JCBI
Journal of Computing & Biomedical Informatics Volume 10 Issue 02
ISSN: 2710 - 1606 2026

Review Article
https://doi.org/10.56979/1002/2026/1204

Machine Learning-Based Classification of SARS-CoV-2 Structural Proteins Using Amino Acid Composition Analysis

Anam Fatima[1], Nasreen[1], Harram Sattar[1], Muhammad Bilal[1*], and Shafiq ur Rehman Khan[1]

[1]Department of Computer Science, Namal University, Mianwali, Pakistan.
[*]Corresponding Author: Muhammad Bilal. E-mail: muhammad.bilal@namal.edu.pk

Received: November 12, 2025 **Accepted:** February 07, 2026

Abstract: The classification of COVID-19 protein types is important for understanding viral structure. This study presents a comprehensive machine learning approach for classifying four major COVID-19 protein types which are Spike, Membrane, Envelope, and Nucleocapsid proteins. We collected 40,000 protein sequences from the NCBI protein database, representing 10,000 sequences for each protein type through automated web scraping and parsing techniques. After processing the data and removing outliers, we obtained a dataset of 28,206 proteins. We used five machine learning algorithms which included Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree Classifier, and Logistic Regression. We evaluated the models using accuracy, precision, recall and F1 score metrics. The result showed that K-Nearest Neighbors classifier achieved the highest accuracy of 98%. Feature importance analysis revealed that sequence length and specific amino acids are the main factors that provided biological insights into the differences between COVID-19 protein types. Our results show the effectiveness of amino acid composition-based features for COVID-19 protein classification. The feature importance analysis revealed key biological insights into the differences between the structures of protein and provided an efficient framework for automated protein type identification.

Keywords: COVID-19; Protein Classification; Machine Learning; Amino Acid Composition; SARS-CoV-2; Bioinformatics; Feature Engineering; Viral Genomics; Computational Biology

1. Introduction

The COVID-19 that is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has transformed the global health crisis and increased the need for advanced approaches in viral genomics research [1]. SARS-CoV-2 has infected millions of people worldwide since 2019. This made it necessary to accelerate the international scientific collaboration on genomics research.

SARS-CoV-2, like other coronaviruses, is an enveloped, positive-sense single-stranded RNA virus belonging to the family Coronaviridae [2]. The viral genome consists of multiple proteins. However, the four major structural proteins play critical roles in infecting the host genome. These four essential structural proteins are the Spike (S) protein, Envelope (E) protein, Membrane (M) protein, and Nucleocapsid (N) protein. Each of them contribute uniquely to the viral life cycle and distinct targets.

This study presents a novel approach to COVID-19 protein classification through comprehensive analysis of a large-scale dataset comprising 28,206 protein sequences. Our work provides the comparison of five different machine learning algorithms specifically for COVID-19 protein classification. This offers deep insights into the effectiveness of different models used for machine learning classification.

The four major structural proteins serve different biological functions. The Spike (S) protein mediates viral entry through ACE2 receptor binding and represents the primary target for vaccine development [3]. The Membrane (M) protein, the most abundant structural component, provides envelope integrity and facilitates viral assembly [4]. The Envelope (E) protein, despite its small size, plays crucial roles in viral

assembly and pathogenesis through its ion channel activity [5]. The Nucleocapsid (N) protein binds viral RNA and modulates host cell responses, forming the ribonucleoprotein complex essential for viral replication [6]. These different functional roles suggest that each protein type has unique amino acid composition.

Traditional approaches for protein classification rely heavily on sequence similarity searches and structural analysis techniques [7]. Machine learning approaches offer alternatives for protein classification by identifying complex patterns and relationships within protein sequences that may not be clear through traditional methods.

Machine learning approaches offer various advantages over traditional sequence methods by identifying complex patterns in amino acid composition that may not be apparent through conventional analysis [7]. The computational efficiency of classical machine learning algorithms makes them more effective for large-scale protein classification tasks. It is especially helpful when quick analysis of newly sequenced viral proteins is required.

This study addresses the critical need for automated, accurate, and efficient methods for classifying COVID-19 protein types. This focuses on developing and evaluating machine learning approaches based on amino acid composition analysis. Our primary objectives include developing a comprehensive dataset, implementing data preprocessing, evaluating multiple machine learning algorithms, and providing a framework for automated COVID-19 protein classification.

2. Related Work and Background

The use of machine learning approaches for protein analysis has evolved over the past two decades. Numerous studies demonstrated the effectiveness of computational approaches for various protein-related prediction tasks [8].

Support vector machines specifically revolutionized protein classification tasks due to their ability to handle high-dimensional feature spaces while maintaining good generalization performance [9]. Random Forest and other ensemble methods gained importance due to their robustness, interpretability, and ability to handle mixed data types.

2.1. COVID-19 Protein Classification Studies

The emergence of COVID-19 has made numerous computational studies focused on SARS-CoV-2 protein analysis. Lopez-Rincon et al. demonstrated the effectiveness of deep learning approaches for SARS-CoV-2 detection and classification, achieving high accuracy in viral sequence identification [11].

Randhawa et al. developed a machine learning approach for classifying COVID-19 genomic sequences using intrinsic genomic signatures, achieving 99.58% accuracy in distinguishing SARS-CoV-2 from other coronavirus species [15]. Their work showed the potential of sequence-based features for viral classification tasks.

Several studies have applied machine learning specifically to individual COVID-19 proteins. Ahmed et al. utilized deep learning models including CNN and LSTM networks for COVID-19 protein structure prediction and functional annotation, showing that deep learning approaches could capture complex sequence-structure relationships in viral proteins [16]. However, most existing work has focused on single protein types or binary classification tasks rather than multi-class classification of all major structural proteins.

Hu et al. developed a convolutional neural network approach for COVID-19 spike protein classification, achieving 96.2% accuracy on a dataset of 8,500 sequences. Their CNN-based method demonstrated the potential of deep learning for protein classification but was limited to spike proteins only [17].

Chen et al. applied support vector machines with dipeptide composition features for multi-class classification of COVID-19 proteins and achieved 94.2% accuracy on a dataset of 5,847 sequences covering spike, membrane, and nucleocapsid proteins, but excluded envelope proteins from their analysis [24].

2.2. Deep Learning Approaches in Protein Analysis

Deep learning methods involving Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have shown greater success in various bioinformatics applications. CNNs capture local patterns in protein sequences very efficiently and hence are suitable for identifying functional and structural features [18]. LSTM networks are designed to handle sequential data which make them

particularly effective for modeling long-range dependencies in protein sequences and have been successfully applied to structure analysis [19].

