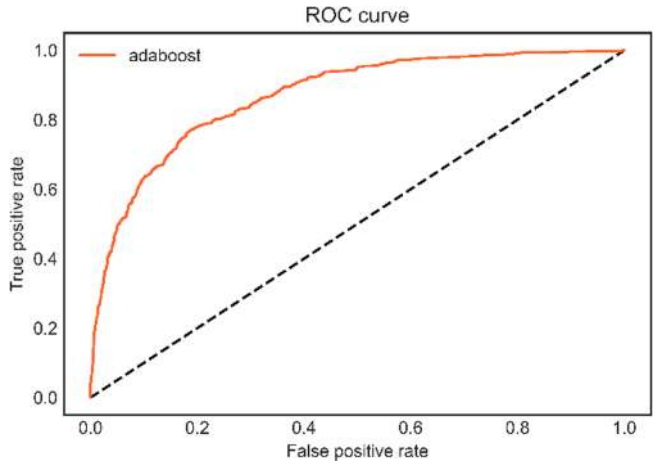

Fig. 19. ROC curve of AdaBoost.

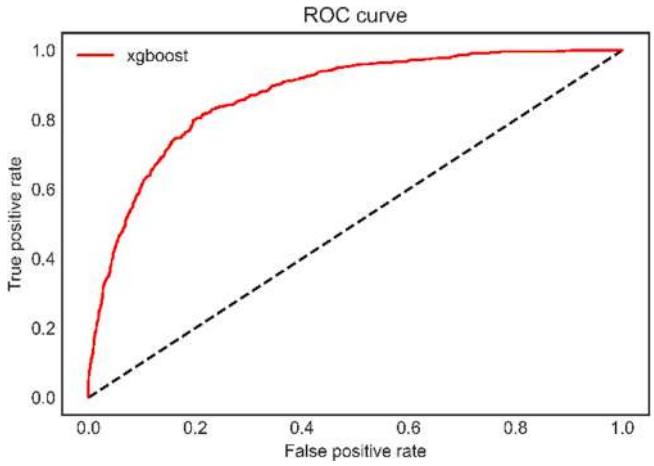

Fig. 21. ROC curve of XGBoost.

that the model can perform satisfying results.

Figs. 30 and 31 shows the confusion matrix and the ROC curve of KNN models. The ROC curve represents that the model predicts accurate results and can be used to predict efficient results. KNN can be used to make predictions of the SARS-CoV-2 virus as the results clearly show the accurate and proficient confusion matrix and a good enough ROC curve with an AUC score of 0.859.

Figs. 32 and 33 shows another baseline model that is used to predict the SARS-COV-2 virus. The results show an AUC score of 0.627, which is the same as that obtained by the SVM model. The ROC curve is not good enough but can be used in making predictions on small datasets.

After applying all the baseline models, inbuilt ensembles were applied to the dataset. Figs. 34 and 35 represents the confusion matrix and ROC curve using random forest, an inbuilt ensemble on the combined dataset to predict the SARS-CoV-2 virus. The output obtained is the best among all the other models and algorithms applied to the dataset. The ROC tilted towards the top left corner, signified its second-highest accuracy and can be confidently used in predicting the SARS-CoV-2 virus.

Figs. 36 and 37 shows the implementation results of another inbuilt model, gradient boosting. It represents the confusion matrix and ROC curve obtained after the application of gradient boosting on the dataset. The output obtained is good enough to consider it as a predicting model. The ROC curve has an AUC score of 0.873, which is enough to make an accurate prediction.

Figs. 38 and 39 shows the confusion matrix and the ROC curve of logistic regression, a baseline model. The ROC obtained from the logistic regression model is just a satisfactory outcome. It could only apply to make predictions over a small dataset.

On analyzing the results of various baseline models and inbuilt ensemble, AdaBoost was also implemented to make predictions. Figs. 40 and 41 discusses the implementation of the confusion matrix and ROC curve obtained after the performance of AdaBoost on the dataset. The ROC curve clearly shows a high cut-off value and a greater sensitivity in comparison.

XGBoost was also used to make predictions of the SARS-CoV-2 virus, causing coronavirus disease. Like AdaBoost, XGBoost also performed the same and gave nearly the same output. Figs. 42 and 43 shows the precise results that are just analyzing above in AdaBoost.

After applying all the machine learning models and algorithms, a

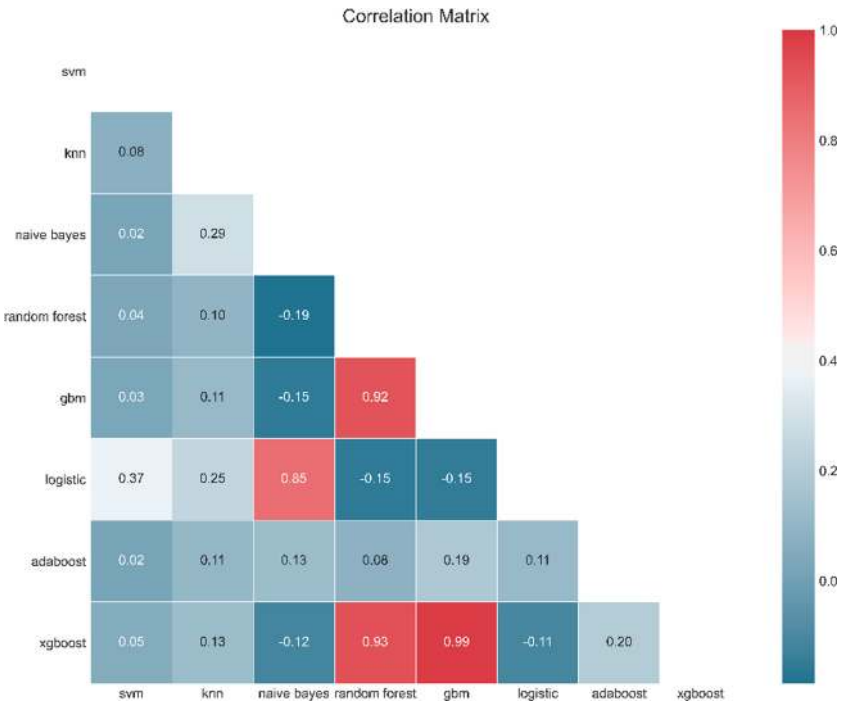

Fig. 22. Correlation matrix for ensemble selection.

