

Zoon0PredV: Potential Virus Species Crossover Prediction Using Convolutional Neural Networks and Viral Protein Sequence Patterns

Bioinformatics and Biology Insights

Volume 20: 1–16

© The Author(s) 2026

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/11779322251415123

journals.sagepub.com/home/bbi



Rudolph Abel Serage¹, Clement Nthambazale Nyirenda^{2,3},
Taiwo Gabriel Omomule¹ , Alan Gilbert Christoffels^{1,4} and
Dominique Elizabeth Anderson¹

Abstract

Biomedical science has made substantial progress toward diagnosing, understanding the pathogenesis, and treating various causative agents of infectious disease. However, novel microbial pathogens continue to emerge, and existing pathogens continue to evolve alternative strategies to thrive in ever-changing environments. Various infectious disease etiological agents originate from animal reservoirs, and several have, over time, acquired the ability to cross the species barrier, altering their host range. Computational approaches in biomedical science capable of analyzing large datasets are invaluable for predicting and monitoring disease outbreaks and their effectiveness is greatly enhanced when integrated with machine learning techniques. The goal of this study is to develop a machine learning model for the prediction of potentially zoonotic organisms, using viral surface proteins that facilitate host cell entry as input data. Sequence data and metadata were obtained from UniProtKB, transformed into a machine-readable format, using frequency chaos game representation and a convolutional neural network model was developed to identify sequence patterns consistent with viruses which infect humans. The model achieves generalized performance of 96.78% accuracy, 0.97 F1 score, and 0.93 MCC (Matthews Correlation Coefficient) on unseen data. The model potentially provides a robust framework for application in early identification of emerging viral threats, supporting public health surveillance and risk mitigation.

Keywords

Frequency chaos game representation, machine learning, species cross-over, viral protein sequences, viral zoonosis

Received: 14 May 2025; accepted: 28 December 2025

Introduction

Recently, there has been a marked increase in the emergence and re-emergence of infectious diseases, posing a significant threat to public health.^{1,2} It is believed that pathogenic etiological agents responsible for the majority of infectious disease outbreaks are zoonotic in origin³ and zoonotic viruses in particular, are of substantial concern, due to their abundance, diversity, and being notoriously difficult to manage and treat.⁴ The virus discovery curve indicates that there is a significant number of viruses which are yet to be discovered.^{5,6} There is limited knowledge about which of the newly discovered organisms may be pathogenic, and capable of triggering zoonotic events with pandemic potential.⁵⁻⁷

Humans, animals, and pathogens live in a dynamic, interactive, and interconnected environment whereby the health of one affects the other.¹ Human population growth,

coupled with activities that result in the encroachment of habitats and perturbation of ecological niches, ultimately facilitate closer interaction between biological entities resulting in close contact between humans and animals

¹SA MRC Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, Cape Town, South Africa

²Research Office, University of the Western Cape, Cape Town, South Africa

³Department of Computer Science, University of the Western Cape, Cape Town, South Africa

⁴African Centres for Disease Control and Prevention, African Union Commission, Addis Ababa, Ethiopia

Corresponding author:

Dominique Elizabeth Anderson, SA MRC Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, Private Bag X17, Bellville, Cape Town 7535, South Africa.
Email: danderson@uwc.ac.za



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

(wild and domestic), which may be reservoirs of infectious pathogens.⁸⁻¹⁰ This proximity relationship is postulated to be one of the drivers of increased emergence and re-emergence of infectious diseases through zoonosis.^{2,8,11}

Zoonotic diseases have been a major economic burden and public health concern on a global scale^{2,12} and there is a growing need for the development of robust prediction systems to mitigate or even prevent epidemic and pandemic events from occurring. Taking into consideration that zoonosis is a complex phenomenon, it requires multifaceted approaches and the need for predictive methods, to complement current surveillance methods.¹³ In addition to traditional surveillance efforts used throughout infectious disease outbreaks,^{7,14} various models, which harness statistical and ML tools, have been developed to predict cross-species spill-over and transmission dynamics of novel and re-emerging pathogens.^{15,16} These models incorporate ecological, demographic, and biogeographic data as features for robust algorithm development, to predict potentially zoonotic pathogens, identify existing reservoirs and new potential hosts, and predict zoonotic hotspots using knowledge from disciplines such as ecology and molecular biology.^{15,17}

A number of studies have reported the applications of ML techniques for the analysis of pathogen and host interaction networks. Wardeh et al¹⁷ developed a ML framework using random forests and gradient boosting, trained on mammalian viral traits and network features to predict potential mammalian hosts of known viruses. Shared pathogen networks were also incorporated through graph-based learning to enhance zoonotic reservoir prediction. The study highlighted the role of host phylogeny in pathogen sharing and quantified overlaps between humans and other mammals. Eng et al¹⁸ applied support vector machines to host tropism signatures from avian and human influenza strains to model zoonotic emergence. Qiang and Kou¹⁹ employed deep learning with protein sequence embeddings to predict avian influenza interspecies transmission. Han et al¹⁵ used boosted regression trees and logistic regression with biogeographic and ecological variables to predict rodent reservoir species of undiscovered zoonoses, achieving accuracies in the 90th percentile and identifying over 150 novel hyper-reservoir species. While ecological models assess species crossover events at a macro scale, they lack molecular resolution and often emphasize few hosts or pathogens, limiting generalizable prediction.^{15,17,20} These limitations motivate the consideration of deep learning models in this study, particularly convolutional neural networks (CNNs), which can automatically extract hierarchical features from genomic and protein data and capture complex nonlinearities in host-pathogen interactions.^{21,22} Capitalizing on the predictive capabilities of CNNs thus provides a promising avenue for improving zoonotic risk prediction and forms the basis of the present work.

Furthermore, for the analysis of cross-species events, Virus-receptor PPI models have been developed to predict these events.^{23,24} However, studies are limited by the availability of experimentally derived and validated PPIs, thereby influencing the amount of available input data for

ML models to produce robust, translatable, and reproducible models.²³⁻²⁵ In addition, PPI studies rely on defined pre-existing interactions, which may be unable to predict viral host switching in which a previously unknown host receptor is targeted for entry into host cells.^{26,27} Despite these limitations, various ML approaches have been developed for the analysis of pathogen-host protein interaction networks to predict cross-species events.^{24,27}

Machine learning has also been applied in several genomic surveillance initiatives, with many recent activities focusing on SARS-CoV-2 and Influenza viruses to not only identify emerging variants with the potential future spread, but also predict host range and species susceptibility.²⁸⁻³⁰ Gussow et al³¹ conducted an in-depth molecular analysis of coronaviruses to assess enhanced pathogenicity. Using comparative genomics and ML techniques, the authors identified signatures present in key genomic regions, such as the nucleocapsid protein and the spike glycoprotein, which appear to be associated with higher case fatality rates and host switching. Similarly, coronavirus spike protein sequences were used by Qiang et al³² to aid the prediction of species cross-over from non-human hosts of this viral taxa, suggesting that SARS-CoV-2 taxonomic relatives may indeed be of concern and should potentially be monitored. UniBind, an artificial intelligence-based framework, integrated protein structure and binding affinity, resulting not only in the efficient prediction of the effects of the binding affinity of SARS-CoV-2 variants to the host receptor but also in the prediction of host susceptibility to these viral variants.³³ A summary of relevant machine-learning models for viral host and spillover prediction is provided in Table 1.

The study presented in this work advances zoonosis prediction beyond traditional ML approaches that rely on ecological, biogeographic, or trait-based features by focusing directly on viral surface protein sequences. The main contribution of this work is a presentation of a proof-of-concept framework that encodes these sequences using FCGR and applies a CNN classifier to predict whether a virus has zoonotic potential. By demonstrating that FCGR-encoded protein features can capture molecular signatures of zoonosis, this work positions CNNs as a powerful alternative to conventional ML methods, which often struggle to generalize across host-pathogen systems. This contribution provides a scalable, data-driven foundation for more accurate and broadly applicable tools for zoonotic surveillance.

Methods

Data acquisition

Data collected for this study was derived from the UniProtKB Knowledge Database⁵⁰ and accessed through the UniProtKB website (<https://www.uniprot.org/uniprot/>, accessed September 21, 2021). Relevant data table fields were selected and corresponding protein sequences for the data entries, were obtained in FASTA format and the dataset is referred to as KW-1160 throughout. The data consist of a total of 358333 data entries, with exploratory data

Table 1. Summary of related machine-learning approaches for viral host prediction and potential spillover assessment.

