Heliyon 9 (2023) e13646

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

PREHOST: Host prediction of coronaviridae family using machine learning

Anusha Chaturvedi, Kushal Borkar, U Deva Priyakumar, P.K. Vinod *

International Institute of Information Technology, Hyderabad, Telangana, 500032, India

ARTICLE INFO

Keywords:
Zoonosis
Coronaviridae
Host specificity
Feature identification
Random forest
Biological sequences
SARS-CoV-2

ABSTRACT

Coronavirus, a zoonotic virus capable of transmitting infections from animals to humans, emerged as a pandemic recently. In such circumstances, it is essential to understand the virus's origin. In this study, we present a novel machine-learning pipeline *PreHost* for host prediction of the family, Coronaviridae. We leverage the complete viral genome and sequences at the protein level (spike protein, membrane protein, and nucleocapsid protein). Compared with the current state-of-the-art approaches, the random forest model attained high accuracy and recall scores of 99.91% and 0.98, respectively, for genome sequences. In addition to the spike protein sequences, our study shows membrane and nucleocapsid protein sequences can be utilized to predict the host of viruses. We also identified important sites in the viral sequences that help distinguish between different host classes. The host prediction pipeline *PreHost* will cater as a valuable tool to take effective measures to govern the transmission of future viruses.

1. Introduction

Emerging Infectious Disease (EID) [1] has been affecting individuals worldwide. The World Health Organization reported in 1970 that EIDs were increasing at a rate never determined before. Over the past 40 years, scientists discovered 40 infectious diseases, including the Ebola virus, MERS, SARS, Zika virus, and swine flu. One EID that recently resulted in a pandemic is Severe Acute Respiratory Syndrome Coronavirus - 2 (SARS CoV-2) [2]. SARS-CoV-2, a zoonotic virus [3], capable of interspecies transmission [4] belongs to the Coronaviridae: a family of enveloped, positive, single-stranded RNA viruses 26–32 kb long. The family is organized into two subfamilies: Coronavirinae or Orthocoronavirinae and Letovirinae. Coronavirinae is further classified into 4 genera namely: Alphacoronavirus, Betacoronavirus, Deltacoronavirus and Gammacoronavirus. Alphacoronavirus and Betacoronavirus affect mammals. On the other hand, Deltacoronavirus and Gammacoronavirus are known to infect both mammals and birds. Alphacoronavirus and Betacoronavirus that infect the human population include HCoV 229E, HCoV - NL63, HCoV-OC43, HCoV-HKU1, SARS-CoV, MERS-CoV and SARS-CoV-2.

Over the past two decades, the evolution of various coronaviruses from animals to human populations raises one of the most pressing research questions: Who was the original carrier of the coronavirus? [5]. Although major research efforts are being made to predict viral escape, the effect of mutations on the RNA structure, forecasting the next outbreak, mortality prediction of patients, the questions of when, where, and how viruses appear are still subject to considerable uncertainty [6],[7],[8].

The family of RNA viruses, namely the Coronaviridae, engenders diverse symptoms that lead to a range of respiratory infections.

* Corresponding author.
E-mail addresses: deva@iiit.ac.in (U.D. Priyakumar), vinod.pk@iiit.ac.in (P.K. Vinod).

https://doi.org/10.1016/j.heliyon.2023.e13646
Received 11 February 2022; Received in revised form 5 February 2023; Accepted 6 February 2023
Available online 11 February 2023

The coronavirus evolved into a pandemic that threatened public health and infected 675 million people worldwide, resulting in 6.76 million deaths (as reported in February 2023). With so many lives at risk, identifying the original host of the virus in advance can prevent the further spread of the virus. With this knowledge, we can take effective measures to control the outbreak of the virus. It can help segregate the host of the coronavirus from the human population, for example, by preventing these recognized hosts from entering human-populated areas.

