ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Online Predictor Using Machine Learning to Predict Novel Coronavirus and Other Pathogenic Viruses

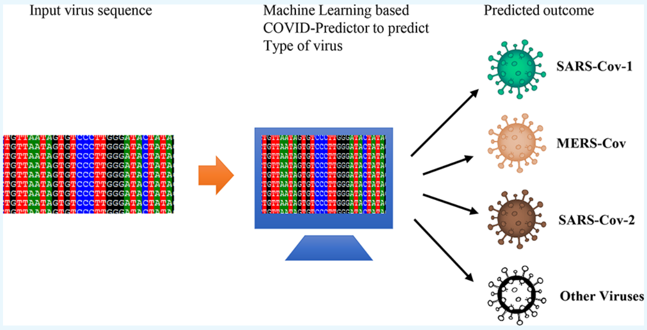
Jnanendra Prasad Sarkar,[∇] Indrajit Saha,[*,∇] Nimisha Ghosh, Debasree Maity, and Dariusz Plewczynski

Cite This: *ACS Omega* 2022, 7, 23069−23074

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: The problem of virus classification is always a subject of concern for virology or epidemiology over the decades. In this regard, a machine learning technique can be used to predict the novel coronavirus by considering its sequence. Thus, we are proposing a machine learning-based novel coronavirus prediction technique, called COVID-Predictor, where 1000 sequences of SARS-CoV-1, MERS-CoV, SARS-CoV-2, and other viruses are used to train a Naive Bayes classifier so that it can predict any unknown sequences of these viruses. The model has been validated using 10-fold cross-validation in comparison with other machine learning techniques. The results show the superiority of our predictor by achieving an average 99.7% accuracy on an unseen validation set of viruses. The same pre-trained model has been used to design a web-based application where sequences of unknown viruses can be uploaded to predict the novel coronavirus.



■ INTRODUCTION

On the 31st of December 2019, the World Health Organization (WHO)[1] was informed about a few pneumonia cases that had been detected in Wuhan City, Hubei Province of China with unknown etiology. Subsequently, on the 7th of January 2020, Chinese authorities identified a novel virus as a cause of this disease. Later, on the 11th of February 2020, the WHO and International Committee on Taxonomy of Viruses declared the name of this virus as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) or novel coronavirus while the disease is called COVID-2019.[1,2] Moreover, the pandemic was declared on the 11th of March 2020 by the WHO. As a precautionary measure, the decision of lockdown was taken around the globe for different time frames. On the other hand, genetic research shows that the novel coronavirus belongs to the family of coronavirus. In this family, Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS) are also present. The medical research community suspects that SARS-CoV-2 is more transmissible but comparatively less fatal than SARS. According to a lot of evidence,[3,4] the transmission rate of SARS-CoV-2 from a human to another human seems higher than SARS and the virus might be of bat or pangolin origin.[2,5,6] It is also suggested that mostly the transmission of this virus is via droplets. It means that once an infected person coughs or sneezes and emitted droplets come into contact with another individual over a short distance, the second individual might get infected.

As of the 30th of April 2022, more than 512 million positive cases have been registered across the world while more than 6 million people have died.[1] Generally, coronavirus can infect multiple organs in different hosts such as animals and humans. It mainly attacks the respiratory system in humans like the other two viruses, SARS-CoV-1 and MERS-CoV, in the same family. The genetic features like potential etiological agents of the SARS-CoV-2 have been identified after metagenomic analysis using next-generation sequencing (NGS)[1]. Moreover, another study[7] shows that the spike protein receptor-binding domain (RBD) of SARS-CoV-2 binds with host receptor angiotensin-converting enzyme 2 (ACE2). It generally helps to regulate the transmission of COVID-19 in cross-species and humans. It is now known that the virus can sustain itself by mutating and creating divergent variants.[9−11] Thus, the early prediction of the disease caused by SARS-CoV-2 is an important but challenging task[8] because SARS-CoV-1, MERS-CoV, and SARS-CoV-2 belong to the same family. The similarities in their sequence identities are also high as found in a study on topological analysis for sequence variability.[12] In this regard, several studies[13,14] suggest deep

Received: January 11, 2022
Accepted: June 10, 2022
Published: June 28, 2022

learning-based classification models to predict coronavirus using X-ray and Computed Tomography (CT) scan images. In most of the cases, the accuracy is achieved in a range of 86% to 98%. Collection of huge medically validated images for such deep-learning based applications is practically another challenge.

