Available online freely at www.isisn.org

Bioscience Research

Print ISSN: 1811-9506 Online ISSN: 2218-3973

Journal by Innovative Scientific Information & Services Network

RESEARCH ARTICLE BIOSCIENCE RESEARCH, 2021 18(SI-1): 47-58. **OPEN ACCESS**

Identification of Novel Protein Sequencing SARS CoV-2 Coronavirus Using Machine Learning

Ali Ghulam[1], Mukhtiar Memon[2], Mansoor Hyder[2], Zulfikar Ahmed Maher[2], Ahsanullah Unar[3], Zar Nawab Khan Swati[4], Dhani Bux Talpur[5], Rahu Sikander[6], Ikram Ullah[7], Ali Farman[8]

[1]Computerization and Network Section, Sindh Agriculture University, Tandojam, **Pakistan**
[2]Information Technology Centre, Sindh Agriculture University, Tandojam, **Pakistan**
[3]School of life Science, University of Science and Technology of **China.**
[4]Department of Computer Science, Karakoram International University Gilgit, **Pakistan**
[5]School of Information and Communication Engineering, Guilin University of Electronic Technology, Guilin **China**
[6]School of Computer Science, Xidian University, Xi'an, **China**
[7]College of Landscaping and Horticulture, Yunnan Agriculture University, Kunming 650201, C**hina**
[8]Nanjing University of Science and Technology, **China**

***Correspondence:** garahu@sau.edu.pk **Received 04-07-2021, Revised: 19-08-2021, Accepted: 20-08-2021 e-Published: 21-08-2021**

The World Health Organization (WHO) declared Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) infection as a global pandemic in March 2020 causing COVID-19 (Coronavirus Disease-19). Till date, more than 173 million people have been infected worldwide, whereas more than 3.7 million deaths have already been reported caused by COVID-19. Protein-to-protein (PPI) interaction plays an important role in the cellular process of SARS-CoV-2 virus infection in the human body. Although the recent emergence of SARS-CoV-2 has prompted a push for deeper understanding of SARS-CoV-2 and development of effective treatment. However, understanding SARS-CoV-2 is even more critical. It was previously discovered that the proteome of the virus was known, and thus it was possible to derive some of the protein structures by experimentation and others by model-based prediction approaches. The results are later verified by experiments. Considerable research attention has been directed toward DEE deploy features extraction algorithm, amino acid composition AAC and pseudo amino acid composition PseAAC algorithms. We have proposed AdaBoost classification models and compared them with other two machine learning classifiers, such as K-Nearest Neighbor and Random Forest. This paper is not intended to be a comprehensive evaluation of AdaBoost, K-Nearest Neighbor, and Random Forest, rather we have used these models to create an ensemble classifier with excellent performance metrics such as accuracy, precision, specificity, recall, and F1 score. Based on the ensemble model, 1326 total human target proteins are predicted to be potential SARS-CoV-2 viral proteins.

Keywords: SARS-CoV- 2 Machine learning, ML Classifications, Proteins sequences, Human Virus Proteins, Protein peptides

INTRODUCTION

This study is concerned with coronavirus disease (COVID-19) which is one of the most complex and deadly diseases of the present time according to the World health organization (WHO). The reason being the novel strain of coronavirus, known as the several acute respiratory syndromes coronavirus (ARS-CoV-2). According to WHO, more than 173 million people have been infected in

200 countries, whereas more than 3.7 million deaths have already been reported caused by COVID-19 (Lai, C. C., et al. 2020). In worldwide healthcare systems the disease has created huge pressure and tension. The first case of the new coronavirus infection was reported in Wuhan City of China by the end of 2019. It has now threatens the entire world from Asia to Europe and America (Ucar F, Korkmaz D. et al. 2020, Ibrahim IM, et al. 2020). The Genome analysis demonstrated the phylogenetic association of SARS-CoV-2 with SARS-like bat viruses. The potential source of viral replication could be bats (Ibrahim IM, et al. 2020), therefore. In addition, pangolins are also found to be the possible intermediate host for novel corona viruses (Kassani SH, et al. 2020). Pneumonia, breathlessness, cold, fever and multiple organ failure (Kumar R, et al. 2020) are the typical symptoms of COVID-19. In order to combat this virus, SARSCoV-2 genetic characteristics should be well understood. It is a single-strand RNA virus with a specific size of approximately 27e32 kb and a diameter of 65 to 125 nm. The international health systems are striving to find and try various vaccines that can reduce the disease spread. In addition, the infected patients are separated in isolation as emergency treatment and medical attention with some general medicine.

