Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

StackEnPred Framework for Enhancing Antimicrobial Peptides Prediction with Sequence-Based Features and Ensemble Machine Learning

Deepthi K S[1], Dr. Ann Baby[2]
[1]*Research Scholar, Department of Computer Science, Rajagiri College of Social Sciences (Autonomous), Kalamassery, Kochi, Kerala*
phdcs2301@rajagiri.edu
[2]*Assistant Professor, Department of Computer Science, Rajagiri College of Social Sciences (Autonomous), Kalamassery, Kochi, Kerala*
ann@rajagiri.edu

ARTICLEINFO

Received: 26 Dec 2024

Revised: 14 Feb 2025

Accepted: 22 Feb 2025

ABSTRACT

Antimicrobial peptides (AMPs) are typically short length peptides that are important for many biological processes and exhibit various functions against different types of organisms. Antibiotics have been used as a cornerstone, effectively against bacterial infections. As such the overuse of antibiotics against the pathogens made them drive evolution and dissemination of microbial resistance mechanisms. This necessitates the innovative strategies to speed up the discovery of Antimicrobial Peptides (AMPs) that act as a promising candidate to traditional antibiotics. Experimental identification of AMPs is costly and time consuming. Machine learning based computational algorithms can be employed to identify the AMP sequences to expedite the discovery of AMPs. This research introduces StackEnPred, a stacked ensemble learning framework that combines sequence-based feature encoding techniques Amino Acid Composition (AAC) and Dipeptide Composition (DPC) to predict AMPs. The model is trained on Deep-AmPEP30 dataset consisting of 1,777 sequences after the preprocessing techniques are provided as input to the proposed StackEnPred model. StackEnPred, consists of two layers. The base learner layer combines Stochastic Gradient Descent (SGD), K-Nearest Neighbors (KNN), Random Forest (RF) and Support Vector Machine (SVM). The meta learner layer consists of MultiLayer Perceptron (MLP), capable of capturing nonlinear interactions for final classification. StackEnPred achieves an accuracy of 83%, AUC-ROC of 0.89, and Matthews Correlation Coefficient (MCC) of 0.6484, outperforming standalone models (SVM: 82% accuracy; RF: 81%) and deep learning architectures (CNN: 79%).

Keywords: Antimicrobial Peptides, AAC, DPC, SVM, RF, KNN.

INTRODUCTION

Humans prioritize their health by taking steps to prevent infectious diseases. Antimicrobial resistance has become a global threat and a potential pandemic (Naghavi et al., 2024; Yan et al., 2020). The excessive use of antibiotics and other heavy-dose medicines have increased in recent years, leading to a decrease in antimicrobial resistance against various infectious diseases in humans. Without effective measures, fatalities are expected to reach 10 million by 2050 (Naghavi et al., 2024). As a solution to handle this issue, researchers have been paved their research into the area of Antimicrobial Peptides (AMPs).

AMPs, also referred to as Host Defense Peptides (HDPs), are generally short length protein sequences that are present in mammals, plants, insects, and microbes (Mookherjee et al., 2020; Yadav & Chauhan, 2024). Natural AMPs have essential role in the human immune system by exhibiting broad-spectrum of antimicrobial properties. AMP's potent activity against microbes, fungi, viruses, and even cancer cells has drawn the attention a lot to design novel sequences (Bahar & Ren, 2013; Zanetti, 2003). Because of their fast action and minimal potential to cause resistance, AMPs are good substitutes for conventional antibiotics. Additionally, AMPs can regulate immunological responses, facilitate wound healing, and serve as carriers for therapeutic substances (Patrulea et al., 2020). Due to their numerous applications , AMPs are ideal candidates for formulation of medicines (H. Zhang et al., 2025)

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Machine learning models reduces the time spend in wet laboratory processes such as screening of peptide sequences, synthesizing peptide library and evaluating their efficiency. Identification through computational methods also entails some challenges. Challenges being underlined include the limited availability of validated dataset, complexity in feature selection and representation, the lack of generalizability and interpretability of prediction models. Major barrier faced in the computational discovery is scarcity and imbalance of labelled dataset, which results in overfitting and biased prediction results.

Ensemble approach is a machine learning technique that efficiently combines several algorithms, often called as base learners to enhance the prediction performance. The core concept of an ensemble learning model is that the combined effect of several models produces more accurate and stable prediction results compared to the individual prediction of the machine learning algorithms. Bagging, boosting and stacking are the common ensemble learning approach techniques. Bagging technique is efficient in variance reduction and prevention of overfitting. Boosting technique reduces the biases in the model by building models one after the other. Ensemble based approaches are high efficient in managing multidimensional data and noise as such they are commonly used in the antimicrobial peptide predictions (Lv et al., 2022; Lertampaiporn, Vorapreeda, Hongsthong, & Thammarongtham, 2021; Caprani, Healy, Slattery, & O'Keeffe, 2021; Ahmad, Akbar, Tahir, Hayat, & Ali, 2022).

Stacking, or stacked generalization (S, 2025; Kanwal, Arif, Ahmed, & Kabir, 2024b), is a powerful machine learning methodology that enhances the efficacy and robustness in AMP prediction tasks. This approach utilizes the strength of each base learner involved, which compensate for their individual weaknesses. The meta-learner involved increases the generalizability of the model by combining the base learning models involved in a systematic way. The key concept of the stacked ensemble learning algorithm is to predictions of different machine learning algorithms from the preceding layer as input features for the subsequent layer (Pavlyshenko, 2018). One of the most significant benefits of using stacking algorithm is that it improves both generalizability and performance of the model (Lu et al., 2023). Traditional machine deep learning algorithmic models have been utilized for AMP prediction; however, ensemble-based approaches remain limited.

SCOPE AND OBJECTIVES

The research scope is to efficiently identify and evaluate the prediction accuracy of AMPs using sequence based generated features. This objective is attained by extracting dipeptide composition (DPC) and amino acid composition (AAC) from experimentally validated datasets of sequences and then modeling a stack based ensemble learning framework. Enhancement of the accuracy of the prediction models is the main focus of this research. The objectives include:

-Designing of a stacking ensemble architecture that integrates both the dipeptide composition and amino acid composition based feature encoding techniques in an efficient and effective manner.

- Assessing and comparing the evaluation metrics of the prediction algorithmic models using the combined and individual effect of AAC and DPC feature extraction techniques.