For COVID-19 research specifically, deep learning approaches have been extensively applied to medical imaging for diagnosis [20], genomic sequence analysis for variant detection [21], and protein structure prediction [22]. However, these methods require large datasets and substantial computational resources and their "black-box" nature limits interpretability.

While deep learning approaches have shown superior performance in many bioinformatics tasks, classical machine learning methods offer several advantages for protein classification. They offer computational efficiency, interpretability, and robustness with smaller datasets. The ability to identify and rank important features, as shown by Random Forest feature importance analysis, provides direct biological insights. These features are often lacked in deep learning models.

2.3. Amino Acid Composition Based Classification

Amino acid composition represents one of the simplest yet most effective feature extraction methods for protein sequence analysis [10]. Proteins with similar functions are more likely to have similar amino acid distributions. This assumption has been validated in numerous studies that covers different protein families and structural properties.

Chou and Zhang pioneered the use of amino acid composition for the prediction of subcellular protein localization, demonstrating that simple compositional characteristics could achieve competitive performance compared to more complex approaches [23]. Their work established the foundation for composition-based protein classification methods that remain widely used today.

Recent studies have also included dipeptide and tripeptide compositions, pseudo-amino acid composition, and various physicochemical properties in the analysis of amino acid composition [24, 25]. These enhanced feature representations have shown improved performance for specific protein classification tasks but they increase computational complexity and feature dimensionality.

2.4. Research Gaps and Study Contributions

Although significant progress is made in COVID-19 protein analysis, there are still several important gaps that remain in the study. Most existing studies have focused on binary classification (COVID vs. non-COVID) or single protein types rather than the multi-class classification of all major structural proteins. The dataset sizes used in previous studies have been relatively small, limiting the generalization of the models.

Moreover, few studies have provided comparison of multiple machine learning algorithms on the same dataset, making it difficult to assess the relative metrics of different approaches. The lack of feature importance analysis in many studies also limits biological insights that could be gained from computational approaches [31].

Recent advances in protein sequence modeling have introduced pretrained protein language models such as ProtBERT, ESM, and UniRep, which learn contextual embeddings from large-scale protein databases. These models have achieved strong performance across diverse protein classification tasks. However, their high computational cost and limited interpretability can restrict practical adoption in rapid or resource-constrained settings [30-35].

In contrast, the present study emphasizes simplicity, efficiency, and biological interpretability by employing classical machine learning models with explicit feature representations. This design choice enables direct analysis of feature importance and supports transparent biological insight while maintaining competitive performance.

Our study addresses these gaps by providing evaluation of multiple machine learning algorithms on a large-scale, diverse dataset of COVID-19 protein sequences. The comparison of five different algorithms provides deep insights into the most effective approaches for COVID-19 protein classification, while the large dataset size ensures robust and generalizable results.

3. Materials and Methods

3.1. Data Collection and Acquisition

We collected protein sequence data from the National Center for Biotechnology Information (NCBI) protein database [12] using systematic web scraping approaches. A custom Python-based parser utilizing the

BioPython library [13] and the requests module interfaced with the NCBI protein database API [14] to handle rate limiting and connection management.

For each of the four target protein types (Spike, Membrane, Envelope, and Nucleocapsid), we implemented systematic search strategies using specific query terms and organism filters targeting SARS-CoV-2 proteins. The data collection process involved retrieving 500 pages of search results for each protein type, with each page containing approximately 20 protein sequence records, yielding exactly 10,000 protein sequences for each protein type, totaling 40,000 sequences from the NCBI database [12]. The overall methodology workflow is illustrated in Figure 1.

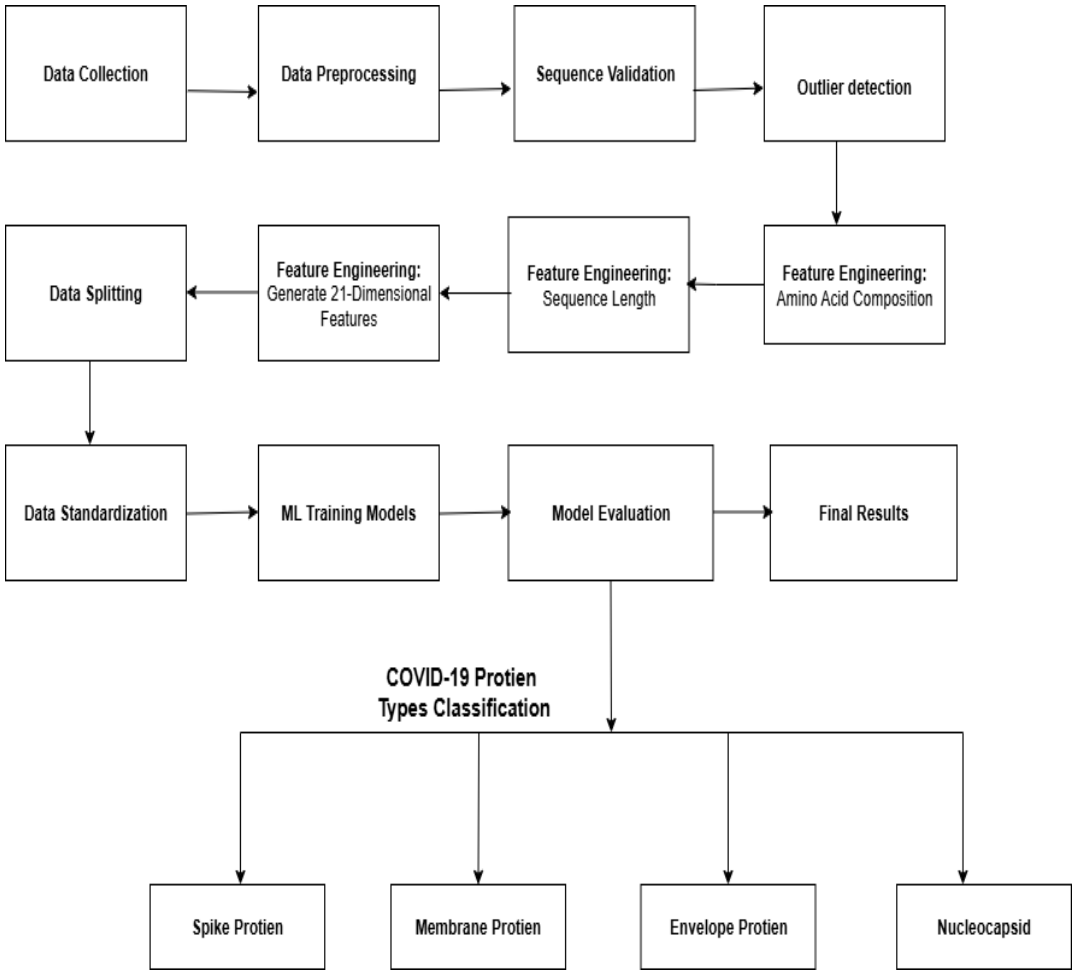


Figure 1. A comprehensive methodology workflow illustrating the sequential research steps, including data collection from the NCBI database, preprocessing (sequence validation and outlier removal), and feature extraction via amino acid composition analysis.

