Journal of Zoonotic Diseases
https://jzd.tabrizu.ac.ir/
Published by the University of Tabriz
Online ISSN: 2717-2910

The machine learning-based predictor to identify putative COVID-19-like host jumping viruses

Varuni Bhardwaj and Mahesh Kulharia*

Centre for Computational Biology and Bioinformatics, Central University of Himachal Pradesh, Dharmshala, India

Article type:
Original article

Keywords:
COVID-19
Logistic regression
Machine learning
SMOTE
Zoonotic viruses

Article history:
Received:
March 8, 2025
Revised:
June 6, 2025
Accepted:
July 18, 2025
Available online:
July 21, 2025

Abstract

Zoonotic viruses, capable of crossing the species barrier from animals to humans, pose significant threats to global health, as demonstrated by outbreaks such as SARS, MERS, and COVID-19. Early identification of these viruses is critical for pandemic preparedness and containment strategies. Machine learning has increasingly been utilized in healthcare and virology to enhance predictive modeling. This study presents a machine learning-based approach for assessing the zoonotic potential of viruses by analyzing key biological features, including protein stability, RNA energy, protein folding success, and codon usage patterns. A curated dataset of viral spike protein sequences was compiled, encompassing both zoonotic and non-zoonotic viruses. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied, ensuring a balanced representation of both categories. The dataset was then normalized using z-score transformation to standardize feature distributions. A logistic regression model was trained and optimized through hyperparameter tuning to achieve an optimal balance between sensitivity, specificity, and accuracy. The model was evaluated using multiple validation strategies, including an independent testing dataset, to assess its robustness and generalizability. Results indicate that the model achieved a prediction accuracy of 78.57%, demonstrating its reliability in distinguishing zoonotic from non-zoonotic viruses. The high specificity ensures that the model effectively minimizes false positives, while sensitivity enables the detection of potential zoonotic threats. The interpretable nature of logistic regression makes the model transparent and applicable to real-world decision-making. By providing a systematic and data-driven approach, this study contributes to the early identification of emerging zoonotic threats, ultimately enhancing global health preparedness and response strategies.

Introduction

The number of zoonotic disease outbreaks has increased rapidly in recent decades, with disastrous repercussions for human health, economics, and society. The emergence of Coronavirus disease 2019 (COVID-19) occurred in Wuhan, China, in

***Corresponding author:** *kulharia@um.ac.ir*
https://doi.org/10.22034/jzd.2025.20118
https://jzd.tabrizu.ac.ir/article_20118.html
Cite this article: Bhardwaj V. and Kulharia M. The machine learning-based predictor to identify putative COVID-19-like host jumping viruses. Journal of Zoonotic Diseases, 2025, 9 (4): 976- 987.

late 2019 and swiftly expanded to become a significant pandemic (1). The single-stranded positive-sense RNA virus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is the causal pathogen of this disease (2, 3). It belongs to the beta coronavirus family and shares a close relation with the human SARS-CoV virus, responsible for the SARS outbreak from 2002 to 2004 (4). As on the 31st of March 2023, about 760 million people have been infected and over 6.9 million deaths attributed to COVID-19 so far (5).

Coronaviruses (CoVs) belong to the family Coronaviridae of the order Nidovirales are large (genome size: 27-32 Kb), enveloped, non-segmented, positive single-strand RNA viruses that have been responsible for many respiratory illness outbreaks including SARS and MERS (6–10). They are called 'corona viruses' because of their characteristic protein spikes (11). By attaching to the ACE2 receptor on the surface of the host cell, the SARS-CoV-2 spike protein functions as the primary viral attachment protein, enabling the virus to enter human cells (11–13). This interaction is critical for viral infectivity, making the spike protein a primary target for vaccine development.

The COVID-19 pandemic, however, has brought into sharp focus the critical demand for timely virus discovery and diagnosis of zoonotic viruses with high potential of cross-species jump from animal to humans (14). Predicting viruses with zoonotic potential is essential for improving response to emerging infectious diseases and effective global health preparedness (15). Logistic regression and other machine learning algorithms can help identify patterns and characteristics associated with zoonotic potential by examining several biological and structural properties of viruses (16). Logistic regression is a statistical method that can predict binary events, such as whether a given virus could jump from animals to humans (17).

In this study, logistic regression is applied for the classification of viruses having COVID-like threat potential. Specifically, it determines whether a virus has traits that increase its likelihood of jumping from animals to humans. To accomplish this, a dataset was compiled consisting viral spike protein sequences of SARS-CoV-2 and viruses belonging to coronaviridae family that have successfully infected humans and spike protein sequences of viruses that have not demonstrated human infectivity. Key biological and structural characteristics, including Protein stability, RNA energy, the Absolute Contact order (AbsCO) score, which indicates success of protein folding and the Root Mean Square Deviation (RMSD) of codon usage were analysed. These features were selected based on their importance in viral function, structure, and evolutionary history.