Many scientists are engaged in research on using Artificial Intelligence (AI) and Machine Learning (ML) in the area for various purpose including proteins functions, DNA sequence predictions, and genome sequence classification. We found the studies for the treatment of medical issues addressing the outbreak prediction of the pandemic SARS-CoV-2 in (Bullock J, et al. 2020, Ardabili SF,et al. 2020). SARS-CoV-2 human protein sequence dataset has been used and computerized tomography (CT) analyzed by Kassani et al. are publicly available (Barstugan M, et al. 2020). Machine learning methods are used to automatically classify SARS-CoV-2 affected protein sequence. In addition, the current research by Horry et al. (Horry MJ, et al. 2020) indicates that machine learning models to follow the same path as the SARS-CoV-2 human protein sequence. However, (Barstugan et al.), have used a 10-fold cross-validation with vector support machine (SVM) classifier only for CT images for the COVID-19 classification. The problem of over-servicing of machine learning can be posed because of the very few images available for model-fitting or training. These approaches have certain inconsistencies and limitations. For example, they have done the classification of more difficult cases with ambiguous, low contrast boundaries and artefact presence (Wang L, et al., 2020). These methods are extremely time-consuming, requiring additional space, computing resulting in cost and time overhead. The problem of portability is that enough human protein and CT images are obtained from SARS-CoV-2 patients during the learning process. Recent research has tended to show the proposed fusion and deep ranking of SARS-CoV-2 detection features (Ozkaya U, et al. 2020). Two (16x16 and 32x32) subsets of 150 CT images were produced in this work. SVM was added to the classification following the improvement of performance by deep functional fusion and ranking method. In addition to image analysis and prediction, work has been carried out on the promising SARS-CoV-2 drug discovery. Edison *et al,.* presented the newly developed Vaxign-ML vaccine learning tool for predicting SARS-CoV-2 candidates (Ong E, et al. 2020). In (Gysi DM, et al. 2020), the authors have employed COVID-19 network-led tools for retrieving the primary pulmonary manifestations of the lung virus and observed cardiovascular disease-related comorbidities. They predicted that the virus may develop through network proximity, diffusion, and AI-based techniques in some special tissues such as the reproductive system and areas of the brain. Similar work has been proposed in (Ge Y, et al. 2020, Batra R, et al. 2020) on the finding of the framework for data-driven drug repositioning using machine-learning and analytical statistical approaches. The research in (Desautels T, et al. 2020) has systematically integrated wide-range knowledge graph, literature, and transcription data in order to identify potential drug candidates for SARS-CoV2. In (D.M. et al.2020) a new machinery training, bioinformatics and supercomputing combination has been utilized to predict antibody structures that are able to target the receptor-binding domain SARS-CoV-2. On the other hand, (Y. Ge, T. et al.2020) has found potential results that indicate their antibody mutilators are likely to bind to RBD SARS-CoV-2 and nullify the virus. The knowledge of small molecular treatment against COVID-19 has been investigated by Batra and others (R. Batra, H. et al. 2020). It also supplied a pipeline for high-performance computer modeling and ensemble docking simulations for COVID-19 therapeutic agent screening.

This is a novel disease that affects the entire world and presents a difficult challenge for both developed and developing countries. While several previous cases have proven to be lower than predicted, that gives us hope that we will finally

stop this deadly virus before it spreads throughout the globe *(*Rubab, S., et al. 2021*)*. Patients with chronic illnesses experience COVID-19's pandemic breakout more frequently than those with other disorders. Strategies in the future should be devised for individuals with chronic illnesses who find themselves dealing with an epidemic *(*Ahmed, H. G., et al.2021*)*. Based on this study, it is determined that the current content may be utilised as a reference in the treatment of COVID-19 potentials. An even greater amount of clinical study is required to examine the effectiveness and safety of phytodrugs, as well as providing evidence for their efficacy (Iqbal, M., et al. 2020).

We did experiments to predict the human protein sequence functions using features extraction SARS-CoV-2 protein dependent on its protein sequences amino acid composition mixture, non-amino computer acid composition and joint triad features techniques for learning. The issue was presented as a double class issue of classification where both groups fit to the proteins that interact and non-interact both the virus and host. Current studies appear to support the notion that novel method of work as our experience indicates the AI environment. In the beginning, we used vector Dipeptide deviation from expected mean (DDE) technology pre-processing stage range. And, instead, we used some common supervised feature reduction learning algorithms like Random Forest (RF), K-Nearest Neighbor (KNN) together with the deep multi-layer model and functional ensemble (AdaBoost classifier) for classification and prediction results. We used 10-fold cross-validation, supervised strategy repeated process of learning. We achieved the accuracy of the test dataset result, the methodology of the compared classification ensemble more than one algorithm. We have therefore predicted SARS-CoV-1326 potential new human protein objectives 2 virus with algorithm ensemble. Ontology of chromosomes for these expected interactions investigated. We have also registered some re-purposable drugs that target the interactions predicted.

MATERIALS AND METHODS

Datasets

In this study, the benchmark datasets for 267 acute human-severe respiratory syndrome (SARS-CoV-2) were used again from the previous study (NCBI). The benchmark data is available. In order to take advantage of the learning process, however, more data have been gathered from the Uniprot database (Gammazza, et al. 2020) (http:/www.uniprot.org/) and a negative dataset of 167 protein sequence has been used. In https:/www.ncbi.nlm.nih.gov/ (Thompson, J. D. et al.) , we have downloaded data sets from NCBI DB to promote research into 297 proteins, which are a human-severe respiratory syndrome SARS-CoV-2 association, and then 167 Non- human-severe respiratory syndrome, SARS-CoV-2 association (Boopathi, S. , et al. 2020, Andersen, K. G. et al.) proteins. We omitted these protein sequences to ensure the suitability of suitable feature extraction methods. In fact, these sequences with the popular CD-HIT tool based on 30%, which are more than 30% percent similarity, have been removed equal to previous studies. The CD-HIT (Y. Huang, 2010) is given on http:/weizhong-lab.ucsd.edu/cdhit web-server. No replication between the SARSCoV-dataset and non- (SARSCoV-2) data set has been made non-human-severe acute respiratory syndrome coronavirus (SARSCoV-2), all non-redundant.

Table1: Data collection and normalization

	Original data	Similarity <30%	Used in experiment datasets
Positive SARS CoV-2	504	130	127
Negative SARS CoV-2	620	167	140
Total	1124	297	267

Feature's extraction techniques

The methods used to extract protein sequence-function based on Dipeptide Deviation from Expected Mean (DDE). Every protein can, therefore, be interpreted as a vector feature profile matrix. The following are the methods for extracting the functionality. The joint triad is a three-frequency distribution that encodes each protein sequence. We conducted all analyses using improvements including the development of human-severe acute respiratory syndrome coronavirus (SARSCoV-2), proteins sequence with the optimization of the DDE from expected mean (V. Saravanan, et al. 2015) and the 400- dimensional identification of SAP peptide with dipeptide. In the development of prediction models, such as (DDE) (V. Saravanan,et al. 2018), various forms of amino acid composition (AAC) of peptide sequences have also been introduced. A protein sample was used for sequencing details to the DDE based composition instead of the traditional AAC. There are 20 standard amino acids in a variety of proteins in an organism, and thus 20x20 = 400 potential DDE

features vectors.

