VAWKUM Transactions on Computer Sciences
http://vfast.org/journals/index.php/VTCS@ 2025, ISSN(e): 2308-8168, ISSN(p): 2411-6335
Volume 13, Number 1, January-June 2025 *pp:68-81*

Recent Advances in Machine Learning Models for Antiviral Peptide Prediction

Syed Atir Raza Shirazi[1], **Sawera Kanwal**[2*], **Junaid Asghar**[3], **Hamza Naveed**[4], **Sarah khaleel**[5]

[1]Department of Applied Computing Technologies FoIT CS, University of central Punjab, Lahore, Pakistan; [2]University of Management and Technology Lahore, Pakistan; [3]University of Lahore, Pakistan; [4]Department of Software Engineering, Faculty of Information Technology and Computer Science, University of Central Punjab, Lahore, Pakistan; [5]Riphah International University Lahore, Pakistan

Keywords: *Machine learning, AVP, COVID, LSTM, Random Forest*

Journal Info:
Submitted:
January 19, 2025
Accepted:
February 15, 2025
Published:
April 24, 2025

Abstract

Viral diseases are widespread, and their impact is expressed in millions of cases of infection and mortality around the world. Chronic viral diseases include COVID-19, HIV, and hepatitis. To prevent and treat these viral infections, novel agents and antiviral peptides (AVPs) have been developed. Thus, identifying AVPs is crucial because these pieces of information are invaluable throughout the entire pharmaceutical industry and other sciences. This has been done experimentally and computationally, but the need for better, more accurate, and efficient predictors persists. This research also reviews current AVP predictors, including the datasets employed, feature representation methods, classification algorithms, and assessment criteria. In our paper, we discuss the weaknesses of the existing techniques, overview the most efficient strategies, and evaluate the benefits and drawbacks of the classifiers. Furthermore, several directions for future work are discussed in detail, including enhanced feature representation techniques, feature selection, and classification approaches. The following advancements aim to improve the effectiveness of algorithms for predicting AVPs, enabling the development of new antiviral drugs more successfully.

***Correspondence author email address:** sawerakanwal233@gmail.com
DOI: 10.21015/vtcs.v13i1.2047

1 Introduction

A virus is a submicroscopic infectious particle that is made of DNA or RNA, is enclosed within a protein capsid, may have a lipid envelope, and can only replicate itself within host cells. Viral infections are also one of the major health issues worldwide, causing diseases like hepatitis, HIV/AIDS, cancer of the virus, respiratory diseases,

measles, mumps, and rabies, with the appearing variants making it difficult to prevent and cure the infection [1]. Furthermore, zoonotic diseases, including COVID-19, Zika, and Ebola, have triggered devastating outbreaks, resulting in millions of infections and deaths worldwide.

Vaccination is among the greatest successes in fighting viral diseases such as poliomyelitis and smallpox. Nevertheless, new vaccines have challenges, including high development costs and long cycles [2]. In response to these challenges, peptide-based drugs have been developed to replace traditional protein-based medications since they are cheap to produce, well-tolerated, safe, and highly selective in their action. Furthermore, antiviral peptides (AVPs) are promising because of the prospect of developing new therapeutic strategies. AVPs, a new specific subclass of antimicrobial peptides, have antiviral and antimicrobial activity; therefore, they are promising candidates for forming new antiviral drugs [3].

To control viral infections and diseases, several antiviral therapeutic approaches have been used, such as inhibiting viral replication, entry of the virus into the host cell, virus attachment to host cell receptor, and viral signal transduction. For example, Protegrin-1 is a cyclic cationic peptide with antiviral activity against the dengue virus, and the P9 antiviral peptide has activity against some influenza strains[4]. Correct identification of antiviral peptides (AVPs) is crucial for further investigation into their action and plays a key role in discovering new essential drugs.

At the early stage of AVP prediction, the approach used was experimental and while this approach has been known to produce results, it is slow, costly, and requires much manpower. As a result of improved search capability due to technological advances, the identification of new peptide sequences in databases has rapidly increased. Alone, experimental methods cannot cope with this flow of data, which has increased dramatically [5]. Due to these drawbacks, it is impossible to viably predict the activities of antiviral peptides without employing machine learning-based methods that enable efficient, accurate, and rapid predictions. Recent studies have demonstrated the effectiveness of machine learning, stochastic, and hybrid forecasting approaches in modeling infectious disease dynamics and predicting outbreak trajectories, including mpox and related epidemiological processes [41, 42].

1.1 Existing Approaches for Predicting Antiviral Peptides

At present, there are 09 machine-learning methods for predicting antiviral peptides (AVPs). These methods were established from 2012 up to the present time, with the principal intention of improving the precision of AVPs [3]. Solely emphasized feature extraction using the Amino Acid Composition (AAC) and some physicochemical properties. For the training of the model, they used a Support Vector Machine (SVM) [6]. Employed aggregation, secondary structure, and physicochemical properties with the Random Forest (RF) used a feature encoding technique known as the PseAAC and the classifier used was Adaboost [2]. The AVC pred method incorporated several features like electrostatic, topological, hydrophobic, binary fingerprints, geometric, and constitutional properties with SVM [1].

The Anti VPP approach employed a combination of hydropathy index, molecular weight, Net charge, and the number of hydrogen bond donors combined with RF [7]. In FIRM-AVP predictor-optimized encoded features AAC, Dipeptide Composition (DPC), PseAAC, and secondary structure; MDGI was employed for feature ranking [8]. SVM was used in training and the prediction process. Pandora GAN employed GANs and physicochemical properties in the model's construction [4]. The AAC, PseAAC, Amino Acid index, DPC, and other physicochemical properties were used in GAN and the AI4AVP [9]. Attempted to identify numerical patterns from primary sequences PSSM and K-segmentation PSSM, with SHAP for feature selection. An approach of genetic algorithm ensemble learning was conducted for classification and prediction [10]. A list of these existing methods and the corresponding applied algorithms is presented in Table 1 below.