The coronavirus encodes 4 major structural proteins: spike protein (S) [9], membrane protein (M) [10], nucleocapsid protein (N) [11], envelope protein(E) [12], and 16 non-structural proteins. The spike (S) protein of SARS-CoV-2 plays an important role in receptor binding. The total length of SARS-CoV-2 S protein is 1273 aa and consists of a signal peptide (amino acids 1–13) located at the N-terminus, the S1 subunit (14–685 residues), and the S2 subunit (686–1273 residues). The S1 subunit contains a receptor-binding domain (RBD) that recognizes and binds to the host receptor angiotensin-converting enzyme 2 (ACE2), while the S2 subunit mediates viral cell membrane fusion by forming a six-helical bundle via the two-heptad repeat (HR) domain [13]. The S1 subunit consists of an N-terminal domain (14–305 residues) and a receptor-binding domain (RBD, 319–541 residues). The S2 subunit consists of a fusion peptide (FP) (788–806 residues), heptapeptide repeat 1 (HR1, 912–984 residues), heptapeptide repeat 2 (HR2, 1163–1213 residues), TM domain (1213–1237 residues), and cytoplasmic domain (1237–1273 residues) [13]. The spike protein binds to the ACE2 receptor [14] to enter the host cells and plays an important role in interspecies transmission. For this reason, previous studies mainly focused on predicting the host of the virus by analyzing the spike protein sequences of the coronavirus [15][16] [17].

The structural proteins of the coronavirus family include the membrane glycoprotein (M), which is capable of binding to all other structural proteins. In the human coronavirus NL63, the M protein acts as the binding site for the host cell and not the S protein. In addition, the binding of Nprotein to Mprotein helps to stabilize Nprotein [18]. Because the M protein co-operates with the spike protein, mutations in this region are able to affect host cell attachment and viral entry. This property encouraged us to predict the virus' host from membrane protein sequences. The total length of the membrane protein ranges from 218 to 263 aa. It consists of three domains: N-terminal ectodomain (1–19 residues), triple-spanning transmembrane (TM) domain (TMI-(20–40 residues), TMII-(51–71 residues), TMIII-(80–100 residues)) and C terminal endodomain (101–222 residues) [18]. The length of the nucleocapsid protein ranges from 419 to 422 aa. The nucleocapsid protein consists of three domains. The central domain acts as an RNA binding site while the remaining two domains, which are acidic in nature, are believed to play a key role in protein-protein interactions [19].

Tang et al. [15] proposed the current state-of-the-art approach to identify hosts of the Coronaviridae family. 730 spike gene sequences associated with 6 host classes were considered. Mock et al. [16] presented a deep-learning approach to identify the host using RNA sequences as input to the model. Their study is limited to Rotavirus, Rabies lyssavirus, and Influenza A virus. The recent work by Kuzmin et al. [17] focused on the spike protein sequences of SARS-CoV-2 for three classification tasks: [i] the human-related CoVs (463 entries) vs. other CoVs (775 entries); [ii] the CoVs whose hosts are avians (300 entries) vs. the CoVs whose hosts are swine (367 entries); [iii] the CoVs whose hosts are mammals (938 entries) vs. the CoVs with all other hosts, which are all avians (300 entries). Although they identified important sites in the viral sequences, their study was confined to a binary classification task on a relatively smaller dataset.

All the studies mentioned above are limited to spike protein or RNA sequences. In this study, we considered four sequence types: genome sequences, spike protein, membrane protein, and nucleocapsid protein sequences to identify hosts associated with six host classes using a larger dataset. Our study further focused on identifying the important regions in the viral sequences that help distinguish between different host classes. For evaluation, we reported precision, recall, and F1 score along with accuracy to address the class imbalance. The machine learning pipeline *PreHost* proposed in our study shows better performance in the virus-host prediction task compared with the existing approaches.

2. Results

To assess how different regions of viral sequences can be used to identify the hosts, we applied the *PreHost* pipeline to four different viral sequences of the Coronaviridae family, namely: genomic sequences, spike protein (S), membrane protein (M), and nucleocapsid protein (N) sequences (see Methods). We used four machine learning algorithms to predict the host of the viral sequences. We trained the machine learning algorithms on 80% of the dataset for all four datasets.

2.1. Genome sequences

The genome sequence dataset consists of 50,700 viral sequences ranging in length from 26 kb to 32 kb belonging to the family, Coronaviridae. Random Forest, KNN, and Decision Tree yielded prediction accuracy greater than 99% for 4-grams (Table 1). We trained our model with different n-grams, where n ranges from 1 to 4. The highest accuracy is obtained when n is set to 4. This suggests

Table 1
Comparison of performance of different models using genome sequences.

Algorithm	Accuracy	F1-Score	Precision	Recall
Naive Bayes	92.21	0.61	0.57	0.75
Random Forest	**99.91**	**0.99**	**0.99**	**0.98**
KNN	99.60	0.91	0.88	0.98
Decision Tree	99.74	0.95	0.95	0.94

that the performance is not merely owing to the classifier used but is also influenced by the 4-gram input preparation technique. Unigram and bigram lack context information. On the other hand, trigram and 4-gram give better results because they take contextual information into account. In addition, as n increases, higher-order n-grams may have a data sparsity problem that makes them less informative, making it difficult to capture the real data distribution without more data.