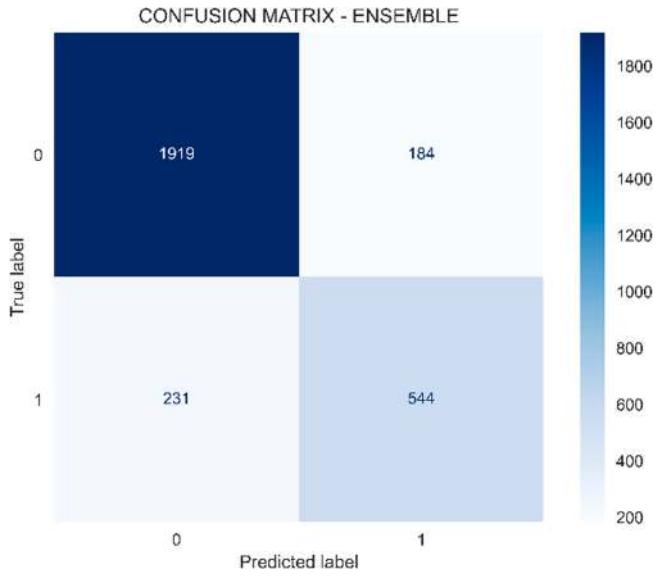

Fig. 23. Confusion matrix of Ensemble.

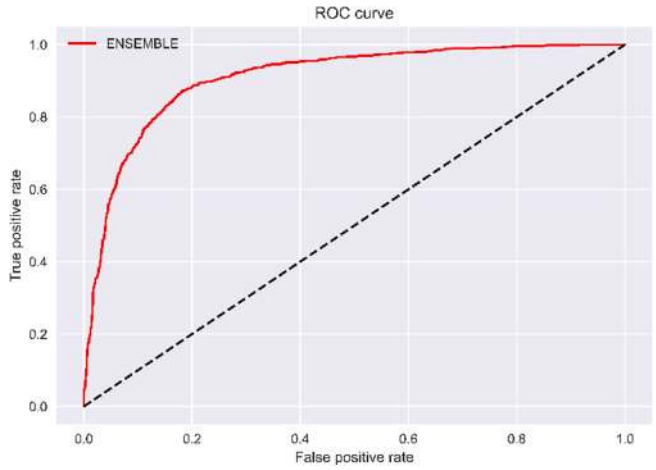

Fig. 24. ROC curve of Ensemble.

correlation matrix for the models was also plotted to check the best combination for making a more accurate and precise prediction of the SARS-CoV-2 virus causing coronavirus disease. Fig. 44 shows the results obtained after plotting the correlation matrix of all the models. The Figure clearly shows that XGBoost, random forest, and gradient boosting are the three algorithms that combine to form a stacked ensemble with XGBoost on the outer layer and random forest, gradient boosting on the inner layer.

After implementing the ensemble formed, Figs. 45 and 46 shows the output obtained, i.e., the confusion matrix and ROC curve as shown. The ROC curve justifies that the ensemble formed outperformed just like a random forest. The ROC curve scored an AUC score of 0.923, which indicates that the model formed is proficient and accurate enough to predict the SARS-CoV-2 virus.

Figs. 47 and 48 shows the confusion matrix and ROC curve on applying neural networks, i.e., MLP-NN on the dataset. The ROC curve shows that the model could be used as a predicting model as it shows better results than other models. The Confusion matrix depicts that the number of positive predictions is well defined and performing with a high accuracy rate.

After analyzing the results obtained from all the machine learning models and algorithms. Fig. 49 shows the comparison of all the ROC curves, which clearly indicates that the ensemble has outperformed all the other models with an AUC score of 0.923 followed by the random forest with an AUC of 0.914 for predicting SARS-CoV-2 virus causing coronavirus disease.

On analyzing all the graphs and matrices, Table 3 represents the

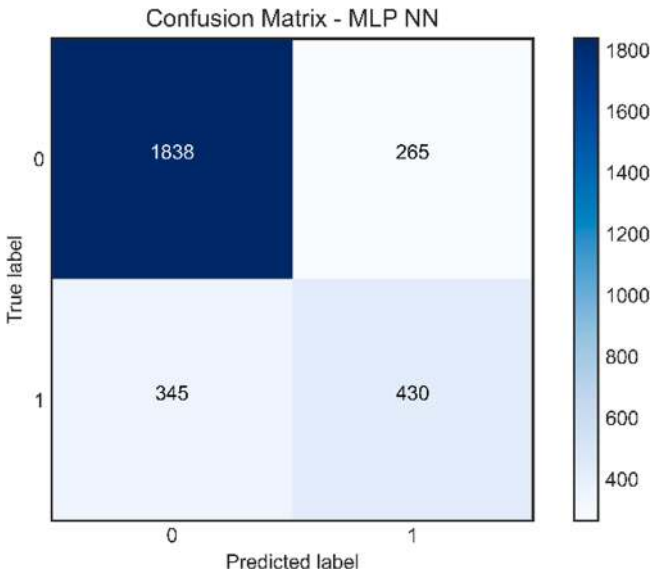

Fig. 25. Confusion matrix of MLP-NN.

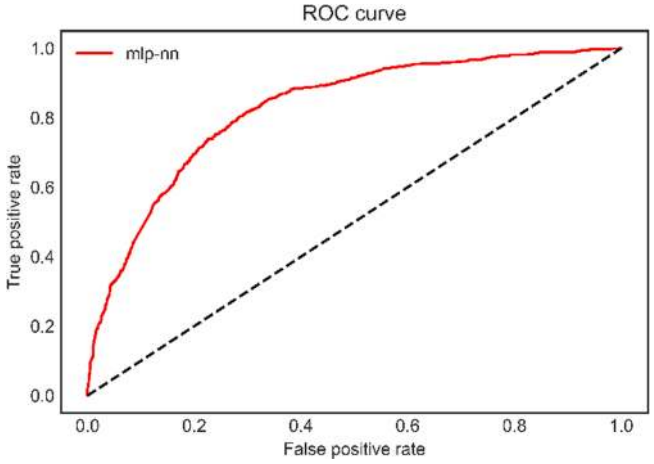

Fig. 26. ROC curve of MLP-NN.