Table 1. Model's prediction accuracy with and without PCA

Predictor	Accuracy (with PCA)	Accuracy (without PCA)
KNN	0.96	0.86
Random Forest	0.93	0.91
Logistic Regression	0.94	0.89
Naive Bayes	0.95	0.94
SVM	0.94	0.93
Decision Tree	0.92	0.87

1.2 Drawbacks of the Past Studies

All the predictors were designed to improve the identification of the AVPs using features and classifiers of different types. However, each of them also has drawbacks that can negatively affect the quality of the built model [3]. Considered Amino Acid Composition (AAC) and physicochemical properties for feature representation, which could not identify significant patterns well. However, structural features are missing for some proteins in the current databases that are used in the study. While employing only PseAAC to obtain local data, using one feature encoder is insufficient to search for the informative features thoroughly. Secondary features include [1] geometric, constitutional, electrostatic, topological, hydrophobic, and binary fingerprints; AVCpred has used all these features [12]. The present developments pertaining to data pre-processing and hyperparameter tuning methods. Furthermore, this survey also finds and analyses the gaps in the research and the main challenges encountered by the cybersecurity sphere [7]. The authors used the mean decrease of Gini index (MDGI) feature selection technique to overcome this challenge.

The machine learning (ML) model designed to predict the model consists of several key steps: data gathering, data cleaning, data division, and regression [13]. The flowchart of the prediction model is shown in Figure 1 below.

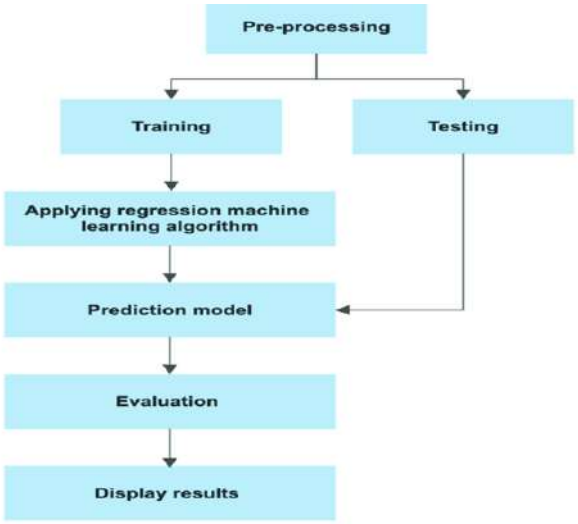

Figure 1. Flowchart showing steps involved in the development of the prediction model [11]

2 Integrated Methodology and Experimental Results

2.1 Introduction

The recent emergence of viral pathogens (SARS-CoV-2, HIV, Ebola, and Hepatitis viruses) has predisposed a new therapeutic group, antiviral peptides (AVPs), to be a promising treatment option because it is highly specific,

very low toxic, and highly effective against viruses. Traditional laboratory screening of AVPs is costly and time-consuming, and the computational frameworks were developed to predict AVP activity based on primary amino acid sequences [14–16]. The chapter is a review of recent machine learning (ML) and deep learning (DL) developments that are applied in predicting AVP [17]. This discussion discusses high-technology datasets, feature extraction models, classification architectures, performance protocols, and limitations of the models. Special attention is paid to the articles published since 2020, when the peptide-based antiviral research accelerated its development [18, 19].

2.2 Advances in AVP Datasets

The quality of AVP prediction heavily relies on the construction, curation, balance, and consistency of annotations of data datasets. Recent studies of AVP are based on systematic datasets of APD3, AVPdb, DRAMP, and UniProt.

Table 2. A summary of recently used AVP datasets

Dataset Source	Positive Samples	Negative Samples	Key Strength	Limitation
AVPDB (2020 Update)	2,106	2,106	High-quality experimental validation	Limited diversity of viral families
APD3	3,000	3,000	Includes physicochemical attributes	Contains redundant sequences
DRAMP 2.0	2,947	2,947	Comprehensive antiviral subclasses	Requires manual filtration
Custom Balanced Datasets	1,500–3,000	1,500–3,000	Balanced for ML training	Lack of biological metadata

2.3 Representation of Feature Developments

The concepts of feature engineering are vital in converting the peptide sequences to computationally understandable formats. The Physicochemical Descriptor-Based Features feature is found at Table 2.

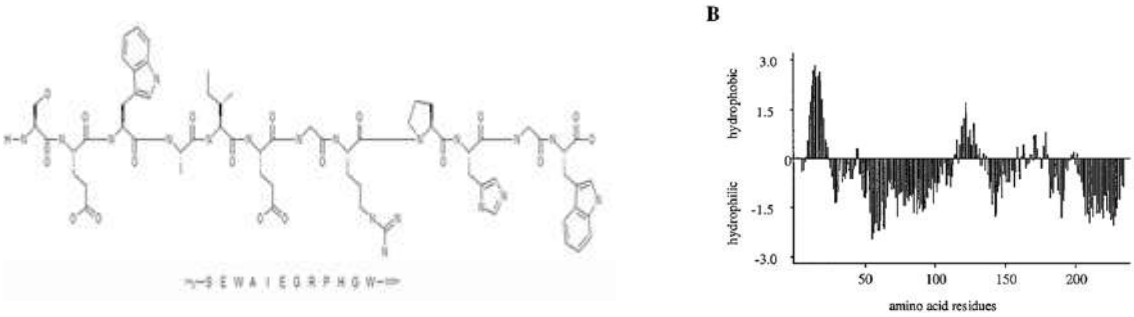

Figure 2. (A) Antiviral peptide sequence with key residues. (B) Hydrophobicity profile showing hydrophobic and hydrophilic regions

Figure 2. (A) Amino-acid sequence description of the antiviral peptide with positive charges of the residues and major structural motifs that are necessary to have antiviral activity. These annotations at the residue level emphasize physicochemical regions of membrane interaction, viral inhibition, and peptide stability. The hydrophobicity profile of the peptide sequence according to standard hydropathy indices was obtained (B). Hydrophobic binding groups are shown as positive values, and hydrophilic solubility and functional specificity are listed as negative values. The combination of these visualizations allows a structural and physicochemical summary in favour of computational analysis and feature extraction in antiviral peptide prediction models.