Since the dataset was highly skewed, we used various evaluation metrics: accuracy, F1 score, precision, and recall. Random Forest achieves the highest prediction accuracy with a mean recall of 0.98. Random Forest was further used to identify important features in the sequence that contribute most to the classification of the host (Fig. 1). The top 4 features that have led to a reduction in mean Gini impurity are ATAA, TGAA, GGGC and AATA. Table 2 shows the cross-validation results on the genomic sequences with Random Forest giving a mean accuracy of 99.81% using 10-fold cross-validation.

2.2. Spike protein sequences

The spike protein sequence dataset consists of 9596 viral sequences belonging to the Coronaviridae family. KNN, Naive Bayes, Random Forest, and Decision tree achieve a high accuracy of greater than 99% and a recall above 0.92 on a test set (Tables 3 and 4). We also used Random Forest algorithm to identify the features that contribute most to the classification of a different host (Fig. 2(A)). The top 4 features that resulted in decreasing the mean Gini Impurity are TLTN, QLNC, TYVK, and IPQN. We further mapped these features to the spike protein sequence (Fig. 2(B)). The feature TYVK corresponds to Feline Coronavirus and is found at sites 1403–1406. The S protein sequence of the Feline coronavirus is of length 1452, which is longer than the SARS-CoV-2 S protein sequence. Fig. 2(B) shows that the feature TLTN maps to the S1 subunit in the NTD region. The epitope of one of the mAbs called 4A8 is NTD of the spike protein and a promising target for therapeutics against COVID-19(13). The other features fall within the S2 subunit, with the IPQN feature in the HR1 region responsible for viral fusion. The various fusion inhibitors targeting the HR1 region are being developed (HR1P, EK1, and EK14C) [13].

2.3. Membrane-M protein sequences

The M protein sequence dataset consists of 846 viral sequences belonging to the Coronaviridae family. Random Forest achieves the highest prediction accuracy of 98.22% with a mean recall of 0.94, whereas Decision Tree achieves a mean recall of 0.88 on a test set (Tables 3 and 4). The important features that contribute the most to the classification of the different hosts are CEGQ, VGKQ, SNMT, and GDSG (Fig. 3(A)). We map these features to the membrane protein sequence (Fig. 3(B)). Feature CEGQ found in Chicken maps to C-terminal endo-domain 153-156, VGKQ present in Cat, is at site 221–224 (C-terminal endo-domain). SNMT, found mainly in Camel, is located at two sites: 214–216 (C-terminal endodomain) and 2–5 (N-terminal ectodomain). GDSG, a feature present in Humans and occasionally in bats, is located at site 189–192 (C- terminal endo-domain). Previous studies suggested that the membrane protein might be responsible for virus interaction with cellular HSPG (heparan sulfate proteoglycans) [18]. The HSPG binding site is located in the C-terminal domain from 153 to 266 aa, which is predicted to be exposed on the virion surface.

2.4. Nucleocapsid protein sequences

The N protein sequence dataset consists of 4021 viral sequences. Random Forest and KNN attain the highest prediction accuracy on a test set compared to the other machine learning algorithms Naive Bayes and Decision Tree (Table 3). KNN and Random Forest achieve an accuracy of 99% (Table 3) with a mean recall of 0.91 and 0.82, respectively (Table 4). The important features identified are, SSPD, IIWV, KDAL, and LEQI (Fig. 4(A)). From Fig. 4(B), we can observe that the trait SSPD present in human falls within the C-terminal endodomain (site 78–81), IIWV found in humans, bats, and cats are located at sites 130–133, 119–122, 124–127, 143–146 and numerous other sites in the NTD region. KDAL, found mainly in swine, is located at two sites: 278–281 (C-terminal domain) and 227–230. LEQI, a feature found in Cat, is at position 318–321 (C-terminal domain). The earlier findings indicate that N-NTD is responsible for RNA binding and N-CTD is responsible for both dimerization and RNA binding. Blocking the RNA binding sites of N-NTD has been proven to be a good approach for antiviral drug development [19]. PJ34, a compound that targets the ribonucleotide

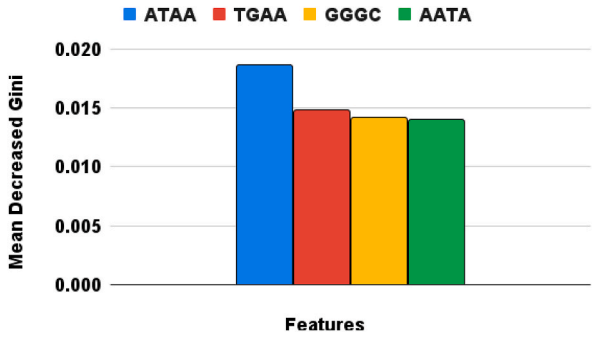

Fig. 1. Important features of genome sequences relevant for host classification.