To prepare the dataset for analysis and ensure that all features are on a comparable scale, Z-score normalization (18) was applied to the dataset. One of the significant challenges in viral threat identification is the class imbalance between viruses that have jumped to humans and those that have not. The Synthetic Minority Oversampling Technique (SMOTE) (19) was applied to address this issue, ensuring an equal representation of both classes. Logistic Regression was the major classification technique because of its simplicity, interpretability, and efficacy in binary classification problems (20). The hyperparameter adjustment of the regularization parameter (21) was performed, which was important for enhancing the model's generalization ability and accuracy.

After fine-tuning, the model was tested on a separate dataset comprising 14 viruses, representing real-world examples to assess its predictive performance. The model accurately predicted 11 viruses, demonstrating its ability to identify zoonotic threats effectively. These results further confirm the utility of the model as a reliable scoring function for predicting zoonotic potential, emphasizing its robustness and scalability for assessing future viral threats.

By integrating key biological features and utilizing machine learning techniques, this study offers a comprehensive and interpretable approach for the early identification of viruses with zoonotic

potential. The final testing findings underscore the model's utility as a valuable tool in the global effort to limit the risks posed by the emerging infectious illness, thereby considerably boosting health preparedness and response strategies.

Methods

Data collection

For this study, two distinct datasets were required: a positive dataset and a negative dataset. The positive dataset was created to include sequences of viruses capable of infecting humans, focusing on proteins required for host cell entry. In contrast, the negative dataset comprised of sequences from viruses that are known not-infect humans.

For the positive dataset, spike protein sequences from different lineages of SARS-CoV-2 were retrieved from the NCBI (National Centre for Biotechnology Information) Virus database (22). Additionally, spike protein sequences of a few other viruses belonging to the coronaviridae family were also added.

The sequences of viruses for the negative dataset were also downloaded from the NCBI database. Due to reduced availability of RNA sequences for some viruses non-infectious to humans, a reverse translation process was employed. The Sequence Manipulation Suite (SMS) was used to convert these viruses' amino acid sequences to their corresponding RNA sequences (23). The Codon Usage Table for each virus from the Codon Usage Database was employed to guide the reverse translation of amino acid sequences into RNA sequences during this process.

Parametric selection

Successful viral host jump is governed by the combination of molecular and structural factors that enable adaptation to new hosts. This study incorporates four key factors: protein structure stability, RNA energy, protein folding and codon conformity. Protein structural stability is essential for maintaining the functional integrity of entry proteins, ensuring their ability to bind effectively to host receptors under diverse physiological conditions. RNA energy, a measure of the stability of RNA secondary structures, regulates viral entry protein expression and replication efficiency, which are pivotal for initiating infection. Proper protein folding ensures that the entry proteins achieve their functional states, as misfolded proteins ay fail to facilitate host receptor binding or evade immune surveillance. Codon usage reveals how well the viral genetic material is optimized for the host's translational machinery, which has a direct impact on the synthesis and functionality of entry proteins. These parameters collectively provide a mechanistic insight into how entry-related viral proteins adapt during host jumps, forming the basis of this study's approach.

Protein Stability Assessment

To assess the stability of predicted protein structures, a two-step computational approach was employed.

Structure Prediction: For the structure prediction of the proteins, the amino acid sequences were submitted to the Phyre2 server (24), which is a powerful protein structure homology modelling platform (24).

Stability calculation: The generated PDB structures were then examined for protein stability using the FoldX software (25). The FoldX Stability determines the Gibbs free energy (ΔG) needed to fold a protein from its unfolded state (26). The software generates an output file that breaks the overall G into different energy components. The total energy is an empirical score derived from several energy terms, which are sidechain H bond, van der Waals, solvation polar, solvation hydrophobic, backbone Hbond, electrostatics, van der Waals clashes, entropy side chain, entropy main chain, sloop entropy, Mloop entropy, cis bond, torsional clash, backbone clash, helix dipole, water bridge, disulfide, electrostatic kon, partial covalent bonds, energy ionization, entropy complex and residue number (25).

RNA energy calculation

RNA energy, an important factor in viral host jump prediction, was determined using the RNAfold

software (27). This tool uses thermodynamics concepts to estimate the secondary structure of RNA molecules based on their primary sequence. The key output of RNAfold is the Minimum Free Energy (MFE) (28), the most stable secondary structural configuration of an RNA molecule under specified thermodynamic conditions. By computing the MFE for each RNA sequence, insights into the RNA molecule's structural stability were obtained. The RNA energy calculation was done for the RNA sequences.

Protein folding success: Absolute Contact Order (AbsCO)

The Absolute Contact Order (AbsCO) was determined for the protein structures, to determine the likelihood of successful protein folding. This metric gives a quantitative assessment of a protein's folding complexity and stability. AbsCO is calculated by identifying pairs of amino acid residues within a specific distance (usually 6 Angstroms) inside a protein structure, representing crucial interactions for stability. The absolute difference between each pair's sequential positions throughout the protein backbone is calculated, and these values are added to get the final AbsCO score. A higher AbsCO suggests a more compact folding pattern, potentially increasing the risk of folding (29).

Codon conformity analysis

Codon usage bias, the non-uniform use of synonymous codons throughout a genome, can have a major impact on protein expression efficiency. To determine the possibility of viral host jump, the codon usage patterns of viral sequences were compared to those of human host.