Feature's The Optimal Feature Selection

The broad range of features undoubtedly includes redundant properties to maximize the performance and strength of the probability score. It is important to take out the related features that help most to make model predictions. The optimum subset function will reduce training and use time, reduce measuring and storage requirements, avoid over-fitting, and boost predictions. Several effective selection techniques for noise reduction or irrelevant effects, as well as for good predictive outputs have been suggested to date such as variance analysis, maximum / max gap, minimum max-relevance redundancy, key component parameter analysis and recurrent removal algorithm.

$$q_j = \frac{m\,j}{M} \quad (1)$$

Feature vector construction

The dipeptide composition of the protein has been used to describe the amino acid information of a given protein (Bhasin, M., et al.) that has been commonly used in various protein function methods (Gupta, S., et al. 2013). The report that the protein relationship was defined by a feature vector by a divergence of amino acid frequencies from respective expected average value.

In this analysis, we have used the dipeptide composition aspect since previous studies to measure a deviation of the dipeptide composition by dipeptide frequencies from the predicted mean values. In this research we have been using a dipeptide composition aspect. For Dipeptide deviation from the predicted mean (V. Saravanan, et al.2015) , we call this measure DDE (V. Saravanan, et al.2018). Three parameters have been constructed by DDE feature vector, i.e. Theoretical mean (Tm), and theoretical variation (Tv), calculation of dipeptide compositions (Dc).

The three above and the DDE parameters are calculated accordingly. DC (i), a measure of the composition of the dipeptide I is given in peptide P.

$$D_{c(i)} = \frac{n_i}{N}. \quad (2)$$

There will be a total of 400 features length for dipeptides (20 × 20 standard amino acids), but not everything will happen at all sequences, of course. Ni is dipeptide i-th frequency, and l-1 is N (i.e., P possible number of dipeptides). TM(i) is given to the theoretical mean.

$$T_{M(i)} = \frac{C_{i1}}{C_N} \times \frac{C_{i2}}{C_N} \quad (3)$$

For the first amino acid codon, Ci1 is the number, while in the given dipeptide 'i,' the second amino acid codons are the number of Ci2 code. CN is the total possible codons, with the exception of three (i.e., 61) stop codons. Since TM(i) does not depend on peptide P, 400 dipeptides were precomputed once in a while. TV(i), dipeptide I theoretical variance is indicated.

$$T_{v(i)} = \frac{T_{M(i)}\left(1 - T_{M(i)}\right)}{N} \quad (4)$$

The theoretical average of I is TM (i), computed with Eq. 2, N 2 The number of peptides in peptide P is again l-1. Again. In the end, the measurement of DDE (i) is

$$DDE_{(t)} = \frac{D_{c(i)} - T_{m(i)}}{\sqrt{T_{V(i)}}} \quad (5)$$

DDE was computed for each of the 400 feature dipeptides proteins. The 400- dimensional feature vector is as follows,

$$DDE_p = \left\{ DDE_{(i)}, \ldots, \ldots DDE_{(n)} \right\}, where, i = 1,2,\ldots,400 \quad (6)$$

.

Proposed AdaBoost Model

The framework models proposed to categories SARSCoV-2 protein sequence structure and function prediction as shown in Fig.1. The process series includes data preprocessing, data labelling and we used machine learning classification algorithms. The first machine learning model for SARS-CoV-2 protein sequence function is SARS-CoV-2 (Severe acute corona viral syndromes based on Ada Boost classifier for SARS-CoV-2 protein sequence function prediction). In this research, AdaBoost classifier was developed and used to reduce redundancy in the prediction model. SARS-CoV-2 AdaBoost consists of the evolution, metastatic and human transmission of a number of

algorithms and multiple acute respiratory syndromes. We have built a 2-stage selection feature protocol, which specifies a correct optimal selection protocol, to remove irrelevant functions on each encoding feature. A probability of five extraction methods was eventually used and then implemented on SARS-CoV-2 AdaBoost encoding models and compared to 3 classification models for machine learning. In addition, various feature combinations have been combined to produce hybrid speed. Following the similarity between AdaBoost and conventional machine learning, we use set K-10-fold cross-validation controls to test the classifier performance. Figure 1 shows the structure of the proposed model.

The knowledge was analyzed by various methods using AdaBoost model. In order to determine algorithms to imitate human brain patterns, the material analyzes have been carried out (Niu, B., et al. 2020). AdaBoost is close to conventional computer learning, but the two networks' training processes differ. Since then, AdaBoost has a true depth. The average gradient and corresponding weight adjustment are calculated before final outputs are measured using the output layer. Production of layer l shall be as follows:

Proposed classifications.