Table 2
Classification results with stratified 10-fold cross-validation using Random Forest.

Fold	Accuracy
Fold 1	99.90
Fold 2	99.80
Fold 3	99.80
Fold 4	99.90
Fold 5	99.40
Fold 6	99.90
Fold 7	99.60
Fold 8	99.90
Fold 9	99.98
Fold 10	99.90
Mean Result	**99.81**
Result on Test Data	**99.81**

Table 3
Comparison of Accuracy results for spike protein (S), membrane protein(M), and nucleocapsid protein(N) sequences using different machine learning algorithms.

Algorithm	S	M	N
Naive Bayes	98.54	92.89	96.63
Random Forest	99.12	98.22	99.12
KNN	99.60	96.64	99.37
Decision Tree	99.21	97.04	98.50

Table 4
Comparison of Recall results for spike protein (S), membrane protein(M), and nucleocapsid protein(N) sequences using different machine learning algorithms.

Algorithm	S	M	N
Naive Bayes	0.93	0.87	0.83
Random Forest	0.93	0.94	0.82
KNN	0.94	0.91	0.91
Decision Tree	0.92	0.88	0.77

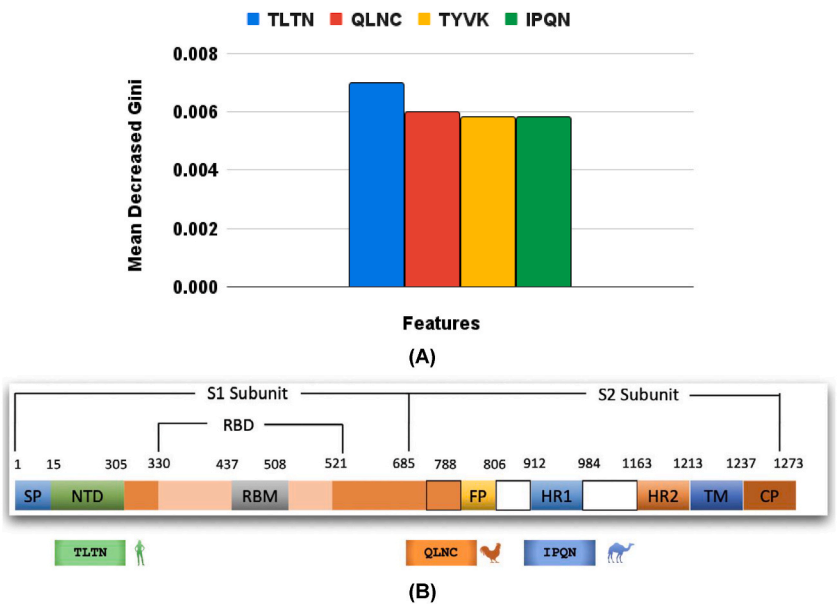

Fig. 2. Feature identification in spike Protein. (A) Important features in spike protein sequences relevant to host classification. (B) Spike protein sequence of SARS-CoV-2 with the important features mapped to their region of occurrence.