To calculate the Root Mean Square Deviation (RMSD) of viral and human codon usage frequency distributions a quantitative approach was used as given in equation 1. The potential adaptability of a virus to the new host was analysed based on its codon conformity profile, where f_i^V is the frequency of codon i in the viral sequence, f_i^H is the frequency of codon i in the human genome and N is the number of codons.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(f_i^V - f_i^H\right)^2} \qquad Eq.1$$

Data preprocessing

The raw data frequently contains inconsistencies, errors, and missing values, which can significantly impact the accuracy and reliability of the results. Therefore, data preprocessing is employed before data analysis since it optimizes the work of machine algorithms. In this step, the raw data is prepared for modelling. It involves cleaning, transforming, and organizing data to ensure its quality and suitability for analysis. Data cleaning is removing noise and inconsistencies from data to improve its quality. To rescale the data methods, such as normalization and standardization are used. This is known as data transformation.

The normalization (30) is a data preprocessing technique that scale numerical features to a common range. This is important as, in the learning process, features with larger magnitudes tend to dominate those at lower scales, resulting in biased models. Normalizing the data such that all features are on a common scale ensures that all features have roughly the same effect on the model's predictions. There are several normalizing methods, including min-max scaling, z-score normalization, and decimal scaling (31). In this study, Z-score normalization was applied to the dataset. The Python script for Z-score normalization is available in the GitHub repository: https://github.com/varunibhardwaj/Supplementary_Code_Zoonotic_Prediction/blob/main/z_score_normalization.py. Z-score normalization standardizes data by subtracting the mean from each data point and dividing by its standard deviation (32). This converts the data into a standard normal distribution

with a mean of 0 and a standard deviation of 1 (33). This method of normalization was chosen to ensure the dataset containing the virus data points, with features protein stability, protein folding success, RNA energy and codon conformity had equal importance. The z-score normalization ensured that the model was not biased towards the features with larger magnitudes. This method enabled a more balanced comparison of the features, enhancing the model's ability to detect patterns.

Resampling is another data preprocessing technique for correcting class imbalance in datasets where one class significantly outnumbers the other (34, 35). It can be done in two ways: under sampling, which reduces the number of samples in the majority class, and oversampling, which increases the number of samples in the minority class (36). The Synthetic Minority Over-Sampling Technique (SMOTE) was employed to solve the dataset's class imbalance. The implementation details can be found in the GitHub repository: https://github.com/varunibhardwaj/Supplementary_Code_Zoonotic_Prediction/blob/main/smote_resampling.py. SMOTE is an over-sampling method that generates synthetic training samples for the minority class using the existing minority class samples and their nearest neighbours (37, 38). By applying SMOTE, the dataset was augmented with new negative data points, so that model can be equally trained on both classes. This procedure resulted in a more balanced representation of both classes, essential for training a robust model that can accurately distinguish between positive and negative classes.

Logistic regression and Hyperparameter tuning

In this study, a Logistic regression model was employed to the dataset. Logistic regression is a type of machine learning algorithm that provides a starting point for many binary classification tasks by modelling the probability that each input belongs to any particular category (39). The regularization strength in logistic regression is controlled by the regularization parameter C. It reduces overfitting by penalizing large coefficients, ensuring the model generalizes well to unseen data.

The model was fine-tuned by adjusting the value of hyperparameter C (21, 40). Various C values were examined to evaluate their impact on model performance.

The five-fold cross-validation was used to see the performance of model changes across different values of C. The logistic regression models were trained and tested for each value of C. By analysing the performance metrics across the five iterations, the optimal C value was determined, balancing underfitting and overfitting to provide most reliable predictive model.

Model evaluation

After tuning the logistic regression model, evaluation measures were calculated. These measures were true positives, true negatives, false positives, and false negatives. These values were used to calculate performance indicators like accuracy, precision, specificity, and sensitivity.

Accuracy (Eq. 2) assesses the model's overall correctness, whereas precision (Eq. 3) represents the proportion represents the proportion of true positives among all predicted positives. Specificity (Eq.4), or true negative rate, assesses the model's capacity to accurately identify negative situations, whereas sensitivity (Eq.5) (recall) shows the model's ability to recognize true positives.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \qquad Eq.2$$

$$Precision = \frac{TP}{TP + FP} \qquad Eq.3$$

$$Sensitivity = \frac{TP}{TP + FN} \qquad Eq.4$$

$$Specificity = \frac{TN}{TN + FP} \qquad Eq.5$$

Model testing

A separate testing dataset of 14 viruses was prepared, to evaluate the model's predictive power further. The fine-tuned logistic regression model was applied to this dataset to independently assess the model's ability to distinguish between viruses with zoonotic potential and those without. The testing procedure involved calculating the classification outputs for each virus and comparing them to their actual labels, which allowed for the assessment of the model's predictive reliability.