Summary	Data input/pathogen focus	Citation
The author presents an alignment-free Chaos Game Representation (CGR) machine learning approach for rapidly classifying novel viral pathogens using only raw genomic sequence data. The model allows for real-time taxonomic predictions of new sequences which have not yet been classified, but does not model host range or spillover risk.	SARS-CoV-2 virus sequences	Randhawa et al ³⁴
This study presents a support vector machine (SVM), based framework to predict adenoviral infection potential for specific hosts using virus-host PPI predictions and taxonomy data. The approach enabled predictions related to infection potential, host specificity, and possible cross-species transmission routes.	Adenovirus	Karabulut et al ³⁵
VIDHOP is a deep neural network (DNN) tool that predicts the likely host species of viruses using viral nucleotide sequences alone. This demonstrates that host-association signatures encoded in viral genomes can be used to infer spillover susceptible species.	Influenza A, Rabies lyssavirus and rotavirus A	Mock et al ³⁶
The authors developed a BERT-based infectivity predictor which leverages LLMs, designed to estimate viral spillover potential across 26 viral families that infect a broad range of vertebrates and arthropods.	26 viral families	Kawasaki et al ³⁷
In this study, machine learning models were developed that are capable of identifying candidate zoonoses using signatures of host range encoded in viral genomes. The models outperformed phylogeny-based methods in predicting the probability that viruses can infect humans, indicating potential zoonotic threats.	86 I RNA and DNA viral species spanning 36 viral families	Mollentze et al ³⁸
A supervised machine learning framework using Gradient boosting machines (GBM) to infer reservoir hosts and arthropod vectors for major groups of human-infective single-stranded RNA (ssRNA) was presented in this study.	ssRNA viruses from 11 families	Babayan et al ³⁹
The authors trained random forest (RF) classifiers on protein sequences from all 11 influenza A proteins, to distinguish viruses originating from avian versus human hosts. The final model accurately differentiated host tropism and could indicate the host range of a newly detected influenza A strain.	Influenza A protein sequences	Eng et al ¹⁸
Using k-mer features from over 9400 viral genomes, 5 machine-learning models were built to distinguish human-infecting viruses from non-human ones. The study demonstrates the feasibility of identifying candidate human pathogens directly from virome sequencing.	Viruses infecting humans and other species	Zhang et al ⁴⁰
Trained on roughly 9500 viruses representing both RNA and DNA genomes the authors presented an interpretable machine learning method to predict whether viruses can directly infect humans, using next-generation raw sequencing reads.	RNA and DNA viruses	Bartoszewicz et al ⁴¹
HostNet, a deep learning framework based on Transformer-CNN-BiGRU architecture to improve prediction of virus-host associations. The approach extracts complex sequence patterns to infer host species for both prokaryotic and eukaryotic viruses.	Prokaryotic and eukaryotic viruses.	Ming et al ⁴²
DeepHoF, a deep learning model that calculates host likelihood scores for 5 broad host classes, including humans, by extracting viral genome features from sequence data. The model quantifies the propensity of viruses, including emerging ones, to infect humans and other hosts.	SARS-CoV-2	Guo et al ⁴³
This study evaluated multiple feature extraction strategies for predicting host susceptibility of RNA viruses using deep neural networks.	RNA virus genomes	Sutanto and Turcotte ⁴⁴
The authors developed gradient-boosted regression tree models to identify viral traits associated with human-to-human transmissibility across 224 known human-infecting virus species. The approach predicts which viruses may possess undocumented potential for sustained transmission and for distinguishing spillover events with epidemic potential.	Viral species known to infect humans	Walker et al ⁴⁵
This study used comprehensive surveillance data ecological and host-range data to estimate which newly discovered viruses pose zoonotic threats. Using data from human, domestic animal, and wildlife surveillance, the authors trained machine-learning classifiers based on gradient-boosted decision trees on over 500 zoonotic and non-zoonotic viruses. Their model predicts both zoonotic potential and likely host associations.	Zoonotic and non-zoonotic viruses infecting avian and mammalian hosts	Pandit et al ⁴⁶
EvoMIL, a deep learning method combining protein language models (PLM) and multiple instance learning (MIL) to infer host species based on viral protein sequences. The model predicts associations for both prokaryotic and eukaryotic hosts.	Prokaryotic and eukaryotic viruses	Liu et al ⁴⁷
This study employed logistic regression to classify coronaviruses based on spike protein sequences to determine whether they possess traits associated with human infection and specifically those associated with increased likelihood of animal-human spillover.	Spike protein sequences of SARS-CoV-2	Bhardwaj and Kulharia ⁴⁸
RNAVirHost, a hierarchical host classification framework, combining virus taxonomy, genomic traits, and sequence homologies, to predict hosts of emergent novel viruses.	RNA viruses	Chen et al ⁴⁹

analysis revealing that the dataset (a) did not exclusively contain entries from viral pathogens and (b) contained 237573 samples with incomplete associated “Virus hosts” metadata. The reporting of this study conforms to the REFORMS statement,⁵¹ the checklist can be found in the supplementary data (Supplemental Table 1). Data used in this study is included in the supplementary information.

Data preprocessing

The KW-1160 dataset preprocessed and cleaned for efficient model development and a Python script was written to automate the cleaning step, requiring 16 CPUs and 32 GB of RAM. The KW-1160 dataset *Taxonomic lineage IDs* and virus species names were standardized to corresponding ontologies used in the NCBI database using the *ete3 toolkit* Python package.⁵² The virus organism taxonomic super kingdom and family was obtained from the NCBI database, using *ete3 toolkit*, and added to the existing dataset in appended columns. Microorganisms other than viruses were removed from the KW-1160 dataset.

Addressing missing values

Missing data is a common data quality issue in statistical analysis and ML. Beyond introducing bias into estimates, missing data can also reduce statistical power, distort parameter estimates, limit generalizability, and, in severe cases, render analyses invalid. To address these issues, a variety of techniques have been developed, ranging from simple approaches such as listwise deletion and mean or median computation to more advanced methods like multiple imputation, maximum likelihood estimation, and ML-based imputers.^{53,54}

For viruses with missing “host” information, metadata imputation from external database records was performed in this study. Additional data used for imputation of missing host information in the KW-1160 dataset was obtained from NCBI Virus,⁵⁵ Enhanced Infectious Disease Database (EID2)⁵⁶ and Virus-Host database,⁵⁷ all of which were accessed on September 21, 2021. The data from the external sources was first standardized to use the NCBI taxonomy names, followed by extraction of corresponding taxonomic IDs. The *host* data was also standardized to match the nomenclature in the KW-1160 dataset, in the format [host name TaxID:ID]. Following standardization, each of the datasets were merged with the KW-1160 dataset, using a left-inner join, such that only samples with matching *taxonomic ID* would be imputed.

The imputation using NCBI-Virus dataset resulted in a 15% reduction of missing host values, while the imputation using EID2 dataset showed a negligible reduction. The most notable impact of imputation activities was observed when the Virus-Host DB data was used, with an approximate 50% reduction of missing values. This significant reduction is attributed to wider federation of additional information sources such as GenBank, ViralZone, literature surveys, in addition to RefSeq and manual curation of the database. Following imputation of data, the preprocessed

KW-1160 dataset used for downstream analysis consists of 317561 samples.

In addition, a column named *Infects human* was added to the dataset and contained binary data indicating whether the taxonomic ID for *Homo sapiens* (9606) was present in the list of viral host names. The rows matching the parameter, labeled “human true,” were considered as positive data, while those which do not match the parameter were considered as negative data and labeled “human false.” The FASTA file containing the protein sequences were mapped to their corresponding samples and the protein names in the KW-1160 dataset were replaced with the protein names in the FASTA headers as a more simplified nomenclature. The headers were then modified to contain the unique entry, protein name, the name of the virus, as well as the infection status, human-true or human-false.

Handling class imbalance and data splitting

To classify the positive and negative dataset, the “Virus hosts” field was designated to indicate whether a viral pathogen was documented to have a human host. Therefore, the primary objective of the model is to predict pathogens with the potential to cross the species barrier and infect humans, and as such, viruses that are reported to successfully infect humans would be classified as positive (with the assumption that they did not originate in the human host) and others as negative. Hence, the problem is formulated as a binary classification problem.

The preprocessed KW-1160 dataset contains class imbalance between the 2 classes, whereby the positive class had substantially more data points (278 791 samples) when compared with the negative class (38 770 samples). A random undersampling (RUS) approach was employed to address the class imbalance problem in the dataset using the *imbalanced-learn* Python package.⁵⁸ Thereafter, the KW-1160 dataset was split into training and test data at a ratio of 80:20 for training and testing purposes, respectively. Also, 20% of the training data was used as a validation set to assess the model generalization before final testing.

Sequence encoding

An R script was written to convert the FASTA protein sequences into machine readable FCGR images using the *kaos* package.⁵⁹ The FCGR chosen was the frequency matrix, with the corners and labels parameters set to false, and default settings are retained for all other parameters. The resulting plots were saved as portable network graphics (PNG) images of 224x224 pixels, at a resolution of 100. Parallel programming coupled with asynchronous programming using the *later* and *parallel R* packages,⁶⁰ respectively on a computational cluster compute node with 32 CPUs and 40 GB of RAM, was employed for efficient execution of the FCGR conversion process.

Model development

A convolutional neural network (CNN) was used to build the classification model as shown in Figure 1. To obtain

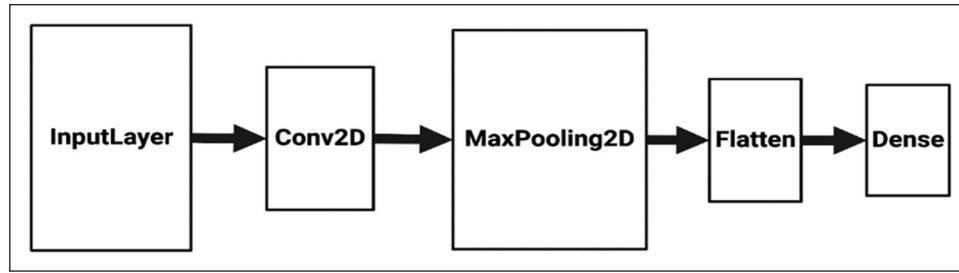


Figure 1. A depiction of the CNN model architecture created. Each box represents a layer in the model architecture. The arrows represent the flow of information between the layers.

Table 2. Selected model hyperparameters and their corresponding values.

Hyperparameters	Values
Number of 2D convolution layer	1-3
Number of units in the convolution layer	48-128
Threshold of evaluation metrics	0.5-0.9
Optimizers	RMSProp or Adam
Optimizer learning rate	0.001-0.019 (step 0.002)
Activation function	ReLu

optimal parameter values for the number and types of layers in the network, hyperparameter tuning was performed. The *Keras-tuner* Python package was used to implement Bayesian hyperparameter search.⁶¹ The Bayesian hyperparameter tuning was employed because it efficiently balances exploration and exploitation by learning from previous evaluations, making it more suitable than other tuning strategies such as grid or random search to obtain optimal parameter values suitable to construct the CNN model on the large and computationally intensive KW-1160 dataset. Based on results from the hyperparameter tuning, an optimized CNN model (ie, *the model with the best hyperparameter values*) with a single convolution and maximum pooling layer was developed using the *Keras* package and the *TensorFlow* package in the Python programming language. The CNN model is expected to have captured the key patterns and features in the protein sequence dataset to predict potentially zoonotic organisms. Table 2 presents the selected hyperparameters and their respective values for the developed CNN model. Along with the hyperparameter values provided in Table 2, the search was implemented for 500 trials, 500 combinations of hyperparameters, 3 kernel and pool sizes, and training for 2 epochs per trial, with validation at the end of each epoch.