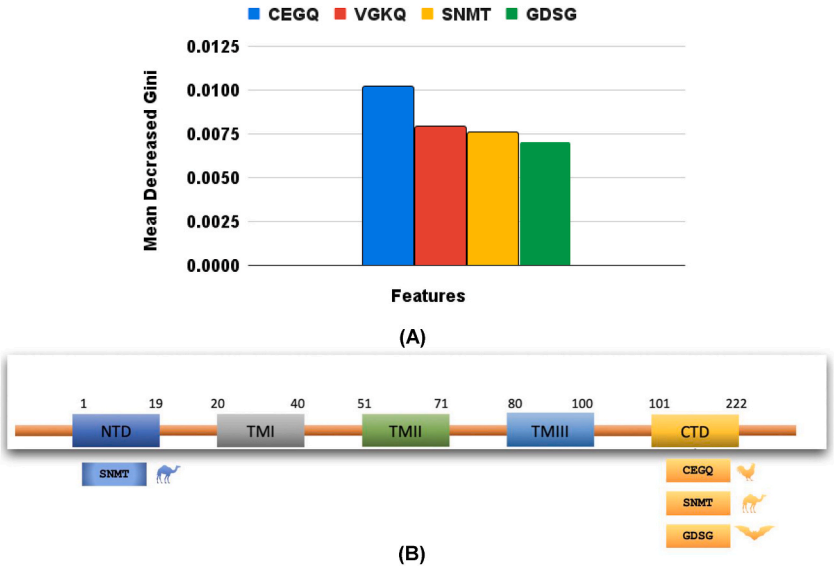

Fig. 3. Feature identification in membrane protein. (A) Important features in membrane protein sequences relevant to host classification. (B) Membrane protein sequence of SARS-CoV-2 with the important features mapped to their region of occurrence.

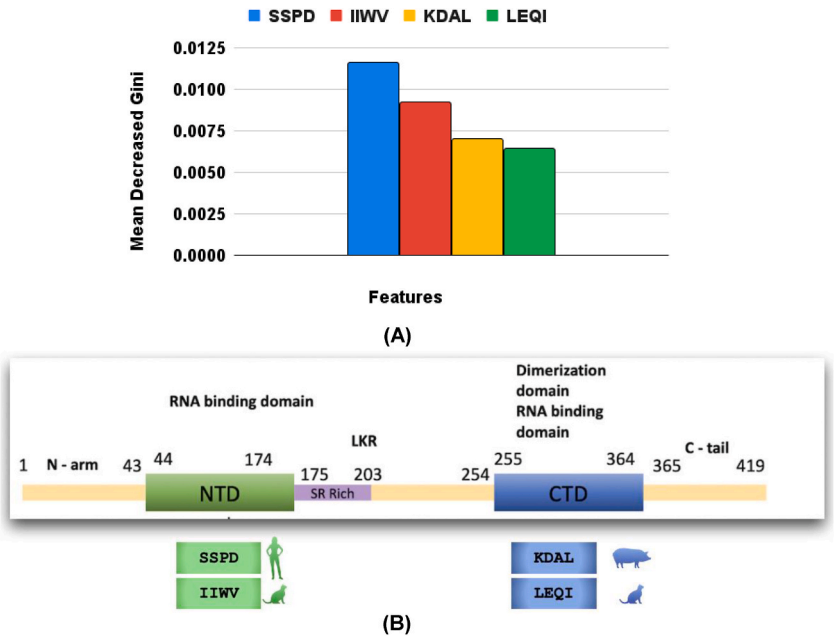

Fig. 4. Feature identification in nucleocapsid protein. (A) Important features in nucleocapsid protein sequences relevant to host classification. (B) Nucleocapsid protein sequence of SARS-CoV-2 with the important features mapped to their region of occurrence.

binding site in N-NTD, can inhibit the RNA binding activity of HCoV-OC43 and its replication [20].

3. Discussion

In this study, we proposed a pipeline *PreHost* to predict the host of viral sequences belonging to the Coronaviridae family. We presented a simple but effective host prediction pipeline that consists of data pre-processing, input preparation, the oversampling technique to tackle class imbalance, and machine learning classifiers. The machine learning algorithms were implemented to identify the potential host by leveraging four different viral sequences - nucleotide sequences, spike protein, M protein, and N protein sequences. The machine learning classifiers using 4-gram fragments as input to the host prediction task showed superior performances.

The biological interpretation of the 4-gram needs further exploration. Compared with the existing approaches, we dealt with a large number of viral sequences and took into account various proteins of coronaviruses. The high prediction accuracy across all four datasets indicates that machine learning models can be employed to predict a host of viral sequences. In addition, we have shown that M and N proteins can also be used to identify the host of the viral sequence. This suggests that both M and N proteins also have features specific to the host. However, the highest accuracy was obtained when genome sequence or spike protein was used, suggesting that these sequences can be used reliably for host prediction. Furthermore, we identified the most relevant features contributing to different host classes and mapped them to their respective domains. This may further aid in the understanding of host-virus interaction.