Results and discussion

Data Collection

973 spike protein sequences were retrieved from the NCBI Virus database for the positive dataset. These sequences reflect different lineages of SARS-CoV-2 and were collected between January 2020 and March 2023. Additionally, spike protein sequences of 4 other viruses belonging to coronaviridae family that have earlier jumped into humans were added into the positive dataset. A total of 977 sequences constituted the positive dataset. A set of 62 virus sequences that had been known for their inability to infect humans, was obtained from the NCBI Virus database for the negative dataset. These included viruses from both coronaviridae and non-coronaviridae families (such as *Flaviviridae*, *Retroviridae*, *Picornaviridae*, and others), providing a degree of taxonomic diversity. Among these, 16 viruses belong to families other than Coronaviridae, thereby enhancing the biological variety of the negative class. As RNA sequences of 12 of these viruses were unavailable, reverse translation was performed using Sequence Manipulation Suite (SMS) and Codon conformity table for each virus. All 12 sequences were converted, resulting in a complete negative dataset.

Protein stability calculation

A total of 1039 protein structures were generated using the Phyre2 server based on homology modelling. The stability of the anticipated structures was evaluated by calculating their Gibbs free energy (ΔG) using FoldX. The ΔG values for viruses infectious to humans had range from 924.47 kcal/mol to 2165.49 kcal/mol. For viruses non-infectious to humans, the range of ΔG values lied between 25.65 kcal/mol to 2629.2 kcal/mol. The Gibbs free energy and energy components provided insights into the stability of the proteins.

RNA energy calculation

RNAfold was used to calculate the RNA energy of 1039 sequences. The Minimum Free Energy (MFE) for each RNA sequence was determined. The RNA energy range for viruses in the positive dataset was -1087.9 Kcal/mol to -595.6 Kcal/mol, while viruses in the negative dataset ranged from -1350.4 Kcal/mol to -66.6 Kcal/mol.

AbsCO score calculation

The Absolute Contact Order (AbsCO) was calculated for all protein structures. AbsCO values were in the range of 31.24 to 40.04 for positive viral proteins, indicating a relatively consistent folding complexity. However, for negative viral proteins, AbsCO score ranged from 1.94 to 61.92, reflecting greater variability in folding.

Codon conformity analysis

The Root Mean Square Deviation (RMSD) between viral and human codon usage patterns was analysed to determine the likelihood of viral host adaptation. The viruses infectious to humans had RMSD value between 0.0108 to 0.0545, showing a high level of codon similarity with the human host. In contrast, viruses non-infectious to humans showed the higher RMSD scores, from 0.0247 to 0.0573, indicating of lower codon compatibility.Supplementary information 1 provides the data of four parameters

for both positive and viruses non-infectious to humans.

Data preprocessing

As the raw data may contain errors and inconsistencies, it was important to curate the raw data for modelling. Data cleaning, normalization and standardization were included in the process of preprocessing of data.

Normalization of the viral datapoints of (a) protein stability, (b) folding success, (c) RNA energy and (d) codon conformity was done by applying Z-score normalization on the dataset. This transformation ensured all features were normalized, with a mean of 0 and a standard deviation of one, allowing for balanced comparison across features. Following normalization, the positive dataset exhibited the following feature ranges: ΔG from -1.6778 kcal/mol to 4.2791 kcal/mol, RNA energy from -0.31298 kcal/mol to 4.6646 kcal/mol, AbsCO from -2.0896 to 0.6004, codon conformity from -0.2481 to 4.3846. For the negative dataset, the normalized ranges were: ΔG from -5.9929 kcal/mol to 6.5084 kcal/mol, RNA energy from -2.9670 kcal/mol to 10.0132 kcal/mol, AbsCO from -11.0465 to 7.892 and codon conformity from 1.0921 to 4.6940. The normalized data of four parameters for the dataset is given in supplementary information 2.

After normalization of the dataset resampling of the dataset was carried out. SMOTE was applied to the dataset to address the class imbalance, resulting in a balanced dataset of 977 positive and 977 negative samples as given in supplementary information 3. This resampling process ensured that both classes were evenly represented as shown in Figure 1.