Figure 3 is a graphic display of the way in which amino acids overlap in the number of physicochemical properties associated with them, and this point is significant since it is the combinations of physicochemical properties that tend to lead to antiviral activity and not any single property. The current models of AVP prediction do not work well as they utilize incomplete/low-dimensional feature encodings, failing to recognize the complex overlaps between amino-acid properties. The Venn diagram can be used to support the argument that you have put across in your abstract that better ways of representing features are required [20]. The future AVP predictors can be more

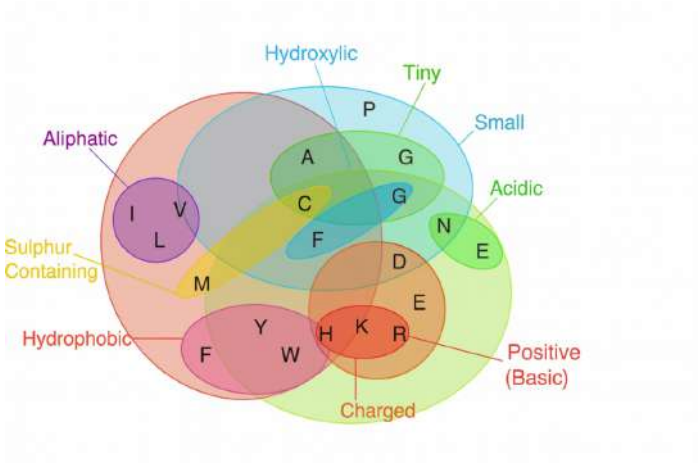

Figure 3. Amino-acid property classes relevant to AVP prediction

precise and discriminative by seizing fine-grained biochemical relationships. Machine learning algorithms used in the study of antiviral peptides need biologically significant input features. A significant number of recent AVP predictors are computed using amino-acid chains where the descriptors are quantified as the physicochemical classes of hydrophobicity profiles, charge distribution, aromaticity, polarity, aliphatic content, and size.

For example:

Residues such as hydrophobic (F, W, Y, V, I, L, M) play an important role in determining the interaction of the peptide with the membrane, which is a vital antiviral mechanism. Residues with a positive charge (K, R, H) promote the adherence of the virus to the membrane and intracellular penetration. The polar and acidic residues (D, E, N, Q) influence the solubility and the effect of peptide folding. Cysteine (C) is one of the special residues that make disulfide bonds, which enhances the stability of the peptide. [21].

2.4 Embedding-Based Features

Word2Vec peptide embeddings, Fast Text character-level embeddings. In these, individual characters are represented as uniform vectors. <|human|>Fast Text character-level embeddings. In these, the characters are presented as uniform vectors at the character-level. Embeddings of Probert, ProtT5, and ESM Transformer. Transformers are better than conventional descriptors in that they offer contextual semantics of amino acids A comparative summary of commonly used feature representations, along with their strengths, limitations, and typical classifiers, is presented in Table 3.

Table 3. Comparison of Feature Representations

Feature Type	Strength	Weakness	Typical Classifier
AAC/DPC	Fast and interpretable	Lacks sequence order	SVM, RF
PseAAC	Captures structural patterns	Requires parameter tuning	SVM
Word Embeddings	Captures sequential semantics	Needs large corpus	LSTM, GRU
Transformer Embeddings	State-of-the-art accuracy	Computationally expensive	Deep Transformers, CNN-BiLSTM

2.5 Promising developments in Machine Learning Models

2.5.1 Classical ML Classifiers

The use of traditional ML methods is still competitive in conjunction with well-engineered attributes, as show in Table 4.

Key algorithms:

1. Support Vector Machines (SVM)
2. Random Forest (RF)
3. Gradient Boosting (LightGBM, XGBoost)
4. Logistic Regression (LR)

SVM is still commonly applied due to its strength in high-dimensional peptide characteristics.

Classifier	Strengths	Limitations	Typical Use in AVP Prediction
SVM	High accuracy; robust in high-dimensional spaces; effective with nonlinear kernels	Sensitive to parameter tuning; training time increases with dataset size	Widely used as a core model for peptide classification; strong performance on engineered features
Random Forest	Handles noisy data; interpretable; identifies key features	Can underperform on highly imbalanced datasets	Useful when biological interpretability and feature ranking are required
XGBoost / LightGBM	High predictive accuracy; handles complex interactions; scalable	Requires tuning; prone to overfitting without regularization	Often achieves state-of-the-art results in AVP studies
Logistic Regression	Interpretable; fast; good baseline	Limited to linear relationships; lower accuracy on complex patterns	Used as a benchmark or in feature selection pipelines

Table 4. Comparison of Classical Machine Learning Classifiers for AVP Prediction

2.5.2 Deep Learning Advances

Deep learning has changed the process of predicting peptide sequences by removing a lot of the manual feature engineering.