To address the above urgent requirement, here, we have developed a machine learning-based technique, called COVID-Predictor, where DNA sequences of three different coronaviruses and other viruses, such as Ebola and Dengue, are used from The National Center for Biotechnology Information (NCBI)[2] and Global Initiative on Sharing All Influenza Data (GISAID)[3] to train a well-known machine learning technique, called the Multinomial Naive Bayes (MNB)[15] classifier, so that it can predict any unknown sequences of these viruses. For this purpose, the *k*-mer algorithm[16,17] is used to create descriptors from the virus sequences. Thereafter, the *n*-gram concept is used to create a Bag-of-Descriptors (BoDs) in order to have count vectors. Such a count vector of sequences is then used to train the MNB. Subsequently, testing is done in the same fashion with 10-fold cross-validation and unseen sequences of coronaviruses from aforementioned databases. The model has also been compared with other two well-known machine learning techniques such as the kernel-based Gaussian Support Vector Machine (GSVM)[18] and Random Forest (RF).[19] The MNB-based model shows the superior performance to other ML-based models by achieving an average of 99.8% accuracy during training using 10-fold cross-validation while 99.7% average accuracy was achieved on unseen virus datasets. The same pre-trained model is used to develop a web application so that scientific and diagnostic communities interested in coronavirus prediction can get the benefit out of this. It is important to mention that the virus encoding to fit for machine learning is the main contribution to this work. Furthermore, to the best of our knowledge, this is the first work that provides online service in the form of a website to predict viruses like SARS-CoV-2, MERS-CoV, SARS-CoV-1, Dengue, and Ebola from their respective sequence to support the medical community.

MATERIALS AND METHODS

In this section, we have discussed about the data preparation, the parameters, and the metrics that are used for COVID-Predictor.

Data Preparation. The datasets of SARS-CoV-1, MERS-CoV, and other kinds of viruses like Ebola and Dengue are downloaded from NCBI while SARS-CoV-2 is downloaded from GISAID in fasta format. Although the proposed predictor does not require sophisticated data prepossessing, it requires complete or near-complete genomes or sequences of viruses. As a result, 344, 291, and 2391 sequences of SARS-CoV-1, MERS-CoV, and SARS-CoV-2, respectively, of length more than 29 kbp while 600 other viruses such as Ebola and Dengue of length more than 10 kbp are considered in our experiment to train and test the proposed predictor. The statistics of the refined consolidated datasets is shown in Table 1. Additionally, more sets of sequences are also collected from NCBI and GISAID for validating the final predictor. The details about additional data are reported in Table 4.

Table 1. Statistics of the Refined Datasets of Corona and Other Viruses

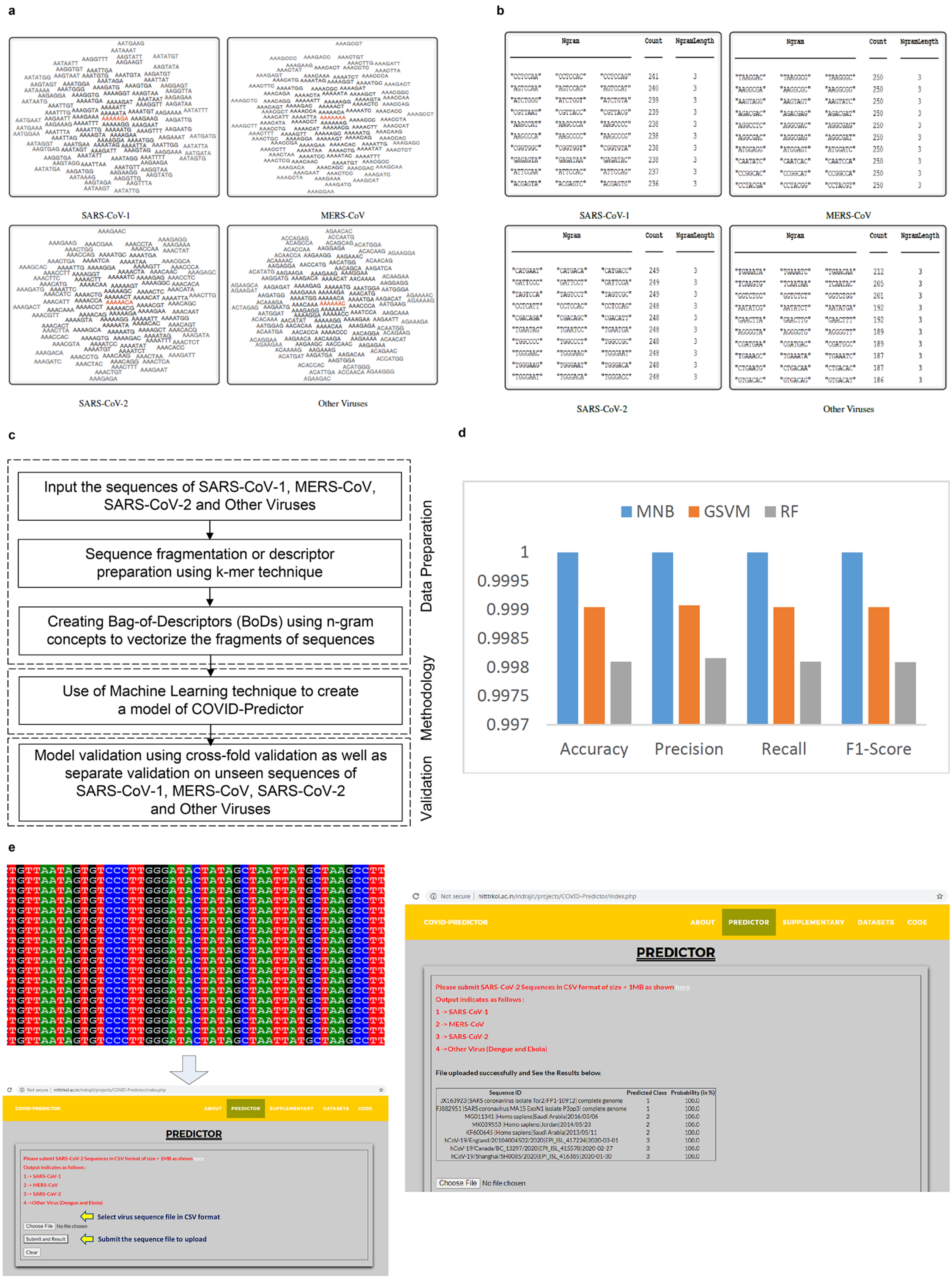
virus name	source of sequence	no. of sequence	max length of sequence	avg length of sequence
SARS-CoV-1	NCBI	340	30,311	29,514
MERS-CoV	NCBI	291	30,150	29,983
SARS-CoV-2	GISAID	2391	29,986	29,512
Other viruses	NCBI	600	19,897	15,316