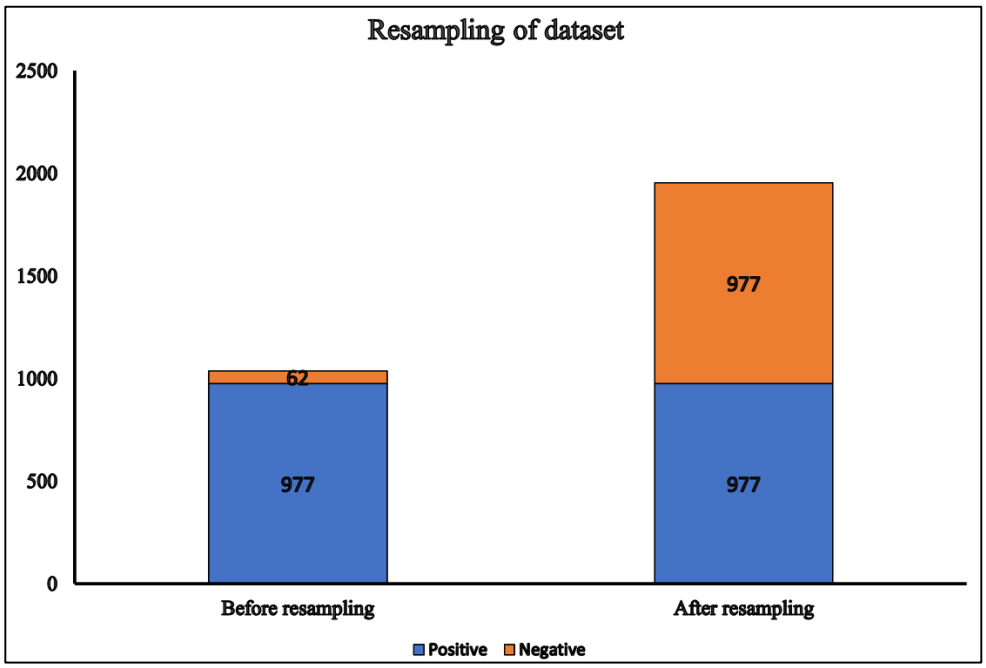

Fig. 1. The dataset before resampling and after resampling

The logistic regression and hyper parameter tuning

Logistic regression model was evaluated using five different datasets. Various values of the regularization parameter C (0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.01, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 1) were tested to assess their impact on model performance.

Model performance evaluation and final model as a scoring function

The performance of the different values of regularization parameter C on the model was assessed and compared. The values obtained are

given in supplementary information 4. Among the tested values, C = 0.001 demonstrated the optimal balance of performance metrics across all folds, as shown in supplementary information 5. The model exhibited high sensitivity from 0.995 to 1.00, indicating its ability to identify true positives accurately. Precision values were consistent, highlighting the model's ability to minimize false positives effectively. Specificity ranged from 0.841 to 0.918, reflecting variability in correctly identifying true negatives across folds. The F1-score and AUC further validated the model's balanced performance.

The finalized model, with regularization parameter C = 0.001, was saved. The code for saving the model is available in the GitHub repository: https://github.com/varunibhardwaj/Supplementary_Code_Zoonotic_Prediction/blob/main/save_model.py. This saved model functions as a scoring tool to assess new viruses for their transformation into "human-infectors" or not by evaluating their protein stability, RNA energy, folding success and codon conformity features. The python script for predicting the zoonotic potential of viruses is available in the GitHub repository: https://github.com/varunibhardwaj/Supplementary_Code_Zoonotic_Prediction/blob/main/predict_zoonotic_risk.py.

Evaluation of the final model

The final saved model was applied to the testing dataset comprising of data of 14 viruses. Of these, 11 viruses belonged to families other than *Coronaviridae*, including Orthomyxoviridae (Influenza A virus), Retroviridae (HIV-1), Filoviridae *(Zaire ebolavirus, Marburg virus*), *Paramyxoviridae* (*Henipavirus nipahens*e, *Henipavirus hendraense), Flaviviridae (Zika virus, Dengue virus, Yellow fever virus), Rhabdoviridae (Lyssavirus rabies, Snakehead rhabdovirus*), and others. This diverse representation was intentionally selected to evaluate the generalizability of the model across a wide range of viral families. The model achieved a prediction accuracy of 78.57%, highlighting its strong ability to identify viruses with zoonotic potential, as shown in Figure 2 and Figure 3. These results further validate the model's effectiveness as a reliable scoring tool for predicting the zoonotic potential of unknown viruses based on key biological features.