- Highlighting the better performance and efficiency of the proposed model.

RELATED WORKS

In the real world problems, Machine learning applications of AMPs poses great importance in the field of therapeutics. Computational techniques developed for the identification of AMPs substantially reduces the effort, time and cost for the experimental discovery of AMPs (Lande et al., 2007). Several computational algorithms have been modeled for the design and identification of AMPs such as AVPpred (Thakur et al., 2012;), BIPEP (Atanaki et al., 2020), AmPEP (Bhadra et al., 2018), ClassAMP (Joseph et al., 2012), DBAASP (Vishnepolsky et al., 2018). in (Zarayeneh & Hanifeloo, 2020), an ensemble algorithm was proposed for predicting antimicrobial peptides using selected features from physiochemical, evolutionary, and structural features. Conventional machine learning algorithmic structures are used to train the algorithm, and prediction is performed using an ensemble model. A multi-tiered stacked ensemble algorithm uses amino acid composition (AAC) and dipeptide composition (DPC) individually for prediction (Suha & Khan, 2024). The stacked ensemble classifier, StackAMP (Karim et al., 2024), achieved high accuracy in prediction using five distinct sequence based feature extraction methods, namely amino

acid composition, dipeptide composition, Moran autocorrelation, Geary autocorrelation, and pseudo amino acid composition. TriStack ensemble model (Han et al., 2024) accurately identified both antimicrobial peptides (AMPs) and anti-inflammatory peptides (AIPs) using a multilayer residual network. Another stack based ensemble learning framework for Antimicrobial Peptides (STAMP) prediction model (Kanwal et al., 2024) was developed, which accurately identified using 84 baseline models, 12 feature encodings, and 7 machine learning algorithms. StackDPPred, a stacked ensemble learning prediction algorithm (Arif et al., 2024) used optimized features for the prediction of properties of defensin peptides (DPs). Another stacked ensemble model predictor, StaBle-ABPpred (Singh et al., 2021) was proposed for the classification of antibacterial peptides. The model used deep learning techniques biLSTM with attention mechanism as the base and ensemble of gradient boosting, logistic regression and random forest at the meta-level.

Feature extraction plays a critical step in the machine learning prediction framework. Yan et al. categorizes encodings into peptide-level features and amino acid-level features. Peptide-level features are again categorized into sequence and structure based features (Yan et al., 2022). Sequence based features represent features based on the amino acid composition or the amino acid groups. Common sequence based features include one hot encoding (Li, Pu, Tang, Zou, & Guo, 2020), general and pseudo amino acid composition (Chou, 2001), reduced amino acid composition (Weathers, Paulaitis, Woolf, & Hoh, 2004) etc. Structural based features include the structural features of the amino acid residues (Chang, Lin, Shih, & Wang, 2015 ;Y. Wang et al., 2012; Sander et al., 2007). Peptide level features are again divided into word and contextual embedding (Veltri, Kamath, & Shehu, 2018). Peptide level feature considers the actual amino acid sequence similar to a word in a sentence (Vaswani et al., 2017; Dallago et al., 2021; Y. Zhang, Lin, Zhao, Zeng, & Liu, 2021). Amino acid composition and dipeptide composition are the most common feature encoding techniques that are identified in literature for effective prediction of AMPs. In research, AMPs are predicted using the combined effect of compositional based features, structural based, physiochemical properties and pseudo amino acid composition techniques. Support Vector Machine (SVM) algorithm is employed to identify the AMP (Meher et al., 2017). This approach identifies antifungal peptides using residue composition, terminal residue and binary profile. SVM algorithm with the compositional based feature encoding technique achieved better performance (Agrawal et al., 2018). This model identifies anticancer peptides based on sequence based features and physiochemical properties. SVM along with RF algorithm models are used in this prediction (Manavalan et al., 2017). This model uses SVM algorithmic model and AAC and binary profile as the feature vectors for the identification of anticancer peptides (Tyagi et al., 2013). In another AMP prediction model, sequence data is efficiently represented using weighted K-nearest neighbor algorithm and predicted using Logistic Regression (Wang et al., 2017).

METHODOLOGY

The steps followed in the methodology of this research are depicted in Figure 1. The first step involves the dataset preparation, then data preprocessing is performed on the dataset, after the preprocessing stage essential features are extracted using feature extraction techniques. In this research sequence based features are used namely amino acid composition (AAC) and dipeptide composition (DPC). Then the model is developed using the stacking ensemble approach and finally the performance is evaluated based on the performance metrics such as accuracy, precision, f1 score and MCC. Also to highlight the robustness of the model, the proposed model is compared with ten other machine learning algorithms. The proposed model is compared against the neural network algorithms such as Deep Neural Networks (DNN), Convolutional Neural Network (CNN) and MultiLayer Perceptron(MLP), traditional machine learning algorithms such as Decision Tree (DT), Logistic Regression (LR), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), K Nearest Neighbor (KNN) and Random Forest (RF), an ensemble model obtained by the combination of SGD, SVM, KNN and RF. The comparison of the proposed model with the other models is done in two concepts. Firstly, the combination of both features AAC and DPC are calculated and compared. In the second concept, either of the feature encoding techniques (either AAC or DPC) are considered and compared.

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

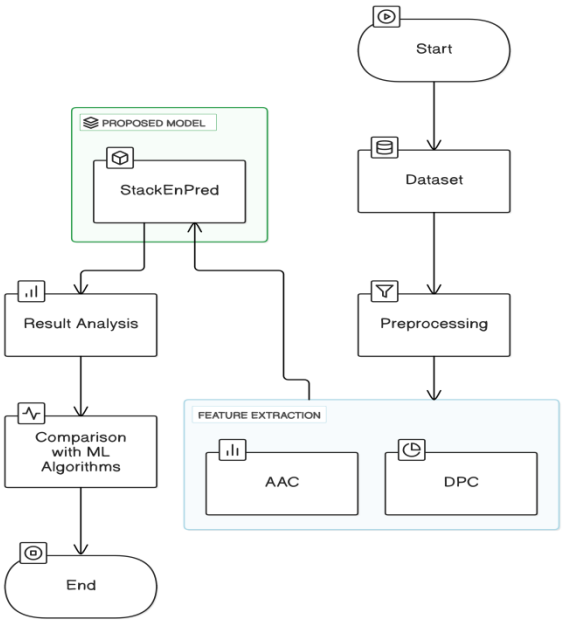

Figure 1. Illustrates the Methodological Flow

Dataset

Deep-AmPEP30 is an openly accessible dataset, which comprises short length (30 amino acids) antimicrobial peptide sequences. This dataset contains an equal number of positive and negative sequences (1529 positive AMPs and 1529 negative sequences). Short length antimicrobial peptide sequences exhibit greater stability, less toxicity and more power in killing microbes (Yan et al., 2020).