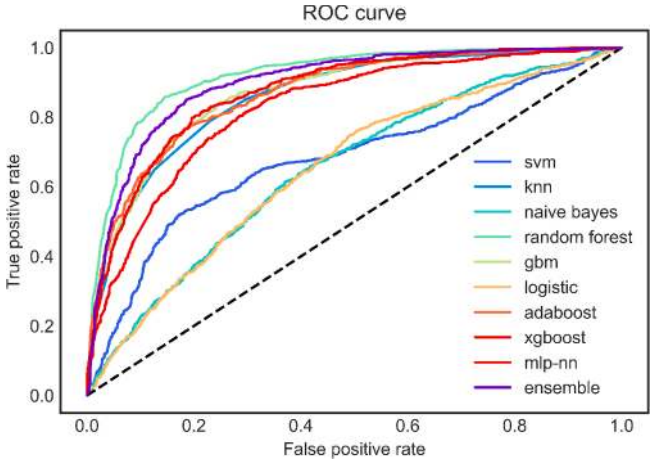

Fig. 27. Comparison of ROC curves.

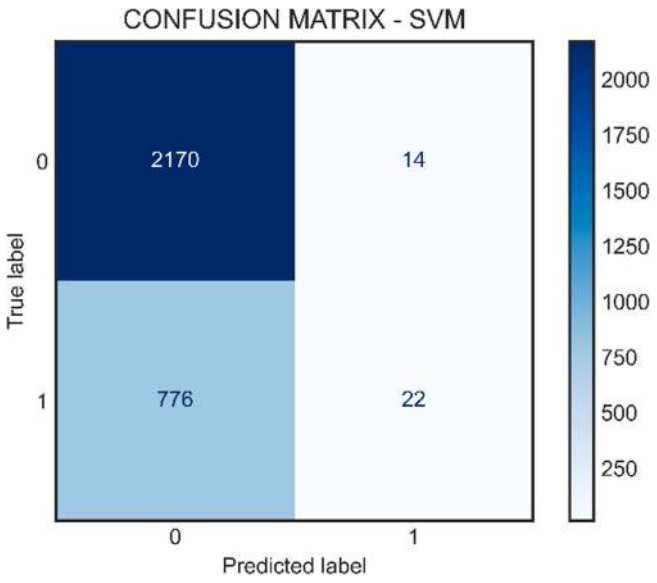

Fig. 28. Confusion matrix of SVM.

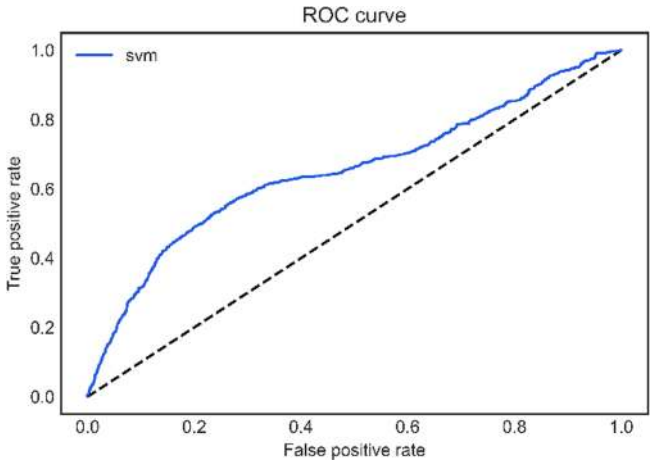

Fig. 29. ROC curve of SVM.

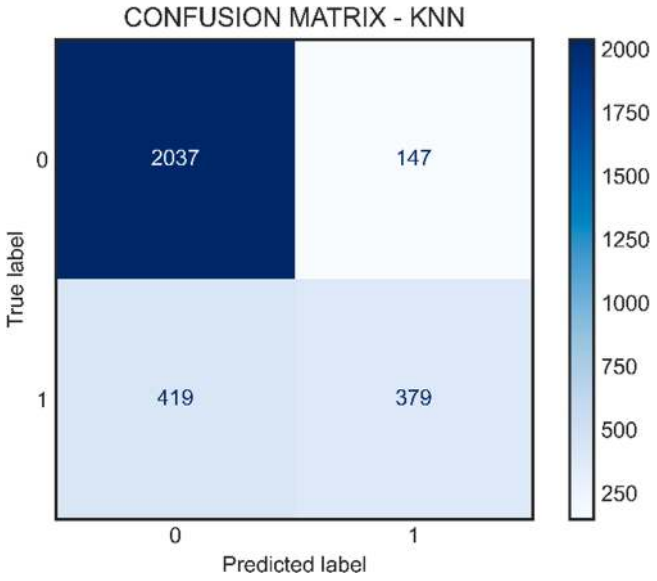

Fig. 30. Confusion matrix of KNN.

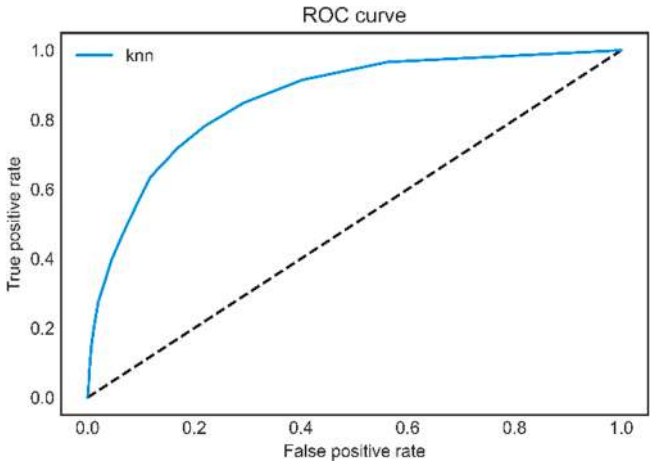

Fig. 31. ROC curve of KNN.