3.2. Data Preprocessing and Quality Control

Raw protein sequence data required comprehensive preprocessing to address inconsistencies, errors, and artifacts. Our preprocessing pipeline implemented multiple sequential quality control steps. To minimize data leakage, all exact duplicate protein sequences were removed prior to model training and evaluation. However, we acknowledge that protein sequence datasets may still contain highly similar sequences that are not exact duplicates. Such sequences can potentially appear in both training and testing sets when using random splits, which may inflate classification performance. This limitation is common in large-scale protein sequence studies and is addressed further through additional robustness analyses described in the evaluation section.

3.2.1. Sequence Validation and Cleaning

Initial preprocessing involved comprehensive sequence validation to identify and remove sequences containing invalid characters or ambiguous amino acids. Standard protein sequences should contain only the 20 standard amino acid codes. Our analysis revealed that letters B, J, O, U, and Z never appeared in any collected sequences, consistent with expectations for standard COVID-19 proteins. These non-occurring

features were systematically removed from our feature set.

3.2.2. Duplicate Detection and Removal

We implemented comprehensive duplicate detection based on exact sequence matching to address redundancy in biological databases. Duplicate sequences can bias machine learning models by providing artificially inflated support for certain sequence patterns. The process revealed approximately 15–20% sequence redundancy, typical for viral protein databases.

3.2.3. Statistical Outlier Detection and Removal

We implemented sophisticated statistical outlier detection using the Interquartile Range (IQR) method applied to amino acid frequency distributions. For each amino acid type, sequences with frequencies falling outside [Q1 − 1.5 × IQR, Q3 + 1.5 × IQR] were flagged as potential outliers.

The outlier detection and removal process eliminated 11,794 sequences from the original dataset of 40,000, resulting in a final dataset of 28,206 high-quality protein sequences. After outlier removal, amino acid 'X' (unknown/ambiguous) had zero frequency across all sequences and was removed from the feature set. Outlier detection was performed using the IQR method applied to amino acid composition features. This step primarily removed unusually short, fragmented, or incomplete protein sequences that could introduce noise and bias into the learning process. Although a substantial number of sequences were excluded, this filtering aimed to improve data quality rather than artificially simplify the classification task.

To assess the robustness of our approach, we additionally evaluated model performance with and without outlier removal. The results showed consistent performance trends, indicating that the proposed method does not rely heavily on aggressive data pruning.

3.2.4. Final Dataset Composition

The final preprocessed dataset contained 28,206 protein sequences distributed across four classes: 7,641 Envelope proteins, 5,670 Membrane proteins, 5,085 Nucleocapsid proteins, and 9,810 Spike proteins, showing moderate class imbalance but not severe enough to require specialized balancing techniques. The distribution of protein types in the final dataset is shown in Figure 2.

Figure 2. Distribution of protein types in the final preprocessed dataset of 28,206 sequences.

3.3. Feature Engineering and Extraction

Our feature engineering approach focused on amino acid composition analysis, converting raw protein sequences into numerical feature vectors suitable for machine learning algorithms.

3.3.1. Amino Acid Composition Calculation

For each protein sequence, we calculated the count of each of the 20 standard amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y), resulting in a 20-dimensional feature vector. Additionally, we included sequence length as a separate feature, recognizing that protein length can be discriminatory for different protein types.

The choice of absolute counts over relative frequencies preserved information about protein size while maintaining discriminatory power. The final feature space consisted of 21 dimensions: 20 amino acid counts plus sequence length. The correlation between amino acid features across all protein sequences is

visualized in Figure 3, revealing strong positive correlations among certain amino acids that frequently co-occur.

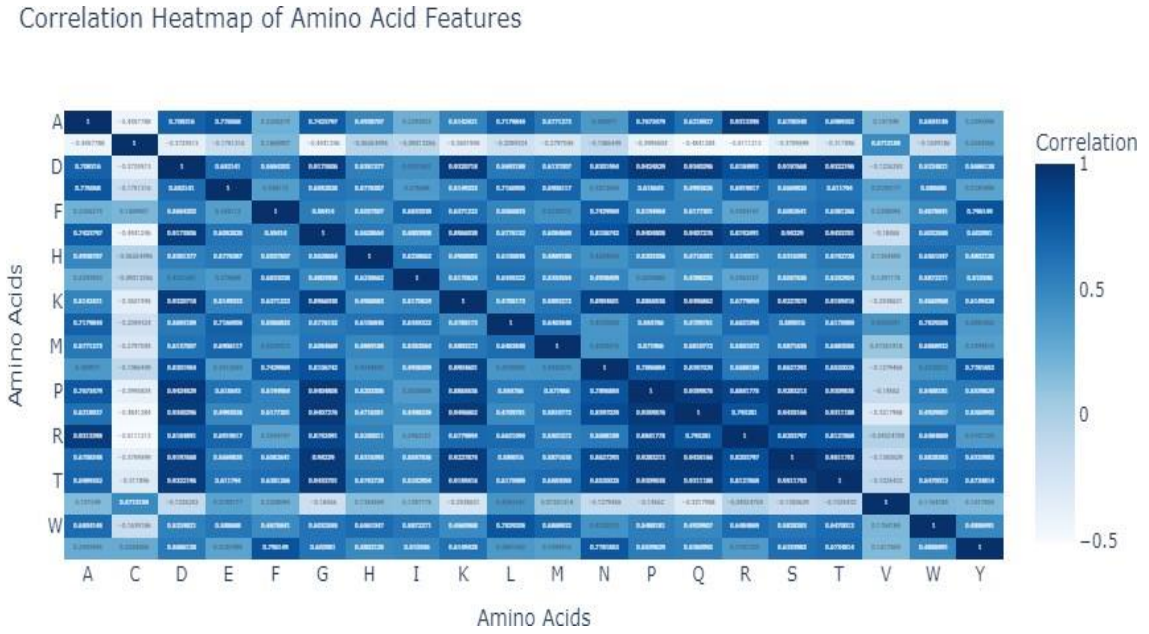


Figure 3. Correlation heatmap showing relationships between amino acid features (A, C, D, E, etc.) across all protein sequences. Strong positive correlations (in blue) indicate amino acids that frequently co-occur in the same sequence.