Parameter Setting and Metrics. The experiments are performed using Python 3.6 and executed on an Intel Core i5-2410M CPU at 2.30 GHz with 8GB of RAM and a Windows 7 operating system. The required input parameters are experimentally set. Such parameters are the number of trees for RF, which is equal to 100, decision for RF is “gini”, the alpha value as a smoothing factor of MNB is 0.1, and the kernel used in GSVM is “rbf”. To evaluate results of COVID-Predictor, the popular performance metrics such as accuracy, precision, recall, F1 score, ROC AUC score, and Matthews Correlation Coefficient (MCC) are used.

COVID-Predictor. The primary objective of the proposed COVID-Predictor is to correctly identify the sequences of coronaviruses. In this regard, the complete or near-complete DNA sequences are split into descriptors using the *k*-mer technique. Such descriptors for four classes of viruses are shown in Figure 1a as a word cloud. Thereafter, the *n*-gram technique is applied to create a feature by considering an *n* number of descriptors. The top 10 *n*-grams for different viruses are shown in Figure 1b. These *n*-grams/features are used to call as a Bag-of-Descriptors (BoDs). Such BoDs are further used to create count vectors for virus sequences. The count vectorization computes the frequencies of *n*-grams in a particular sequence and creates a numeric feature vector that is then used for subsequent machine learning techniques, such as Multinomial Naive Bayes (MNB), the kernel based Support Vector Machine (SVM), and a tree-based technique like Random Forest (RF) to evaluate their performance. The choice of selecting these machine learning techniques can be attributed to their popularity and to accomplishing the task. Independently, all three machine learning techniques are evaluated with features generated by count vectorization after considering different values of the *k*-mer and *n*-gram. Based on the performance of the three machine learning techniques over 10-fold cross-validation on training data, we have finalized MNB as the underlying technique for building COVID-Predictor as used in our web application. The pipeline of the proposed COVID-Predictor is described in Figure 1c.

RESULTS AND DISCUSSION

The dataset consisting of all four types of virus sequences SARS-CoV-1, MERS-CoV, SARS-CoV-2, and other viruses has been divided into two sets, one for a training set and the other for testing purposes. A stratified sampling method is applied to prepare the training dataset to ensure that representatives from all four types of virus classes are present. As a result, 1000 virus sequences are used in training. Moreover, data samples are carefully selected from each category to avoid imbalance class problems. The testing dataset contains those sequences that are not present in the training dataset. The training dataset is used in the three independent machine learning techniques, viz., MNB, GSVM, and RF. For each machine learning technique, the descriptors of virus sequences are created using the *k*-mer method. Thereafter, such descriptors are combined


a
SARS-CoV-1
MERS-CoV
SARS-CoV-2
Other Viruses
b
c
Input the sequences of SARS-CoV-1, MERS-CoV, SARS-CoV-2 and Other Viruses
Sequence fragmentation or descriptor preparation using k-mer technique
Creating Bag-of-Descriptors (BoDs) using n-gram concepts to vectorize the fragments of sequences
Use of Machine Learning technique to create a model of COVID-Predictor
Model validation using cross-fold validation as well as separate validation on unseen sequences of SARS-CoV-1, MERS-CoV, SARS-CoV-2 and Other Viruses
Data Preparation
Methodology
Validation
d
MNB
GSVM
RF
1
0.9995
0.999
0.9985
0.998
0.9975
0.997
Accuracy
Precision
Recall
F1-Score
e
Not secure | nitttrkol.ac.in/indrajit/projects/COVID-Predictor/index.php
COVID-PREDICTOR
ABOUT
PREDICTOR
SUPPLEMENTARY
DATASETS
CODE
PREDICTOR
Select virus sequence file in CSV format
Choose File No file chosen
Submit and Result
Submit the sequence file to upload
Clear