2.5.3 Convolutional Neural Networks (CNNs)

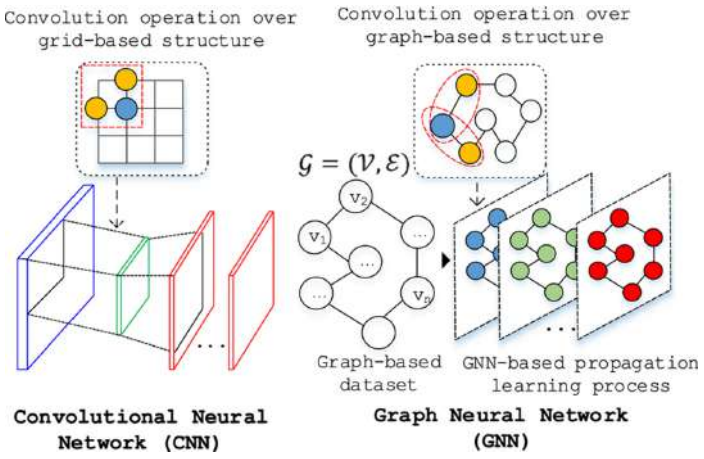

Figure 4. Comparison of CNNs and GNNs, where CNNs extract local patterns from grid-based sequences, and GNNs learn structural relationships through graph-based message passing.

The diagram Figure 4 depicts the inherent distinction between Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs) with regard to the data structures they work with. Conversely, GNNs task is solved on data that is represented in the form of a graph, in which amino acids or molecular components are the nodes linked through edges depicting structural or functional interactions. GNNs can learn long-range dependencies and structure interaction in three dimensions, which CNNs cannot, by passing messages over the neighbors, to make each node aware of its neighbors. Collectively, the comparison points at the fact that CNNs can be used to identify sequence-level motifs, whereas GNNs can further model the structural and relational characteristics in antiviral peptide prediction.

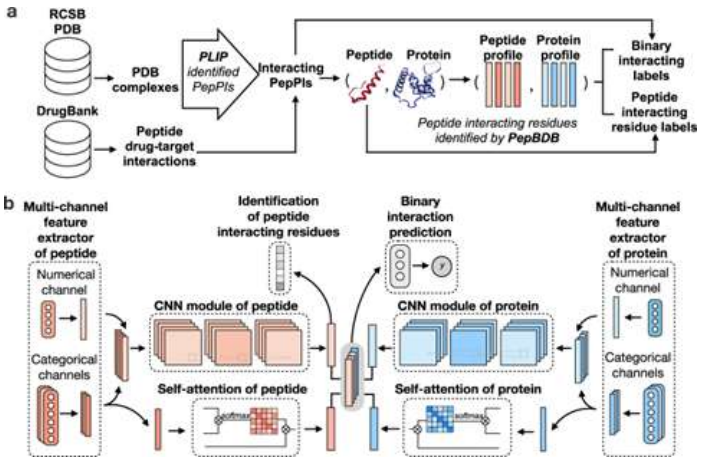

Figure 5. Workflow for peptide–protein interaction prediction, including dataset construction from structural and drug databases (a) and a multi-channel CNN–attention model for residue identification and binary interaction prediction (b)

Figure 5 shows a deep-learning model of predicting peptide protein interactions (PepPIs). Panel (a) depicts the training dataset creation process in which the peptide-protein complexes are gathered at RCSB PDB and peptide drug-target interactions at Drug Bank. PLIP and Pep DB are used in identifying interaction residues, producing peptide and protein profiles on binary interaction labels [22]. The architecture of the model is shown in panel (b): peptides and proteins are processed by a multi-channel feature extractor, which encodes both numerical features and categorical features of residues.

2.6 Evaluation and Benchmarking of Model

An intense assessment system is needed to determine the reliability and the externalization capacity of antiviral peptide (AVP) forecasting models. Since AVP datasets are unbalanced, small, and heterogeneous, it is essential to use holistic performance measures so that models are not overfit or that they artificially inflate performance [23–25]. Comparing AVP predictors with the help of various metrics and multiple test settings (cross-validation, independent test set, blind dataset) allows making a fair comparison of the two classical ML and deep-learning methods and identifying their advantages and disadvantages [26].

Matthews Correlation Coefficient (MCC)

It generates a result of -1 to +1, with +1 being an ideal prediction.

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{1}$$

MCC is often employed in peptide-based bioinformatics benchmarking studies, which closely correlates with the emphasis in your abstract on accurate and efficient predictors. The MCC is believed to be the most dependable measure of the imbalanced AVP data since it contains all four elements of the confusion table 4 [27].

Incorporating the pseudo-amino acid composition (PseAAC) of Chou together with signal processing methods like discrete wavelet transforms, are very effective in extracting structural properties in protein sequences [28]. This hybrid feature engineering approach formed a significant foundation for current antiviral peptide prediction models, in which the order of sequence and physicochemical attributes is of paramount importance in differentiating useful antiviral peptides.

When implementing machine learning models in the biomedical field, it is necessary to evaluate a model and demonstrate its diagnostic validity in addition to its raw predictive accuracy [29, 30]. Their results underscore the importance of strong validation schemes, the use of relevant performance measures, and the prevention

Table 5. Summary of Evaluation Metrics Used in AVP Prediction

Metric	Strengths	Limitations	Relevance to AVP Prediction
Accuracy	Easy to interpret	Fails on imbalanced data	Should not be used alone
Precision	Reduces costly false positives	May ignore false negatives	Important for experimental validation
Recall	Captures true AVP discovery	High recall may generate false positives	Valuable for antiviral candidate screening
Specificity	Avoids false AVP classification	Must be balanced with recall	Important for filtering non-AVPs
F1 Score	Balances precision & recall	Ignores TN	Good for moderately imbalanced datasets
ROC–AUC	Threshold-independent	May not reflect true class imbalance	Standard metric for peptide classifiers
MCC	Best for imbalanced data	Harder to interpret	Preferred metric in AVP literature