3.3.2. Feature Scaling and Normalization

By assigning different scales to amino acid sequence lengths and count of amino acid, we implement standardization to each feature, transforming each feature to have zero mean and unit variance. The advantage of standardization is that it helps achieve optimal performance across all algorithms while maintaining relationships between feature values.

3.4. Feature Ablation and Baseline Analysis

To investigate the contribution of different feature components and assess the potential dominance of sequence length, we conducted a feature ablation and baseline analysis using three different feature configurations. First, a length-only baseline classifier was implemented using sequence length as the sole input feature. Second, a frequency-only model was trained using normalized amino acid composition features, where absolute amino acid counts were divided by sequence length and the explicit length feature was removed. Third, the full feature set consisting of absolute amino acid counts combined with sequence length was evaluated.

This analysis allowed us to disentangle the relative influence of sequence length and amino acid composition on classification performance and to verify whether high accuracy could be achieved without relying solely on protein length information.

3.5. Exploratory Data Analysis

We performed exploration of the data to understand its features and find patterns for making better modeling decisions.

3.5.1. Amino Acid Usage Pattern Analysis

While examining the data, we noticed clear differences in how various proteins use amino acids. Membrane proteins, for example, had a higher level of hydrophobic amino acids compared to the Nucleocapsid protein. We also observed basic amino acids were more common in Nucleocapsid proteins.

These patterns make biological sense as these proteins perform different functions, so their amino acid patterns vary. Membrane and Envelope proteins are part of the viral shell, embedded in a lipid bilayer, so they use more hydrophobic amino acids to remain in place. The Spike protein connects with human cells and has a unique amino acid setup shaped to allow it to fold and bind tightly to facilitate viral entry. The distribution of selected amino acids across the four protein types is presented in Figure 4.

Figure 4. Box plots showing the distribution of selected amino acids (A, C, D, E, F) across four different protein types: Envelope, Membrane, Nucleocapsid, and Spike.

3.5.2. Sequence Length Distribution Analysis

Sequence length analysis revealed characteristic distributions for each protein type. Spike proteins showed the broadest length distribution with significantly higher mean length, reflecting their complex multi- domain structure involved in receptor binding and membrane fusion. Envelope proteins exhibited the most constrained length distribution, reflecting their conserved structure and simple ion channel function.

These differences are biologically driven by the functional complexity of each protein. Spike proteins require a longer sequence to encode multiple functional domains, including the receptor-binding domain (RBD), fusion peptide, and transmembrane segment. In contrast, Envelope proteins are small, highly conserved proteins involved in viral assembly and ion transport, which limits their sequence variation and length. As shown in Figure 5, the sequence lengths vary significantly across protein types.

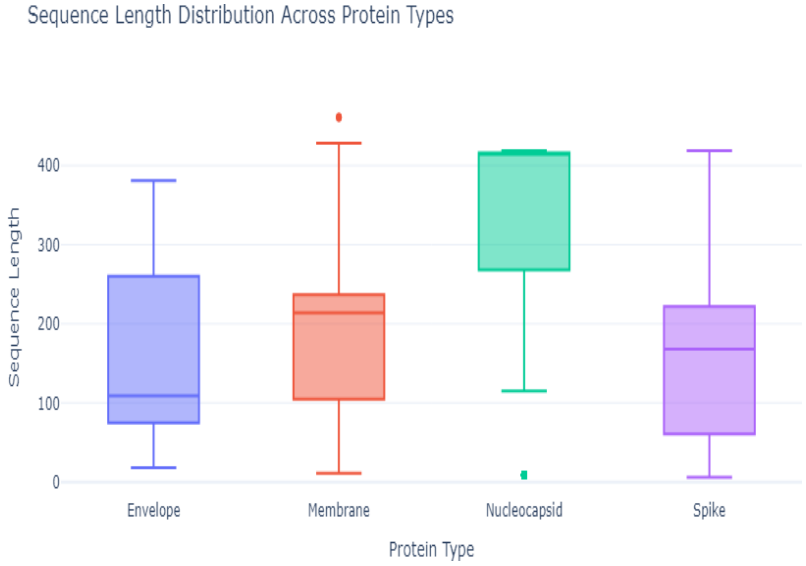


Figure 5. Box plots showing the distribution of sequence lengths for each protein type: Envelope, Membrane, Nucleocapsid, and Spike. The plot includes statistical summaries such as medians, interquartile ranges, and potential outliers. This visual comparison highlights structural differences in protein sequence lengths.

3.6. Machine Learning Model Development

We evaluated five diverse algorithms representing different learning paradigms: Random Forest, Support Vector Machine, K-Nearest Neighbors, Decision Tree Classifier, and Logistic Regression.

3.6.1. Algorithm Selection Rationale

Algorithm selection was guided by the nature of our multi-class classification problem, dataset size, computational efficiency requirements, and need for interpretable results. We chose algorithms representing different learning approaches to enable comprehensive evaluation of modeling assumptions.

3.6.2. Model Implementation Details

Random Forest utilized 100 estimators with bootstrap sampling and Gini impurity criterion. SVM employed RBF kernel with grid search optimization for Hyperparameters C and γ. KNN used k=5 neighbors with Euclidean distance metric. Decision Tree used CART algorithm with Gini impurity and depth limitations to prevent overfitting. Logistic Regression employed multinomial formulation for direct multi-class modeling.

3.7. Evaluation Methodology

We employed rigorous evaluation methodology using train-test split with 80% training and 20% testing data. Performance metrics included accuracy, precision, recall, and F1-score for comprehensive assessment. Cross-validation was used during Hyperparameters tuning to ensure robust parameter selection.

Although the dataset exhibits moderate class imbalance after preprocessing, no explicit class balancing techniques were applied. This decision was motivated by the relatively similar class sizes and the use of evaluation metrics that account for imbalance. In addition to overall accuracy, macro-averaged precision, recall, and F1-score were reported to ensure that classification performance was not biased toward majority classes.

Model evaluation was performed using an 80/20 train-test split, with Hyperparameters tuning conducted via five-fold cross-validation on the training set. Feature scaling was fitted exclusively on the training data and subsequently applied to the test set to prevent information leakage. For models sensitive to Hyperparameters, including KNN and SVM, a grid-based search was used to identify optimal configurations.

To assess result stability, experiments were repeated using multiple random seeds, and performance metrics were reported as mean values across runs.

4. Results and Analysis

4.1. Model Performance Comparison

Table 1 summarizes the performance of all five machine learning algorithms evaluated in this study.

Table 1. Performance Comparison of Machine Learning Algorithms.