The InputLayer serves as the entry point of the network, where input data is fed into the model. The Conv2D layer performs 2-dimensional convolution operations to extract image features, while the MaxPooling2D layer reduces dimensionality and retains the most salient features. The Flatten layer then transforms the feature maps from 2 dimensions into a 1-dimensional vector, preparing the data for classification. Finally, the dense layer acts as the output unit, producing the binary classification result.

The rectified linear unit (ReLU) activation function was used in the CNN model development to improve model efficiency by addressing inherent neural network problems such as the vanishing gradient problem.^{62,63} In addition, sequence similarity was anticipated due to the possibility of conserved regions in transmembrane proteins and high-dimensionality data,⁶⁴⁻⁶⁶ and as such the L2 regularizer was used to prevent the model from overfitting highly correlated data. The sigmoid activation was employed in the final output layer to predict class probabilities for binary classification problems in neural networks.^{64,66,67}

Model training was implemented on an NVIDIA graphics processing unit with 12 GB of memory, which is available in the Ilifu computational cluster. Based on the data splitting strategy, a leave-one-out approach was utilized to train the model, and the trained model was validated for 50 epochs on the training and validation data, respectively. The data was randomly shuffled at the start of each epoch, with performance metrics recorded at the end of each epoch. Model checkpoints were generated, storing the best model weights at the end of each epoch. A summary of the workflow used in this study is shown in Figure 2 and a Nextflow pipeline was created for ease of extension and reproducibility, and is available on GitHub (<https://github.com/Rudolph-afk/Zoon0PredV>).

Model evaluation and proof of concept

A small-scale study consisting of 3 imbalance class distribution scenarios was conducted, and the results obtained were used to examine under which class distributions the developed model achieved the best generalization performance. Three models were trained on these varying proportions derived as follows:

- The first dataset represented the complete KW-1160 dataset with no modifications, that is, (278 791 samples for the positive class and 38 770 entries for the negative class).
- The second dataset was an under-sampled derivative with the majority class being only two-thirds (67%) greater than the minority class (57 862 entries for the positive class and 38 768 samples for the negative class).

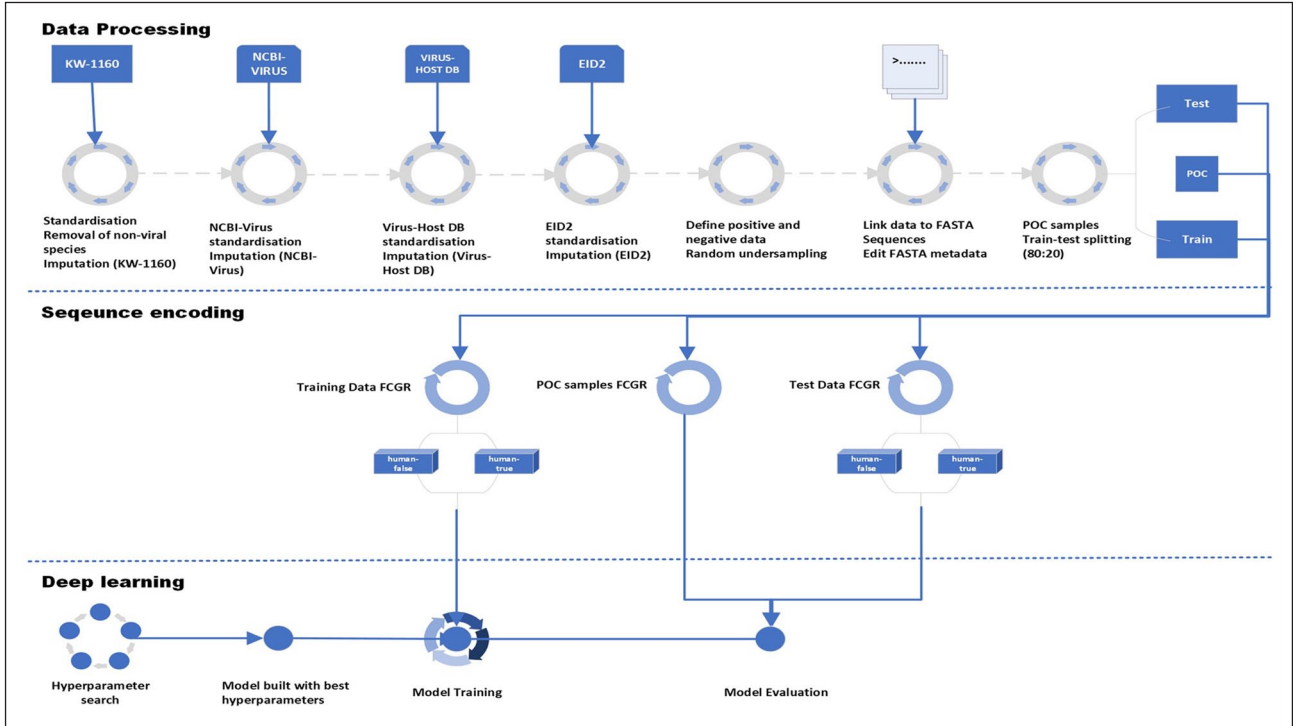


Figure 2. A graphical representation summarizing the workflow used in this study. POC, proof of concept.

- (c) The final under-sampled derivative dataset represented equal proportions of the majority and minority classes, containing 38 768 samples for each.

The model was evaluated using the test data, and the model performance on accuracy, precision, recall, F1 score, area under the receiver operating characteristic curve (ROC-AUC), and the Matthews correlation coefficient (MCC). Accuracy is the ratio between the correctly classified samples and the total number of samples in the test dataset.⁶⁸ It is given as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is calculated as the ratio between correctly classified samples and all samples assigned to that class.⁶⁸ It is computed as

$$Precision = \frac{TP}{TP + FP}$$

The recall, also known as the True Positive Rate or Sensitivity, is calculated as the ratio between correctly classified positive samples and all samples assigned to the positive class.⁶⁹ It is given as

$$Recall = \frac{TP}{TP + FN}$$

The F1 score is the harmonic mean of precision and recall,⁷⁰ and is calculated as

$$F1 = 2x$$

MCC measures the correlation between the true and predicted classes.⁷¹ It is computed as

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

where TP denotes true positive, FP is false positive, FN is false negative, and TN denotes true negative. For accuracy, precision, recall, and F1-score, the ideal model achieves a value of 1.0, while the poorest performance corresponds to 0. For MCC, the range of predictive value is from -1 (total disagreement) to +1 (perfect prediction), with 0 meaning random guessing.

The performance of the model was evaluated using a test dataset of 19 326 protein sequences, comprising 11 398 positive and 7928 negative samples. To provide a comprehensive assessment of the model's performance, a confusion matrix was employed, where other performance metrics were also computed. Furthermore, the model was used for proof-of-concept prediction using the proof-of-concept (POC) samples.

Results

The goal of this study is to develop a CNN model for the prediction of potentially zoonotic organisms, using viral surface proteins that facilitate host cell entry as input data. This section presents the results of the empirical investigation conducted in the study.

Model performance on imbalance classes

Table 3 presents the model results for the selected evaluation metrics for the 3 imbalanced class distributions, under which the best model-skewed class condition is determined. The results in Table 2 show that the model demonstrated distinct performance patterns across the 3 skewed class datasets, which reflect the effect of class distribution on predictive capacity. On the full imbalanced dataset

Table 3. Evaluation metrics for the validation dataset with 3 imbalanced class distributions.

Skewed classes	Accuracy	True positive	False positive	True negative	False negative	MCC	F1 score	ROC-AUC
ZoonosisFull	93.57	94772	1767	12092	5575	0.74	0.96	0.97
ZoonosisTwoThirds	95.37	20274	775	13061	844	0.90	0.96	0.99
ZoonosisOne2One	97.14	13211	129	13917	671	0.94	0.97	1.00

(ZoonosisFull), the model achieved high overall accuracy (93.57%) and a strong ROC-AUC score (0.97), indicating that the model successfully captured important sequence and image-level patterns in the dataset. However, the imbalance led to reduced sensitivity to the minority class, with 5575 false negatives recorded.

As indicated by Chicco and Jurman,⁷² when a model achieves a high predictive accuracy and is accompanied with a low MCC, the resultant effect is the sensitivity of the model to imbalanced classes in a dataset, which is observed in the ZoonosisFull dataset. In addition, while the F1 score (0.96) indicates that the model maintained a balance between precision and recall, the lower MCC of 0.74 highlights its weaker ability to generalize across classes when exposed to disproportionate class sizes. For zoonotic surveillance, this imbalance implies that a substantial number of zoonotic samples may potentially remain undetected.

The two-thirds under-sampled dataset (ZoonosisTwoThirds) offered a better trade-off between accuracy (95.37%), sensitivity, and specificity. False negatives were markedly reduced (844), while false positives remained relatively low (775). These improvements were reflected in the higher MCC (0.90) and ROC-AUC (0.99), which indicate that the reduction of skew allowed the model to more effectively capture discriminative features across both classes, resulting in improved confidence in model predictions.

The one-to-one balanced dataset (ZoonosisOne2One) provided the strongest overall performance, with the highest accuracy (97.14%), F1 score (0.97), MCC (0.94), and a ROC-AUC of 1.00. Both false positives (129) and false negatives (671) were minimized, demonstrating that the model generalized effectively without favoring one class over the other. This outcome illustrates that the balanced training dataset allows the network to fully capitalize on its representational power to generate reliable and interpretable predictions. Importantly, this improved model performance highlights the role of balanced class distribution in shaping the predictive integrity of the model across both classes.