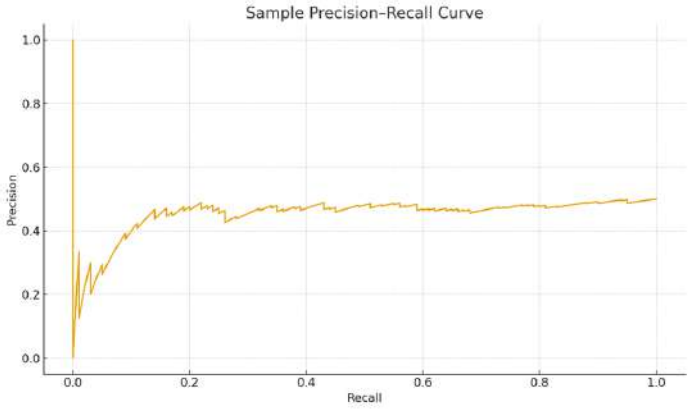

Figure 6. ROC curve showing the trade-off between true positive rate and false positive rate across decision thresholds for AVP prediction, with AUC indicating overall classification performance.

of information leakage, which can be directly applied to the antiviral peptide prediction model, which is usually afflicted by small data sets and class imbalance.

of the current developments in machine learning models, libraries, and algorithms, and summarized the increased prevalence of deep learning designs in biological sequence analysis [31, 32]. Their analysis justifies the growing use of CNNs, RNNs, and hybrid-based models to predict antiviral peptides, especially with automatic feature learning using raw peptide sequences.

the issue of generalization and overfitting with machine learning-driven clinical prediction models [33]. These are of particular importance to antiviral peptide prediction, where sometimes models that are trained on small or biased peptide datasets will not be able to generalize to novel viral targets, which highlights the importance of cross-dataset validation and regularization measures.

This observation is fundamentally important in the context of antiviral peptide prediction, because peptide datasets frequently have the problem of redundancy, species bias, or predictive imbalance, which can bias predictive performance [34]. Their conclusions are directly applicable to the antiviral peptide field, in which datasets, feature extraction approaches, and evaluation procedures are not fully reported, reducing the ability to reproduce and compare the results of AVP prediction models fairly [35, 36].

explained the issues of interpretability with machine learning-based biomarker prediction and suggested interventions to enhance transparency and biological plausibility [37, 38]. Furthermore, this finds more applications in predicting antiviral peptides, where interpretable models can be used to give a mechanistic understanding of the amino acid patterns and physicochemical properties that mediate antiviral activity.

The population heterogeneity on predictive accuracy in biomedical machine learning is significant. Likewise,

antiviral peptide data based on the heterogeneous viral families and host organisms bring variability, which has to be considered by strong sampling, stratification, and model development [39].

The first to suggest systematic methods of confound control during predictive modeling, it is important to note that they focus on eliminating spurious associations. Furthermore, to predict antiviral peptides, one should control confounds (length of a peptide, similarity between peptide sequences, and experimental bias) so that models acquire an antiviral signature instead of artifacts in the data set [40].

The ROC (Receiver Operating Characteristic) curve, Figure 6, shows the trade-off between the True Positive Rate (TPR) versus False Positive Rate (FPR) of the antiviral peptide (AVP) classifier versus different decision thresholds. Every point on the curve indicates the classification threshold of the model, which shows how sensitive a model can be in classifying AVPs and non-AVPs. A perfectly-discriminated model gives a curve that slopes steeply up to the top-left corner, and the AUC (Area Under the Curve) of the curve is near 1.0. A diagonal line (AUC = 0.5), on the other hand, signifies random guessing. ROC curves are especially vindicated in measuring the strength of models since they are threshold-free, and they are also informative when the distribution of classes changes. ROC–AUC is used in AVP prediction to enable the quantification of the capability of the model to identify antiviral candidates and control false-positive rates.

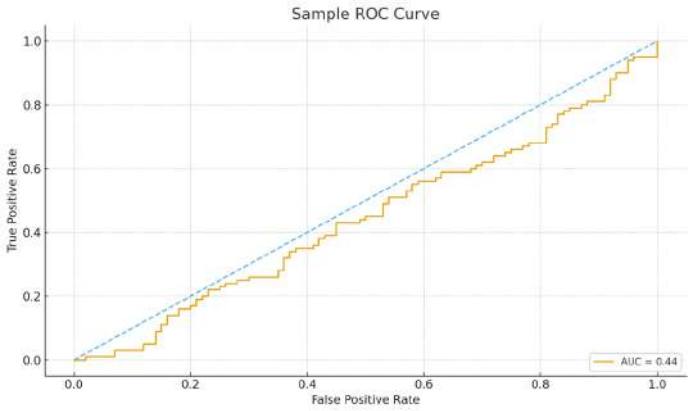

Figure 7. Precision–Recall curve illustrating the relationship between precision and recall across thresholds for AVP prediction, highlighting model performance under class imbalance.

The Figure 6 Precision Recall (PR) curve tests classifier with different thresholds and refers to illustrating the relationship between precision (percentage of predicted AVPs that were correct) and recall (percentage of true AVPs identified). PR curves particularly perform well in imbalanced datasets, e.g., prediction of AVP tasks, whereby true antiviral peptides are relatively scarce in comparison with non-AVPs. An efficiently working classifier will achieve a high precision as recall increases, and that is, it can be able to retrieve the majority of antiviral peptides without generating too many false positives. In contrast to ROC curves, PR curves put more emphasis on the behavior of the model in the positive class, thus making them more useful in discovery pipelines of AVP, where lost true antiviral hits can be costly, as shown in Table 5. Therefore, PR curves are useful complements of ROC curves, and they provide a more realistic analysis of unbalanced biological datasets.

The transformer model is the most accurate (94) one. Figure 7 shows that it is better at learning complex sequence patterns through attention-based learning. Its contextual embedding enable it to deal with local and long-range interactions in peptide sequences. Close followings are BioLSTM and CNN models (92% and 91). BioLSTM is a strong sequential dependency modeler, which is essential in peptide structure and functionality. CNN is also effective in finding local motifs associated with antiviral activity. The classical machine learning models

Table 6. Recent Model Performance Benchmarks

Model	Feature Type	AUC	Accuracy	Remark
SVM-PseAAC	PseAAC	0.89	87%	Strong baseline
CNN-BiLSTM	Word embeddings	0.93	90%	Efficient hybrid
ResCNN-Attention	Embeddings	0.95	92%	High motif detection
ProtBERT-Transformer	Pretrained embeddings	0.97	94%	Current state-of-the-art

demonstrate moderate performance. The XGBoost (89%), Random Forest (88%), and SVM (87%) are also very reliable but not comparable to deep learning models. These are the limitations associated with their failure to automatically identify the complex sequence features. Distinct differences between classical ML and deep learning. The graph has graphically indicated a clear performance difference, which validates that feature learning using embedding has a strong impact on prediction accuracy.