This paper is not intended to be a comprehensive analysis of the conversion of the COVID 19 protein sequence, different three classifications are sequentially used to predict protein sequence structure and functions,

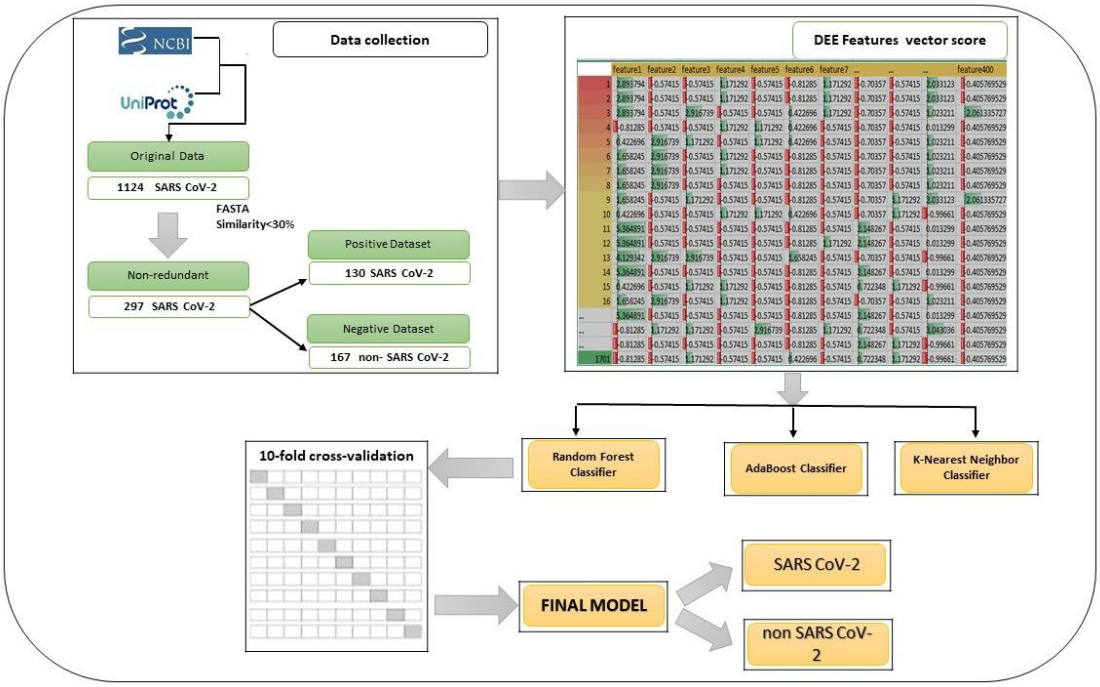

Figure 1. Network Architecture of the Proposed Framework

based on two labeling classes. In this classification model, each amino acid class considered extracted feature. The three classifiers used to label the predicted class are one number format and binary array format. The three classifications are conducted using AdaBoost, K-Nearest Neighbor (KNN), with a variety of neighbor numbers. The main aim of the classification structure is to reduce the learning complexity to examine the unknown samples accurately. The KNN classification is classified according to the K values of a class of its closest neighbors. The KNN classification focused on the density distribution distance measurement that did not relate to calculation of decision boundaries. The RF classification is based on the calculation of probability and the calculation of the maximum probability.

Protein structures related to COVID-19 prediction based on computational approaches.

We mainly focused on structure of a protein is an important resource to understand how it operates, but structure testing can take months or longer, and some prove insecure. There are computational methods developed by researchers to predict the

protein structure from the amino acid sequence. In our previous work, we used AlphaFold (Senior AW, et al. 2020) deep learning system, focuses precisely on the prediction of the protein structure where similar protein structures, called "free modelling," are not available. In current research, we have continued to improve these methods aiming to provide useful predictions, we therefore share predicted protein structures generated using our recently developed methods in SARS-CoV-2.

A recent line of research has established that our system of structure prediction is still in progress and the accuracy of the structures we provide cannot be assured, though we are confident that the system is more exact than our previous CASP13 system. The prediction of the experimentally defined SARS-CoV-2 spike protein structure shared by the Protein Data Bank was confirmed by the system and we were confident that our model predictions on other proteins could be useful. Our models include confidence values per residue to help indicate which structural sections are most likely to be correct. Protein predictions have only been given that are lacking adequate templates or otherwise difficult to model templates. Although the main focus of current theoretic effort is not on these understudied proteins, they may increase researchers' comprehension of SARS-CoV-2.

RESULTS AND DISCUSSION

The ranking of human-severe acute respiratory syndrome coronavirus (SARSCoV-2) and non-human-severe acute respiratory syndrome coronavirus (SARSCoV-2) has been evaluated by the results of 4 performance metric parameters, namely accuracy (Ac), sensitivity (Sn), specificity (Sp), and MCC. The absolute precision of measured accuracy classification of human-severe acute respiratory syndrome coronavirus and non-human-severe acute respiratory syndrome coronavirus (SARSCoV-2). The sensitivity and basic characteristics of the AdaBoost model, indicating the ability to recognize and recognize human-severe acute respiratory syndrome coronavirus (SARSCoV-2) correctly predicted. The following steps are normally written:

$$Sensitivity = \frac{TP}{TP+FN} \quad (7)$$

$$Specificity = \frac{TN}{TN+FP} \quad (8)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (9)$$

For this type of formula, the human-severe acute respiratory syndrome coronavirus (SARSCoV-2) and non-human-severe acute respiratory syndrome coronavirus (SARSCoV-2) numbers are indicated in the formulas above, respectively (True positive) and TN (True negative). Also, FP (false positive) and FN (false negative), respectively, are prediction to be non-human-severe acute respiratory syndrome coronavirus, are the number of non- human-severe acute respiratory syndrome coronavirus reported but predicted as carcinogenesis and the number of known human-severe acute respiratory syndrome coronavirus (SARSCoV-2). Also, with x-axis sensitivity and Y-axis precision, we have established receiver operating characteristic (ROC). The importance of the area under the ROC curve (AUC) is useful to evaluate model efficiency across the whole range of decision values.

Here is Matthew's similitude's versatility, similarity, precision and functionality. The connection between Sn, MCC and Acc and Sp has been calculated to determine Matthew's connection of similarities.