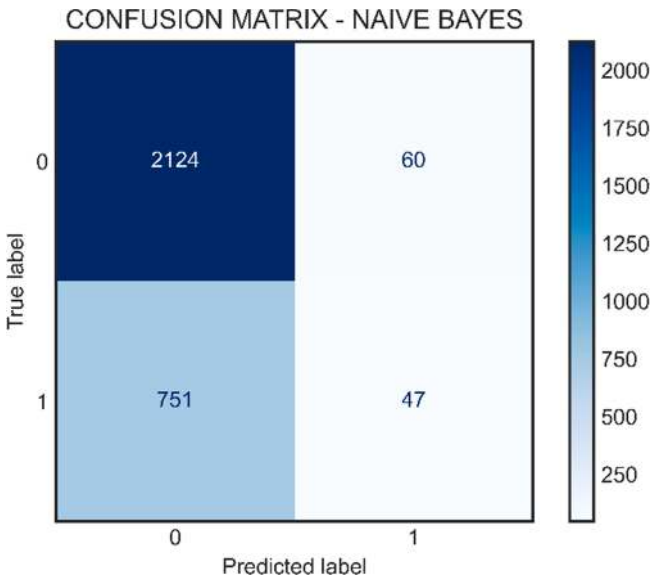

Fig. 32. Confusion matrix of Naïve Bayes.

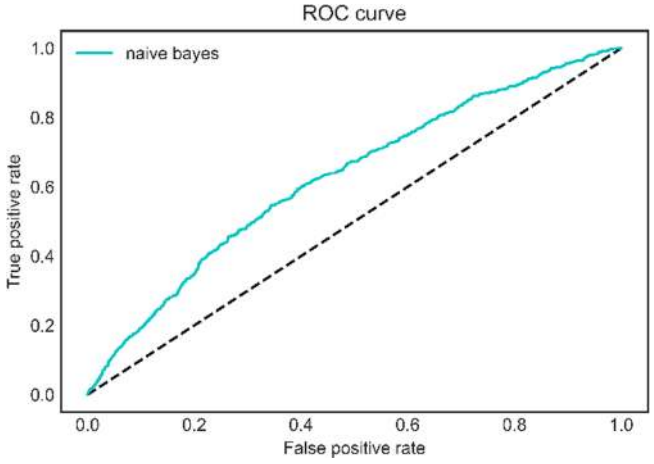

Fig. 33. ROC curve of Naïve Bayes.

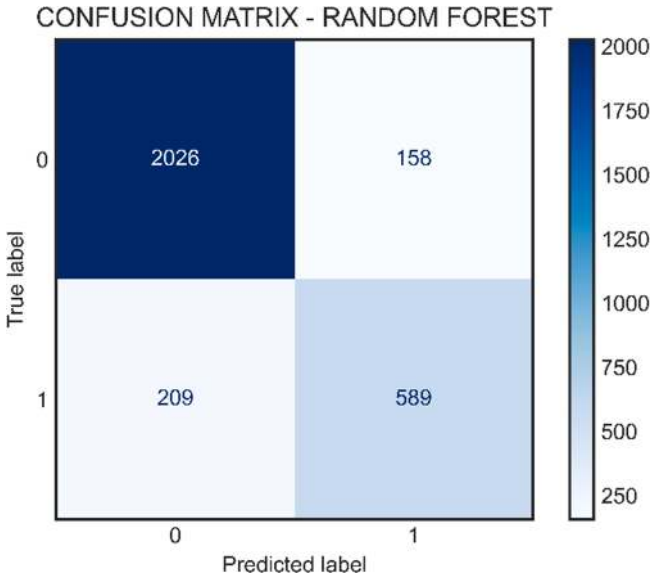

Fig. 34. Confusion matrix of Random forest.

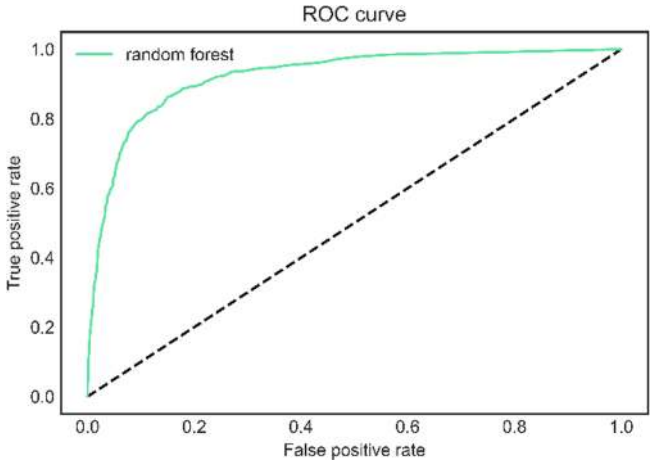

Fig. 35. ROC curve of Random forest.

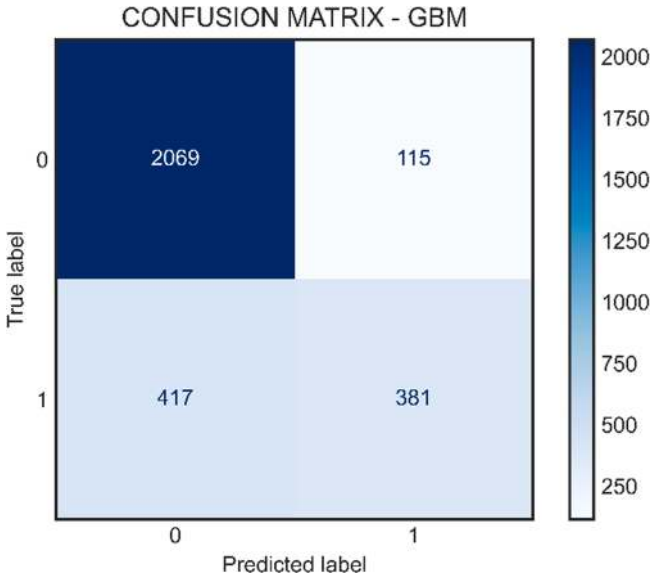

Fig. 36. Confusion matrix of Gradient Boosting.

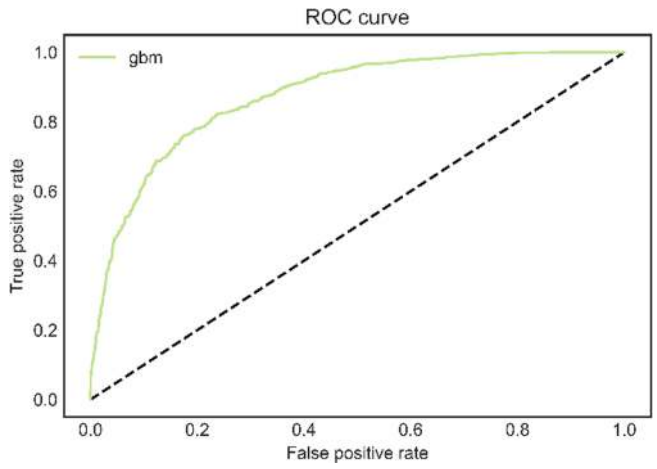

Fig. 37. ROC curve of Gradient boosting.