Preprocessing

Preprocessing step increases the applicability of the model under diverse conditions. The Deep-AmPEP dataset was preprocessed in order to remove similar and redundant sequences. Levenshtein distance method was applied to identify the similar peptides (Berger et al., 2020). Sequences with similarity measure more than 70% were excluded from the dataset to reduce the redundancy factor. After the exclusion, 1777 peptide sequences were obtained (837 positive AMPs and 940 negative AMPs). Dataset is then randomly partitioned into training and testing data in 80:20 ratios.

Feature Generation

Feature generation is the next important step. AAC and DPC are two important sequence based features extracted from the AMP sequences in this research. AAC represents the frequency count of each standard amino acids. AAC provides 20 features. DPC takes the frequency of the adjacent amino acids. It provides 400 features (Zulfiqar et al., 2023).

Proposed StackEnPred Model

The proposed model is a stacking based ensemble, StackEnPred framework. StackEnPred consist, at the base level, four distinct machine learning algorithms and at the meta level, deep learning based neural network. Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), K Nearest Neighbor (KNN), and Random Forest (RF) (Suha & Khan, 2024) machine learning classifiers are used at the base learner layer. These algorithms are known for their computational efficiency and are effective in handling high dimensional data to learn complex or hidden patterns (Potdukhe, 2025; Sharma et al., 2021; Gull et al., 2019; Tripathi & Tripathi, 2019; Xu et al., 2021). Applying an ensemble approach to these powerful machine learning algorithms reduces the limitations of individual learning models and enhances the overall predictive capability.

At the meta-level, a deep neural network algorithm based Multilayer Perceptron (MLP) model is employed. This model recognizes complex relationships among the outputs of the base learners, and improves the accuracy and

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

robustness of the final predictions. MLP is tuned with 400 hidden layers, 300 iterations, RELU activation and adam solver.

Model effectiveness is analysed using the metrics precision, recall, f1 score, and accuracy. MCC values are calculated inorder to get a balanced assessment of the model's efficacy. The architecture of the StackEnPred model is shown in Figure 2.

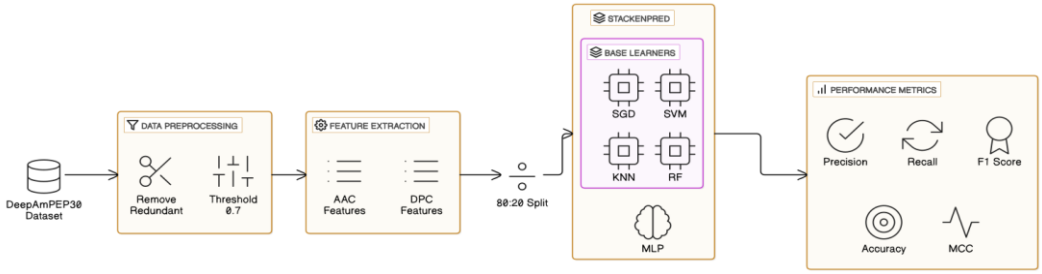

Figure 2. Demonstration of the StackEnPred framework

The efficacy of this machine learning model is measured using six different metrics, such as accuracy (ACC), precision, F1 score, recall, and Mathew's correlation coefficient (MCC). AUROC (Area Under the Receiver Operating Characteristic curve) is an important evaluation metric in machine learning algorithms. High AUC values indicate the best performance.

$$ACC = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FN} \quad (2)$$

$$Recall = \frac{TN}{FP + TN} \quad (3)$$

$$MCC = \frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (FP + TN) \times (TN + FN)}} \quad (4)$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

TP and TN denote the true positives and true negatives. Similarly, FP and FN denote the false positives and false negatives. Precision and Recall denote the model's correctly predicted true positives and true negatives. ACC denotes how correctly the model can predict true positives and true negatives. MCC represents a correlation between the actual and predicted value of the model. It takes a value between -1 and 1. If the MCC value is near 1, then the model's accuracy has better performance. F1 score represents a balance between precision and recall.

The proposed model, StackEnPred, achieved 83% accuracy, 84% precision value, 77% recall and 80% F1 score. Moreover, the AUC-ROC curve for the stacked ensemble model is shown in Figure 4. The AUC value obtained is 0.89, which is of better model performance and reveals that the model can effectively identify both positive and negative antimicrobial sequences. Figure 4 plots the AUC curve of the proposed StackEnPred model.

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

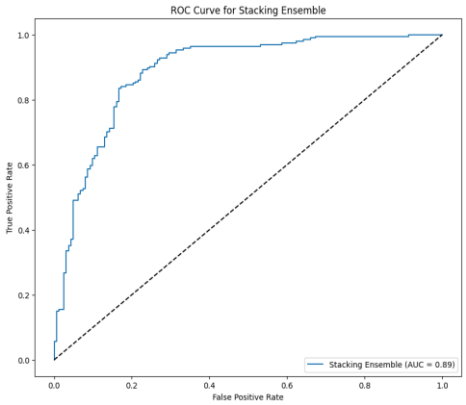

Figure 4. AUC curve of StackEnPred Model

Performance Comparison of StackEnPred with Baseline Machine Learning Models Using AAC and DPC Feature Encodings

The ten various machine learning models such as Logistic Regression (LR) (Uddin et al., 2019), Decision Tree (DT) (A. Karim et al., 2019), Stochastic Gradient Descent (SGD) (Newton et al., 2018), Support Vector Machine (SVM) (Zulfiqar et al., 2022), Random Forest (RF) (Liu & Zhao, 2017), K-Nearest Neighbor (KNN) (Zhang et al., 2017), and deep neural network algorithms such as deep neural network (DNN) (Yi et al., 2016), convolutional neural networks (CNN) (Niu et al., 2021; (Y. Zhang et al., 2020; Bukhari et al., 2020; Patel, 2025, (Peram, 2025) and Multi-layer Perceptron (MLP) (Popescu et al., 2009) and an ensemble approach (ensembled using SGD, SVM, RF and KNN models) are trained individually, and the performances are compared based on the accuracy, precision, F1 score and recall.