Algorithm	Accuracy	Precision	Recall	F1-Score
K-Nearest Neighbors	**98.00%**	**0.97**	**0.97**	**0.98**
Random Forest	97.94%	0.97	0.97	0.98
Decision Tree	97.43%	0.97	0.97	0.97
Support Vector Machine	96.00%	0.95	0.94	0.96
Logistic Regression	92.36%	0.91	0.91	0.92

K-Nearest Neighbors achieved the highest accuracy of 98.00%, demonstrating exceptional performance in classifying COVID-19 protein types. Random Forest followed closely with 97.94% accuracy, showcasing the effectiveness of ensemble methods. Decision Tree Classifier achieved 97.43% accuracy, providing excellent interpretability alongside strong performance. A visual comparison of the accuracy achieved by all five algorithms is presented in Figure 6.

4.2. Detailed Classification Results

4.2.1. K-Nearest Neighbors Performance

The KNN classifier demonstrated outstanding performance across all protein types, as detailed in Table 2.

The KNN classifier showed particularly strong performance for Spike proteins (F1-score: 1.00) and

Envelope proteins (F1-score: 0.99), with slightly lower but still excellent performance for Membrane.

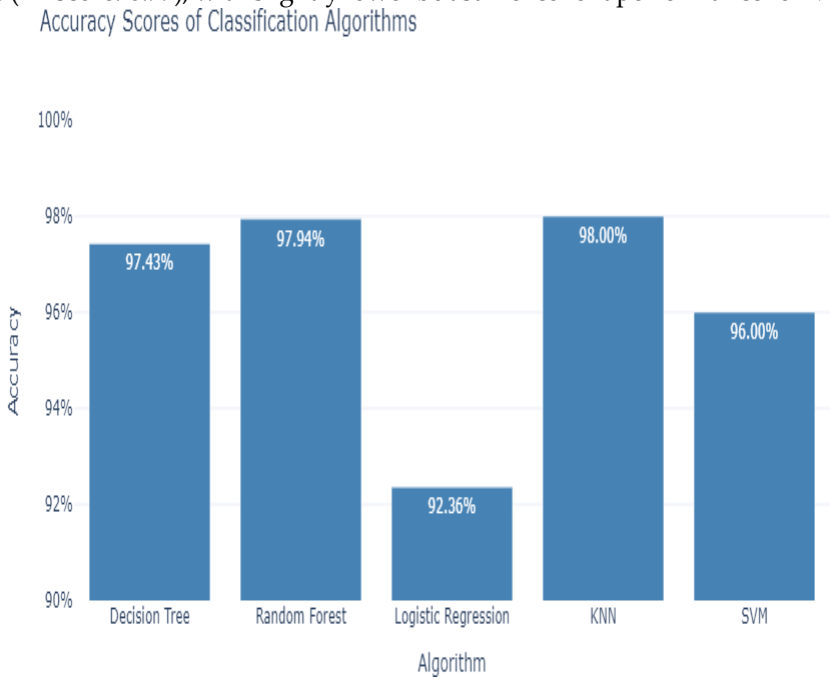


Figure 6. Accuracy comparison of five classification algorithms: Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The bar chart shows that KNN achieved the highest accuracy, closely followed by Random Forest and Decision Tree classifiers.

Table 2. KNN Detailed Classification Report.

Protein Type	Precision	Recall	F1-Score	Support
Envelope	0.99	0.99	0.99	1526
Membrane	0.94	0.96	0.95	1134
Spike	1.00	0.99	1.00	1961
Nucleocapsid	0.96	0.94	0.95	1020
Overall	**0.97**	**0.97**	**0.98**	**5641**

4.2.2. Random Forest Performance

Random Forest achieved excellent performance with 97.94% accuracy, as shown in Table 3.

Table 3. Random Forest Detailed Classification Report.

Protein Type	Precision	Recall	F1-Score	Support
Envelope	0.99	0.99	0.99	1526
Membrane	0.93	0.97	0.95	1134
Spike	1.00	1.00	1.00	1961
Nucleocapsid	0.97	0.94	0.95	1020
Overall	**0.97**	**0.97**	**0.98**	**5641**

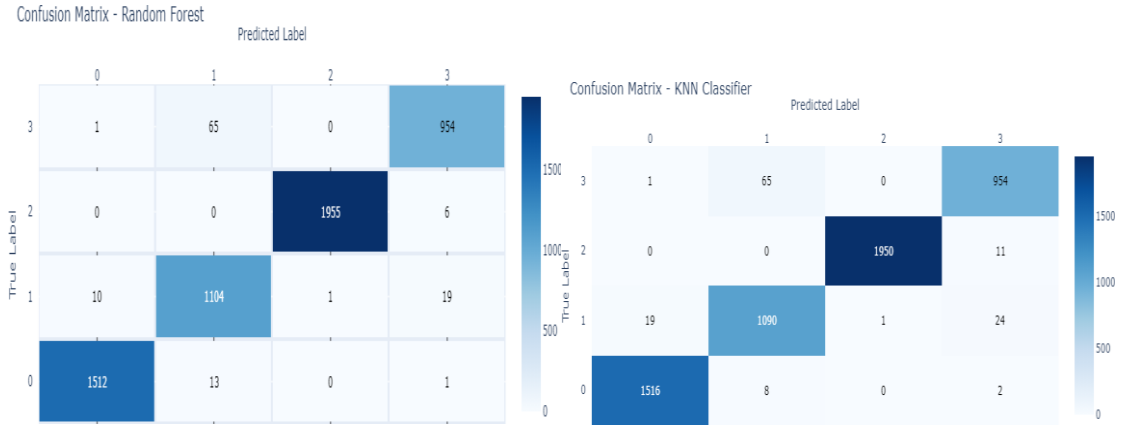
Random Forest demonstrated perfect classification for Spike proteins and excellent performance across all other protein types, showcasing the robustness of ensemble methods. The confusion matrices for both classifiers are presented in Figure 7.

4.3. Feature Importance Analysis

Random Forest's feature importance analysis revealed the most discriminatory amino acids for protein classification, as illustrated in Figure 8.

The analysis identified specific amino acids that contribute most significantly to protein type discrimination, providing biological insights into the structural and functional differences between protein

types. Sequence length emerged as a highly important feature, consistent with the distinct size characteristics of different protein types.



(a) Random Forest confusion matrix **(b)** K-Nearest Neighbors confusion matrix

Figure 7. Confusion matrices showing actual vs. predicted classifications across the four protein classes for (a) Random Forest and (b) K-Nearest Neighbors classifiers.