4. Methods

4.1. General workflow

We propose a machine-learning pipeline for host prediction using viral genomic and protein sequences. The proposed pipeline includes [i] pre-processing of the viral sequences, [ii] dealing with class imbalance using the oversampling technique SMOTE, [iii] host prediction using machine learning (ML) algorithms, and [iv] feature importance to identify the most relevant features for host prediction (Fig. 5).

4.2. Data collection and pre-processing

In this study, we obtained viral sequences (genome sequences, spike protein, membrane protein, and nucleocapsid protein sequences) as data points and the six hosts as classes. We have 50,700 nucleotide sequences, 9596 spike protein sequences, 846 M protein sequences, and 4021 N protein sequences associated with 6 hosts namely: human, bat, pig, chicken, camel, and cat (Table 5 and Table 6) from the viral Pathogen Database downloaded March 1, 2021 [21],[22]. We filtered these sequences to remove the incomplete genome sequences and the duplicate sequences. We split the dataset in an 80:20 ratio for training and testing sets with a balanced number of data points for each class using SMOTE (see below). We also generated an embedding for the biological sequences to be fed as input to the classification model.

4.3. Input preparation

To convert sequences to numerical values we used Tf-idf Vectorizer with N-grams since it splits the viral sequences into n-grams. We used different values of n for the N-gram. Tf-idf Vectorizer converts these N-grams to produce a sparse representation of the normalized counts of each gram. These features are given as input to the machine learning models.

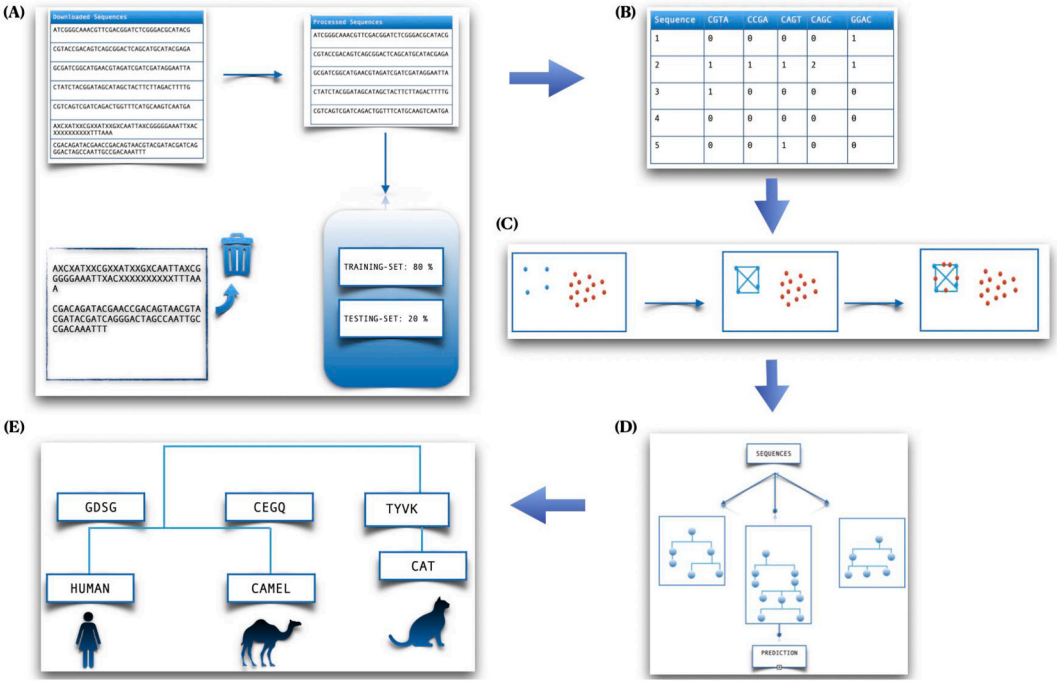

Fig. 5. The *PreHost* Prediction Pipeline. (A) Collection of viral sequences and pre-processing, (B) Generating a count matrix of 4-gram, (C) SMOTE analysis, (D) leveraging ML classifiers to predict host, and (E) analysis of predicted host and the key features.

Table 5
Host distribution for nucleotide/genome sequences.

Host	Number of Samples
Human	48,972
Bat	179
Swine	721
Chicken	381
Camel	355
Cat	92

Table 6
Host distribution for spike protein (S), membrane protein(M), and nucleocapsid protein(N) sequences.