Heat map of the proposed model (when considering both AAC and DPC features) against various machine learning algorithms is depicted below in table 1. From the comparison results, the StackEnPred Model is identified as the top-performing model, showing more effective balance in Precision, Recall, F1 Score, Accuracy, and MCC across all comparisons. The proposed model shows highest Precision at 84, surpassing Support Vector Machine (SVM) at 83 and Random Forest (RF) at 81. In terms of Recall, the Ensemble model (SGD, SVM, KNN, RF) leads with 82, followed by Stochastic Gradient Descent (SGD) at 78, and SVM, K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP) attain the value 77. Notably, the StackEnPred Model maintains a Recall of 77 as well. The Model attains the highest F1 score value of 80, closely followed by SVM and the Ensemble model at 79. When considering Accuracy, the StackEnPred Model again takes the highest value with 83, followed by SVM at 82 and RF at 81. The Matthews Correlation Coefficient (MCC) is highest for the Stacked Ensemble Model at 0.6484, with SVM at 0.6368 and RF at 0.6082.

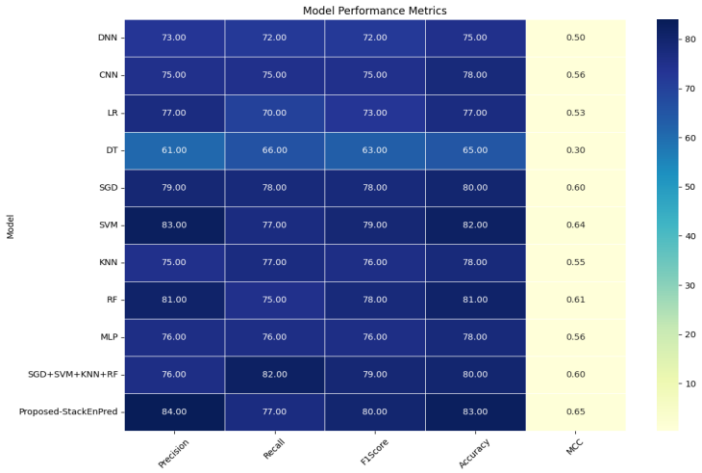

Table 1. Represents the Heatmap to assess the model performances.

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Performance Comparison of StackEnPred using AAC Feature Encoding only

The figure 5 below helps to analyze the performance values of accuracy, precision, recall, F1 score and MCC among other machine learning models. The proposed model, StackEnPred shows high accuracy when compared with other machine learning models when considering only the amino acid composition feature encoding technique only. The model achieved 82% accuracy higher than all the models compared (DNN, CNN, LR, DT, SGD, SVM, KNN, RF, Ensemble model (SGD, SVM, KNN, RF)) with MCC value 0.64. The high accuracy value suggests that the model is a robust and well balanced predictive model. The models CNN, SVM and RF attained 81% of accuracy with MCC values 0.63, 0.63 and 0.51 respectively. The DT model shows a poor accuracy value of 69% among other models. The high precision value demonstrates the capability of the model to predict the correct values. The proposed StackEnPred model achieved a high precision value of 84%, followed by RF and SVM models with 83%. The Decision Tree (DT) model shows the lowest precision value. The ensemble model (SGD, SVM, KNN, RF) demonstrates the highest recall value of 81%, which indicates the model's ability to capture true values with less false values. DT exhibits the lowest recall value of 63%. The StackEnPred model exhibits the highest F1 score value 79%. Lowest F1 score of 65% is shown by the DT model. Conversely, the Decision Tree (DT) consistently exhibits the lowest scores, signaling potential limitations in its capacity to generalize from the training data, possibly due to overfitting or instability with high-dimensional data.