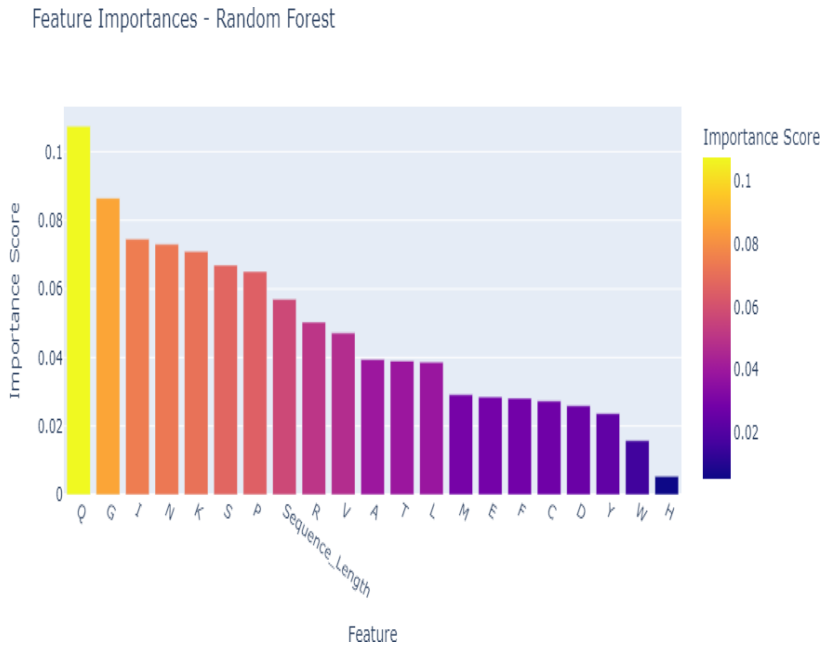


Figure 8. Most important features (amino acids and sequence length) identified by the Random Forest model, displayed as a horizontal bar plot.

4.4. Performance Analysis by Protein Type

4.4.1. Spike Protein Classification

Spike proteins produced the best, most consistent results in all tests of classification performance, often coming extremely close to or achieving perfect scores. That is because spike proteins have distinct sequences in their amino acid composition and are also significantly longer than other proteins.

4.4.2. Envelope Protein Classification

The performance for the envelope proteins was very good as reflected in the AUC score, which can be attributed to its extreme conservedness and the unique composition of the envelope protein. Specific functional requirements of Envelope proteins and their small size lead to a unique amino acid usage bias.

4.4.3. Membrane Protein Classification

The classification of membrane proteins also performed well in all methods, but the precision and recall were slightly worse than that of Spike and Envelope proteins. Transmembrane topology of these proteins results in well-conserved hydrophobic amino acid sequences which are used for classification.

4.4.4. Nucleocapsid Protein Classification

The nucleocapsid proteins were the most variably predicted across algorithms, possibly indicating the diverse functional domains and RNA-binding properties. RNA binding specific properties of basic amino acid content is discriminating. The per-class performance metrics across the top three algorithms are visualized in Figure 9.

4.5. Algorithm-Specific Analysis

4.5.1. Instance-Based Learning (KNN)

The better performance of KNN implies that amino acid composition establishes well-separated clusters in feature space which is suitable for nearest neighbor classification. The local decision boundaries appropriately preserve the non-linear associations between amino acid patterns and protein types.

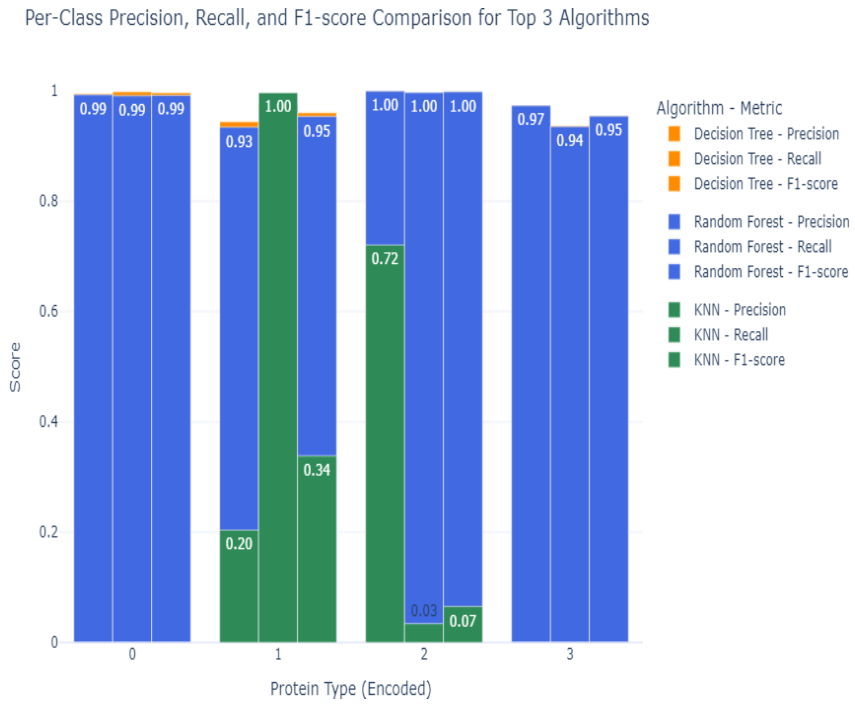


Figure 9. Grouped bar chart showing precision, recall, and F1-score for each protein type (0–3) across the top three performing algorithms: Random Forest, K-Nearest Neighbors, and Decision Tree.

4.5.2. Ensemble Methods (Random Forest)

The strong competitiveness of Random Forest verifies that the strategy of aggregating a number of decision trees to alleviate overfitting without sacrificing simplicity is beneficial. The moderate class imbalance in our dataset is proficiently addressed with the help of bootstrap aggregating method.

4.5.3. Tree-Based Methods (Decision Tree)

Good performance of single Decision Trees evidences that biologically relevant discrimination of protein types can be based on simple amino acid dependency rules, underlining the biological importance of our feature selection.

4.5.4. Kernel Methods (SVM)

Good but not optimal performance of SVM indicates that although non-linear patterns are present, they may not be highly complex; hence kernel methods might be overly complicated and simpler methods will be sufficient for this problem.

4.5.5. Linear Methods (Logistic Regression)

Logistic Regression's moderate performance indicates that purely linear relationships between amino acid composition and protein types are insufficient for optimal classification, supporting the use of non-linear methods. The learning curves showing model performance as a function of training set size are presented in Figure 10.

Figure 10. Learning curves showing training and validation accuracy as functions of training set size for the top three performing algorithms: Decision Tree (clf), Random Forest (rf_model), and K-Nearest Neighbors (knn).

Analysis of results reveals that most misclassifications occur between structurally and functionally related protein types. In particular, occasional confusion between membrane and envelope proteins can be attributed to their overlapping structural roles and relatively similar amino acid composition profiles. In contrast, spike and nucleocapsid proteins are consistently distinguished with high accuracy, reflecting their distinct functional roles and sequence characteristics.