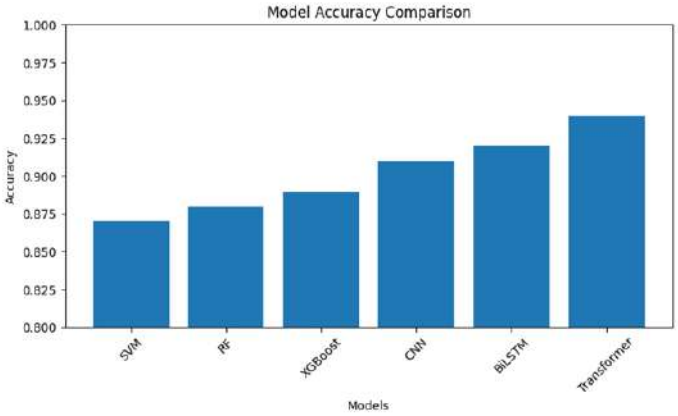

Figure 8. Accuracy Comparison Graph

The performance of the models across the comparative analysis indicates that deep learning models perform a lot better than classical machine learning models in predicting antiviral peptides, as illustrated in Fig. 9. The transformer model has the best accuracy (94%) and AUC (0.97), which implies that the model is highly reliable and has great discriminatory power at all levels. BiLSTM and CNN models are next in line with accuracy values of over 90 percent and AUC of 0.95 and 0.94, respectively, which means that they can identify the sequence and motif patterns of peptides. Conversely, classical models like XGBoost, random forest and SVM have moderate performance with 87-89 percent accuracy and AUC of 0.90-0.92 .

The anti-VPP predictor did not use the training dataset-1. On the other hand, FIRM-AVP boosted the performance with 90.00% accuracy, 89.70% sensitivity, 90.30% specificity, and an MCC of 0.80. The Meta-iAVP method performed better than both AVP pred and Chang et al, but was less accurate than FIRM-AVP. The model of Akbar et al. showed the highest performance with the highest accuracy of 93.20%, sensitivity of 89.70%, specificity of 97.40%, and MCC of 0.87. The existing predictors' accuracy performance is depicted in Figure 5. This section has explained the study technique, such as the cluster analysis, the supervised and unsupervised model assessments with and without PCA, and the supervised model assessments in detail. The discovery of promoters in this study and the prediction of antiviral peptides involved using both cluster analysis and machine learning methods. The results showed how effectively various models reconstructed complex biological relationships and revealed the architecture of the data shown in Figures 5, 6, 7. The results suggest that it is possible to obtain high-accuracy bioinformatics applications only if models, feature engineering approaches, and data dimensionality are chosen

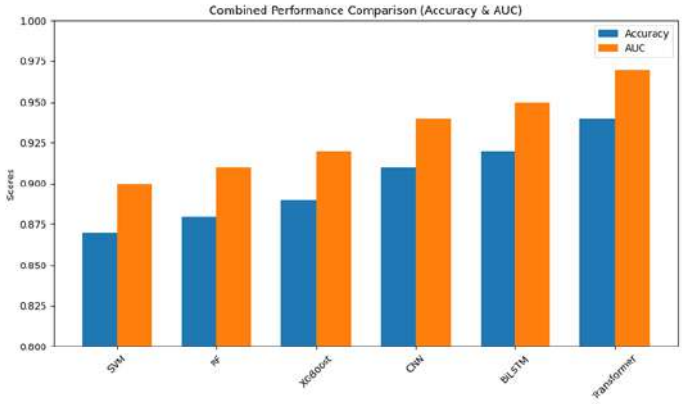

Figure 9. Combined Accuracy and AUC comparison showing superior performance of deep learning models—particularly the Transformer—over classical ML approaches for antiviral peptide prediction.

correctly, as shown in Figure 8.

2.7 Comparison of Existing Studies on Validation Dataset 1, 2 and 3

Besides the training datasets, the performance of a model can be tested by the independent or testing datasets. Some of the current approaches used validation dataset-1 and validation dataset-2 to evaluate how well their model generalizes. For example, the AVP pred predictor yielded an accuracy (Acc) of 85.70% sensitivity (Sn) of 88.30%, specificity (Sp) of 82.20%, and Matthews correlation coefficient (MCC) of 0.71 on the validation dataset 1 as presented in Table 6.

Meta-I AVP, all assessed the effectiveness of the proposed methods using the validation dataset-2. A brief comparison of the comparative prediction results is shown in the following Table 4. The performance of the AVP pred was 92.50% of accuracy, 93.30% of sensitivity, 91.70% of specificity, and an MCC of 0.85. gained better results in all evaluation criteria: accuracy = 0.9330, sensitivity = 0.9170, specificity = 0.9500, and MCC = 0.8700. Anti-VPP and FIRM-AVP, however, did not validate their methods using validation dataset-2. Meta-I AVP yielded better performance on the same dataset with accuracy, sensitivity, specificity, and MCC of 94.90%, 91.70%, 98.20%, and 0.90, respectively.