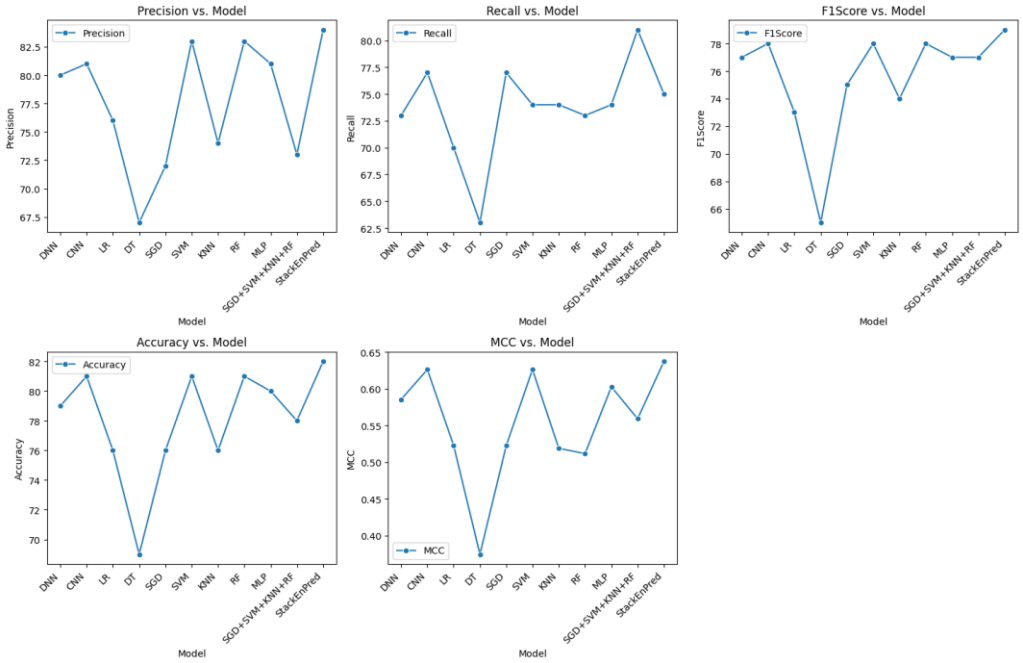

Figure 5. Performance metrics for AAC feature

Performance Comparison of StackEnPred using DPC Feature Encoding only

Figure 6 below shows the comparative analysis of the proposed model, StackEnPred (when considering DPC features only) against ten other machine learning models. This results shows that the proposed model, StackEnPred, the ensemble model (combination of SGD, SVM, RF and KNN models), RF and SVM models achieved the highest accuracy of 80%. All the four models show the same MCC value of 0.60. This value highlights the model's robustness in handling the imbalanced datasets. The StackEnPred, stacked ensemble learning model consistently achieved highest performance values signifying a robust and efficient predictive modeling. Decision Tree (DT) algorithm depicts lowest accuracy and precision value due to its lack of capability in handling complex data. Using DPC feature encoding, Logistic Regression (LR) achieved a high precision value of 83%, while 80% of precision value was achieved by the StackEnPred model. KNN model achieved the highest recall value of 79%, closely followed by Both RF and SVM models (78%). Logistic Regression (LR) model shows the lowest recall value of 56%. CNN and the Ensemble Model (SGD, SVM, KNN, RF) demonstrate the highest F1-scores, both achieving

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

78%, showing a good balance between precision and recall. Decision Tree has the lowest F1-score (61%), reflecting its poorer performance in both precision and recall.

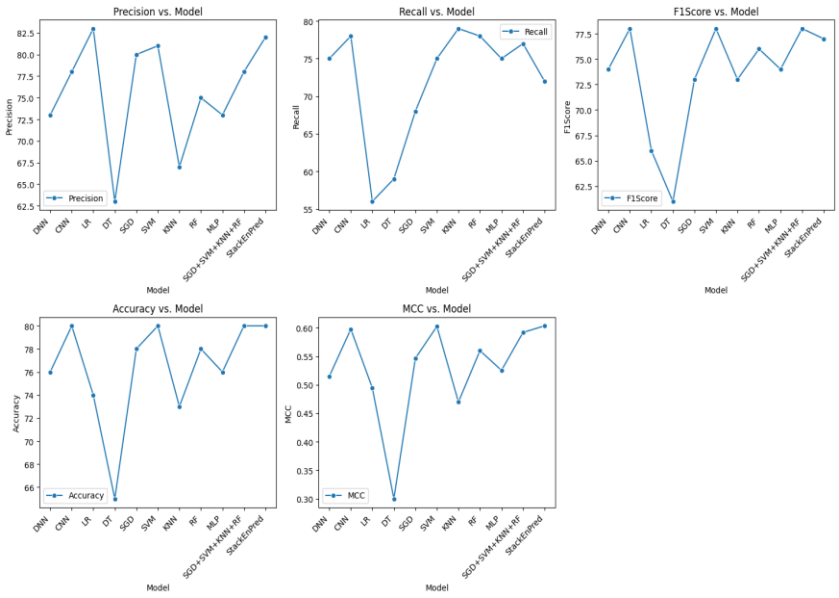

Figure 6. Performance metrics for AAC feature

RESULT ANALYSIS AND DISCUSSION

CNN and SVM demonstrate competitive performance in feature-specific comparisons, especially when utilising only AAC or DPC features. The Decision Tree model always shows suboptimal performance, highlighting issues related to overfitting and generalisation. If Recall is emphasized, the Ensemble model (SGD, SVM, KNN, RF) may serve as an appropriate alternative, though it may entail minor reductions in Precision and MCC. The proposed StackEnPred model shows the highest effectiveness, providing enhanced generalisation and robustness across various feature sets.

Accuracy of the StackEnPred model when using the combined effect of AAC and DPC feature encoding techniques is shown in figure 7. The bar graph shows the accuracies achieved by the proposed model StackEnPred while using the AAC and DPC individually and combined. When considering separately, AAC and DPC, the accuracy attained by the model is 82% and 80% respectively. Highest accuracy of 83% is obtained when both the sequence based features are combined.

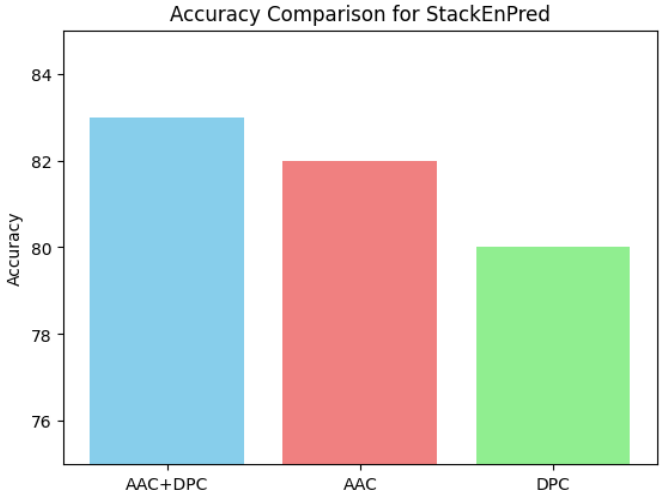

Figure 7. Comparison of accuracies

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

CONCLUSION AND FUTURE WORK

Antimicrobial peptides can be considered as an alternative for antibiotics against antimicrobial resistance. The discovery of antimicrobial peptides has gotten great attention around the world. Wet lab discovery of AMPs is time-consuming. Predicting antimicrobial peptides using machine learning reduces the time in drug discovery. A stacked ensemble-based learning model with diverse base learners (SGD, SVM, KNN, RF) and a neural network (MLP) based meta-learner with a combined effect of amino acid and dipeptide composition feature encoding techniques are proposed. The proposed model achieves greater accuracy value than all other leading models in the field. The proposed model, when considering both the AAC and DPC features, achieved 83% of accuracy which is higher than other machine learning algorithms. The model also achieved high accuracies of value 82% and 80% when considering the features AAC and DPC separately. In any of the cases the model showed high accuracy.

In the future, antimicrobial peptide (AMP) classification can be achieved through multiple enhancements. Expanding the feature set and incorporating hybrid feature representations will enable more comprehensive pattern recognition, leading to improved predictive accuracy. Current proposed model incorporates only sequence based- AAC and DPC feature encoding techniques. These feature encoding techniques can capture essential information from the sequences, but they may not be fully capable of representing the complex structural properties of the sequences. Structural features, predicted through computational methods, can also be included to capture information about the three-dimensional conformation of the peptide. Features derived from evolutionary information, such as position-specific scoring matrices (PSSMs), can trace unknown patterns and motifs within AMP families. Integrating diverse feature representations can be able to learn complex patterns between sequence, structure and function leading to improved accuracy predictions.