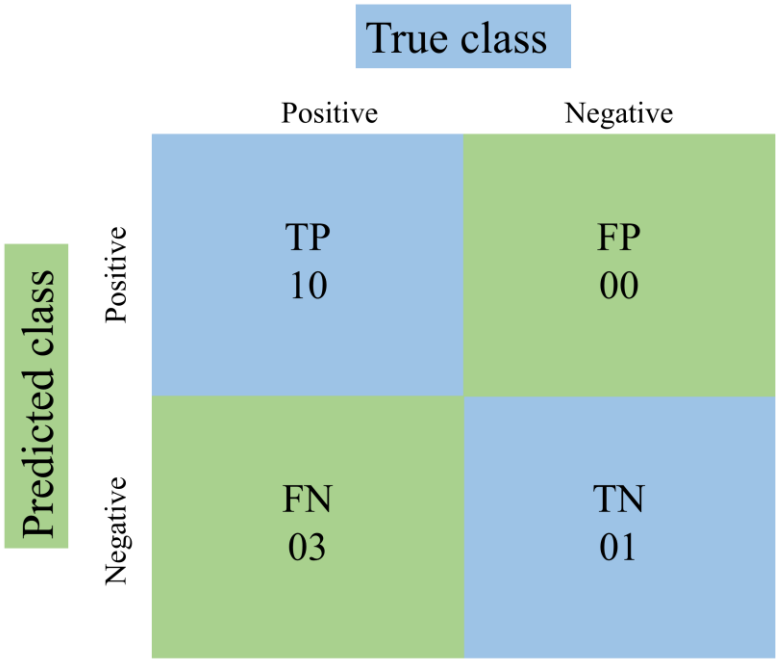

Fig. 2. Final Model Testing: Confusion Matrix Representation

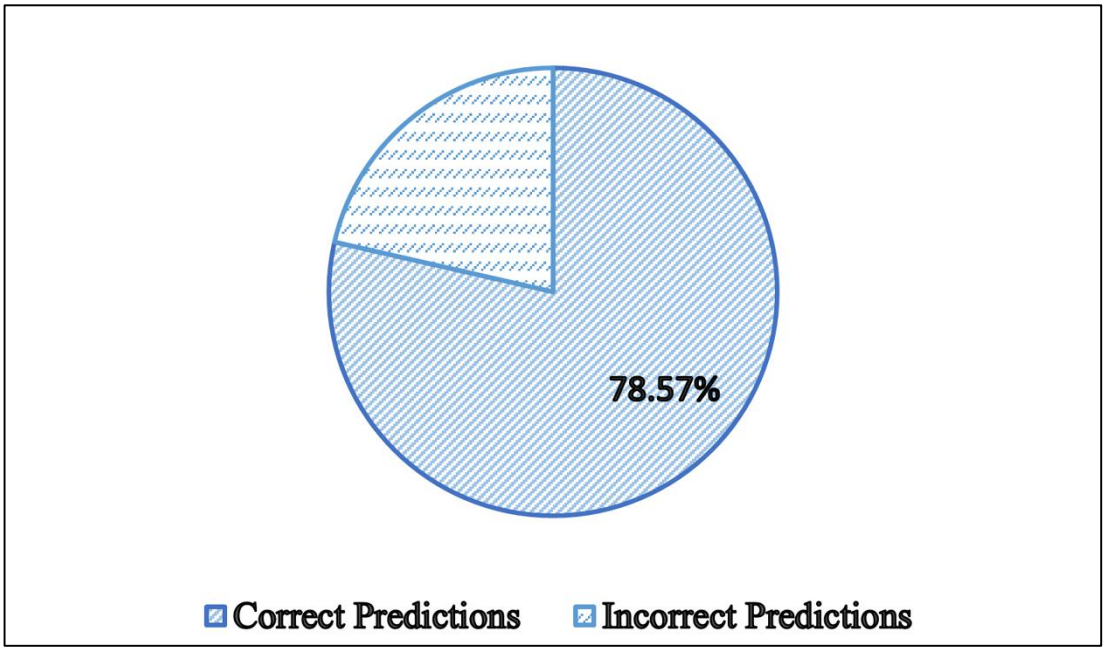

Fig. 3. Classification accuracy of the final model on the testing dataset

Feature Importance Analysis

To understand the contribution of each biological feature in predicting zoonotic potential, the feature importance derived from the logistic regression model was analyzed. The absolute values of the model coefficients were used to assess the influence of each feature.

As illustrated in Figure 4, the most influential feature was RMSD (Codon usage), suggesting that codon adaptation plays a significant role in the ability of a virus to jump into humans. This was followed by the AbsCO Score (Folding success), RNA energy (RNA stability), and Delta G (Protein stability). These results highlight the biological relevance of each feature in modelling host adaptability and support the robustness of the selected predictors. The features were chosen based on their mechanistic importance: Delta G reflects protein stability within the host environment; RNA energy represents the structural stability of viral RNA, influencing replication and translation; AbsCO Score indicates the efficiency of viral protein folding; and RMSD captures codon adaptation to the host's translational machinery.