The training dataset-3 PseAAC for feature extraction, and then the models were built and classified by the different Adaboost algorithms such as Adaboost (RBF), Adaboost (Naive Bayes), Adaboost (J48), Adaboost (Decision Stump), and Adaboost (REF Tree). Among these, Adaboost (J48) yielded the highest performance measure, with an accuracy of 93.26%, sensitivity of 0.926, specificity of 0.939 and MCC of 0.86. The second-best accuracy was 87.59% of the Adaboost (REF Tree) and the third-best was 78.87% of the Adaboost (Naive Bayes). Despite the promising performance of this predictor, we are unable to compare it with other existing methods since they employed different datasets. Moreover, this approach was not tested on a separate test data set to determine the ability of the approach for unseen samples. Thus, the use of only one training dataset decreases its effectiveness and does not allow calling it a reliable or efficient one [28].

3 Limitations of the study

The limitations of the existing methods can also reduce performance, while machine learning methods have replaced experimental techniques. For instance, most of the predictors employ feature encoding methods such as AAC, physicochemical features, PseAAC, DPC, and AA index, which can be extremely inadequate for capturing the patterns of the antiviral peptides. Feature selection is always important when improving a model; however,

only FIRM-AVP [6] and [9] have incorporated these techniques. Moreover, these methods do not involve deep learning frameworks and produce lower performance. An idea of an online web server could be a useful addition for predicting AVPs and would greatly extend the usage of the model. However, the existing approaches have not been designed to include such web servers.

4 Conclusion and Future Direction

Antiviral peptides (AVPs) are significant in the development of vaccines and antiviral drugs because they are less expensive, more tolerable, have lower toxicity, and are more selective than most conventional drugs. Currently, there are many computational predictors suggested to enhance the detection of AVPs. Among them, some of the methods demonstrated the best performance on training dataset-1 and validation dataset-2, but were not tested on the opposite arrangement, i.e training on dataset-2 and testing on dataset-1. Conversely, the Meta-I AFP predictor had the best results in all four datasets, which shows it is more robust. Besides, a number of the current studies applied only a part of the existing datasets and did not conduct extensive training and validation on all of these datasets. Thus, it is possible to say that the Meta-I AFP is a promising predictor to distinguish between AVPs and non-AVPs.

Even with these developments, AVP identification continues to be a difficult issue in bioinformatics and drug design, and high-predictive precision is an issue of importance. Achieving better prediction performance with the integration of higher-level feature representations, e.g., biological subwords, fastText embeddings, transformer-based bidirectional encoder representations (BERT), has proven useful in prediction.

Author Contributions

Syed Atir Raza Shirazi: Conceptualization, Literature Review, Writing—Original Draft Preparation. **Sawera Kanwal**: Methodology Development, Supervision, Technical Validation, Data Curation. **Junaid Asghar**: Formal Analysis, Visualization, Writing—Review and Editing. **Hamza Naveed**: Project Administration, Quality Assurance of Content. **Sarah khaleel**: Data Collection, Resource Management, Verification of Sources.

Compliance with Ethical Standards

It is declared that all authors don't have any conflict of interest. It is also declare that this article does not contain any studies with human participants or animals performed by any of the authors. Furthermore, informed consent was obtained from all individual participants included in the study.

References

[1] M. Nawaz, H. Yu, F. Akhtar, T. Ma, and H. Zheng, "Deep learning in the discovery of antiviral peptides and peptidomimetics: Databases and prediction tools," *Mol. Diversity*, pp. 1–36, 2025.

[2] N. Lefin, L. Herrera-Belén, J. G. Farias, and J. F. Beltrán, "Review and perspective on bioinformatics tools using machine learning and deep learning for predicting antiviral peptides," *Mol. Diversity*, vol. 28, no. 4, pp. 2365–2374, 2024.

[3] M. M. A. Shirazi, S. Haghighat, Z. Nikbakht, E. Salimkia, and A. Kiumarsy, "Next-generation antiviral peptides: AI-driven design, translational delivery platforms, and future therapeutic directions," *Virus Res.*, p. 199642, 2025.

[4] N. Periwal *et al.*, "Antiprotozoal peptide prediction using machine learning with effective feature selection techniques," *Heliyon*, vol. 10, no. 16, 2024.

[5] J. Yan, J. Cai, B. Zhang, Y. Wang, D. F. Wong, and S. W. Siu, "Recent progress in the discovery and design of antimicrobial peptides using traditional machine learning and deep learning," *Antibiotics*, vol. 11, no. 10, p. 1451, 2022.

[6] M. Daniyal, M. Qureshi, R. R. Marzo, M. Aljuaid, and D. Shahid, "Exploring clinical specialists' perspectives on the future role of AI: Evaluating replacement perceptions, benefits, and drawbacks," *BMC Health Serv. Res.*, vol. 24, no. 1, p. 587, 2024. doi: 10.1186/s12913-024-10848-7.

[7] T. Talaei Khoei and N. Kaabouch, "Machine learning: Models, challenges, and research directions," *Future Internet*, vol. 15, no. 10, p. 332, 2023.

[8] G. Varoquaux and O. Colliot, "Evaluating machine learning models and their diagnostic value," in *Machine Learning for Brain Disorders*, pp. 601–630, 2023.

[9] S. Tufail, H. Riggs, M. Tariq, and A. I. Sarwat, "Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms," *Electronics*, vol. 12, no. 8, p. 1789, 2023.

[10] S. Surana *et al.*, "PandoraGAN: Generating antiviral peptides using generative adversarial networks," *SN Comput. Sci.*, vol. 4, no. 5, p. 607, 2023.

[11] F. Ali *et al.*, "DBP-DeepCNN: Prediction of DNA-binding proteins using wavelet-based denoising and deep learning," *Chemom. Intell. Lab. Syst.*, vol. 229, p. 104639, 2022.

[12] D. Pandey, K. Niwaria, and B. Chourasia, "Machine learning algorithms: A review," *Mach. Learn.*, vol. 6, no. 2, 2019.

[13] J. F. B. Lissabet, L. H. Belén, and J. G. Farias, "AntiVPP 1.0: A portable tool for prediction of antiviral peptides," *Comput. Biol. Med.*, vol. 107, pp. 127–130, 2019.