Protein	Human	Bat	Camel	Chicken	Cat	Swine
S	8826	66	93	283	16	318
M	512	54	20	170	8	86
N	3503	70	41	237	14	169

4.4. SMOTE

The dataset considered in this study is highly imbalanced. Hence, we used the over-sampling technique SMOTE [23] to address the problem of class imbalance to generate synthetic data points for the minority classes. The minority class is over-sampled by taking each minority class sample and introducing synthetic samples along the line segments joining any/all of the k nearest neighbors of the minority class.

4.5. Classification

We predicted six host classes (Human, Bat, Camel, Cat, Swine, and Chicken) based on genome and protein sequences. To perform this task, we considered four well-known classifiers: Naive Bayes [24], KNN [25], Random Forest [26], and Decision Tree [27]. We used Python's Scikit learn [28] toolbox for the classification tasks. The parameters were set to default values in the sci-kit learn. The parameters used when training the Random Forest are n_estimators = 100, which is the number of trees built before the mean prediction was made, and max_features, which is the sqrt of the total number of features for the best split. To measure the quality of the split, we used Gini Impurity as our criterion. The nodes were expanded until all the leaves were pure. The metrics used for evaluation are Accuracy, F1-score, Recall, and Precision.

Accuracy is the classification metric well suited for multi-class classification. It is defined as the ratio of true predictions and the total number of samples (Eq [1]).

$$Accuracy = (TP + TN) \;/\; (TP + FP + FN + TN) \tag{1}$$

Precision is the classification metric used when we need to be certain about a prediction. It is defined as the ratio of correctly predicted positive samples and the total predicted positive samples (Eq [2]).

$$Precision = (TP) \;/\; (TP + FP) \tag{2}$$

Recall is the evaluation metric used to identify the total number of actual positives predicted. It is useful when we want to capture maximum positives (Eq [3]).

$$Recall = (TP) \;/\; (TP + FN) \tag{3}$$

F1-Score is the metric used to have a model with both good precision and recall. It is defined as the harmonic mean of precision and recall (Eq [4]).

$$F1 = 2 \;*\; (Precision \;*\; Recall) \;/\; (Precision + Recall) \tag{4}$$

4.6. Feature importance

In the host prediction task, it is also essential to locate the significant sites in distinguishing one host class from another. Since Random Forest has built-in feature importance, we used Random Forest to identify the relevant sites contributing to the different host classes. We used the Gini impurity for feature identification, considering the 4 grams as the features into which the viral sequences were split. The features that contribute most to reducing mean Gini impurity are deemed the most relevant.

Training a decision tree requires iteratively splitting the current data into two branches. For instance, if we have 10 data points, the perfect split would be two branches with 5 data points each. *Gini impurity* is a metric used to evaluate how good a split is quantitatively. It gives us the probability of misclassifying a data point. Mathematically, Gini impurity can be expressed as

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i)) \tag{5}$$

where C is the total number of classes and p(i) is the probability of selecting a data point of class i, for i in range (1, 6) and belongs to a class: Human, Cat, Camel, Bat, Chicken, Swine. The optimal split is calculated by subtracting the weighted impurities of the branches from the original impurities, also known as Gini Gain. The goal is to maximize Gini Gain.

Declarations

Authors contributions

Conceived and designed the experiments: U Deva Priyakumar and P.K.Vinod; Performed the experiments: Anusha Chaturvedi and Kushal Borkar; Analyzed and Interpreted the data: Anusha Chaturvedi and Kushal Borkar; Contributed reagents, materials, analysis tools or data: U Deva Priyakumar and P.K.Vinod; Wrote the paper: Anusha Chaturvedi, Kushal Borkar, U. Deva Priyakumar, P.K. Vinod.

Financial statement

U Deva Priyakumar and P K Vinod are financially supported by iHUB-Data, IIIT Hyderabad.

Data availability statement

Data included in article/supp. Material/referenced in article.

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.

Acknowledgement

We acknowledge the support provided by Dr. Manasa Kondamadugu in overall coordination.