Biological insight into FCGR image classification

Frequency Chaos Game Representation, employed in this study, generated greyscale images of 224x224 pixels (examples of the generated features for 4 entries are shown in Figure 3). The FCGR image is a large icosagon, which contains 20 edges and 20 icosagons,⁷³ with the edges representing each of the 20 standard amino acids, translated from nucleic acid sequences in the protein database.^{50,74}

Protein sequences derived from methods such as chromatography, mass spectrometry, and X-ray crystallography⁷⁵ may occasionally result in errors whereby an amino acid is not clearly identified.⁷⁵⁻⁷⁷ For example, a precise distinction between aspartic acid or asparagine, or glutamic acid or glutamine may result in the presence of different letter representations—B (aspartic acid [A] or asparagine [N]), J (leucine [U] or isoleucine [W]), X (unknown amino acid), Z (glutamic acid [G] or glutamine [Q])—in the sequenced proteins. This is an important consideration when using FCGR, as it only recognizes the standard 20 amino acids and it is assumed that these different letter representations are omitted in the FCGR software.

In addition, 2 recently discovered amino acids are considered part of the proteinogenic code; selenocysteine, and pyrrolysine, represented by the letters U and O, respectively.⁷⁸ It is not clear how FCGR would deal with these amino acids should they occur in a given protein sequence, and while this was not of concern in this study, it is an important consideration when FCGR is applied to studies involving proteins containing these unique amino acids.

Model test results

The model performance on the test data, evaluated across 50 epochs, is shown in Figure 4.

The training and validation accuracy of the model exhibited 3 distinct phases over the course of training. During the initial phase (epochs 0-8), both training and validation accuracy remained around 60%, indicating that the network was in the early stages of learning and had not yet fully captured discriminative image features from the dataset. In the transition phase (epochs 8-12), an increase in accuracy was observed for both sets, reaching approximately 95% to 100%. This rapid improvement indicates that the model successfully identified key patterns in the dataset necessary for accurate virus host classification. During the stabilization phase (epochs 12-50), accuracy for both training and validation converged near 96.8% (*the best model was obtained on the 48th epoch, which achieved 96.80% accuracy and 0.92 MCC on validation data*) and remained stable, demonstrating effective generalization and minimal overfitting. The near-overlap of the training and validation accuracy curves throughout the training process reflects a robust model that performs consistently across both datasets.

Furthermore, the loss curves are similar to the trend observed in accuracy, which illustrates the effectiveness of the weight optimization process during model development. In the initial phase (epochs 0-8), training loss was high (~2.0) while the validation loss was recorded around

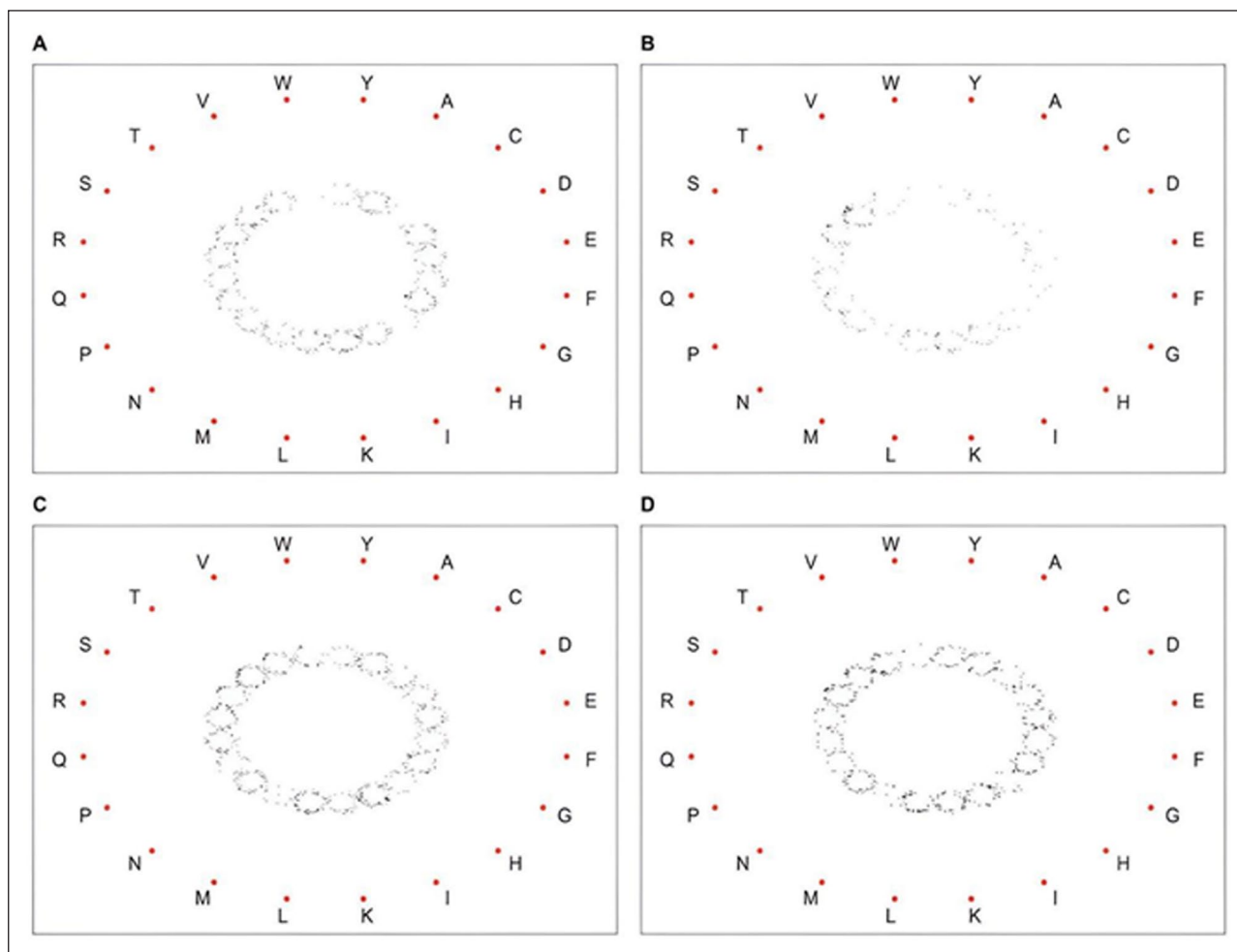


Figure 3. The frequency chaos game representation (FCGR) of 4 virus surface proteins: (A) influenza B virus nucleoprotein (560aa), (B) human orthopneumovirus major surface glycoprotein G (315aa), (C) simian immunodeficiency virus envelope glycoprotein gp160 (865aa), and (D) influenza A virus hemagglutinin (566aa).

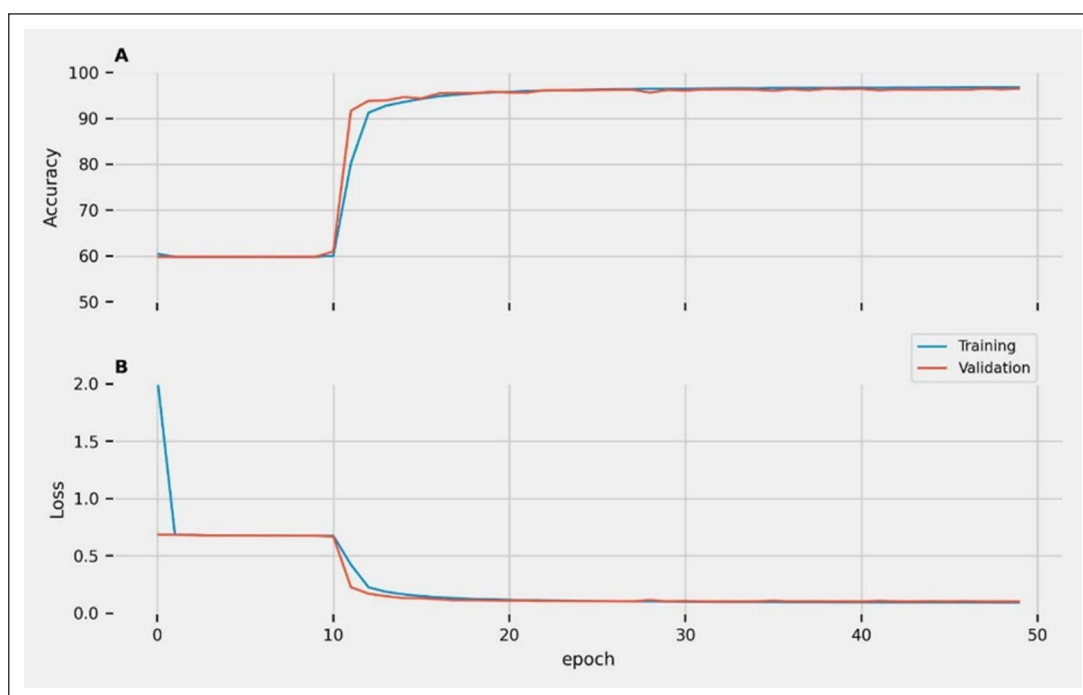


Figure 4. The model performance during training and validation on each epoch: (A) training and validation accuracy through the iterative model training and (B) training and validation error/loss through the iterative model training.

0.6, indicating poor early and unstable performance. During the transition phase (epochs 8-12), loss values decreased steadily for both training and validation sets, corresponding to an increase in model accuracy. In the stabilization phase (epochs 12-50), training and validation losses converged to very low values (0.3-0.4) and remained nearly identical, indicating that the model continued to optimize effectively while maintaining strong generalization. The close alignment of the loss curves further confirms that the model avoided underfitting or overfitting.

The performance of the model at predicting the minority (ie, positive class) is also presented in Figure 5.

The analysis of binary classification metrics from the confusion matrix in Figure 5 reveals a robust performance of the model across various indicators, with accuracy at 96.83%, precision at 97.75%, recall at 96.96%, F1 score at 97.35%, and specificity at 96.65%. The high accuracy signifies that the model correctly classifies approximately 97% of all samples, which illustrates solid overall performance despite class imbalance in the test dataset. Notably, the precision rate indicates that when the model predicts a positive sample, it is accurate 97.75% of the time, indicating a low false positive rate. Similarly, the high recall shows the capability of the model to identify 96.96% of actual positive samples, emphasizing the robustness of the model in critical scenarios where missing positive samples could have serious implications.⁷⁹

Furthermore, while the model achieved high precision and recall, the sensitivity of these results in model evaluation warrants consideration. Precision focuses on minimizing false positives, while recall aims to capture all true positives. For imbalanced datasets, the F1 score and specificity together provide a comprehensive evaluation of the performance of a classifier across both positive and negative samples.⁷² Although the model demonstrates excellent performance, there remains room for improvement, particularly in addressing the current counts of false positives and false negatives, which are 259 and 353, respectively. Furthermore, the model achieved a MCC of 0.93, which, as noted by Chicco and Jurman,⁷² is a more reliable metric than both the F1 score and accuracy for evaluating binary classification performance. In addition, the ROC-AUC score of the model was 0.99, illustrating its robustness and exceptional performance on the test dataset.