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (10)$$

Impact of extraction algorithm

For formulating the protein sequences using with AdaBoost, based on DDE feature extraction techniques we obtained the high output of our human severe acute respiratory syndrome coronavirus (SARSCoV-2) is complicated. These same data sets are used for the same data points for different classifications. The key selection technique (DEE) for the condensation of the hybrid space of the function was eventually introduced. This highlights the utility of the following different types of individual and hybrid features. the analyzes were conducted using the AdaBoost classification process. AdaBoost model results based on DEE approach have chosen the optimal parameter for the model growth with excellent

performance. The content analysis used as a metric to evaluate parameters of the correct function for predictability of human severe acute respiratory syndrome coronavirus (SARSCoV-2) is used to evaluate various parameter values. The values of O in l are set to 1, 2,3,4,5 and 6 respectively and the corresponding (SARSCoV-2) precision value is provided in Table 2 as the synthetic amino acid O in the dataset sample is provided.

The results are 10-fold cross-validation, our proposed AdaBoost classifier model performance is achieved (Chou, K. C., 2001, Zhou, G. P., 2003). This study used quantitative techniques to analyze based on 10-fold cross-validation with the difference between positive and negative data sets. The statistical analysis showed that our prediction performance was relatively excellent with an overall accuracy of 89.88%. And even the ROC curve is shown in Figure 2. This shows that the AUC approximates 92%, which indicates the very good estimation of our model. The data were analyzed using a DEE-extracting feature model, the human severe acute respiratory syndrome coronavirus (SARSCoV-2) achieves high ROC curve performance (auc=0.92%) and then the ROC curve also performs well.

Table 2: AdaBoost Identifying optimum parameter for various models.

	Cross-validation				Independent			
Classifiers	ACC	Sensitivity	Specificity	MCC	ACC	Sensitivity	Specificity	MCC
AdaBoost	89.88 %	87.14%	92.62%	79.41 %	0.9287 %	0.9090%	0.9484%	0.8744 %

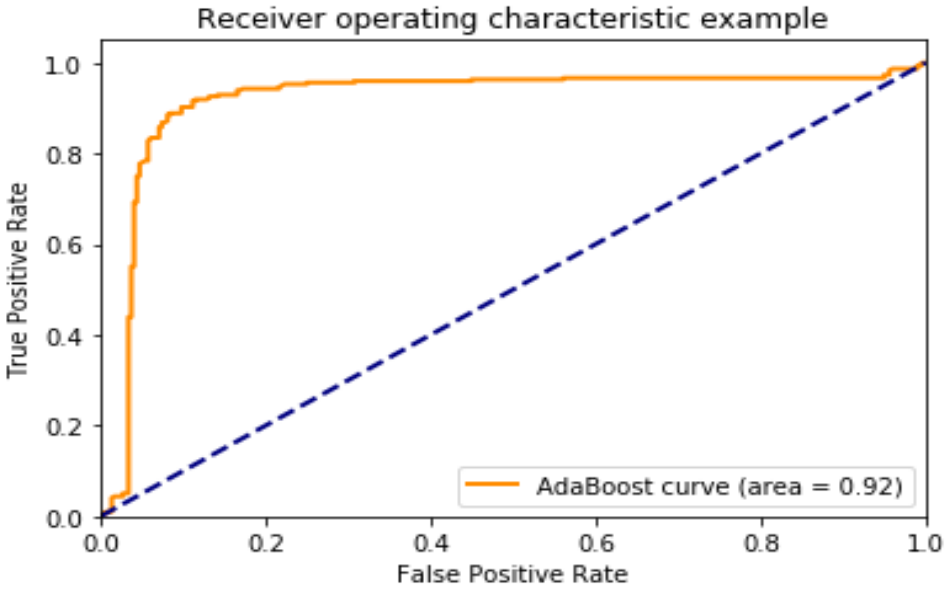

Figure 2: Proposed model ROC- auc score predictive models'

Proposed methods compare with 3 ML classification models.

The results analysis consists of comparison different classifier with stages. Results were considered significant and intuitively the validity of Figure 3. We have tested three 3 classification algorithms and found that the AdaBoost worked better for the nucleocapsid protein and spike protein model than for the other algorithms, while the random forest RF, are the best 2′-o-ribose algorithm. Table 3 summarizes the performance parameters for each classifier model result. Our

prediction models have been used to evaluate your classes with the top 100 compounds based on the lowest binding power (LBE) from every virtual screening output of three proteins. Our findings can be compared to results of earlier studies that human severe acute respiratory syndrome coronavirus (SARSCoV-2) process ROC-AUC curve involves ROC curves similar to other classifier approaches and the precision of AUC values are achieved; 87.14%, and ROC AUCs. ROC-AUC is the product of the procedure. Specificity: 92.62%. As shown in Figure 3. The findings for our proposed AdaBoost classifier model are complemented by 3 classification systems, such as the KNN classifier (KNN) (P. Cunningham, 2007), Random Forest Classifier (RFC) (Breiman, L. 2001), the AdaBoost Classifier (ABC) (Jin, Y. H., et al. 2008).

Classifying results using the hybrid features extraction models.

Table 3. Demonstrate the classifier's predictive effectiveness using the room for the DDE hybrid characteristics. But the AdaBoost classification provided excellent results in this mixed comparison of features. Following the use of data foreseen in the AdaBoost DEE model classification, a score of 93.53% was greater than that of the DEE expected AdaBoost model classification. Furthermore, as shown in table 3, we presented a comparison result with 3 MLCs.