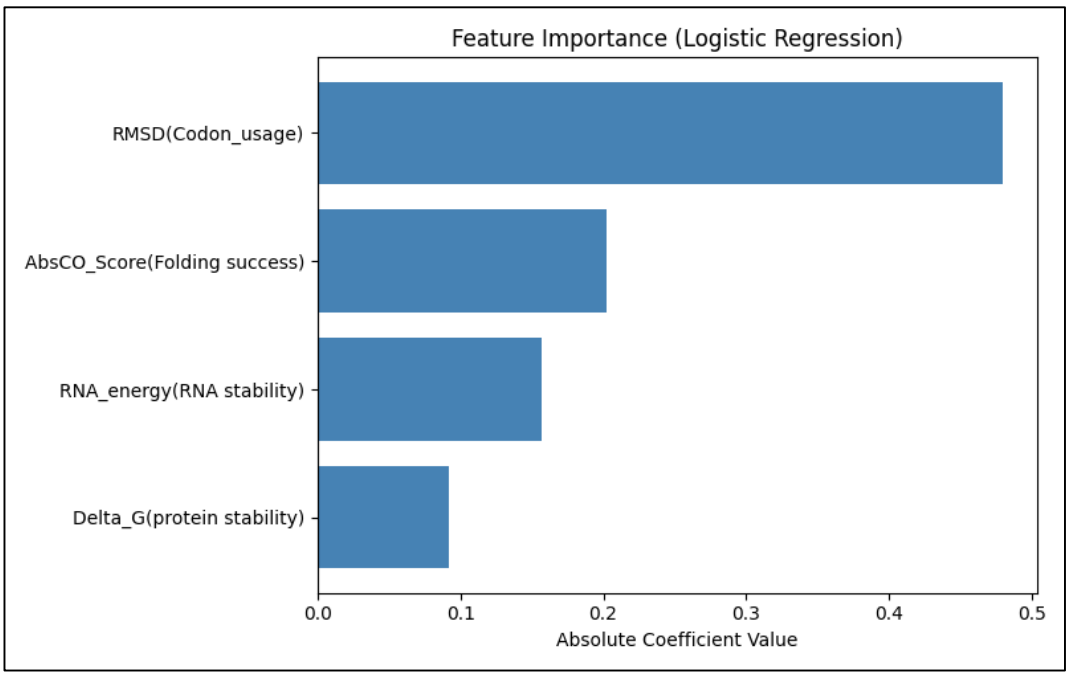

Fig. 4. Feature importance based on logistic regression coefficients

Conclusion

The development of reliable tools to predict viral spillover events and mitigate future pandemics has become necessary due to the growing threat of zoonotic diseases. This study developed and evaluated a machine learning based model to predict viruses that may jump onto humans as hosts, using key biological features such as protein stability, RNA energy, protein folding success and codon conformity. By addressing significant challenges like data normalization, class imbalance through SMOTE, and hyperparameter optimization, the model achieved an optimal balance of sensitivity, specificity, and accuracy. Final testing on a separate dataset of 14 real-world viruses yielded a prediction accuracy of 78.57%, underscoring the model's reliability in identifying zoonotic threats. These findings underscore the model's potential to serve as an important scoring function in viral risk assessments, contributing to the early identification of viruses with pandemic potential. Moreover, the model has the potential to be part of global health surveillance systems, thus supporting efforts to mitigate pandemic risks. Future studies should focus on refining the model by incorporating a broader range of viral features and expanding the training dataset, which could improve its generalizability and accuracy.

Acknowledgements

The authors would like to thank the Central University of Himachal Pradesh for providing the computational facilities and UGC non-NET fellowship for financial support.

Conflict of interests

The authors have no conflict of interest.

Ethical approval

Not applicable.

References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China. 2019. N Engl J Med. 2020;382(8):727-733. https://doi.org/10.1056/nejmoa2001017
2. Tahan S, Parikh BA, Droit L, Wallace MA, Burnham CAD, Wang D. SARS-CoV-2 E gene variant alters analytical sensitivity characteristics of viral detection using a commercial reverse transcription-PCR assay. J Clin Microbiol. 2021;59(7):e0007521. https://doi.org/10.1128/jcm.00075-21

3. Cao Y, Wang J, Jian F, Xiao T, Song W, Yisimayi A, et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. Nature. 2022;602(7898):657-663. https://doi.org/10.1038/s41586-021-04385-3
4. Markov P V, Ghafari M, Beer M, Lythgoe K, Simmonds P, Stilianakis NI, et al. The evolution of SARS-CoV-2. Nat Rev Microbiol. 2023;21(6):361-379. https://doi.org/10.1038/s41579-023-00878-2
5. Word Health Organization (WHO).WHO COVID-19 dashboard, COVID-19 Cases, World. [Internet]. Available from: https://data.who.int/dashboards/covid19/cases. [17th September 2024].
6. Pal M, Berhanu G, Desalegn C, Kandi V. Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2): An Update. Cureus. 2020;12(3):e7423. https://doi.org/10.7759/cureus.7423
7. Payne S. Family Coronaviridae. Viruses. 2017:149–58. https://doi.org/10.1016/B978-0-12-803109-4.00017-9
8. Deng X, Baker SC. Coronaviruses: Molecular Biology (Coronaviridae). Encyclopedia of Virology. 2021:198–207. https://doi.org/10.1016/B978-0-12-814515-9.02550-9
9. Sharma HN, Latimore COD, Matthews QL. Biology and Pathogenesis of SARS-CoV-2: Understandings for Therapeutic Developments against COVID-19. Pathogens. 2021;10(9):1218. https://doi.org/10.3390/pathogens10091218
10. Zhou Z, Qiu Y, Ge X. The taxonomy, host range and pathogenicity of coronaviruses and other viruses in the Nidovirales order. Anim Dis. 2021;1(1):5. https://doi.org/10.1186/s44149-021-00005-9
11. Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S, et al. A novel coronavirus associated with severe acute respiratory syndrome. N Engl J Med. 2003;348(20):1953-66. https://doi.org/10.1056/NEJMoa030781
12. Jackson CB, Farzan M, Chen B, Choe H. Mechanisms of SARS-CoV-2 entry into cells. Nat Rev Mol Cell Biol. 2022;23(1):3-20. https://doi.org/10.1038/s41580-021-00418-x
13. Tang X, Liu Y, Wang J, Long T, Yee Mok BW, Huang Y, et al. Identifications of novel host cell factors that interact with the receptor-binding domain of the SARS-CoV-2 spike protein. J Biol Chem. 2024;300(6):107390. https://doi.org/10.1016/j.jbc.2024.107390
14. Ilkhani H, Hedayat N, Farhad S. Novel approaches for rapid detection of COVID-19 during the pandemic: A review. Anal Biochem. 2021;634:114362. https://doi.org/10.1016/j.ab.2021.114362
15. Mollentze N, Streicker DG. Predicting zoonotic potential of viruses: where are we? Curr Opin Virol. 2023;61:101346. https://doi.org/10.1016/j.coviro.2023.101346
16. Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM. Classification and prediction of diabetes disease using machine learning paradigm. Health Inf Sci Syst. 2020;8(1):7. https://doi.org/10.1007/s13755-019-0095-z
17. Lombardo L, Cama M, Conoscenti C, Märker M, Rotigliano E. Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy). Nat Hazards. 2015;79(3):1621–48. https://doi.org/10.1007/s11069-015-1915-3
18. Suarez-Alvarez MM, Pham DT, Prostov MY, Prostov YI. Statistical approach to normalization of feature vectors and clustering of mixed datasets. Proc Math Phys Eng Sci. 2012;468(2145):2630–51. https://doi.org/10.1098/rspa.2011.0704
19. Cateni S, Colla V, Vannucci M. A method for resampling imbalanced datasets in binary classification tasks for real-world problems. Neurocomputing. 2014;135:32–41. https://doi.org/10.1016/j.neucom.2013.05.059
20. Céline S, Celine S, Maria Dominic M, Savitha Devi M. Logistic Regression for Employability Prediction. Int J Innov Technol Explor Eng. 2020; 9(3):2471-2478. https://doi.org/10.35940/ijitee.C8170.019320
21. Wojciuk M, Swiderska-Chadaj Z, Siwek K, Gertych A. Improving classification accuracy of fine-tuned CNN models: Impact of hyperparameter optimization. Heliyon. 2024;10(5):e26586. https://doi.org/10.1016/j.heliyon.2024.e26586
22. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford). 2020;2020:baaa062. https://doi.org/10.1093/database/baaa062