Additionally, as AMPs exhibit a broad spectrum of biological activities, refining the model to predict their multifunctional capabilities will enhance its applicability.

Deep learning architectures can be integrated to extract more complex sequence patterns, further boosting classification performance. CNN and Recurrent Neural Networks (RNN) are two powerful deep learning architectures that are capable of extracting complex patterns and relationships from the sequence data. Combined use of both neural networks can be used to enhance the strength of both approaches.

Attention mechanisms can be included in the deep learning architectures to enhance the prediction accuracy and also to highlight the most suitable prediction regions in the sequence for the classification.

Generalizability of the model is another limitation that can be addressed by increasing the diversity and quality of datasets. Generalizability ensures the robustness and reliability of various AMP classifications. Data Augmentation can be applied for the addition of synthetically generated sequences into the dataset as such data imbalance and scarcity problem is tackled.

An emerging approach, Explainable AI (XAI), that is being used to provide insights about the model predictions can be employed to understand key features and patterns that lead to the prediction and activity of AMP. SHapley Additive explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are the two techniques involved to attain XAI.

Inclusion of multi omics data is another research direction to enhance the prediction accuracy. Multi omics data such as proteomics, genomics and transcriptomics provide greater insights into the action and their interactions with the biological systems.

By addressing these challenges, researchers can develop more accurate, robust, and interpretable models for AMP discovery, directing the process for the development of novel antimicrobial agents to combat antimicrobial resistance.

REFRENCES

[1] Agrawal, P., Bhalla, S., Chaudhary, K., Kumar, R., Sharma, M., & Raghava, G. P. S. (2018). In silico approach for prediction of antifungal peptides. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.00323

[2] Ahmad, A., Akbar, S., Tahir, M., Hayat, M., & Ali, F. (2022). iAFPs-EnC-GA: Identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and ensemble learning approach. *Chemometrics and Intelligent Laboratory Systems*, *222*, 104516. https://doi.org/10.1016/j.chemolab.2022.104516

[3] Arif, M., Musleh, S., Ghulam, A., Fida, H., Alqahtani, Y., & Alam, T. (2024). StackDPPred: Multiclass prediction of defensin peptides using stacked ensemble learning with optimized features. *Methods*, *230*, 129–139. https://doi.org/10.1016/j.ymeth.2024.08.001

[4] Atanaki, F. F., Behrouzi, S., Ariaeenejad, S., Boroomand, A., & Kavousi, K. (2020). BIPEP: sequence-based prediction of biofilm inhibitory peptides using a combination of NMR and physicochemical descriptors. *ACS Omega*, *5*(13), 7290–7297. https://doi.org/10.1021/acsomega.9b04119

[5] Bahar, A., & Ren, D. (2013). Antimicrobial peptides. *Pharmaceuticals*, *6*(12), 1543–1575. https://doi.org/10.3390/ph6121543

[6] Berger, B., Waterman, M. S., & Yu, Y. W. (2020). Levenshtein distance, sequence comparison and biological database search. *IEEE Transactions on Information Theory*, *67*(6), 3287–3294. https://doi.org/10.1109/tit.2020.2996543

[7] Bhadra, P., Yan, J., Li, J., Fong, S., & Siu, S. W. I. (2018). AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-19752-w

[8] Bukhari, S. a. S., Razzaq, A., Jabeen, J., Khan, S., & Khan, Z. (2020). Deep-BSC: Predicting raw DNA binding pattern in Arabidopsis thaliana. *Current Bioinformatics*, *16*(3), 457–465. https://doi.org/10.2174/1574893615999200707142852

[9] Caprani, M. C., Healy, J., Slattery, O., & O'Keeffe, J. (2021). Using an ensemble to identify and classify macroalgae antimicrobial peptides. *Interdisciplinary Sciences Computational Life Sciences*, *13*(2), 321–333. https://doi.org/10.1007/s12539-021-00435-6

[10] Chang, K. Y., Lin, T., Shih, L., & Wang, C. (2015). Analysis and prediction of the critical regions of antimicrobial peptides based on conditional random fields. *PLoS ONE*, *10*(3), e0119490. https://doi.org/10.1371/journal.pone.0119490

[11] Chou, K. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Structure Function and Bioinformatics*, *43*(3), 246–255. https://doi.org/10.1002/prot.1035

[12] Dallago, C., Schütze, K., Heinzinger, M., Olenyi, T., Littmann, M., Lu, A. X., . . . Rost, B. (2021). Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets. *Current Protocols*, *1*(5). https://doi.org/10.1002/cpz1.113

[13] Gull, S., Shamim, N., & Minhas, F. (2019). AMAP: Hierarchical multi-label prediction of biologically active and antimicrobial peptides. *Computers in Biology and Medicine*, *107*, 172–181. https://doi.org/10.1016/j.compbiomed.2019.02.018

[14] Han, J., Chen, Q., Su, J., Kong, T., Song, Y., Long, S., & Liu, J. (2024). TriStack enables accurate identification of antimicrobial and anti-inflammatory peptides by combining machine learning and deep learning approaches. *Future Generation Computer Systems*, *161*, 259–268. https://doi.org/10.1016/j.future.2024.07.024

[15] Joseph, S., Karnik, S., Nilawe, P., Jayaraman, V. K., & Idicula-Thomas, S. (2012). ClassAMP: a prediction tool for classification of antimicrobial peptides. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *9*(5), 1535–1538. https://doi.org/10.1109/tcbb.2012.89

[16] Kanwal, S., Arif, R., Ahmed, S., & Kabir, M. (2024). A novel stacking-based predictor for accurate prediction of antimicrobial peptides. *Journal of Biomolecular Structure and Dynamics*, 1–12. https://doi.org/10.1080/07391102.2024.2329298

[17] Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., & Alazab, M. (2019). A comprehensive survey for intelligent spam email detection. *IEEE Access*, *7*, 168261–168295. https://doi.org/10.1109/access.2019.2954791

[18] Karim, T., Shaon, M. S. H., Ali, M. M., Ahmed, K., Bui, F. M., & Chen, L. (2024). STAckAMP: Stacking-Based Ensemble Classifier for Antimicrobial Peptide identification. *IEEE Transactions on Artificial Intelligence*, *5*(11), 5666–5675. https://doi.org/10.1109/tai.2024.3421176