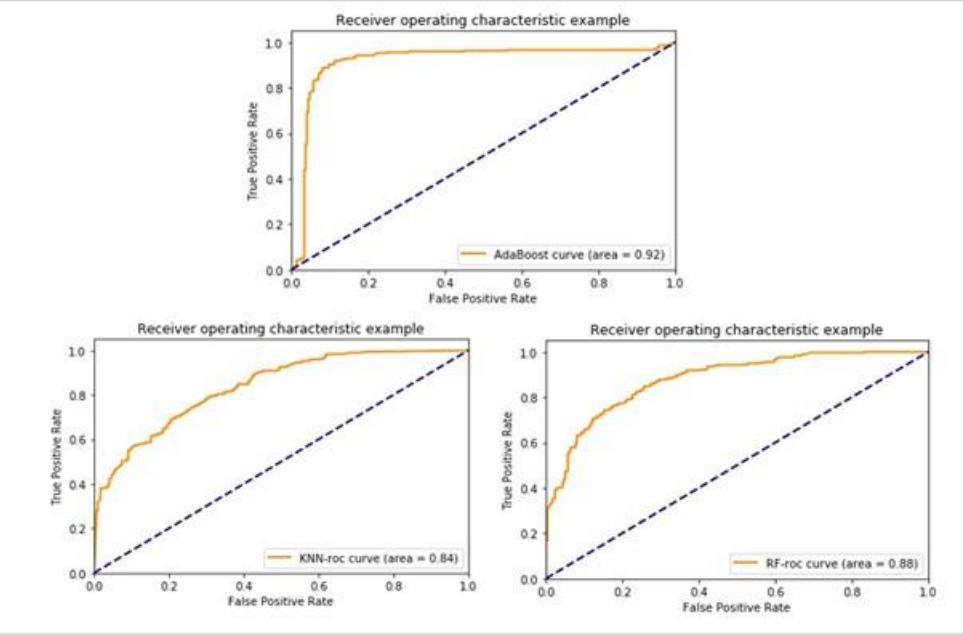

Figure 3: Proposed model compare with other classifiers

Table 3: The efficiency of different methods on the basis of classifiers on various feature extraction schemes.

	Cross-validation				Independent			
	ACC	Sensitivity	Specificity	MCC	ACC	Sensitivity	Specificity	MCC
AdaBoost	89.88%	87.14%	92.62%	79.41%	0.9287%	0.9090%	0.9484%	0.8744%
RF	78.62%	75.29%	81.95%	59.37%	0.8621%	0.8174%	0.9068%	0.7555%
KNN	74.05%	71.14%	76.95%	49.18	0.8746%	0.8257%	0.9234%	0.7798%

Identifying a comparison of combined 3 three classification based on feature extraction approaches.

In particular ROC AUC curves score that suit human severe acute respiratory syndrome coronavirus (SARSCoV-2) which contain ROC corresponding to one feature extraction approaches are DEE and then used AdaBoost classification achieved ROC (auc=0.92%) score. In the figure 4 are seen comparison results, as that our proposed AdaBoost classification performance is better than other two machine learning classifiers according to these ROC-AUC performance of AdaBoost classifier model based on (DDE) features profile vector matrix as shown in figure 4.

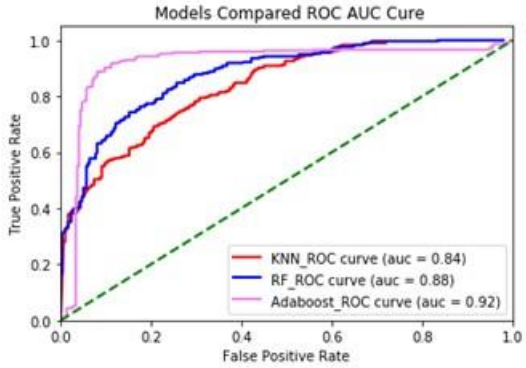

Figure 4: Proposed model ROC-auc score result.

Identifying a comparison of combined 3 three classification based on feature extraction approaches.

The best result for the DDE features profile vector matrix used with AdaBoost classifier model and then two other classification systems, such as the KNN and Random forest classier. Then a precision score was achieved which indicates that our test overall AdaBoost classification is initially and then secondly the KNN classification according to our experimental tests and 3rd one is RF classifiers as shown in Figure 5.

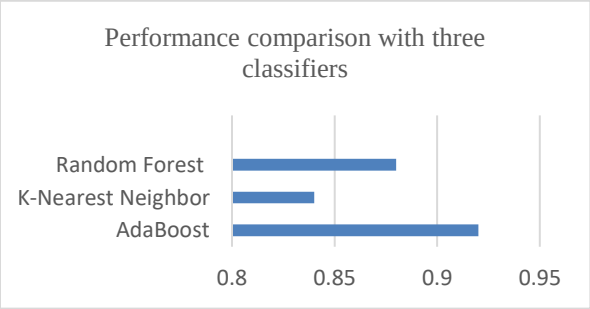

Figure 5. Performance comparison of different classifiers

Prediction of Protein structure model (SARSCoV-2)

Additional variables were derived from protein structure or feature or characteristics corresponding or similar functions; particularly derived from an organism of the same species. The data were normalized by SARS-Cov-2 protein sequence structure models have been predicted and/or refined. Content analysis was undertaken proteins that were difficult to predict by homology modeling due to the lack of experimentally determined homologous protein structures. Data on several variables were used to three models, models based on our pipeline for structural forecast, refined models. Additional prediction were made based upon our sequence and other protein models were further developed. We finally have revised our predictions or chosen some models from the available models for 4 membrane proteins that are more probable and refined these models in the ER membrane bilayer as their physiological oligomeric forms.

DISCUSSION

The dataset is divided into 80% for training databases and 20% for tests in SARS-Cov-2 classification-based positive and negative datasets with the python computing environment. The prediction of the SARS-Cov-2 protein positive and negative sets is based on two different class labeling methods. On SARS-Cov-2 protein sequences, the inclusive comparison of the machine learning processes is studied. The adjustable factor is AdaBoost model, and then K in the classifier KNN, and the estimator number in the RF classifier.