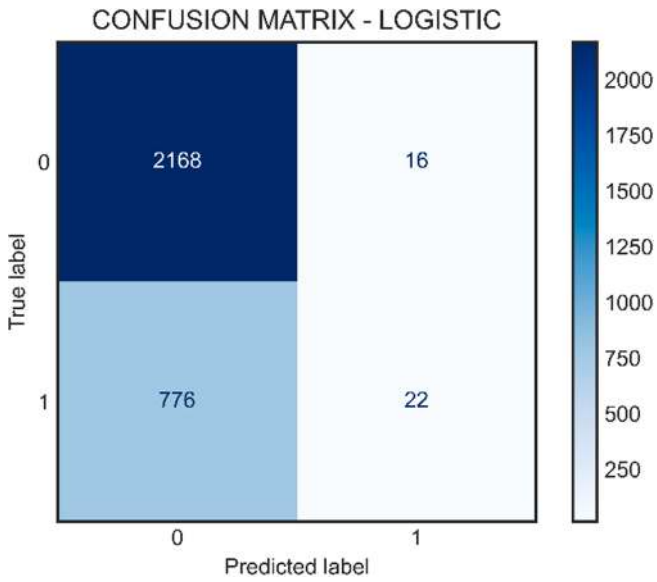

Fig. 38. Confusion matrix of Logistic regression.

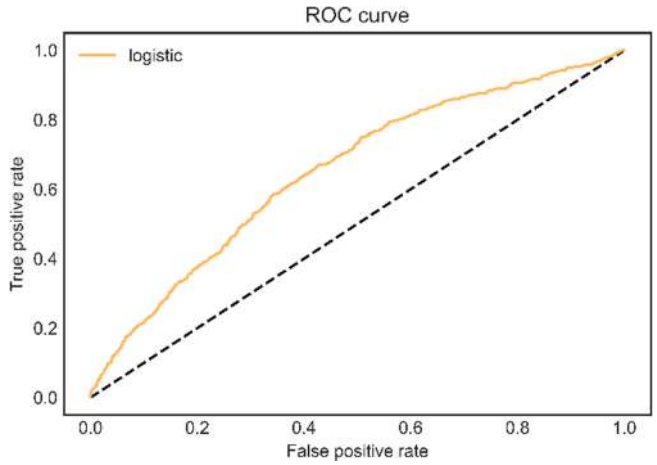

Fig. 39. ROC curve of Logistic regression.

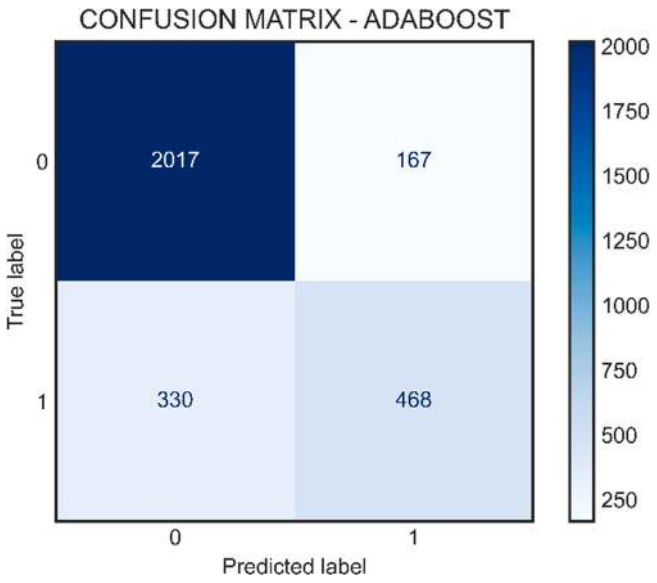

Fig. 40. Confusion matrix of AdaBoost.

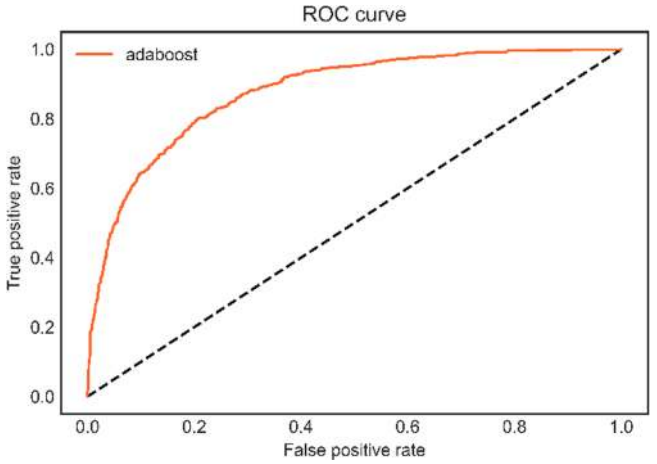

Fig. 41. ROC curve of AdaBoost.

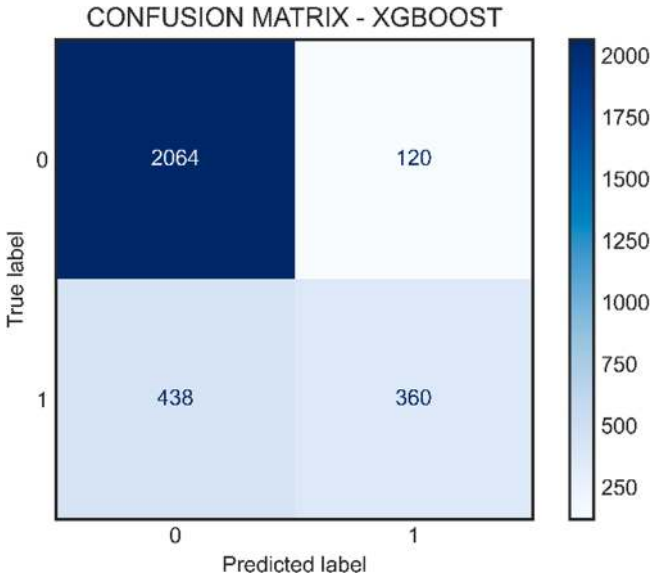

Fig. 42. Confusion matrix of XGBoost.