For proof-of-concept testing, 4 entries from A0A1W5YKT3 (Bat coronavirus, Spike glycoprotein), A0A0P0KH07 (Human coronavirus 229E, Spike glycoprotein), Q5EED8 (Human immunodeficiency virus 1, Envelope glycoprotein), and A0A0M4Q8U3 (Influenza D virus, Nucleoprotein) were tested on the model. The Bat coronavirus and Influenza D virus were correctly predicted as non-human infecting viruses with probability scores of 0.0010 and 0.00042, respectively, which are below a selected threshold of 0.5 in the study. The significantly low probability scores indicate that the proteins do not have signatures associated with sequences from viruses which have been reported to infect humans. The scores also indicate that these viruses potentially require substantial sequence evolution to permit future species barrier cross-over. The

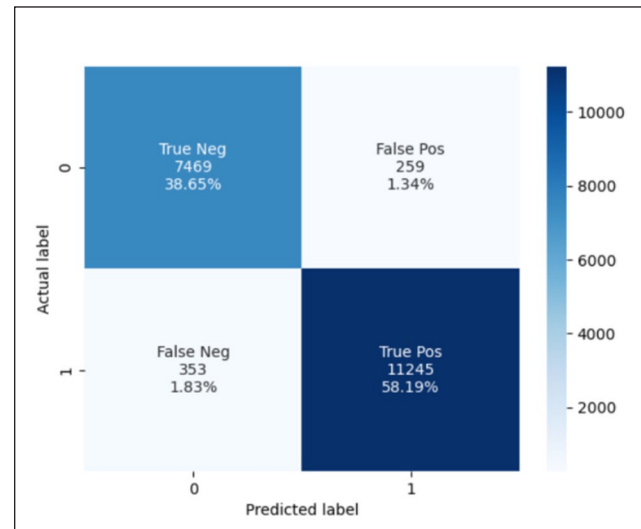


Figure 5. Confusion matrix illustrating the performance of the model with 11 245 true positives, 7469 true negatives, 259 false positives, and 353 false negatives.

Bat coronavirus is indicated to have 101 hosts in the KW-1160 dataset, and the low probability scores obtained from the model, coupled with the wide host range of this virus, illustrate the complexity and rareness of zoonotic events, thus possibly supporting the pinhole model.³ However, the selected entry for the proof-of-concept may be a strain which has not undergone mutations to allow species crossover.

Discussion

Several epidemics and pandemics are linked to host switching by viral pathogens, originally established in an animal host or reservoir.⁸⁰ Epizootic and zoonotic diseases are driven by spill-over of a pathogen to a previously unexposed, non-susceptible host, and when these events occur, the resultant outbreaks can have devastating consequences.⁸⁰ Despite the clear threats to public health and biosecurity which are caused by the emergence and re-emergence of zoonotic diseases, many host crossover events are not detected or reported, and the modeling of infectious disease to predict spill-over remains constrained by several challenges.^{81,82} Public health research priorities toward emerging infectious diseases are largely focused on the detection and surveillance of EIDs, as well as the identification of factors driving transmission, to intervene for public safety and mitigate the effects of disease.^{7,83} Detection efforts are focused on deployment of analytical, laboratory-based methods for identification of microorganisms, ranging from traditional culturing to modern molecular and “-omics” techniques.^{7,83,84} In addition to traditional surveillance efforts, various statistical and ML models, which make use of different features and prediction targets, have been developed to predict cross-species spill-over of novel and re-emerging infectious agents, as well as transmission dynamics once an outbreak has occurred.^{15,16,85}

In this study, a machine learning approach was used to develop a model that predicts the zoonotic potential of pathogenic species by learning protein sequence patterns of viral pathogens. Considering that host specificity is critically dependent on viral interaction with host cells, receptor binding (and changes thereof) inevitably plays a vital role,⁸⁰ and as such, viral proteins involved in pathways of host cell entry were used to train, validate, and evaluate the model. From the dataset perspective, the positive samples used in this study consisted of viral pathogens known to infect human hosts, while those documented not to have a human host formed the negative samples. The trained model could then be used to predict if an unknown virus would be capable of infecting a human host cell, based on the consistency of protein sequence patterns learned during model training.

The Chaos Game Representation (CGR) is a sequence representation scheme inspired by chaos theory in physics, originally proposed by Jeffrey⁵⁹ as a visual representation scheme for DNA sequences. Frequency Chaos Game Representation (FCGR) is an adaptation of the original CGR method and has been modified to accommodate protein sequences.⁷³ Another variant of CGR proposed by Mu et al,⁸⁶ called DCGR, incorporates amino acid physiochemical attributes, which are important determinants of protein structure, interaction, and function. Several encoding methods are available which use mathematical transformations as well as pre-computed embeddings,⁸⁷⁻⁹¹ however, FCGR is an underrepresented feature encoding method, shown to achieve good metrics in our study.

Convolutional neural networks were used in this study due to their exceptional image classification capability, particularly for the FCGR images. There is no standard convention for building CNN models due to varying performances of different model architectures.⁹² This is often further complicated by the presence of a multitude of parameters which require tuning^{93,94} and can include the number of layers to use, the number of nodes within each layer, a specific optimizer, the learning rate of the optimizer, the activation function, and others.⁹⁴

The high accuracy obtained with the developed CNN model in this study shows the excellent capability of convolutional neural networks, coupled with the FCGR features.

It is noteworthy to mention the hyperparameter optimization approach resulted in significant benefits during the CNN model development. The model provided optimal performance with a single layer and multiple nodes within its architecture due to the characteristics and complexity of the dataset used in this study. Based on the complexity of the dataset, with reference to large sample sizes, the model reached faster convergence and utilized lower computational time during training, which highlights the stability of the model in achieving high performance across the selected metrics.

Furthermore, previous models often focus on virus-host interactions¹⁵ and analysis of host receptor similarity,^{23,25} which tend to limit the utility of the latter model, particularly if a virus emerges and uses a different host receptor to

those already known. However, the developed model achieved high performance because of the significantly large dataset used in this study (since the model is exposed to a sufficient number of samples for efficient training and better generalization), when compared with the quantity of data used in previous studies, such as the 10 host receptor protein sequences in Bae and Son,²⁵ 211 interaction pairs in Yan et al,²⁴ and 277 host receptor protein sequences in Cho and Son.²³ In addition, the training data used in this study consist of a highly diverse dataset, which included viruses reported to infect plants as well as those reported to infect non-eukaryote organisms.

Also, CGR has been applied to viral proteins, with tools such as PhaVIP,⁹⁵ which classifies phage virion proteins, and Spike2CGR, which models coronavirus spike proteins.⁹⁶ These approaches convert sequences into images and employ convolutional neural networks for classification tasks, which aligns with our approach. In a related context, deep learning frameworks such as VIDHOP, tested on influenza A virus, rabies lyssavirus and rotavirus A, achieving AUC of between 0.93 and 0.98 for each viral species,³⁶ HostNet,⁴² and Virus2Vec, tested on real-world coronavirus spike and rabies virus sequence data,⁹⁷ employ embedding strategies or neural architectures to classify viral hosts with high accuracy. Compared with these embedding-driven models, our FCGR approach provides a holistic frequency representation that does not rely on motif or k-mer context but instead emphasizes global compositional structure.

From protein language modeling (pLMs), recent algorithmic advances such as ESM2 and related frameworks have demonstrated remarkable performance in tasks such as host tropism prediction, escape mutant detection, and structural inference.^{88,98} Further advances in protein and genomic language modeling demonstrate the capacity of sequence-only learning approaches to capture biologically meaningful signals related to viral evolution and pathogenicity. Virus specific generative models such as SpikeGPT⁹⁹ and SARITA¹⁰⁰ have shown that large language models (LLMs), not only generate realistic SARS-CoV-2 spike protein sequences, but can also retrospectively anticipate the emergence of mutations associated with altered transmissibility, and be used to examine pathogen evolution.^{99,100} At a broader scale, models such as Evo 2, trained on DNA and RNA sequences spanning all domains of life, can enable accurate prediction of mutational effects and pathogenicity.¹⁰¹

While these models highlight the power of large-scale language modeling, they are primarily designed as generalist or pathogen-specific and are focused on variant effect prediction and evolutionary forecasting, rather than direct zoonotic risk assessment. Large language models and pLMs generally require substantial computational resources, however, our study offers a computationally efficient alternative tool for early detection of host switching in emerging viruses, through a targeted, protein sequence approach which focuses on viral surface proteins involved in host cell entry.

Interestingly, 5 phage portal protein (PP) samples from bacterial and plant hosts were observed in the false positive

predictions, namely, A0A0K2FHA1 (Achromobacter phage phiAxp-2), A7TWJ1 (Staphylococcus virus tp310-2), I7HHN4 (Helicobacter virus KHP30), I7KR94 (Yersinia virus R1RT), and M4QNN7 (Tetraselmis viridis virus S20). Portal proteins have a low sequence similarity but highly conserved functionality, playing a role in bidirectional viral DNA passage.¹⁰² These phage portal proteins are being considered as potential antiviral drug targets in herpes simplex virus infections.¹⁰³ The “plasticity” of phage PP may explain the erroneous classification by the model, due to the presence of signatures consistent with proteins involved in viral entry into human host cells. This observation may indeed be of interest for further investigation, as false positives in this dataset may contain samples which could be considered for therapeutic experimentation, as in Dedeo et al.¹⁰³ The other false positives may be as a result of protein similarity. However, this does not eliminate the possibility that some of the false positives may be of future concern, having the capability to bind to human host cells, but still lacking machinery for sustained infection and replication.