SARS-CoV-1

Ngram			Count	NgramLength
"CCTCCAA"	"CCTCCAC"	"CCTCCAG"	241	3
"AGTCCAA"	"AGTCCAG"	"AGTCCAT"	240	3
"ATCTGGG"	"ATCTGGT"	"ATCTGTA"	239	3
"CGTTAAT"	"CGTTACC"	"CGTTACG"	239	3
"AAGCCAT"	"AAGCCCA"	"AAGCCCC"	238	3
"AAGCCCA"	"AAGCCCC"	"AAGCCCG"	238	3
"CGGTGGC"	"CGGTGGT"	"CGGTGTA"	238	3
"GAGAGTA"	"GAGATAA"	"GAGATAC"	238	3
"ATTCCAA"	"ATTCCAC"	"ATTCCAG"	237	3
"ACGAGTA"	"ACGAGTC"	"ACGAGTG"	236	3

MERS-CoV

Ngram			Count	NgramLength
"TAAGGAC"	"TAAGGCG"	"TAAGGGC"	250	3
"AAGGCGA"	"AAGGCGC"	"AAGGCGG"	250	3
"AAGTAGG"	"AAGTAGT"	"AAGTATC"	250	3
"AGACGAC"	"AGACGAG"	"AGACGAT"	250	3
"AGGCCTC"	"AGGCGAC"	"AGGCGAG"	250	3
"AGGCGAC"	"AGGCGAG"	"AGGCGAT"	250	3
"ATGGAGG"	"ATGGAGT"	"ATGGATC"	250	3
"CAATATC"	"CAATCAC"	"CAATCCA"	250	3
"CCGGCAC"	"CCGGCAT"	"CCGGCCA"	250	3
"CCTACGA"	"CCTACGG"	"CCTACGT"	250	3

SARS-CoV-2

Ngram			Count	NgramLength
"CATGAAT"	"CATGACA"	"CATGACC"	249	3
"GATTCCC"	"GATTCCT"	"GATTCGA"	249	3
"TAGTCCA"	"TAGTCCT"	"TAGTCGC"	249	3
"CCTCATT"	"CCTCCAC"	"CCTCCAG"	248	3
"CGACAGA"	"CGACAGC"	"CGACATT"	248	3
"TGAATAG"	"TGAATCC"	"TGAATGA"	248	3
"TGGCCCC"	"TGGCCCT"	"TGGCCGC"	248	3
"TGGGAAC"	"TGGGAAG"	"TGGGAAT"	248	3
"TGGGAAG"	"TGGGAAT"	"TGGGACA"	248	3
"TGGGAAT"	"TGGGACA"	"TGGGACC"	248	3

Other Viruses

Ngram			Count	NgramLength
"TGAAATA"	"TGAAATC"	"TGAACAA"	212	3
"TCAAGTG"	"TCAATAA"	"TCAATAC"	205	3
"GGTCTCC"	"GGTCTCT"	"GGTCTGG"	201	3
"AATATCG"	"AATATCT"	"AATATGA"	192	3
"GAACTTA"	"GAACTTG"	"GAACTTT"	192	3
"AGGGGTA"	"AGGGGTG"	"AGGGGTT"	189	3
"CGATGAA"	"CGATGAC"	"CGATGCC"	189	3
"TGAAAGC"	"TGAAATA"	"TGAAATC"	187	3
"CTGAATG"	"CTGACAA"	"CTGACAC"	187	3
"GTGACAC"	"GTGACAG"	"GTGACAT"	186	3

Please submit SARS-CoV-2 Sequences in CSV format of size < 1MB as shown here

Output indicates as follows :

1 -> SARS-CoV-1

2 -> MERS-CoV

3 -> SARS-CoV-2

4 ->Other Virus (Dengue and Ebola)

File uploaded successfully and See the Results below.