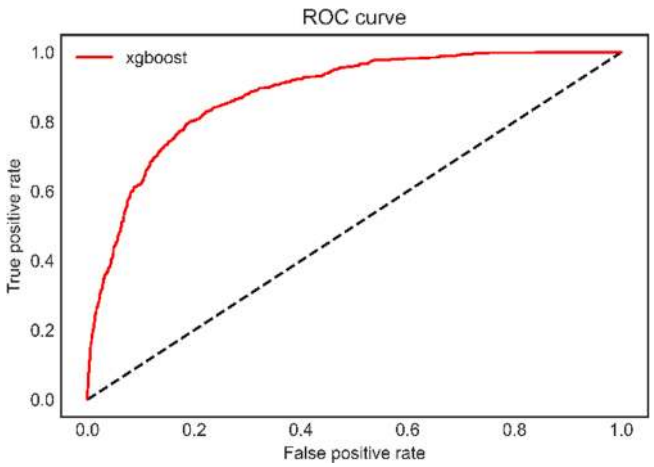

Fig. 43. ROC curve of XGBoost.

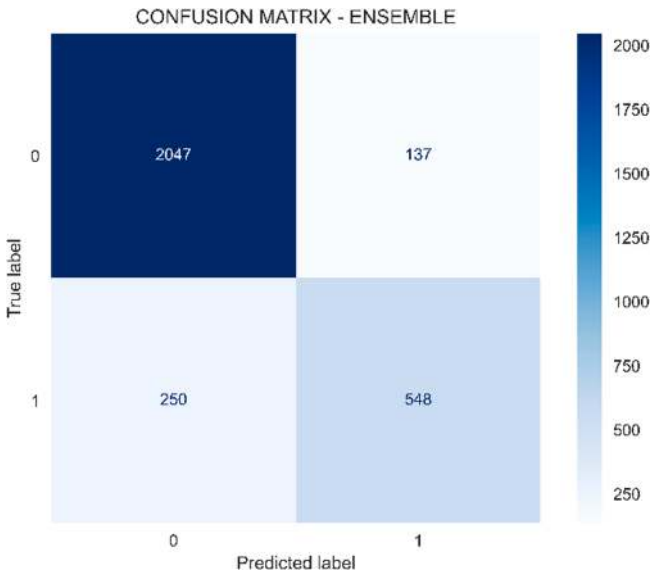

Fig. 45. Confusion matrix of Ensemble.

complete data of the predictions made for predicting SARS-CoV-2 virus causing coronavirus disease. The Table contains the validation AUC score, validation accuracy, MAE, MSE, RMSE of all the models, concluding that the ensemble performs the best among all the other models with a validation accuracy of 87.7934 and an AUC score of 0.923.

Conclusion

Coronavirus is a majorly spreading disease all across the world. In such a pandemic situation, predicting and analyzing if a patient is suffering from SARS-CoV and SARS-CoV-2, causing coronavirus disease, would ease the number of tests and gatherings as well. Thus, predicting these viruses could be considered beneficial for the researchers by getting the complete description and significant symptoms, causing coronavirus to make vaccines as soon as possible. Based on the mentioned hypothesis, the paper implements various models and algorithms on different combinations of B-cell datasets and many others to predict SARS-CoV and SARS-CoV-2, causing coronavirus disease. The paper briefly discusses all the baseline and inbuilt machine learning models used to predict the viruses. The paper also aims at a new stacked ensemble, which was also implemented to make predictions on the dataset. After applying all the machine learning models and algorithms, the best results came out to be of random forest with an AUC score of 0.919 followed by the ensemble with an AUC score of 0.908 for predicting the SARS-CoV virus. These are the validation results obtained after applying various models on the B-cell dataset. After analyzing the validation results, the labeled SARS-CoV dataset was used to find the test

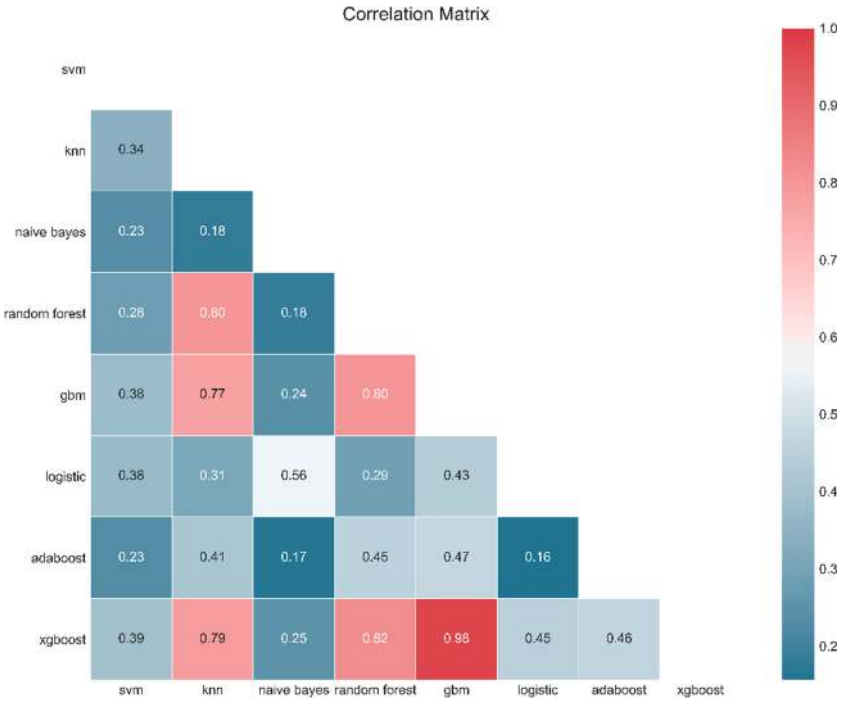

Fig. 44. Correlation matrix for model selection.