A surprising observation in the false negative class was the erroneous classification of 31 Human Immunodeficiency Virus (HIV) entries. This was unexpected, as HIV is an established, long-term endemic virus with characteristic signatures of viruses with reported human hosts. Investigation of some of the HIV samples, such as A0A2P1DQ38, Q7SPP5, and A0A2P1DR91 showed the warning “Lacks conserved residue(s) required for the propagation of feature annotation,” according to UniProtKB. Computationally derived feature annotation is reliant on existing knowledge and annotations based on sequence homology, resulting in errors which are propagated in databases and give rise to contradictory interpretations of the data.^{104,105}

Thus, we demonstrated the capability of generating a robust model with good performance metrics. The insights generated from the developed model indicate the existence of patterns in the sequences of virus surface proteins that interact with host cells at the initial stage of infection. The insights may also be indicative of zoonotic potential, and it could possibly aid in identifying zoonotic viruses, using sequence data extracted from pathogen surveillance programs, as input into the model. Taken together, the results from this study showed the presence of consistent patterns in surface proteins of viruses reported to infect humans which differ from surface proteins of viruses which do not infect humans. From a biological perspective, this is expected, as host range is determined by successful infection,^{17,106,107} and virus-host cellular protein-protein interactions are a key mechanism.^{27,108,109} Furthermore, it is plausible that a similar approach could be adopted to design a model which predicts epizootic events for hosts other than humans, and particularly for animals of domestic and agricultural importance. In addition, the approach is flexible enough to support multi-category classification, with a simple modification of the final layer in the model architecture, such that a single model could potentially predict cross-species likelihood for several hosts rather than for a single host. Such a model would be a valuable application

Table 4. Accuracy and AUC comparison between Zoon0PredV and other models in literature.

Model	Accuracy	AUC	Citation
IILLS	Not reported	0.90	Yan et al ²⁴
ViCIPR	83.3%	1.00	Cho and Son ²³
GBM	Not reported	0.773	Mollentze et al ³⁸
Zoon0PredV	96.78%	0.99	This study

of ML to the One Health initiative, moving the focus from solely humans to other host organisms.

To further validate the effectiveness of the developed model in line with the work done in literature, Table 4 presents a comparison between our Zoon0PredV model and other models.

From Table 4, compared with prior approaches, Zoon0PredV demonstrates superior predictive performance. While IILLS²⁴ and GBM³⁸ did not report accuracy, their AUC values (0.90 and 0.773, respectively) are notably lower than that of Zoon0PredV (0.99). ViCIPR²³ reported an accuracy of 83.3% with an AUC of 1.00, but its accuracy lags significantly behind the 96.78% achieved by Zoon0PredV. Taken together, these results indicate that our model delivers a more balanced and consistently high performance, combining both excellent accuracy and discriminatory power (AUC). In addition, the Zoon0PredV learned patterns present in viral surface proteins such that even if a new virus emerges, targeting an uncommon host receptor, the viral protein patterns will still be detected. To our knowledge, no previous study identified in our literature search has combined FCGR of viral surface proteins with CNNs to develop a machine learning model aimed at predicting viral species cross-over events.

A number of limitations are acknowledged in this study. Although this study presents promising findings, the model has not been compared with other models that utilize common dataset, and it may contain biases that have not been identified in this analysis. Furthermore, we included limited viral sequences in our POC dataset (n=4), and while biologically plausible predictions were observed, the inclusion of additional sequences (derived from databases or synthetically generated) would provide larger scale validation of the biological relevance and predictive performance of our model.

Several areas of research are investigating pattern analysis for biological inference, and as such, comparison and ablation studies using classical and advanced sequence embedding tools such as those performed by Lin et al⁸⁸ and Jiao et al⁹¹ are needed to understand model performance and to examine if a hybrid approach to the task can yield more optimal results. Inherent data bias is an additional consideration that could potentially affect the model performance and Generalizability. For instance, a specific strain may have the capacity to infect several hosts, but in a database, may appear to infect a single host species. This phenomenon may be due to research priorities, based on perceived host “value” (human vs horse). In this way, even if the virus can infect additional hosts, systemic bias in data

representation and data priority exists.¹¹⁰ It is envisaged that with the increased research in One Health, that research priority will become less skewed.

Conclusions

The rise in epidemic and pandemic events of zoonotic origin has prompted the need to efficiently predict and mitigate future incidents. This study aimed to produce a proof-of-concept approach to predicting the zoonotic potential of viruses. A zoonosis prediction model was created using, as input, FCGR-encoded sequences of virus surface proteins, which facilitate viral entry into host cells. The model developed in this study showed the existence of patterns in the sequences of virus surface proteins that interact with host cells at the initial stage of infection and are indicative of zoonotic potential. It should, however, be noted that inferences about host-virus associations using sequence data alone may not capture the biotic and abiotic factors, which play important roles in host tropism.⁴² The CNN binary classification model obtained a 96% accuracy on the test data (ie, generalization performance), outperforming other approaches found in literature. However, the approach we used may benefit from using data with clear evidence of zoonosis to produce a more robust model. In addition, the study developed a binary classification model which focuses on cross-species prediction to human hosts, and as such, we suggest that future studies include other host organisms by building a multi-categorical model representing the varying host species, in line with a holistic One Health approach. The use of natural language processing tools such as ESM2, the largest language model trained to date for a variety of protein-related tasks, could also be used with our specific dataset to further refine our methodology and strengthen the current model, thereby increasing the reliability and subsequent use of ML technologies in public health-related research and pathogen surveillance.

Future research should incorporate systematic benchmarking across different encoding methods to evaluate their impact on zoonosis prediction performance. The study will also be extended to include more recent deep learning architectures, such as ResNet,¹¹¹ EfficientNet,¹¹² and Vision Transformers (ViTs),¹¹³ to assess their relative effectiveness in molecular zoonosis prediction. In addition, the dataset, currently based on UniProtKB 2021, will be updated to the latest releases to improve data recency, increase the number of available sequences, and enhance generalizability. These updates will further enable evaluation of the robustness of the model against newly discovered viral sequences. Collectively, these efforts aim to enhance both the predictive accuracy and practical applicability of the proposed approach in real-world scenarios.

Acknowledgments

The authors wish to acknowledge Mr Peter Van Heusden and Dr Nasr Eshibona from the South African National Bioinformatics Institute for code evaluation and guidance.

ORCID iDs

Taiwo Gabriel Omomule  <https://orcid.org/0000-0002-0293-7003>

Dominique Elizabeth Anderson  <https://orcid.org/0000-0002-4337-8009>

Ethical Considerations

Ethical approval was not required for this study.

Author Contributions

Rudolph Abel Serage: Investigation; Writing – original draft; Methodology; Validation; Writing – review & editing; Visualization; Formal analysis.

Clement Nthambazale Nyirenda: Writing – review & editing; Visualization; Formal analysis; Investigation.

Taiwo Gabriel Omomule: Writing – review & editing.

Alan Gilbert Christoffels: Investigation; Funding acquisition; Writing – review & editing.

Dominique Elizabeth Anderson: Conceptualization; Writing – original draft; Writing – review & editing; Supervision.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding for this project was supplied by the DSI/NRF Research Chair in Bioinformatics, Grant number 64751.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Use of AI

No scientific data has been generated or modified using AI tools such as ChatGPT.

Supplemental Material

Supplemental material for this article is available online.

References

- Calistri P, Iannetti S, Danzetta ML, et al. The components of “one world—one health” approach. *Transbound Emerg Dis*. 2013;60 suppl 2:4-13. doi:10.1111/tbed.12145
- Dallas TA, Carlson CJ, Poisot T. Testing predictability of disease outbreaks with a simple model of pathogen biogeography. *R Soc Open Sci*. 2019;6:190883. doi:10.1098/rsos.190883
- Warren CJ, Sawyer SL. How host genetics dictates successful viral zoonosis. *PLoS Biol*. 2019;17:e3000217. doi:10.1371/journal.pbio.3000217
- Carrasco-Hernandez R, Jácome R, López Vidal Y, Ponce De, León S. Are RNA viruses candidate agents for the next global pandemic? A review. *ILAR J*. 2017;58:343-358. doi:10.1093/ilar/ilx026
- Anthony SJ, Epstein JH, Murray KA, et al. A strategy to estimate unknown viral diversity in mammals. *mBio*. 2013;4:e00598-e00613. doi:10.1128/mBio.00598-13
- Woolhouse MEJ, Howey R, Gaunt E, Reilly L, Chase-Topping M, Savill N. Temporal trends in the discovery