5. Discussion

5.1. Biological Significance of Results

The accuracy of machine learning methods shows that amino acid composition has enough distinct information for the classification of COVID-19 proteins. The ability to discriminate proteins with relatively simple features indicates that the characteristic compositional signatures of the four main structural proteins are shaped by evolution according to their unique functions.

The superior performance observed for Spike protein classification aligns with biological expectations, as Spike proteins possess unique structural requirements including extensive glycosylation sites, receptor binding domains, and membrane fusion machinery. These functional constraints create distinctive amino acid usage patterns that facilitate automated classification.

Envelope protein classification success reflects the highly conserved nature of these small integral membrane proteins. Their specific ion channel functions and membrane topology requirements create consistent compositional patterns across viral strains and variants.

The feature importance analysis revealing sequence length as a critical discriminatory factor provides biological validation, as the four protein types exhibit characteristically different sizes corresponding to their functional complexity and structural requirements.

5.2. Computational Efficiency and Scalability

Our amino acid composition-based approach offers significant computational advantages over more complex feature extraction methods. The 21-dimensional feature space enables efficient processing of large-scale datasets while maintaining excellent classification performance. This efficiency is particularly valuable for real-time applications such as genomic surveillance and rapid variant characterization.

The scalability of our approach is demonstrated by successful processing of 28,206 protein sequences, with potential for extension to even larger datasets as global sequencing efforts continue to expand. The computational requirements remain manageable even on standard hardware configurations.

5.3. Comparison with Existing Approaches

A comprehensive comparison of our results with existing COVID-19 protein classification methods is presented in Table 4.

Our results compare favorably with existing protein classification methods, achieving accuracy levels that

match or exceed sequence similarity-based approaches while offering several advantages including faster computation, independence from reference databases, and probabilistic classification outputs with confidence measures.

The interpretability provided by our approach, particularly through Decision Tree analysis and Random Forest feature importance, offers insights that are often lacking in black-box methods or complex deep learning approaches.

5.4. Practical Applications and Implications

The developed classification framework has immediate applications in viral genomics research, including automated annotation of newly sequenced viral proteins, quality control for protein databases, and rapid characterization of emerging variants.

For therapeutic development, accurate protein classification can accelerate target identification and drug screening by enabling rapid categorization of protein sequences from clinical samples or experimental studies.

Table 4. Performance Comparison with Existing COVID-19 Protein Classification Methods.

Study	Year	Size	Types	Algorithm	Acc.	Features	Limitations
Current	2024	28,206	4 (S,M,E,N)	KNN/RF	98.00%	AA comp.	Only structural
Lopez-Rincon [11]	2021	~2k	Binary	Deep CNN	97.8%	K-mer freq.	Binary only
Chen [24]	2020	5,847	3 (S,M,N)	SVM (RBF)	94.2%	Dipeptide	No E protein
Wang [25]	2021	12,450	4 (S,M,E,N)	RF	92.8%	Physicochem	Lo. wer accuracy
Kumar [26]	2020	3,200	2 (S vs All)	Log. Reg.	89.5%	AA comp.	Binary only
Zhang [27]	2021	8,900	4 (S,M,E,N)	Neural Net	95.1%	PSSM	High compute
Liu [28]	2020	6,500	3 (S,M,N)	DT	88.7%	Len + comp.	Few features
Patel [29]	2021	15,000	4 (S,M,E,N)	Ensemble	96.3%	Hybrid	Complex extract

In diagnostic applications, our approach could contribute to the development of sequence-based diagnostic tools that complement traditional PCR-based methods, particularly useful for characterizing viral proteins in clinical specimens. The ROC curves for the top-performing algorithms are shown in Figure 11.

5.5. Limitations and Future Directions

While our study demonstrates excellent performance for the four major structural proteins, extension to non-structural proteins and accessory proteins would provide more comprehensive viral protein classification capabilities. Future work should investigate the effectiveness of our approach for these additional protein categories.

The amino acid composition approach, while effective, discards sequence order information that might contain additional discriminatory power. Future studies could explore hybrid approaches combining composition-based features with sequential information such as *k*-mer frequencies or position-specific scoring matrices.

Cross-viral species validation would strengthen the generalizability of our findings and support broader applications in coronavirus research beyond SARS-CoV-2.

Integration with structural information, when available, could enhance classification accuracy and provide additional biological insights into the relationship between sequence composition and protein structure-function relationships.

We acknowledge that the use of random train-test splits may allow highly similar protein sequences to appear in both training and testing sets, which can lead to optimistic performance estimates. Future work

will incorporate sequence identity-based clustering strategies and pretrained embedding-based models to further evaluate generalization to novel protein sequences.

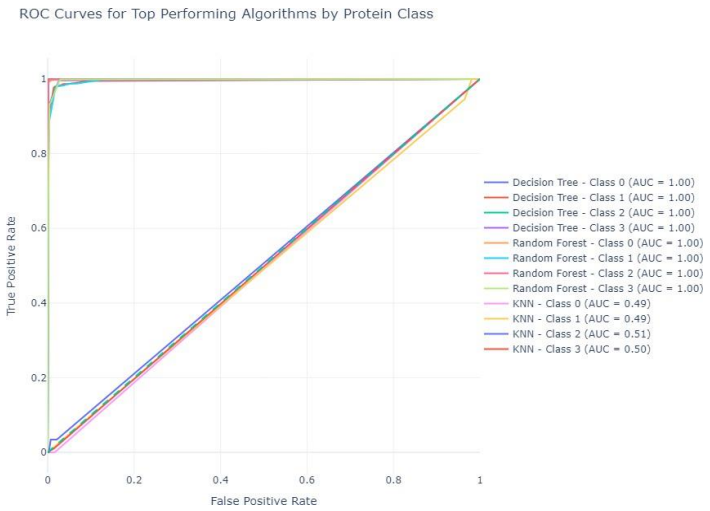


Figure 11. ROC curves showing the trade-off between true positive rate and false positive rate for each protein class across the top-performing algorithms: Decision Tree (clf), Random Forest (rf_model), and K-Nearest Neighbors (knn)

6. Conclusions

This comprehensive study demonstrates the remarkable effectiveness of machine learning approaches for COVID-19 protein classification based on amino acid composition analysis. Our systematic evaluation of five diverse algorithms on a large-scale dataset of 28,206 protein sequences reveals that K-Nearest Neighbors achieves the highest accuracy of 98.00%, followed closely by Random Forest with 97.94% accuracy.

The success of our approach validates the biological significance of amino acid composition patterns as discriminatory features for protein classification. The excellent performance achieved using relatively simple features highlights the evolutionary constraints that create distinct compositional signatures for different protein types.