Sequence ID	Predicted Class	Probability (in %)
JX163923 |SARS coronavirus isolate Tor2/FP1-10912| complete genome	1	100.0
FJ882951 |SARS coronavirus MA15 ExoN1 isolate P3pp3| complete genome	1	100.0
MG011341 |Homo sapiens|Saudi Arabia|2016/03/06	2	100.0
MK039553 |Homo sapiens|Jordan|2014/05/23	2	100.0
KF600645 |Homo sapiens|Saudi Arabia|2013/05/11	2	100.0
hCoV-19/England/20104004502/2020|EPI_ISL_417224|2020-03-01	3	100.0
hCoV-19/Canada/BC_13297/2020|EPI_ISL_415578|2020-02-27	3	100.0
hCoV-19/Shanghai/SH0085/2020|EPI_ISL_416385|2020-01-30	3	100.0

Figure 1. (a) Word cloud of descriptors generated using *k*-mer techniques from sequences of SARS-CoV-1, MERS-CoV, SARS-CoV-2, and other viruses. (b) Top 10 *n*-grams of descriptors generated using *k*-mer techniques from sequences of SARS-CoV-1, MERS-CoV, SARS-CoV-2, and other viruses. (c) Pipeline of the proposed COVID-Predictor. (d) Performance measures for $k = 7$ and $n = 3$. (e) Screenshots of the web-based COVID-Predictor to select a virus sequence file as csv and results after prediction.

using the *n*-gram technique to create a count vector that is used to train the classifiers. In our experiments, the value *k* of *k*-mer varies between 2 and 7, while the value of the *n*-gram varies between 2 and 5. Each machine learning method is evaluated with 10-fold cross-validation followed by further validation on an unseen dataset taken from NCBI and GISAID.

The performance metrics of each machine learning technique with 10-fold cross-validation for different values of the *k*-mer and *n*-gram are reported in Tables 2 and 3. Four quantitative metrics of Table 2 are further consolidated as a

Table 2. Prediction Performance of Different Machine Learning Techniques after Performing 10-Fold Cross Validation with Different Values of *k*-mer and *n*-gram on 1000 Genome Sequences of SARS-CoV-1, MERS-CoV, SARS-CoV-2, and Other Virus Sequences

		n-gram = 2				n-gram = 3				n-gram = 4				n-gram = 5				Aggregated
Method	k-mer	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Score
MNB	2	0.99810	0.99817	0.99810	0.99810	0.99810	0.99817	0.99810	0.99810	0.99810	0.99817	0.99810	0.99810	0.99905	0.99910	0.99905	0.99905	0.99835
GSVM		0.94857	0.95725	0.94857	0.94952	0.96762	0.97151	0.96762	0.96795	0.98190	0.98324	0.98190	0.98191	0.99238	0.99276	0.99238	0.99237	0.97359
RF		0.99429	0.99458	0.99429	0.99428	0.99429	0.99458	0.99429	0.99428	0.99429	0.99457	0.99429	0.99428	0.99619	0.99632	0.99619	0.99618	0.99482
MNB	3	0.99810	0.99817	0.99810	0.99810	0.99810	0.99817	0.99810	0.99810	0.99905	0.99910	0.99905	0.99905	0.99810	0.99817	0.99810	0.99810	0.99835
GSVM		0.96762	0.97151	0.96762	0.96795	0.98190	0.98324	0.98190	0.98191	0.99238	0.99276	0.99238	0.99237	0.99905	0.99909	0.99905	0.99905	0.98561
RF		0.99429	0.99458	0.99429	0.99428	0.99524	0.99548	0.99524	0.99523	0.99810	0.99816	0.99810	0.99809	0.99714	0.99725	0.99714	0.99714	0.99623
MNB	4	0.99810	0.99817	0.99810	0.99810	0.99905	0.99910	0.99905	0.99905	0.99810	0.99817	0.99810	0.99810	1.00000	1.00000	1.00000	1.00000	0.99882
GSVM		0.98190	0.98324	0.98190	0.98191	0.99238	0.99276	0.99238	0.99237	0.99905	0.99909	0.99905	0.99905	1.00000	1.00000	1.00000	1.00000	0.99344
RF		0.99429	0.99456	0.99429	0.99428	0.99714	0.99725	0.99714	0.99714	0.99714	0.99725	0.99714	0.99714	0.99810	0.99817	0.99810	0.99810	0.99670
MNB	5	0.99905	0.99910	0.99905	0.99905	0.99810	0.99817	0.99810	0.99810	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.99929
GSVM		0.99238	0.99276	0.99238	0.99237	0.99905	0.99909	0.99905	0.99905	1.00000	1.00000	1.00000	1.00000	0.99905	0.99908	0.99905	0.99905	0.99765
RF		0.99619	0.99633	0.99619	0.99618	0.99714	0.99725	0.99714	0.99714	0.99810	0.99816	0.99810	0.99809	0.99810	0.99816	0.99810	0.99809	0.99740
MNB	6	0.99810	0.99817	0.99810	0.99810	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.99905	0.99908	0.99905	0.99905	0.99929
GSVM		0.99905	0.99909	0.99905	0.99905	1.00000	1.00000	1.00000	1.00000	0.99905	0.99908	0.99905	0.99905	0.99714	0.99724	0.99714	0.99714	0.99882
RF		0.99714	0.99725	0.99714	0.99714	0.99810	0.99816	0.99810	0.99809	0.99810	0.99816	0.99810	0.99809	0.99714	0.99724	0.99714	0.99714	0.99764
MNB	7	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.99905	0.99908	0.99905	0.99905	0.99905	0.99908	0.99905	0.99905	0.99953
GSVM		1.00000	1.00000	1.00000	1.00000	0.99905	0.99908	0.99905	0.99905	0.99714	0.99724	0.99714	0.99714	0.99619	0.99632	0.99619	0.99619	0.99811
RF		0.99810	0.99816	0.99810	0.99809	0.99810	0.99816	0.99810	0.99809	0.99714	0.99724	0.99714	0.99714	0.99810	0.99816	0.99810	0.99809	0.99787