- of human viruses. *Proc R Soc B*. 2008;275:2111-2115. doi:10.1098/rspb.2008.0294
7. Carroll D, Daszak P, Wolfe ND, et al. The global virome project. *Science*. 2018;359:872-874. doi:10.1126/science.aap7463
 8. Brierley L, Fowler A. Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning. *PLoS Pathog*. 2021;17:e1009149. doi:10.1371/journal.ppat.1009149
 9. Faburay B. The case for a “one health” approach to combating vector-borne diseases. *Infect Ecol Epidemiol*. 2015;5:28132. doi:10.3402/iee.v5.28132
 10. Taylor LH, Latham SM, Woolhouse MEJ. Risk factors for human disease emergence. *Phil Trans R Soc Lond B*. 2001;356:983-989. doi:10.1098/rstb.2001.0888
 11. French RK, Holmes EC. An ecosystems perspective on virus evolution and emergence. *Trends Microbiol*. 2020;28:165-175. doi:10.1016/j.tim.2019.10.010
 12. Smith KF, Goldberg M, Rosenthal S, et al. Global rise in human infectious disease outbreaks. *J R Soc Interface*. 2014;11:20140950. doi:10.1098/rsif.2014.0950
 13. Kesselring J, Zinsstag J, Schelling E, et al. *One Health: The Theory and Practice of Integrated Health Approaches*. Swiss Archives of Neurology Psychiatry and Psychotherapy; 2021.
 14. Baum SE, Machalaba C, Daszak P, Salerno RH, Karesh WB. Evaluating one health: are we demonstrating effectiveness? *One Health*. 2017;3:5-10. doi:10.1016/j.onehlt.2016.10.004
 15. Han BA, Schmidt JP, Bowden SE, Drake JM. Rodent reservoirs of future zoonotic diseases. *Proc Natl Acad Sci USA*. 2015;112:7039-7044. doi:10.1073/pnas.1501598112
 16. Royce K, Fu F. Mathematically modeling spillovers of an emerging infectious zoonosis with an intermediate host. *PLoS ONE*. 2020;15:e0237780. doi:10.1371/journal.pone.0237780
 17. Wardeh M, Blagrove MSC, Sharkey KJ, Baylis M. Divide and conquer—machine-learning integrates mammalian, viral, and network traits to predict unknown virus-mammal associations. *Nat Commun*. 2021;12:3954. doi:10.1101/2020.06.13.150003
 18. Eng C, Tong J, Tan T. Predicting zoonotic risk of influenza A viruses from host tropism protein signature using random forest. *IJMS*. 2017;18:1135. doi:10.3390/ijms18061135
 19. Qiang X, Kou Z. Predicting interspecies transmission of avian influenza virus based on wavelet packet decomposition. *Comput Biol Chem*. 2019;78:455-459. doi:10.1016/j.compbiolchem.2018.11.029
 20. Olival KJ, Hosseini PR, Zambrana-Torrel C, Ross N, Bogich TL, Daszak P. Host and viral traits predict zoonotic spillover from mammals. *Nature*. 2017;546:646-650. doi:10.1038/nature22975
 21. Pan X, Shen HB. RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC Bioinformatics*. 2017;18:136. doi:10.1186/s12859-017-1561-8
 22. Rao T, Li X, Zhang H, Xu M. Multi-level region-based convolutional neural network for image emotion classification. *Neurocomputing*. 2019;333:429-439. doi:10.1016/j.neucom.2018.12.053
 23. Cho M, Son HS. Prediction of cross-species infection propensities of viruses with receptor similarity. *Infect Genet Evol*. 2019;73:71-80. doi:10.1016/j.meegid.2019.04.016
 24. Yan C, Duan G, Wu FX, Wang J. ILLS: predicting virus-receptor interactions based on similarity and semi-supervised learning. *BMC Bioinformatics*. 2019;20:651. doi:10.1186/s12859-019-3278-3
 25. Bae SE, Son HS. Classification of viral zoonosis through receptor pattern analysis. *BMC Bioinformatics*. 2011;12:96. doi:10.1186/1471-2105-12-96
 26. Deng L, Nie W, Zhao J, Zhang J. A hybrid deep learning framework for predicting the protein-protein interaction between virus and host. *Published online June 8, 2021*. doi:10.21203/rs.3.rs-506156/v1
 27. Kösesoy İ, Gök M, Kahveci T. Prediction of host-pathogen protein interactions by extended network model. *Türk J Biol*. 2021;45:138-148. doi:10.3906/biy-2009-4
 28. Alberts F, Berke O, Rocha L, Keay S, Maboni G, Poljak Z. Predicting host species susceptibility to influenza viruses and coronaviruses using genome data and machine learning: a scoping review. *Front Vet Sci*. 2024;11:1358028. doi:10.3389/fvets.2024.1358028
 29. Elste J, Saini A, Mejia-Alvarez R, et al. Significance of artificial intelligence in the study of virus–host cell interactions. *Biomolecules*. 2024;14:911. doi:10.3390/biom14080911
 30. Rancati S, Nicora G, Prosperi M, Bellazzi R, Salemi M, Marini S. Forecasting dominance of SARS-CoV-2 lineages by anomaly detection using deep AutoEncoders. *Brief Bioinform*. 2024;25:bbae535. doi:10.1093/bib/bbae535
 31. Gussow AB, Auslander N, Faure G, Wolf YI, Zhang F, Koonin EV. Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proc Natl Acad Sci USA*. 2020;117:15193-15199. doi:10.1073/pnas.2008176117
 32. Qiang XL, Xu P, Fang G, Liu WB, Kou Z. Using the spike protein feature to predict infection risk and monitor the evolutionary dynamic of coronavirus. *Infect Dis Poverty*. 2020;9:33. doi:10.1186/s40249-020-00649-8
 33. Wang G, Liu X, Wang K, et al. Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution. *Nat Med*. 2023;29:2007-2018. doi:10.1038/s41591-023-02483-5
 34. Randhawa GS, Soltysiak MPM, El Roz H, De Souza CPE, Hill KA, Kari L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *PLoS ONE*. 2020;15:e0232391. doi:10.1371/journal.pone.0232391
 35. Karabulut OC, Karpuzcu BA, Türk E, Ibrahim AH, Süzek BE. ML-AdVInfect: a machine-learning based adenoviral infection predictor. *Front Mol Biosci*. 2021;8:647424. doi:10.3389/fmolb.2021.647424
 36. Mock F, Viehweger A, Barth E, Marz M. VIDHOP, viral host prediction with deep learning. *Bioinformatics*. 2021;37:318-325. doi:10.1093/bioinformatics/btaa705
 37. Kawasaki J, Suzuki T, Hamada M. Hidden challenges in evaluating spillover risk of zoonotic viruses using machine learning models. *Commun Med*. 2025;5:187. doi:10.1038/s43856-025-00903-w
 38. Mollentze N, Babayan SA, Streicker DG. Identifying and prioritizing potential human-infecting viruses from their genome sequences. *PLoS Biol*. 2021;19:e3001390. doi:10.1371/journal.pbio.3001390
 39. Babayan SA, Orton RJ, Streicker DG. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. *Science*. 2018;362:577-580. doi:10.1126/science.aap9072
 40. Zhang Z, Cai Z, Tan Z, et al. Rapid identification of human-infecting viruses. *Transbound Emerg Dis*. 2019;66:2517-2522. doi:10.1111/tbed.13314

41. Bartoszewicz JM, Seidel A, Renard BY. Interpretable detection of novel human viruses from genome sequencing data. *NAR Genom Bioinform.* 2021;3:lqab004. doi:10.1093/nar-gab/lqab004
42. Ming Z, Chen X, Wang S, et al. HostNet: improved sequence representation in deep neural networks for virus-host prediction. *BMC Bioinformatics.* 2023;24:455. doi:10.1186/s12859-023-05582-9
43. Guo Q, Li M, Wang C, et al. Predicting hosts based on early SARS-CoV-2 samples and analyzing the 2020 pandemic. *Sci Rep.* 2021;11:17422. doi:10.1038/s41598-021-96903-6
44. Sutanto K, Turcotte M, eds. Extracting and evaluating features from RNA virus sequences to predict host species susceptibility using deep learning. In: *2021 13th International Conference on Bioinformatics and Biomedical Technology.* ACM; 2021:81-89.
45. Walker JW, Han BA, Ott IM, Drake JM. Transmissibility of emerging viral zoonoses. *PLoS ONE.* 2018;13:e0206926. doi:10.1371/journal.pone.0206926
46. Pandit PS, Anthony SJ, Goldstein T, et al. Predicting the potential for zoonotic transmission and host associations for novel viruses. *Commun Biol.* 2022;5:844. doi:10.1038/s42003-022-03797-9
47. Liu D, Young F, Lamb KD, Robertson DL, Yuan K. Prediction of virus-host associations using protein language models and multiple instance learning. *PLoS Comput Biol.* 2024;20:e1012597. doi:10.1371/journal.pcbi.1012597
48. Bhardwaj V, Kulhari M. The machine learning-based predictor to identify putative COVID-19-like host jumping viruses. *J Zoonotic Dis.* 2025;9:979-987. doi:10.22034/jzd.2025.20118
49. Chen G, Jiang J, Sun Y. RNAVirHost: a machine learning-based method for predicting hosts of RNA viruses through viral genomes. *Gigascience.* 2024;13:giae059. doi:10.1093/gigascience/giae059
50. The UniProt Consortium Bateman A, Martin MJ, et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 2021;49:D480-D489. doi:10.1093/nar/gkaa1100
51. Kapoor S, Cantrell EM, Peng K, et al. REFORMS: consensus-based recommendations for machine-learning-based science. *Sci Adv.* 2024;10:eadk3452.
52. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol.* 2016;33:1635-1638. doi:10.1093/molbev/msw046
53. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. *J Big Data.* 2019;6:27. doi:10.1186/s40537-019-0192-5
54. Liu X, ed. Methods for handling missing data. In: *Methods and Applications of Longitudinal Data Analysis.* Elsevier; 2016:441-473.
55. Hatcher EL, Zhdanov SA, Bao Y, et al. Virus variation resource—improved response to emergent viral outbreaks. *Nucleic Acids Res.* 2017;45:D482-D490. doi:10.1093/nar/gkw1065
56. Wardeh M, Risley C, McIntyre MK, Setzkorn C, Baylis M. Database of host-pathogen and related species interactions, and their global distribution. *Sci Data.* 2015;2:150049. doi:10.1038/sdata.2015.49
57. Mihara T, Nishimura Y, Shimizu Y, et al. Linking virus genomes with host taxonomy. *Viruses.* 2016;8:66. doi:10.3390/v8030066
58. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res.* 2017;18:1-5. doi:10.48550/ARXIV.1609.06570
59. Jeffrey HJ. Chaos game representation of gene structure. *Nucl Acids Res.* 1990;18:2163-2170. doi:10.1093/nar/18.8.2163
60. Zhao P. R with parallel computing from user perspectives. R-bloggers 2016. Accessed December 19, 2023. https://www.r-bloggers.com/2016/09/r-with-parallel-computing-from-user-perspectives/#google_vignette
61. O'Malley T, Bursztein E, Long J, et al. KerasTuner. Github 2019. Accessed January 2, 2026. <https://github.com/keras-team/keras-tuner>
62. Alaeddine H, Jihene M. A comparative study of popular CNN topologies used for imagenet classification. In: Suresh A, Udendhran R, Vimal S, eds. *Advances in Bioinformatics and Biomedical Engineering.* IGI Global; 2020:89-103. doi:10.4018/978-1-7998-3591-2.ch007
63. Lin G, Shen W. Research on convolutional neural network based on improved Relu piecewise activation function. *Procedia Computer Science.* 2018;131:977-984. doi:10.1016/j.procs.2018.04.239
64. Ghojogh B, Crowley M. The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. arXiv. Published online May 28, 2019. doi:10.48550/ARXIV.1905.12787
65. Humayoo M, Cheng X. Parameter estimation with the ordered l2 regularization via an alternating direction method of multipliers. *Appl Sci.* 2019;9:4291. doi:10.3390/app9204291
66. Korotcov A, Tkachenko V, Russo DP, Ekins S. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. *Mol Pharmaceutics.* 2017;14:4462-4475. doi:10.1021/acs.molpharmaceut.7b00578
67. Vargas R, Mosavi A, Ruiz R. Deep learning: a review. *Math Comput Sci.* Published online October 10, 2018. doi:10.20944/preprints201810.0218.v1
68. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep.* 2022;12:5979. doi:10.1038/s41598-022-09954-8
69. Dehmer M, Basak SC, eds. *Statistical and Machine Learning Approaches for Network Analysis.* 1st ed. Wiley; 2012.
70. Emmert-Streib F, Moutari S, Dehmer M. *Elements of Data Science, Machine Learning, and Artificial Intelligence Using R.* Springer International Publishing; 2023.
71. Lantz B. *Machine Learning with R: Learn How to Use R to Apply Powerful Machine Learning Methods and Gain an Insight into Real-World Applications.* Packt Publishing; 2013.
72. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* 2020;21:6. doi:10.1186/s12864-019-6413-7
73. Löchel HF, Eger D, Sperlea T, Heider D. Deep learning on chaos game representation for proteins. *Bioinformatics.* 2020;36:272-279. doi:10.1093/bioinformatics/btz493
74. Steward K. Essential amino acids: chart, abbreviations and structure. Technol Networks 2019. Accessed December 19, 2023. <https://www.technologynetworks.com/applied-sciences/articles/essential-amino-acids-chart-abbreviations-and-structure-324357>