[14] A. S. Chowdhury *et al.*, "Better understanding and prediction of antiviral peptides through primary and secondary structure feature importance," *Sci. Rep.*, vol. 10, no. 1, p. 19260, 2020.

[15] T.-T. Lin *et al.*, "AI4AVP: An antiviral peptide predictor using deep learning with generative adversarial network data augmentation," *Bioinform. Adv.*, vol. 2, no. 1, p. vbac080, 2022.

[16] M. Qureshi, H. Iftikhar, and M. Daniyal, "Clinical application of machine learning models for early-stage chronic kidney disease detection," *Diagnostics, 15(20), 2610.*, 2025.

[17] S. Gupta *et al.*, "IL17eScan: A tool for the identification of peptides inducing IL-17 response," *Front. Immunol.*, vol. 8, p. 1430, 2017.

[18] Z. Chen *et al.*, "iFeature: A Python package and web server for feature extraction and selection from protein and peptide sequences," *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, 2018.

[19] Z. N. K. Swati *et al.*, "Content-based brain tumor retrieval for MR images using transfer learning," *IEEE Access*, vol. 7, pp. 17809–17822, 2019.

[20] F. Ali *et al.*, "DBPPred-PDSD: A machine learning approach for prediction of DNA-binding proteins using discrete wavelet transform," *Chemom. Intell. Lab. Syst.*, vol. 182, pp. 21–30, 2018.

[21] Y. Xu *et al.*, "iSNO-PseAAC: Predicting cysteine S-nitrosylation sites in proteins," *PLoS One*, vol. 8, no. 2, p. e55844, 2013.

[22] A. N. Sarangi, M. Lohani, and R. Aggarwal, "Prediction of essential proteins in prokaryotes using Chou's pseudo amino acid composition," *Protein Pept. Lett.*, vol. 20, no. 7, pp. 781–795, 2013.

[23] S. Ahmed *et al.*, "Improving secretory protein prediction in Mycobacterium tuberculosis using unbiased dipeptide composition with SVM," *Int. J. Data Min. Bioinform.*, vol. 21, no. 3, pp. 212–229, 2018.

[24] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[25] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins*, vol. 43, no. 3, pp. 246–255, 2001.

[26] M. Arif *et al.*, "TargetCPP: Accurate prediction of cell-penetrating peptides using optimized multiscale features and gradient boosting," *J. Comput.-Aided Mol. Des.*, vol. 34, pp. 841–856, 2020.

[27] Y.-L. Chen and Q.-Z. Li, "Prediction of apoptosis protein subcellular localization using hybrid pseudo-amino acid composition," *J. Theor. Biol.*, vol. 248, no. 2, pp. 377–381, 2007.

[28] X.-Y. Sun *et al.*, "Identifying protein quaternary structural attributes using Chou's PseAAC and discrete wavelet transform," *Mol. BioSyst.*, vol. 8, no. 12, pp. 3178–3184, 2012.

[29] M. Qureshi, K. Ishaq, M. Daniyal, H. Iftikhar, M. Z. Rehman, *et al.*, "Forecasting cardiovascular disease mortality using artificial neural networks in Sindh, Pakistan," *BMC Public Health*, vol. 25, no. 1, p. 34, 2025. doi: 10.1186/s12889-024-19505-1.

[30] G. Varoquaux and O. Colliot, "Evaluating machine learning models and their diagnostic value," *Machine Learning for Brain Disorders*, pp. 601–630, 2023.

[31] S. Tufail, H. Riggs, M. Tariq, and A. I. Sarwat, "Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms," *Electronics*, vol. 12, no. 8, p. 1789, 2023.

[32] G. Varoquaux and O. Colliot, "Evaluating machine learning models and their diagnostic value," *Machine Learning for Brain Disorders*, pp. 601–630, 2023.

[33] J. M. Kernbach and V. E. Staartjes, "Foundations of machine learning-based clinical prediction modeling: Part II—Generalization and overfitting," in *Machine Learning in Clinical Neuroscience: Foundations and Applications*, pp. 15–21, 2021.

[34] A. Khan, M. Qureshi, M. Daniyal, and K. Tawiah, "A novel study on machine learning algorithm-based cardiovascular disease prediction," *Health Soc. Care Community*, vol. 31, no. 1, p. 1406060, 2023. doi: 10.1111/hsc.1406060.

[35] I. Hussain, M. Qureshi, M. Ismail, H. Iftikhar, J. Zywiołek, and J. L. López-Gonzales, "Optimal feature selection in high-dimensional data based on robust techniques: Application to different health databases," *Heliyon*, vol. 10, no. 17, p. e29564, 2024. doi: 10.1016/j.heliyon.2024.e29564.

[36] A. W. K. Yeung, S. More, J. Wu, and S. B. Eickhoff, "Reporting details of neuroimaging studies on individual traits prediction: A literature survey," *NeuroImage*, vol. 256, p. 119275, 2022.

[37] R. Jiang, C. W. Woo, S. Qi, J. Wu, and J. Sui, "Interpreting brain biomarkers: Challenges and solutions in interpreting machine learning-based predictive neuroimaging," *IEEE Signal Process. Mag.*, vol. 39, no. 4, pp. 107–118, 2022.

[38] R. Wang, P. Chaudhari, and C. Davatzikos, "Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 120, no. 6, p. e2211613120, 2023.

[39] O. Benkarim *et al.*, "Population heterogeneity in clinical cohorts affects the predictive accuracy of brain imaging," *PLoS Biol.*, vol. 20, no. 4, p. e3001627, 2022.

[40] D. Chyzhyk, G. Varoquaux, M. Milham, and B. Thirion, "How to remove or control confounds in predictive models, with applications to brain biomarkers," *GigaScience*, vol. 11, p. giac014, 2022.