Table 3. ROC AUC Score and MCC of Different Machine Learning Techniques after Performing 10-Fold Cross Validation with Different Values of *k*-mer and *n*-gram on 1000 Genome Sequences of SARS-CoV-1, MERS-CoV, SARS-CoV-2, and Other Virus Sequences

		n-gram = 2		*n*-gram = 3		*n*-gram = 4		*n*-gram = 5	
method	*k*-mer	ROC-AUC-score	MCC	ROC-AUC-score	MCC	ROC-AUC-Score	MCC	ROC-AUC-Score	MCC
NB	2	0.99998	0.99746	0.99810	0.99746	0.99998	0.99746	1.00000	0.98987
GSVM		0.99997	0.93321	0.99994	0.95759	0.99998	0.97610	0.99998	0.98985
RF		0.99795	0.93321	0.99998	0.95761	0.99872	0.97608	0.99940	0.99873
NB	3	0.99810	0.99746	0.99999	0.99746	1.00000	0.99873	0.99872	0.99874
GSVM		0.99994	0.95759	0.99998	0.97610	0.99998	0.98985	1.00000	0.99874
RF		0.99998	0.95760	0.99872	0.97612	0.99940	0.98987	1.00000	0.99746
NB	4	0.99872	0.99746	1.00000	0.99873	1.00000	0.99746	1.00000	0.99946
GSVM		0.99998	0.97610	0.99998	0.98985	1.00000	0.99874	1.00000	0.99874
RF		0.99997	0.97608	0.99940	0.98985	0.99872	0.99874	1.00000	0.99876
NB	5	0.99940	0.99873	0.99872	0.99746	0.99811	0.99874	1.00000	0.99876
GSVM		0.99998	0.98985	1.00000	0.99874	0.99974	0.99495	1.00000	0.99873
RF		1.00000	0.98987	1.00000	0.99875	0.99798	0.99591	1.00000	0.99874
NB	6	0.99872	0.99746	1.00000	0.99897	1.00000	0.99998	0.99938	0.99873
GSVM		1.00000	0.99874	0.99874	0.98885	1.00000	0.99873	1.00000	0.99620
RF		1.00000	0.99876	1.00000	0.98973	1.00000	0.99877	1.00000	0.99624
NB	7	1.00000	0.99877	1.00000	1.00000	0.99938	0.99873	0.99938	0.99873
GSVM		1.00000	0.99873	1.00000	1.00000	1.00000	0.99620	1.00000	0.99493
RF		1.00000	0.99873	1.00000	1.00000	1.00000	0.99623	0.99998	0.99491

Table 4. Prediction Performance of COVID-Predictor on Validation Data

source	data samples	accuracy	precision	recall	F1 score	ROC-AUC-score	MCC
NCBI	493 sequences (only SARS-CoV-2)	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
NCBI + GISAID	1090 sequences (90 SARS-CoV-1, 200 MERS-CoV, 200 SARS-CoV-2, 600 other viruses)	0.99908	0.99908	0.99908	0.99908	0.99912	0.99852
NCBI + GISAID	2043 sequences (103 SARS-CoV-1, 41 MERS-CoV, 1599 SARS-CoV-2, 300 other viruses)	0.98217	0.98991	0.98217	0.98602	0.98432	0.98291
NCBI + GISAID	3143 sequences (90 SARS-CoV-1, 41 MERS-CoV, 2152 SARS-CoV-2, 860 other viruses)	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
NCBI + GISAID	3500 sequences (90 SARS-CoV-1, 250 MERS-CoV, 2410 SARS-CoV-2, 750 other viruses)	0.99971	0.99971	0.99971	0.99971	0.99952	0.99940
NCBI + GISAID	4000 sequences (90 SARS-CoV-1, 220 MERS-CoV, 3030 SARS-CoV-2, 2639 other viruses)	0.99975	0.99975	0.99975	0.99974	0.99949	0.99937
GISAID	4747 sequences (only SARS-CoV-2)	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

single aggregated score for ease of comparison; however, the metrics of Table 3 are examined separately. The aggregated score is computed simply by taking the average of all the scores following a similar approach to what was used elsewhere.[20] The boundary of the aggregated score is [0,1], where a higher value signifies a better result. It is evident from Table 2 that MNB-based COVID-Predictor produces a higher aggregated score, i.e., 0.99953, for the value of the *k*-mer equal to 7. This is also depicted in Figure 1d. Similar results are also observed for MNB-based COVID-Predictor for other values of the *k*-mer.