The data from NCBI, and UniProt database were taken from the datasets studied here. The first data set contains 267 protein domains of 127 protein which acute human-severe respiratory syndrome, 140 non- acute human-severe respiratory syndromes. The sample of a protein domain was used for the amino acid composition. Thus, in a 400-dimensional space, every AdaBoost input actually equals a vector. The 10-fold cross-validations that are considered the most rigorous and objective test procedure in statistical prediction were used by calculations in order to show AdaBoost Learner's power and are widely used by more and more investigators to test the strengths of various predictors.

Table 4. classification results for a number labelling method were discussed.

Classifiers	ACC
AdaBoost	0.9287%
RF	0.8621%
KNN	0.8746%

The results of SARS-Cov-2 protein-assisted sequence detection for the binary labelling method were discussed in Table 4. The preferable results confirmed that the AdaBoost classification achieved an exactify of 0.9287% for a single numbering method, while for the binary array labeling method, RFC classifier achieved 0.8621% accuracy and KNN classifications obtained 0.8746% accuracy.

In addition, 0.9287% percent of accuracy for a single number labeling method was achieved with the best K value >=50. The best K >=400 value for the binary array labeling method is achieved by 0.8746% precision. The selection of K values has some impact on the precise classification of a given number labelling method but has a strong impact on the precision of the classification for binary labelling.

CONCLUSION

According to the proposed technique, which selected successfully 127 and 140 proteins collected and then extracted 400 features length 400 featured features extracted by inception from collected datasets while also improving classification accuracy, this approach was successful. A further complication for the present hypothesis is that order to increase the performance of the proposed method, DDE Features extraction and ranking approach have been implemented. Overall, these studies provide support for the validity of feature profile vector matrix based on Dipeptide Deviation from Expected Mean (DDE). Proteins amino acid pair tendency, and composition and transition of physiochemical parameters, the proposed dipeptide deviations (DDE -based feature vector for accurate prediction of epitopes) were reported.

The proposed DDE features protein sequence function vector matrix has been found to effectively differentiate between the specific SARS CoV-2 and non- SARS CoV-2. A possible interpretation of this finding is that results suggested that in predicting exact SARS CoV-2 effected protein sequence. The proposed AdaBoost classifier model was better performance than other two ML classifier. In conclusion, we assume that the python framework is built as an efficient tool for predicting the exact python epitope.

In response to the recent SARS-CoV-*2* outbreak, the scientific community has been galvanised based on decades of fundamental research characterized by the virus family. An extended discussion of SARS-CoV-2 is beyond the scope of this paper. Current research seems to indicate that SARS-CoV-2 outbreak reply share viral genomes in open-access databases, enabling researchers to quickly develop tests on this novel pathogen. other laboratories have experimentally determined, and computerized structures of some viral proteins shared, and epidemiological data have been shared by others. We hope that we can contribute to scientific work by releasing structure predictions of several studied SARS-CoV-2 proteins, the virus that causes COVID-19, using the latest versions of the AlphaFold system. These predictions of proteins sequence structure are not experimentally verified, but hopefully they will help the scientific community to ask how the virus works and serve as a platform for the hypothesis production of future research on therapeutics. We are obliged to the work of many other laboratories: without the work of researchers from around the globe who responded with incredible agility to the COVID 19 outbreak, this work would not be possible.

The AdaBoost Learner is an extremely handy classification system. In prediction of the protein domain structural category for the previous two datasets. The AdaBoost classifier is expected to also predict other protein attributes, such as subcellular localization type of membrane, class of enzyme and subfamily, active enzymes, classification of the coupled G-protein receptor and type of quaternary protein structure among many others.

CONFLICT OF INTEREST

The authors declared that present study was performed in absence of any conflict of interest.

ACKNOWLEGEMENT

We are delighted to have given all available resources and their important suggested recommendations for the effective completion of the study project to Dr. Mukhtiar Memon, the Associate Professor, Information Technology Centre.

AUTHOR CONTRIBUTIONS

Ali Ghulam, Mukhtiar Memon, Mansoor Hyder and Rahu Sikander contributed equally to this work.) Mukhtiar Memon conceptualized and analyzed the manuscript. Ali Ghulam wrote the initial manuscript. Mansoor Hyder, Rahu Sikander, Zulfikar Ahmed Maher, Ahsanullah Unar helped design the method and the code. Zulfikar Ahmed, Zar Nawab Khan Swati, Dhani Bux Talpur, Ikram Ullah and Ali Farman revised the manuscript and polished the expression of English. All of the authors have read and approved the final manuscript.

REFERENCES

Andersen, K. G. *et al*., "Correspondence: The proximal origin of SARS-CoV-2. Nature Medicine". 26 (4), 450–452. https://doi.org/10.1038/s41591-020-0820-9

Ardabili SF, *et al*. 2020, "COVID-19 outbreak prediction with machine learning". 2020. Available at SSRN 3580188.

Ahmed, H. G., et al., 2021, "Impact of COVID-19 outbreak on patients with chronic illnesses during the lockdown." Bioscience Research (2021): 445-454.

Barstugan M, *et al*. 2020, "Coronavirus (covid-19) classification using CT images by machine learning methods".2020. arXiv preprint arXiv:200309424.

Batra R, *et a*l. 2020, "Screening of therapeutic agents forCOVID-19 using machine learning and ensemble dockingsimulations", 2020. arXiv preprint arXiv:200403766.

Bhasin, M., *et al*., "ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST", Nucleic acids research, 32(suppl_2), W414-W419.