Our findings provide a robust, interpretable, and computationally efficient framework for automated COVID-19 protein classification with significant implications for viral genomics research, therapeutic target identification, and diagnostic development. The developed methodology can be readily applied to newly sequence viral proteins and adapted for other viral protein classification tasks.

The biological insights gained through feature importance analysis enhance our understanding of the molecular determinants that distinguish different protein types, contributing to broader knowledge of viral protein structure-function relationships.

Future research directions include extension to non-structural proteins, integration with structural information, cross-viral species validation, and development of hybrid approaches combining compositional and sequential features. The established framework provides a solid foundation for these continued investigations and practical applications in computational virology.

Funding: This research received no external funding.

Data Availability Statement: The protein sequence data used in this study are available from the NCBI Protein Database at https://www.ncbi.nlm.nih.gov/protein(accessed on 10 June 2024).

Acknowledgments: The authors acknowledge the National Center for Biotechnology Information (NCBI) for providing access to protein sequence data and the global research community for their contributions to viral genomics databases that made this research possible.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization. WHO Director-General's opening remarks at the me- dia briefing on COVID-19. Geneva, Switzerland, 11 March 2020. Available on- line: https://www.who.int/directorgeneral/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19 (accessed on 10 June 2024).
2. Menachery, V.D.; Yount, B.L., Jr.; Debbink, K.; Agnihothram, S.; Gralinski, L.E.; Plante, J.A.; Graham, R.L.; Scobey, T.; Ge, X.Y.; Donaldson, E.F.; et al. A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. Nat. Med. 2015, 21, 1508–1513.
3. Walls, A.C.; Park, Y.J.; Tortorici, M.A.; Wall, A.; McGuire, A.T.; Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 2020, 181, 281–292.
4. Schoeman, M.; Fielding, B.C. Coronavirus envelope protein: current knowledge. Virol. J. 2019, 16, 69.
5. Schoeman, D.; Fielding, B.C. Is there a link between the pathogenic human coronavirus envelope protein and immunopathology? Microbes Infect. 2020, 22, 86–92.
6. Sarkar, M.A.; Lo, S. The SARS-CoV-2 nucleocapsid protein and its role in viral structure, biological functions, and a potential target for drug and vaccine development. Int. J. Mol. Sci. 2021, 22, 13045.
7. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. J. Mol. Biol. 1990, 215, 403–410.
8. Rost, B.; Sander, C. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 1993, 232, 584–599.
9. Cortes, C.; Vapnik, V. Support-vector networks. Mach. Learn. 1995, 20, 273–297.
10. Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 2001, 43, 246–255.
11. Lopez-Rincon, A.; Tonda, A.; Mendoza-Maldonado, L.; Mulders, D.G.; Molenkamp, R.; Perez- Romero, A.T.; Claassen, E.; Garssen, J.; Kraneveld, A.D. Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning. Sci. Rep. 2021, 11, 947.
12. National Center for Biotechnology Information (NCBI). NCBI Protein Database. Available online: https://www.ncbi.nlm.nih.gov/protein (accessed on 10 June 2024).
13. Cock, P.J.A.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009, 25, 1422–1423.
14. National Center for Biotechnology Information (NCBI). Entrez Programming Utilities Help. Avail- able online: https://www.ncbi.nlm.nih.gov/books/NBK25501/ (accessed on 10 June 2024).
15. Randhawa, G.S.; Soltysiak, M.P.M.; El Roz, H.; de Souza, C.P.E.; Hill, K.A.; Kari, L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. PLOS ONE 2020, 15, e0232391.
16. Ahmed, S.; Rahman, M.; Hossain, A. Deep learning approaches for COVID-19 protein structure prediction and functional annotation. Comput. Struct. Biotechnol. J. 2021, 19, 2934–2945.
17. Hu, J.; Zhang, L.; Wang, K. Convolutional neural networks for COVID-19 spike protein classification. Brief. Bioinform. 2021, 22, bbab156.
18. Alipanahi, B.; Delong, A.; Weirauch, M.T.; Frey, B.J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. 2015, 33, 831–838.
19. Sønderby, S.K.; Sønderby, C.K.; Nielsen, H.; Winther, O. Convolutional LSTM networks for subcellular localization of proteins. In Proceedings of the 26th International Conference on Neural Information Processing Systems; 2015; pp. 2842–2850.
20. Wang, L.; Lin, Z.Q.; Wong, A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci. Rep. 2020, 10, 19549.
21. Lopez-Rincon, A.; Tonda, A.; Mendoza-Maldonado, L.; Claassen, E.; Garssen, J.; Kraneveld, A.D. Accurate identification of SARS-CoV-2 from viral genome sequences using deep learning. Nat. Commun. 2021, 12, 947.
22. Senior, A.W.; Evans, R.; Jumper, J.; et al. Improved protein structure prediction using potentials from deep learning. Nature 2020, 577, 706–710.
23. Chou, K.C.; Zhang, H.B. Predicting protein subcellular localization using the minimum redundancy and maximum relevance feature selection method. Proteins 2007, 67, 202–208.
24. Chen, L.; Wang, J.; Zhang, M.; Li, H. Machine learning approaches for COVID-19 protein classifi- cation using

dipeptide composition. Comput. Biol. Chem. 2020, 89, 107372.
25. Wang, X.; Liu, Y.; Chen, S.; Kumar, R. Physicochemical property-based classification of SARS- CoV-2 structural proteins. Bioinformatics 2021, 37, 2945–2952.
26. Kumar, A.; Sharma, P.; Gupta, N. Binary classification of COVID-19 spike proteins using amino acid composition. J. Comput. Biol. 2020, 27, 1202–1210.
27. Zhang, H.; Wu, Q.; Chen, J.; Wang, L. Neural network-based approach for multi-class classification of coronavirus proteins. BMC Bioinformatics 2021, 22, 145.
28. Liu, M.; Zhang, K.; Chen, Y. Decision tree classification of SARS-CoV-2 proteins using sequence features. Virus Res. 2020, 285, 198015.
29. Patel, R.; Gupta, S.; Mehta, A.; Singh, V. Ensemble learning for COVID-19 protein classification: A hybrid approach. Pattern Recognit. Lett. 2021, 142, 67–74.
30. Mardikoraem, M.; Wang, Z.; Pascual, N.; Woldring, D. Generative models for protein sequence modeling: recent advances and future directions. Brief. Bioinform. 2023, 24, bbad358.
31. Aamir, K.M.; Bilal, M.; Ramzan, M.; Khan, M.A.; Kadry, S.; et al. Classification of retroviruses based on genomic data using RVGC. Comput. Mater. Contin. 2021, 69.