Additionally, ROC-AUC scores and MCC values are reported in Table 3 for different values of the *k*-mer and *n*-gram for MNB-, GSVM-, and RF-based predictors. It is found that the ROC-AUC score and MCC value of MNB-based predictor are 1 and 1, respectively, for the *k*-mer equal to 7 and *n*-gram equal to 3. Thus, according to the results, we have prepared the pre-trained model of COVID-Predictor with 1000 genomic sequences of four virus classes for values of the *k*-mer and *n*-gram of 7 and 3, respectively. To gain further confidence, we have used an additional validation set of sequences as reported in Table 4. While validating with 2043 samples, it is observed that 3 cases are false positive considering prediction of SARS-CoV-2 is positive. After further investigation, it has been found that these 3 sequences are for SARS-CoV-1 and are misclassified by COVID-Predictor as SARS-CoV-2. As our primary objective is to predict SARS-CoV-2, we further wanted to examine the rate of false negatives as well. For this purpose, an additional two sets of pure SARS-CoV-2 sequences and four sets of mixed sequences are used separately. One of the pure SARS-CoV-2 sequence sets with 493 samples and the other with 4747 samples are collected from NCBI and GISAID, respectively. Similarly, the sets of mixed sequences are also collected from NCBI and GISAID. In both the pure SARS-CoV-2 cases, COVID-Predictor predicted SARS-CoV-2 with 100% accuracy, whereas prediction accuracy of the rest of the sets is found to be very close to 100%. Moreover, the values of the ROC AUC score and MCC in all cases are found to be significantly good and impressive. This experiment establishes that COVID-Predictor with the proposed feature building approach has potential to predict SARS-CoV-2 with higher accuracy. The same pre-trained model is used to build the web-based application where the unknown sequences can be uploaded to predict the class of coronavirus. The screenshots of the web-based predictor are shown in Figure 1e.

■ CONCLUSIONS

In the current context, it has become very much essential to predict coronavirus as early as possible because SARS-CoV-2 infection has been one of the worst pandemics, where both the infection and death rate ravaged the entire globe. While vaccination is a preventive measure, early detection is also equally important. As a contribution to the society, in this study, we have proposed COVID-Predictor for predicting the coronaviruses, viz., SARS-CoV-1, MERS-CoV, and SARS-CoV-2 based on their sequences. The same is also provided as a web application so that scientific and diagnostic communities related to coronavirus prediction can get the benefit out of this. In order to achieve better performance, we have carefully performed data preprocessing, paying careful attention to building efficient feature vectors, leveraging the appropriate machine learning technique by performing 10-fold cross-validation. Experimentally, the Multinomial Naive Bayes technique is finalized for building a web-based predictor as MNB performed better on different datasets of sequences, which are collected from NCBI and GISAID. As a further scope of research, our study will focus on developing an efficient predictor to detect the particular variant, new dominant or emerging lineages of SARS-CoV-2.

■ AUTHOR INFORMATION

Corresponding Author

Indrajit Saha – *Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata 700106 West Bengal, India;* Email: indrajit@nitttrkol.ac.in

Authors

Jnanendra Prasad Sarkar – *Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032 West Bengal, India;* orcid.org/0000-0002-6644-8620

Nimisha Ghosh – *Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha 751030, India; Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw 02-097, Poland;* orcid.org/0000-0002-0697-6368

Debasree Maity – *Department of Electronics and Communication Engineering, MCKV Institute of Engineering, Howrah, West Bengal 711204, India*

Dariusz Plewczynski – *Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, 02-097 Warsaw, Poland; Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-927 Warsaw, Poland*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c00215

Author Contributions

$^{\nabla}$Equally Contributed

Funding

This work was carried out during the tenure of an ERCIM 'Alain Bensoussan' Fellowship Program awarded to N. G.. The work has been partially supported by a CRG short-term research grant on COVID-19 (CVD/2020/000991) from the Science and Engineering Research Board (SERB), Department of Science and Technology, Govt. of India. However, they do not provide any fee for journal publication. This work is co-funded by Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) programme. This work has also been co-supported by the Polish National Science Centre (2019/35/O/ST6/02484 and 2020/37/B/NZ2/03757).