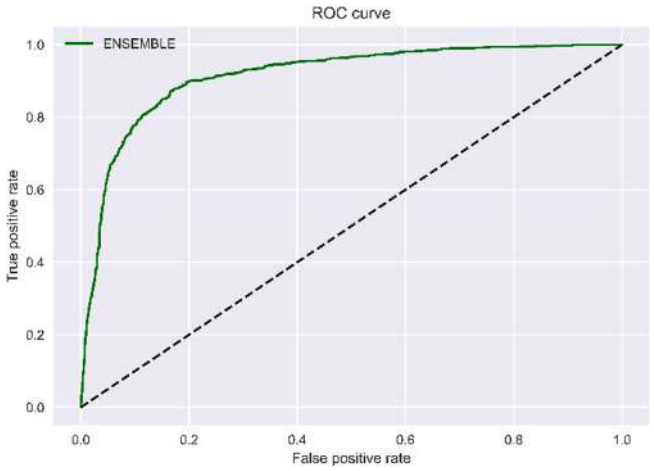

Fig. 46. ROC curve of Ensemble.

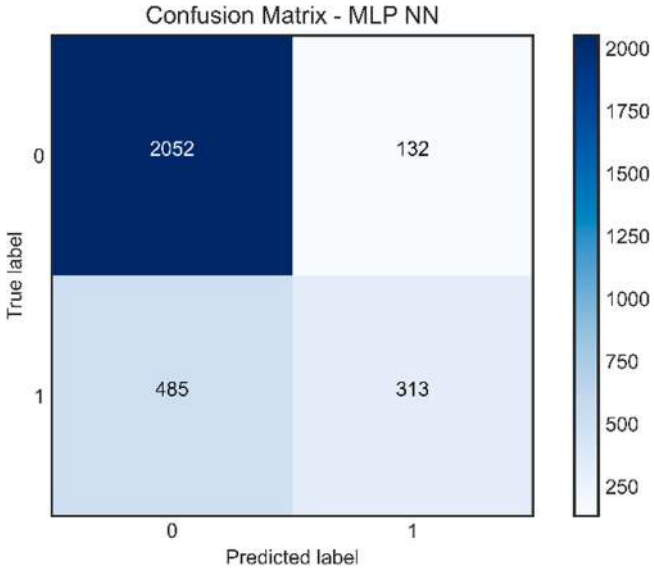

Fig. 47. Confusion matrix of MLP-NN.

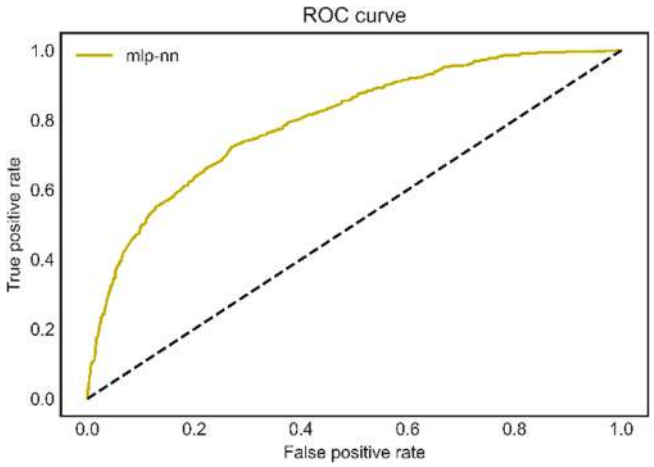

Fig. 48. ROC curve of MLP-NN.

results on the models implemented. Though the results are not proficient enough because the SARS-CoV dataset contains a smaller number of positive values in the dataset and the dataset is small and not precise

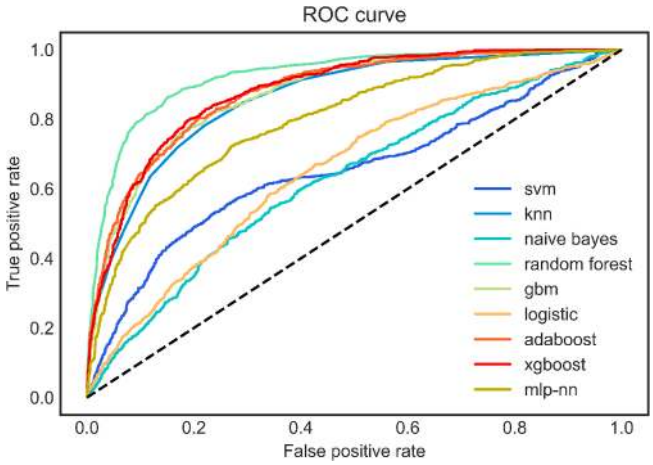

Fig. 49. Comparison of ROC curves.

Table 3
SARS-CoV-2 validation results.

Model	Validation AUC	Validation accuracy (%)	MSE	RMSE	MAE
SVM	0.652	73.7425	0.2626	0.5124	0.2626
KNN	0.859	81.0195	0.1898	0.4357	0.1898
Naive-Bayes	0.627	72.8035	0.2719	0.5215	0.2719
Random Forest	0.914	87.0221	0.1220	0.3494	0.1220
GBM	0.873	82.1596	0.1784	0.4224	0.1784
Logistic	0.654	73.4406	0.2656	0.5135	0.2656
AdaBoost	0.877	83.3333	0.1667	0.4082	0.1667
XGBoost	0.880	81.2877	0.1871	0.4326	0.1871
Ensemble	0.923	87.7934	0.1298	0.3602	0.1298
MLP-NN	0.810	79.1412	0.2086	0.4567	0.2086

enough to obtain the test results to justify all the models implemented compared to the B-cell dataset. Thus, after acquiring the desired results for predicting the SARS-CoV virus, the B-cell dataset, and SARS-CoV dataset were combined to make predictions for the SARS-CoV-2 virus. Various models and algorithms and a new stacked ensemble were used to predict the SARS-CoV-2 virus. The ensemble outperformed with an AUC of 0.923, followed by a random forest with an AUC of 0.914. Thus, predictions are made for both the viruses and results to be accurate enough to make predictions. An improvement can be observed if the SARS-CoV dataset contains more labeled data and a considerable number of positive values. This could improve the accuracy of predicting the SARS-CoV-2 virus and contribute to verifying the test results for predicting the SARS-CoV virus. Also, labeling the COVID dataset would ease the verification of the models and ensembles applied, improving the accuracy of the different machine learning models and algorithms.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Davis RE, Ngo VN, Lenz G, Tolar P, Young RM, Romesser PB, et al. Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. Nature 2010;463 (7277):88–92.
[2] Pieper K, Grimbacher B, Eibel H. B-cell biology and development. J Allergy Clin Immunol 2013;131(4):959–71. https://doi.org/10.1016/j.jaci.2013.01.046.
[3] Huang C, Liu WJ, Xu W, Jin T, Zhao Y, Song J, Wen H. A bat-derived putative cross-family recombinant coronavirus with a reovirus gene. PLoS Pathogens 12(9); 2016: e1005883.