[19] Lande, R., Gregorio, J., Facchinetti, V., Chatterjee, B., Wang, Y., Homey, B., . . . Gilliet, M. (2007). Plasmacytoid dendritic cells sense self-DNA coupled with antimicrobial peptide. *Nature*, *449*(7162), 564–569. https://doi.org/10.1038/nature06116

[20] Lertampaiporn, S., Vorapreeda, T., Hongsthong, A., & Thammarongtham, C. (2021). Ensemble-AMPPred: Robust AMP Prediction and Recognition Using the Ensemble Learning Method with a New Hybrid Feature for Differentiating AMPs. *Genes*, *12*(2), 137. https://doi.org/10.3390/genes12020137

[21] Li, J., Pu, Y., Tang, J., Zou, Q., & Guo, F. (2020). DeepAVP: a Dual-Channel deep neural network for identifying Variable-Length antiviral peptides. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 3012–3019. https://doi.org/10.1109/jbhi.2020.2977091

[22] Liu, Y., & Zhao, H. (2017). Variable importance-weighted Random Forests. *Quantitative Biology*, *5*(4), 338–351. https://doi.org/10.1007/s40484-017-0121-6

[23] Lu, M., Hou, Q., Qin, S., Zhou, L., Hua, D., Wang, X., & Cheng, L. (2023). A stacking ensemble model of various machine learning models for daily runoff forecasting. *Water*, *15*(7), 1265. https://doi.org/10.3390/w15071265

[24] Lv, H., Yan, K., Guo, Y., Zou, Q., Hesham, A. E., & Liu, B. (2022). AMPpred-EL: An effective antimicrobial peptide prediction model based on ensemble learning. *Computers in Biology and Medicine*, *146*, 105577. https://doi.org/10.1016/j.compbiomed.2022.105577

[25] Manavalan, B., Basith, S., Shin, T. H., Choi, S., Kim, M. O., & Lee, G. (2017). MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget*, *8*(44), 77121–77136. https://doi.org/10.18632/oncotarget.20365

[26] Meher, P. K., Sahu, T. K., Saini, V., & Rao, A. R. (2017). Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Scientific Reports*, *7*(1). https://doi.org/10.1038/srep42362

[27] Mookherjee, N., Anderson, M. A., Haagsman, H. P., & Davidson, D. J. (2020). Antimicrobial host defence peptides: functions and clinical potential. *Nature Reviews Drug Discovery*, *19*(5), 311–332. https://doi.org/10.1038/s41573-019-0058-8

[28] Naghavi, M., Vollset, S. E., Ikuta, K. S., Swetschinski, L. R., Gray, A. P., Wool, E. E., . . . Aziz, S. (2024). Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050. *The Lancet*. https://doi.org/10.1016/s0140-6736(24)01867-1

[29] Newton, D., Pasupathy, R., & Yousefian, F. (2018). RECENT TRENDS IN STOCHASTIC GRADIENT DESCENT FOR MACHINE LEARNING AND BIG DATA. *2018 Winter Simulation Conference (WSC)*, 366–380. https://doi.org/10.1109/wsc.2018.8632351

[30] Niu, M., Lin, Y., & Zou, Q. (2021). sgRNACNN: identifying sgRNA on-target activity in four crops using ensembles of convolutional neural networks. *Plant Molecular Biology*, *105*(4–5), 483–495. https://doi.org/10.1007/s11103-020-01102-y

[31] Patel, N. M. (2025). Image based chronic renal disease diagnosis using Convolution Neural Network Deep Learning approach. *Journal of Information Systems Engineering & Management*, *10*(10s), 80–89. https://doi.org/10.52783/jisem.v10i10s.1347

[32] Patrulea, V., Borchard, G., & Jordan, O. (2020). An update on antimicrobial peptides (AMPs) and their delivery strategies for wound infections. *Pharmaceutics*, *12*(9), 840. https://doi.org/10.3390/pharmaceutics12090840

[33] Pavlyshenko, B. (2018). Using Stacking Approaches for Machine Learning Models. *IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, 255–258. https://doi.org/10.1109/dsmp.2018.8478522

[34] Peram, N. M. M. V. N. S. (2025). Computational Approaches for Drug-Protein Interaction Analysis in Cancer: Machine Learning and Structural Bioinformatics Perspectives. *Journal of Information Systems Engineering & Management*, *10*(10s), 231–247. https://doi.org/10.52783/jisem.v10i10s.1368

[35] Popescu, M., Balas, V. E., Perescu-Popescu, L., & Mastorakis, N. (2009). Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems Archive*, *8*(7), 579–588. https://doi.org/10.5555/1639537.1639542

[36] Potdukhe, N. K. (2025). Heart disease prediction and detection using machine learning. *Journal of Information Systems Engineering & Management*, *10*(20s), 373–381. https://doi.org/10.52783/jisem.v10i20s.3130

[37] S, N. M. (2025). Diagnostic Predictive Approaches for Liver Disease Detection using Stacked Ensemble Model with Data Augmentation. *Journal of Information Systems Engineering & Management*, *10*(13s), 750–760. https://doi.org/10.52783/jisem.v10i13s.2157

[38] Sander, O., Sing, T., Sommer, I., Low, A. J., Cheung, P. K., Harrigan, P. R., . . . Domingues, F. S. (2007). Structural Descriptors of gp120 V3 Loop for the Prediction of HIV-1 Coreceptor Usage. *PLoS Computational Biology*, *3*(3), e58. https://doi.org/10.1371/journal.pcbi.0030058

[39] Sharma, R., Shrivastava, S., Singh, S. K., Kumar, A., Saxena, S., & Singh, R. K. (2021). AniAMPpred: artificial intelligence guided discovery of novel antimicrobial peptides in animal kingdom. *Briefings in Bioinformatics*, *22*(6). https://doi.org/10.1093/bib/bbab242

[40] Singh, V., Shrivastava, S., Singh, S. K., Kumar, A., & Saxena, S. (2021). StaBle-ABPpred: a stacked ensemble predictor based on biLSTM and attention mechanism for accelerated discovery of antibacterial peptides. *Briefings in Bioinformatics*, *23*(1). https://doi.org/10.1093/bib/bbab439