References

[1] D.B. McArthur, Emerging infectious diseases, Nurs. Clin. 54 (2) (2019) 297–311.
[2] Chih-Cheng Lai, Tzu-Ping Shih, Wen-Chien Ko, Hung-Jen Tang, Hsueh Po-Ren, Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): the epidemic and the challenges, Int. J. Antimicrob. Agents 55 (3) (March 2020), 105924.
[3] José Millán, Alfonso Rodriguez, German Camacho, Mendoza Henry, A new emerging zoonotic virus of concern: the 2019 novel Coronavirus (SARS CoV-2), Infectio 24 (3) (2020) 187–192.
[4] R. de Groot, S. Baker, R. Baric, Virus Taxonomy: 2019 Release, July 2019. Technical report.
[5] Kit-San Yuen, Zi-Wei Ye, Sin-Yee Fung, Chi-Ping Chan, Dong-Yan Jin, SARS-CoV-2 and COVID-19: the most important research questions, Cell Biosci. 10 (1) (December 2020) 1–5.
[6] Cyrus M. Maher, Istvan Bartha, Steven Weaver, Predicting the mutational drivers of future SARS-CoV-2 variants of concern, Sci. Transl. Med. 14 (633) (February 2022), eabk3445.
[7] Brian Hie, D Zhong Ellen, Bonnie Berger, Learning the language of viral evolution and escape, Science 371 (6526) (2021) 284–288.
[8] Priyanka Mehta, Shanmukh Alle, Anusha Chaturvedi, Clinico genomic analysis reveals mutations associated with COVID-19 disease severity: possible modulation by RNA structure, Pathogens 10 (9) (August 2021).
[9] Fang Li, Structure, function, and evolution of coronavirus spike proteins, Annual review of virology 3 (1) (2016) 237.
[10] Peter J.M. Rottier, The coronavirus membrane glycoprotein, in: *The Coronaviridae.*, the Viruses, Springer, Boston, MA, 1995.
[11] Ruth McBride, Marjorie Van Zyl, C Fielding Burtram, The coronavirus nucleocapsid is a multifunctional protein, Viruses 6 (8) (August 2014) 2991–3018.
[12] Dewald Schoeman, C Fielding Burtram, Coronavirus envelope protein: current knowledge, Virol. J. 16 (1) (December 2019) 1–22.
[13] Yuan Huang, Chan Yang, Xin-feng Xu, Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19, Acta Pharmacol. Sin. 41 (9) (September 2020) 1141–1149.
[14] Jun Lan, Jiwan Ge, Jinfang Yu, Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor, Nature 581 (2020) 215–220.
[15] Qin Tang, Yulong Song, Mijuan Shi, Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition, Sci. Rep. 5 (1) (November 2015) 1–8.
[16] Florian Mock, Adrian Viehweger, Emanuel Barth, Manja Marz, H.O.P. Vid, Viral host prediction with deep learning, Bioinformatics 37 (3) (February 2021) 318–325.
[17] Kiril Kuzmin, Ayotomiwa Ezekie Adeniyi, Arthur Kevin DaSouza Jr., Machine learning methods accurately predict host specificity of coronaviruses based on spike sequences alone, Biochem. Biophys. Res. Commun. 533 (3) (December 2020) 553–558.
[18] Rumana Mahtarin, Shafiqul Islam, Md Jahirul Islam, Structure and dynamics of membrane protein in SARS-CoV-2, J. Biomol. Struct. Dyn. 40 (10) (June 2022) 4725–4738.
[19] Mei Yang, Suhua He, Xiaoxue Chen, Structural insight into the SARS-CoV-2 nucleocapsid protein C-terminal domain reveals a novel recognition mechanism for viral transcriptional regulatory, Front. Chem. 8: 624765 (2021).
[20] Wanchao Yin, Chunyou Mao, Xiaodong Luan, Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir, Science 368 (6498) (June 2020) 1499–1504.
[21] Brett Pickett, Douglas Greer, Yun Zhang, Virus pathogen Database and analysis resource(ViPR): a comprehensive bioinformatics Database and analysis resource for the coronavirus research community, Viruses 4 (11) (November 2012) 3209–3226.

[22] Brett Pickett, Eva Sadat, and Yun Zhang. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic acids research, 40(D1): D593–D598.
[23] V Chawla Nitesh, K.W. Bowyer, Lawrence Hall, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (June 2002) 321–357.
[24] Irina Rish, An Empirical Study of the Naive Bayes Classifier, vol. 3, August 2001, pp. 41–46.
[25] Gongde Guo, Hui Wang, David Bell, KNN Model-Based Approach in Classification, Springer, Berlin, Heidelberg, November 2003, pp. 986–996.
[26] Leo Breiman, Random forests, Mach. Learn. 45 (1) (October 2001) 5–32.
[27] S.R. Safavian, D. Landgrebe, A survey of decision tree classifier methodology, IEEE Transactions on Systems, Man, and Cybernetics 21 (3) (1991) 660–674.
[28] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Scikit learn: machine learning in Python, J. Mach. Learn. Res. 12 (November 2011) 2825–2830.