23. Stothard P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques. 2000;28(6):1102, 1104. https://doi.org/10.2144/00286ir01
24. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10(6):845-58. https://doi.org/10.1038/nprot.2015.053.
25. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res. 2005;33:W382-8. https://doi.org/10.1093/nar/gki387
26. Guerois R, Nielsen JE, Serrano L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. J Mol Biol. 2002;320(2):369-87. https://doi.org/10.1016/S0022-2836(02)00442-4
27. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26. . https://doi.org/10.1186/1748-7188-6-26
28. Tripathi A, Mishra KK, Tiwari S, Vashist PC. Nature inspired optimization algorithm for prediction of "minimum free energy" "RNA secondary structure." J Reliab Intell Environ. 2019;5(4):241–57. https://doi.org/10.1007/s40860-019-00091-0
29. Shi Y, Zhou J, Arndt D, Wishart DS, Lin G. Protein contact order prediction from primary sequences. BMC Bioinformatics. 2008;9:255. https://doi.org/10.1186/1471-2105-9-255
30. Singh D, Singh B. Investigating the impact of data normalization on classification performance. Appl Soft Comput. 2020;97:105524. https://doi.org/10.1016/j.asoc.2019.105524
31. Jain A, Nandakumar K, Ross A. Score normalization in multimodal biometric systems. Pattern Recognit. 2005;38(12):2270–85. https://doi.org/10.1016/j.patcog.2005.01.012
32. Latha L, Thangasamy S. Efficient approach to Normalization of Multimodal Biometric Scores. Int J Comput Appl. 2011;32(10):975–8887.
33. Alshdaifat E, Alshdaifat D, Alsarhan A, Hussein F, El-Salhi SMFS. The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms' Performance. Data 2021, Vol 6, Page 11. 2021;6(2):11. https://doi.org/10.3390/data6020011
34. Liu S, Zhang J, Xiang Y, Zhou W, Xiang D. A study of data pre-processing techniques for imbalanced biomedical data classification. Int J Bioinform Res Appl. 2020;16(3):290–318. https://doi.org/10.48550/arXiv.1911.00996
35. Marqués AI, García V, Sánchez JS. On the suitability of resampling techniques for the class imbalance problem in credit scoring. J Oper Res Soc. 2013;64(7):1060–70. https://doi.org/10.1057/jors.2012.120
36. Wongvorachan T, He S, Bulut O. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. Information 2023, Vol 14, Page 54. 2023;14(1):54.https://doi.org/10.3390/info14010054
37. Elreedy D, Atiya AF, Kamalov F. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Mach Learn. 2024;113(7):4903–23. https://doi.org/10.1007/s10994-022-06296-4
38. Elreedy D, Atiya AF. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. Inf Sci (N Y). 2019;505:32–64. https://doi.org/10.1016/j.ins.2019.07.070
39. de Menezes FS, Liska GR, Cirillo MA, Vivanco MJF. Data classification with binary response through the Boosting algorithm and logistic regression. Expert Syst Appl. 2017;69:62–73. https://doi.org/10.1016/j.eswa.2016.08.014
40. Krithiga R, Ilavarasan E. Hyperparameter tuning of AdaBoost algorithm for social spammer identification. Int J Pervasive Comput Commun. 2020;17(5):462–82. https://doi.org/10.1108/IJPCC-09-2020-0130