Notes

The authors declare no competing financial interest.
The datasets of SARS-CoV-1, MERS-CoV, SARS-CoV-2 and other viruses, code, online COVID-Predictor and supplementary are available at "http://www.nitttrkol.ac.in/indrajit/projects/COVID-Predictor/". Moreover, all the virus genomes used in this work are publicly available at NCBI and GISAID.

■ ACKNOWLEDGMENTS

We thank all those who have contributed sequences to NCBI and GISAID databases.

■ ADDITIONAL NOTES

[1] https://www.worldometers.info/coronavirus/.
[2] https://www.ncbi.nlm.nih.gov/.
[3] https://www.gisaid.org/.

■ REFERENCES

(1) Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; et al. A Novel Coronavirus from

Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **2020**, *382*, 727−733.

(2) Zhou, P.; Yang, X. L.; Wang, X. G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H. R.; Zhu, Y.; Li, B.; Huang, C. L.; et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **2020**, *579*, 270−273.

(3) Chan, J. F.-W.; Yuan, S.; Kok, K.-H.; To, K. K.-W.; Chu, H.; Yang, J.; Xing, F.; Liu, J.; Yip, C. C.-Y.; Poon, R. W.-S.; et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* **2020**, *395*, 514−523.

(4) Li, Q.; Guan, X.; Wu, P.; Wang, X.; Zhou, L.; Tong, Y.; Ren, R.; Leung, K. S. M.; Lau, E. H. Y.; Wong, J. Y.; et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N. Engl. J. Med.* **2020**, *382*, 1199−1207.

(5) Andersen, K. G.; Rambaut, A.; Lipkin, W. I.; Holmes, E. C.; Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med.* **2020**, 450.

(6) Zhang, Y.-Z.; Holmes, E. C. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. *Cell* **2020**, 223.

(7) Wan, Y.; Shang, J.; Graham, R.; Baric, R. S.; Li, F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. *J. Virol.* **2020**, *94*, e00127.

(8) Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan. China. *Lancet* **2020**, *395*, 497−506.

(9) Saha, I.; Ghosh, N.; Maity, D.; Sharma, N.; Sarkar, J. P.; Mitra, K. Genome-wide analysis of Indian SARS-CoV-2 genomes for the identification of genetic mutation and SNP. *Infect., Genet. Evol.* **2020**, *85*, 104457.

(10) Greaney, A. J.; Starr, T. N.; Barnes, C. O.; Weisblum, Y.; Schmidt, F.; Caskey, M.; Gaebler, C.; Cho, A.; Agudelo, M.; Finkin, S. et al. Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nat. Commun.* **2021**, *12,* 10.1038/s41467-021-24435-8.

(11) Dejnirattisai, W.; Huo, J.; Zhou, D.; Zahradnik, J.; Supasa, P.; Liu, C.; Duyvesteyn, H. M.; Ginn, H. M.; Mentzer, A. J.; Tuekprakhon, A.; et al. SARS-CoV-2 Omicron-B. 1.1.529 leads to widespread escape from neutralizing antibody responses. *Cell* **2022**, *185*, 467−484.e15.

(12) Sarkar, J. P.; Saha, I.; Seal, A.; Maity, D.; Maulik, U. Topological Analysis for Sequence Variability: Case Study on more than 2K SARS-CoV-2 sequences of COVID-19 infected 54 countries in comparison with SARS-CoV-1 and MERS-CoV. *Infect., Genet. Evol.* **2021**, *88*, 104708.

(13) Salehi, A. W.; Baglat, P.; Gupta, G. Review on machine and deep learning models for the detection and prediction of Coronavirus. *Mater. Today: Proc.* **2020**, *33*, 3896−3901.

(14) Gupta, G.; Salehi, A. W.; Sharma, B.; Kumar, N.; Vaidya, S.; Vaidya, P. COVID-19: Automated Detection and Monitoring of Patients Worldwide Using Machine Learning. In: Azar, A.T., Hassanien, A.E., Eds. *Studies in Systems, Decision and Control: Modeling, Control and Drug Development for COVID-19 Outbreak Prevention;* Springer 2021, *366*, 731−761.

(15) George, H.; Langley, J. P. Estimating Continuous Distributions in Bayesian Classifiers. *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence,1995;* Morgan Kaufmann Publishers Inc. 1995, *69*, 338−345.

(16) Solis-Reyes, S.; Avino, M.; Poon, A.; Kari, L. An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. *PLoS One* **2018**, *13*, No. e0206409.

(17) Manekar, S. C.; Sathe, S. R. A benchmark study of k-mer counting methods for high-throughput sequencing. *GigaScience* **2018**, *7*, 1−13.

(18) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273−297.

(19) Breiman, L. Random Forests. *Mach. Learn.* **2005**, *45*, 5−32.

(20) Nepusz, T.; Yu, H.; Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *N. Engl. J. Med.* **2012**, *9*, 471−472.