[41] Suha, S. A., & Khan, A. H. (2024). Predicting short antimicrobial peptides utilizing multi-tiered stacked ensemble machine learning technique. *2017 IEEE Region 10 Symposium (TENSYMP)*, 1–6. https://doi.org/10.1109/tensymp61132.2024.10751811

[42] Thakur, N., Qureshi, A., & Kumar, M. (2012). AVPpred: collection and prediction of highly effective antiviral peptides. *Nucleic Acids Research*, *40*(W1), W199–W204. https://doi.org/10.1093/nar/gks450

[43] Tripathi, V., & Tripathi, P. (2019). Detecting antimicrobial peptides by exploring the mutual information of their sequences. *Journal of Biomolecular Structure and Dynamics*, *38*(17), 5037–5043. https://doi.org/10.1080/07391102.2019.1695667

[44] Tyagi, A., Kapoor, P., Kumar, R., Chaudhary, K., Gautam, A., & Raghava, G. P. S. (2013). In silico models for designing and discovering novel anticancer peptides. *Scientific Reports*, *3*(1). https://doi.org/10.1038/srep02984

[45] Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1). https://doi.org/10.1186/s12911-019-1004-8

[46] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017). Attention is All you Need. *arXiv (Cornell University)*, *30*, 5998–6008. Retrieved from https://arxiv.org/pdf/1706.03762v5

[47] Veltri, D., Kamath, U., & Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, *34*(16), 2740–2747. https://doi.org/10.1093/bioinformatics/bty179

[48] Vishnepolsky, B., Gabrielian, A., Rosenthal, A., Hurt, D. E., Tartakovsky, M., Managadze, G., . . . Pirtskhalava, M. (2018). Predictive Model of Linear Antimicrobial Peptides Active against Gram-Negative Bacteria. *Journal of Chemical Information and Modeling*, *58*(5), 1141–1151. https://doi.org/10.1021/acs.jcim.8b00118

[49] Wang, P., Ge, R., Liu, L., Xiao, X., Li, Y., & Cai, Y. (2017). Multi-label learning for predicting the activities of antimicrobial peptides. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-01986-9

[50] Wang, Y., Ding, Y., Wen, H., Lin, Y., Hu, Y., Zhang, Y., . . . Lin, Z. (2012). QSAR modeling and design of cationic antimicrobial peptides based on structural properties of amino acids. *Combinatorial Chemistry & High Throughput Screening*, *15*(4), 347–353. https://doi.org/10.2174/138620712799361807

[51] Weathers, E. A., Paulaitis, M. E., Woolf, T. B., & Hoh, J. H. (2004). Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Letters*, *576*(3), 348–352. https://doi.org/10.1016/j.febslet.2004.09.036

[52] Xu, J., Li, F., Leier, A., Xiang, D., Shen, H., Lago, T. T. M., . . . Song, J. (2021). Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides. *Briefings in Bioinformatics*, *22*(5). https://doi.org/10.1093/bib/bbab083

[53] Yadav, N., & Chauhan, V. S. (2024). Advancements in peptide-based antimicrobials: A possible option for emerging drug-resistant infections. *Advances in Colloid and Interface Science*, *333*, 103282. https://doi.org/10.1016/j.cis.2024.103282

[54] Yan, J., Bhadra, P., Li, A., Sethiya, P., Qin, L., Tai, H. K., . . . Siu, S. W. (2020). Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning. *Molecular Therapy — Nucleic Acids*, *20*, 882–894. https://doi.org/10.1016/j.omtn.2020.05.006

[55] Yan, J., Cai, J., Zhang, B., Wang, Y., Wong, D. F., & Siu, S. W. I. (2022). Recent progress in the discovery and design of antimicrobial peptides using traditional machine learning and deep learning. *Antibiotics*, *11*(10), 1451. https://doi.org/10.3390/antibiotics11101451

[56] Yi, N. H., Shiyu, N. S., Xiusheng, N. D., & Zhigang, N. C. (2016). A study on Deep Neural Networks framework. *IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, 1519–1522. https://doi.org/10.1109/imcec.2016.7867471

[57] Zanetti, M. (2003). Cathelicidins, multifunctional peptides of the innate immunity. *Journal of Leukocyte Biology*, *75*(1), 39–48. https://doi.org/10.1189/jlb.0403147

[58] Zarayeneh, N., & Hanifeloo, Z. (2020a). Antimicrobial peptide prediction using ensemble learning algorithm. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2005.01714

[59] Zarayeneh, N., & Hanifeloo, Z. (2020b). Antimicrobial peptide prediction using ensemble learning algorithm. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2005.01714

[60] Zhang, H., Lv, J., Ma, Z., Ma, J., & Chen, J. (2025). Advances in antimicrobial peptides: mechanisms, design innovations, and biomedical potential. *Molecules*, *30*(7), 1529. https://doi.org/10.3390/molecules30071529

[61] Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for kNN Classification. *ACM Transactions on Intelligent Systems and Technology*, *8*(3), 1–19. https://doi.org/10.1145/2990508

[62] Zhang, Y., Lin, J., Zhao, L., Zeng, X., & Liu, X. (2021). A novel antibacterial peptide recognition algorithm based on BERT. *Briefings in Bioinformatics*, *22*(6). https://doi.org/10.1093/bib/bbab200

[63] Zhang, Y., Yan, J., Chen, S., Gong, M., Gao, D., Zhu, M., & Gan, W. (2020). Review of the Applications of Deep Learning in Bioinformatics. *Current Bioinformatics*, *15*(8), 898–911. https://doi.org/10.2174/1574893615999200711165743

[64] Zulfiqar, H., Ahmed, Z., Ma, C., Khan, R. S., Grace-Mercure, B. K., Yu, X., & Zhang, Z. (2022). Comprehensive prediction of lipocalin proteins using artificial intelligence strategy. *Frontiers in Bioscience-Landmark*, *27*(3). https://doi.org/10.31083/j.fbl2703084

[65] Zulfiqar, H., Guo, Z., Grace-Mercure, B. K., Zhang, Z., Gao, H., Lin, H., & Wu, Y. (2023). Empirical comparison and recent advances of computational prediction of hormone binding proteins using machine learning methods. *Computational and Structural Biotechnology Journal*, *21*, 2253–2261. https://doi.org/10.1016/j.csbj.2023.03.024