Boopathi, S., *et al*. 2020, "Novel 2019 coronavirus structure, mechanism of action, antiviral drug promises and rule out against its treatment. Journal of Biomolecular Structure and Dynamics". https://doi.org/10.1080/07391102.2020.1758788

Breiman, L. 2001, "Random Forests. Machine Learning", 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Bullock J, *et al.* 2020, Mapping the landscape of artificial intelligence applications against COVID-19". 2020. arXiv preprint arXiv:200311336.

Chou, K. C., 2001, "Proteins: Structure, Function, and Genetics", Erratum: ibid., 2001, Vol.44, 60), 43, 246-255.

D.M. *et al.* 2020, "Network medicine framework for identifying drug repurposing opportunities for covid-19" [Preprint] (2020),[cited 2020 Aug. 9]. arXiv:200407229

Desautels T, *et al.* 2020, "Rapidin silico design of antibodies targeting SARS-CoV-2using machine learning and supercomputing". BioRxiv, 2020.

Gammazza, *et al.* 2020, "Human molecular chaperones share with SARS-CoV-2 antigenic epitopes potentially capable of eliciting autoimmunity against endothelial cells: possible role of molecular mimicry in COVID-19." Cell Stress and Chaperones, 25.5 (2020): 737-741.

Ge Y, *et al.* 2020), "A data-drivendrug repositioning framework discovered a potentialtherapeutic agent targeting COVID-19". bioRxiv 2020.

Gupta, S., *et al.* 2013, "Open-Source Drug Discovery Consortium. (2013). In silico approach for predicting toxicity of peptides and proteins". PloS one, 8(9), e73957.

Gysi DM, *et al.* 2020, "Network medicine framework for identifying drugrepurposing opportunities for covid-19", 2020. arXiv preprintarXiv:200407229.

Horry MJ, *et al.* 2020, "X-Ray image based COVID-19 detection using pretraineddeep learning models". engrXiv 2020.

Ibrahim IM, *et al.* 2020, "COVID- 19 spike-host cell receptor GRP78 binding site prediction". J Infect, 2020;80(5):554e62.

Iqbal, M., et al., 2020, "Potential phytomedicines against COVID-19: A review." Bioscience Research (2020): 2417-2422.

Jin, Y. H., *et al.* 2008, "Predicting subcellular localization with AdaBoost Learner", Protein and Peptide Letters, 2008, 15(3), 286-289.

Kassani SH, *et al*. 2020, "Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: a machine learning-based approach". 2020.arXiv preprint arXiv:200410641.

Kumar R, *et al.* 2020, “Accurate prediction of COVID-19 using chest X-Ray images through deep feature learning model with SMOTE and machine learning Q4 classifiers”. medRxiv 2020.

Lai, C. C., *et al.* 2020, “severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges”. International journal of antimicrobial agents, 2020, 55(3), 105924.

NCBI, virus: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Severe%20acute%20respiratory%20syndrome%20coronavirus%202%20(SARSCoV2), %20taxid:2697049

Niu, B., *et al.* 2020, “Predicting protein structural class with AdaBoost learner”, Protein and peptide letters, 13(5), 489-492.

Ong E, et al. 2020, “COVID-19 coronavirus vaccine design using reverse vaccinology and machinelearning”. BioRxiv 2020.

Ozkaya U, *et al.* 2020, “Coronavirus (COVID-19) classification using deep features fusion and ranking technique”. 2020. arXiv preprint arXiv:200403698.

P. Cunningham, 2007, S.J. Delany, “k-Nearest neighbour classifiers”, Multiple Classier System, 2007, pp. 1–17

R. Batra, H. *et al.* 2020, “Screening of therapeutic agents for COVID-19 using machine learning and ensemble docking simulations” J Phys Chem Lett, 11 (2020), pp. 7058-7065

Rubab, S., *et al.,* 2021, "Challenging COVID-19 and its outbreak in Pakistan and across the borders." Bioscience Research (2021): 284-294.

Senior AW, *et al.* 2020, “Improved protein structure prediction using potentials from deep learning”, Nature. 2020 Jan;577(7792):706-710. doi: 10.1038/s41586-019-1923-7. Epub 2020 Jan 15. PMID: 31942072.

Thompson, J. D. *et al*., "CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice” Nucleic Acids Res, 22 (22), 4673–4680. https://doi.org/10.1093/nar/22.22.4673

Ucar F, Korkmaz D. *et al.* 2020, “COVIDiagnosis-Net: deep Bayes- SqueezeNet based diagnostic of the coronavirus disease 2019 (COVID-19) from X-Ray images” Med Hypotheses, 2020:109761.

V. Saravanan, *et al.* 2015, "Harnessing computational biology for exact linear B-Cell epitope prediction: A novel amino acid composition based feature descriptor," OMICS, A J. Integrative Biol., vol. 19, no. 10, pp. 648658, Oct. 2015, doi: 10.1089/omi.2015.0095.

V. Saravanan, *et al.,* 2018, “BCIgEPREDA dual-layer approach for predicting linear IgE epitopes," Mol. Biol., vol. 52, no. 2, pp. 285293, Mar. 2018, doi: 10.1134/S0026893318020127.

Wang L, *et al*., 2020 “COVID-Net: a tailored deep convolutionalneural network design for detection of COVID-19 cases fromchest radiography images”. 2020. arXiv preprintarXiv:200309871.

Y. Ge, T. *et al*., 2020, “A data-driven drug repositioning framework discovered a potential therapeutic agent targeting COVID-19”, [Preprint] (2020), [cited 2020 March 12]. BioRxiv, 2020.03.11.986836

Y. Huang, 2010, “CD-HIT suite: A Web server for clustering and comparing biological sequences", Bioinformatics, vol. 26, no. 5, pp. 680682, Mar. 2010.

Zhou, G. P., 2003, and Doctor, K. “Proteins: Structure”, Function, and Genetics, (2003), 50, 44-48.