[4] Ji L-N, Chao S, Wang Y-J, Li X-J, Mu X-D, Lin M-G, Jiang R-M. Clinical features of pediatric patients with COVID-19: a report of two family cluster cases. World J Pediatr 2020;16(3):267–70. https://doi.org/10.1007/s12519-020-00356-2.
[5] Pinto D, Park YJ, Beltramello M, Walls AC, Tortorici MA, Bianchi S, et al. Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. Nature 2020:1–6.
[6] Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med 2020;26(4):450–2. https://doi.org/10.1038/s41591-020-0820-9.
[7] Lu X, Zhang L, Du H, Zhang J, Li YY, Qu J, Zhang W, Wang Y, Bao S, Li Y, Wu C, Liu H, Liu Di, Shao J, Peng X, Yang Y, Liu Z, Xiang Y, Zhang F, Silva RM, Pinkerton KE, Shen K, Xiao H, Xu S, Wong GWK. SARS-CoV-2 infection in children. N Engl J Med 2020;382(17):1663–5. https://doi.org/10.1056/NEJMc2005073.
[8] van Doremalen N, Bushmaker T, Morris DH, Holbrook MG, Gamble A, Williamson BN, Tamin A, Harcourt JL, Thornburg NJ, Gerber SI, Lloyd-Smith JO, de Wit E, Munster VJ. Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. N Engl J Med 2020;382(16):1564–7. https://doi.org/10.1056/NEJMc2004973.
[9] Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Peters B. The immune epitope database (IEDB): 2018 update. Nucl Acids Res 47(D1); 2019: D339–D343.
[10] Longacre Jr A. U.S. Patent No. 5,233,169. Washington, DC: U.S. Patent and Trademark Office; 1993.
[11] Poran A, Harjanto D, Malloy M, Arieta CM, Rothenberg DA, Lenkala D, van Buuren MM, Addona TA, Rooney MS, Srinivasan L, Gaynor RB. Sequence-based prediction of SARS-CoV-2 vaccine targets using a mass spectrometry-based bioinformatics predictor identifies immunogenic T cell epitopes. Genome Med 2020;12(1). https://doi.org/10.1186/s13073-020-00767-w.
[12] Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. Cell Host Microbe; 2020.
[13] Ivanov D. Predicting the impacts of epidemic outbreaks on global supply chains: a simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case. Transp Res Part E: Logist Transp Rev 2020;136:101922.
[14] Bullard J, Dust K, Funk D, Strong JE, Alexander D, Garnett L, et al. Predicting infectious SARS-CoV-2 from diagnostic samples. Clin Infect Dis 2020.
[15] Nomi T, Inoue S, Fujita H, Sadamitsu K, Sakaguchi M, Tenma A, et al. Epitope prediction of antigen protein using attention-based LSTM network. BioRxiv 2020.
[16] Chen YW, Yiu C-P, Wong K-Y. Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CLpro) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. F1000Res 2020;9:129. https://doi.org/10.12688/f1000research.22457.2.
[17] Dietterich TG. Ensemble learning. Handbook Brain Theory Neural Netw 2002;2:110–25.
[18] Jordan MI, Rumelhart DE. Forward models: supervised learning with a distal teacher. Cogn Sci 16(3); 1992: 307–354.
[19] Schlkopf B, Smola AJ, Bach F. Learning with kernels: support vector machines, regularization, optimization, and beyond. The MIT Press; 2018.
[20] García-Laencina PJ, Sancho-Gómez JL, Figueiras-Vidal AR, Verleysen M. K nearest; 2009.
[21] Tzanos G, Kachris C, Soudris D. Hardware acceleration on gaussian naive bayes machine learning algorithm. In 2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST). IEEE; 2019. pp. 1–5.
[22] Liaw A, Wiener M. Classification and regression by randomForest. R News 2002;2(3):18–22.
[23] Xu Q, Xiong Y, Dai H, Kumari KM, Xu Q, Ou HY, Wei DQ. PDC-SGB: Prediction; 2017.
[24] Kleinbaum DG, Dietz K, Gail M, Klein M, Klein M. Logistic regression. New York: Springer-Verlag; 2002.
[25] Liu H, Tian H-q, Li Y-F, Zhang L. Comparison of four Adaboost algorithm based artificial neural networks in wind speed predictions. Energy Convers Manage 2015;92:67–81. https://doi.org/10.1016/j.enconman.2014.12.053.
[26] Chen T, He T, Benesty M, Khotilovich V, Tang Y. Xgboost: extreme gradient boosting. R Package Version 2015;(4-2):1–4.
[27] Zare M, Pourghasemi HR, Vafakhah M, Pradhan B. Landslide susceptibility mapping at Vaz Watershed (Iran) using an artificial neural network model: a comparison between multilayer perceptron (MLP) and radial basic function (RBF) algorithms. Arab J Geosci 2013;6(8):2873–88. https://doi.org/10.1007/s12517-012-0610-x.
[28] Zheng A, Casari A. Feature engineering for machine learning: principles and techniques for data scientists. O'Reilly Media, Inc.; 2018.
[29] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning; 2006. pp. 233–240.
[30] Atangana E, Atangana A. Facemasks simple but powerful weapons to protect against COVID-19 spread: can they have sides effects? Results Phys 2020;19:103425. https://doi.org/10.1016/j.rinp.2020.103425.
[31] Atangana A. Modelling the spread of covid-19 with new fractal-fractional operators: can the lockdown save mankind before vaccination? Chaos Soliton Fractals 2020;136(1):01–18.
[32] Atangana A. Extension of rate of change concept: from local to nonlocal operators with applications. Results Phys 2020;19:103515. https://doi.org/10.1016/j.rinp.2020.103515.
[33] Atangana A, Araz SI. Nonlinear equations with global differential and integral operators: existence, uniqueness with application to epidemiolog. Results Phys 103593.