75. Pietrzyk AJ, Bujacz A, Jaskolski M, Bujacz G. Identification of amino acid sequences via X-ray crystallography: a mini review of case studies. *Biotechnologia*. 2014;94:9-14. doi:10.5114/bta.2013.46427
76. Searle BC, Dasari S, Turner M, et al. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. *Anal Chem*. 2004;76:2220-2230. doi:10.1021/ac035258x
77. Vyatkina K, Wu S, Dekker LJM, et al. De novo sequencing of peptides from top-down tandem mass spectra. *J Prot Res*. 2015;14:4450-4462. doi:10.1021/pr501244v
78. Lopez MJ, Mohiuddin SS. *Biochemistry, Essential Amino Acids*. Statpearls Publishing; 2024.
79. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag*. 2009;45:427-437. doi:10.1016/j.ipm.2009.03.002
80. Parrish CR, Holmes EC, Morens DM, et al. Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol Mol Biol Rev*. 2008;72:457-470. doi:10.1128/MMBR.00004-08
81. Glennon EE, Jephcott FL, Restif O, Wood JLN. Estimating undetected Ebola spillovers. *PLoS Negl Trop Dis*. 2019;13:e0007428. doi:10.1371/journal.pntd.0007428
82. Roberts M, Dobson A, Restif O, Wells K. Challenges in modelling the dynamics of infectious diseases at the wildlife-human interface. *Epidemics*. 2021;37:100523. doi:10.1016/j.epidem.2021.100523
83. Temmam S, Davoust B, Berenger JM, Raoult D, Desnues C. Viral metagenomics on animals as a tool for the detection of zoonoses prior to human infection? *IJMS*. 2014;15:10377-10397. doi:10.3390/ijms150610377
84. Sandle T, ed. Microbial identification. In: *Pharmaceutical Microbiology*. Elsevier; 2016:103-113.
85. Eid FE, ElHefnawi M, Heath LS. DeNovo: virus-host sequence-based protein-protein interaction prediction. *Bioinformatics*. 2016;32:1144-1150. doi:10.1093/bioinformatics/btv737
86. Mu Z, Yu T, Qi E, Liu J, Li G. DCGR: feature extractions from protein sequences based on CGR via remodeling multiple information. *BMC Bioinformatics*. 2019;20:351. doi:10.1186/s12859-019-2943-x
87. Jing X, Dong Q, Hong D, Lu R. Amino acid encoding methods for protein sequences: a comprehensive review and assessment. *IEEE/ACM Trans Comput Biol Bioinform*. 2020;17:1918-1931. doi:10.1109/TCBB.2019.2911677
88. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. 2023;379:1123-1130. doi:10.1126/science.ade2574
89. Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA*. 2021;118:e2016239118. doi:10.1073/pnas.2016239118
90. ElAbd H, Bromberg Y, Hoarfrost A, Lenz T, Franke A, Wendorff M. Amino acid encoding for deep learning applications. *BMC Bioinformatics*. 2020;21:235. doi:10.1186/s12859-020-03546-x
91. Jiao S, Li B, Wang L, et al. Beyond ESM2: graph-enhanced protein sequence modeling with efficient clustering. *arXiv*. Published online April 24, 2024. doi:10.48550/arXiv.2404.15805
92. Lu Z, Chen L, Chiang CK, Sha F. Hyper-parameter tuning under a budget constraint. In: *Proceedings of the Twenty-eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization; 2019:5744-5750.
93. Klein A, Falkner S, Bartels S, Hennig P, Hutter F. Fast Bayesian hyperparameter optimization on large datasets. *Electron J Statist*. 2017;11:4945-4968. doi:10.1214/17-EJS1335SI
94. Sarawagi S, Ganguli R. Deep neural network surrogates for optimal design of helicopter rotor. *Trans Indian Natl Acad Eng*. 2021;6:653-664. doi:10.1007/s41403-021-00227-w
95. Shang J, Peng C, Tang X, Sun Y. PhaVIP: phage Vlrion protein classification based on chaos game representation and vision transformer. *Bioinformatics*. 2023;39(suppl 1):i30-i39. doi:10.1093/bioinformatics/btad229
96. Murad T, Ali S, Khan I, Patterson M. Spike2CGR: an efficient method for spike sequence classification using chaos game representation. *Mach Learn*. 2023;112:3633-3658. doi:10.1007/s10994-023-06371-4
97. Ali S, Bello B, Chourasia P, et al. Virus2Vec: viral sequence classification using machine learning. *Proc Mach Learn Res*. 2023;209:6-18. doi:10.48550/ARXIV.2304.12328.
98. Lytras S, Lamb KD, Ito J, et al. Pathogen genomic surveillance and the AI revolution. *J Virol*. 2025;99:e01601-e01624. doi:10.1128/jvi.01601-24
99. Dhodapkar RM. A deep generative model of the SARS-CoV-2 spike protein predicts future variants. *Bioinformatics*. Published online January 18, 2023. doi:10.1101/2023.01.17.524472
100. Rancati S, Nicora G, Bergomi L, et al. SARITA: a large language model for generating the S1 subunit of the SARS-CoV-2 spike protein. *Brief Bioinform*. 2025;26:bbaf384. doi:10.1093/bib/bbaf384
101. Brixi G, Durrant MG, Ku J, et al. Genome modeling and design across all domains of life with Evo 2. *Genomics*. Published online February 21, 2025. doi:10.1101/2025.02.18.638918
102. Lokareddy RK, Sankhala RS, Roy A, et al. Portal protein functions akin to a DNA-sensor that couples genome-packaging to icosahedral capsid maturation. *Nat Commun*. 2017;8:14310. doi:10.1038/ncomms14310
103. Dedeo CL, Cingolani G, Teschke CM. Portal protein: the orchestrator of capsid assembly for the dsDNA tailed bacteriophages and herpesviruses. *Annu Rev Virol*. 2019;6:141-160. doi:10.1146/annurev-virology-092818-015819
104. Holliday GL, Davidson R, Akiva E, Babbitt PC. Evaluating functional annotations of enzymes using the gene ontology. *Methods Mol Biol*. 2017;1446:111-132. doi:10.1007/978-1-4939-3743-1_9
105. Zaru R, Magrane M, Orchard S, UniProt Consortium. Challenges in the annotation of pseudoenzymes in databases: the UniProtKB approach. *FEBS J*. 2020;287:4114-4127. doi:10.1111/febs.15100
106. Carlson CJ, Zipfel CM, Garnier R, Bansal S. Global estimates of mammalian viral diversity accounting for host sharing. *Nat Ecol Evol*. 2019;3:1070-1075. doi:10.1038/s41559-019-0910-6
107. Wells K, Morand S, Wardeh M, Baylis M. Distinct spread of DNA and RNA viruses among mammals amid prominent role of domestic species. *Glob Ecol Biogeogr*. 2020;29:470-481. doi:10.1111/geb.13045
108. Kerr SA, Jackson EL, Lungu OI, et al. Computational and functional analysis of the virus-receptor interface reveals host range trade-offs in new world arenaviruses. *J Virol*. 2015;89:11643-11653. doi:10.1128/JVI.01408-15

109. Parvez MK, Parveen S. Evolution and emergence of pathogenic viruses: past, present, and future. *Intervirology*. 2017;60:1-7. doi:10.1159/000478729
110. Iuchi H, Kawasaki J, Kubo K, et al. Bioinformatics approaches for unveiling virus-host interactions. *Comput Struct Biotechnol J*. 2023;21:1774-1784. doi:10.1016/j.csbj.2023.02.044
111. He K, Zhang X, Ren S, Sun J, eds. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2016:770-778.
112. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. arXiv. Published online May 28, 2019. doi:10.48550/ARXIV.1905.11946
113. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv. Published online October 22, 2020. doi:10.48550/ARXIV.2010.11929