JOURNAL OF ENGINEERING RESEARCH AND TECHNOLOGY, VOLUME 12, ISSUE 1, MARCH 2025

Received on (04-02-2025) Accepted on (01-07-2025)

Sequence-based Deep Learning Model for AMPs Activity Prediction Against Specific Pathogen Strains

Abdullah Abu Nada[1], and Iyad H. AlShami

1 Faculty of Information Technology, Islamic University of Gaza, Palestine, https://orcid.org/0000-0003-1029-1860

2 Faculty of Information Technology, Islamic University of Gaza, Palestine, eshami@iugaza.edu.ps

https://doi.org/10.33976/JERT.12.1/2025/1

Abstract— The pipeline of antimicrobial peptides (AMPs) discovery and design is costly and time-consuming. So, several machine learning approaches and techniques have been used in the in-silico stage to help discover the new AMPs by indicating their activity before moving them into the in-vitro and in-vivo stages. However, machine learning and statistical methods require a lot of feature engineering and domain experts. Recently, many deep learning approaches have shown their superior performance in several fields, which have inspired researchers to take advantage to reduce the efforts of AMPs discovery and design tasks. However, only some efforts provided a specific model that can determine the active AMPs against specific pathogens strains, which would improve the quality of samples that could pass the laboratory and clinical trials. This paper attempts to propose a sequence-base deep learning model that helps laboratories predict potential active AMPs against specific pathogen strains by preparing up-to-date datasets for activity AMPs based on the latest updated databases and taking the power of combining multiple deep learning architectures to build AMPs activity prediction models. The best AMPs activity prediction models achieved 97%-98% of accuracy and Matthew's correlation coefficient of 0.93 to 0.96.

Index Terms— Antimicrobial Peptide, Deep Learning, AMPs, Activity, CNN, Bi-LSTM, Self-Attention.

I INTRODUCTION

Antimicrobial peptides (AMPs, also known as host defense peptides) are a primary unit in every natural immunity reaction. They are broad-spectrum and potent antibiotics that show potential as novel therapeutic agents, especially short ones [1], [2]. AMPs are short proteins, typically five to one hundred amino acids long, positively charged, and amphipathic, that can inhibit the negatively charged microbial membranes [3].

Based on AMPs as promising novel therapeutic agents, many laboratories have been motivated to discover and design new and novel AMPs. Nevertheless, discovering and designing AMPs in the pharmaceutical industry's research and development (R&D) are costly, time-consuming, exhausting, and require several phases of laboratory and clinical trials [4]. Therefore, an urgent need to develop an in-silico approach (It means that the experiments are performed on the computer) to help discover the potential active peptides.

Due to the cost-efficiency and less-time consumption of the machine learning (ML) approach in recent decades, several machine learning approaches have been widely used in drug discovery, especially in the AMPs identification field [4]-[10]. However, most ML and statistical techniques were used to build general AMP predictive models that distinguish between the AMPs and non-AMPs sequences to help identify the potential candidate's peptides for laboratory and clinical trials [8]. Only some of them proposed specific activity prediction models against specific species [11], to give the activity prediction task accuracy in determining whether the discovered peptide is active against certain strains. In addition, the ML and statistical techniques require a lot of time, effort, and a high level of knowledge in the AMP field to perform many feature selection and extraction experiments to achieve a significant model performance.

More recently, the deep learning (DL) approaches proved their superior performance in many state-of-the-art domains, like object detection, object recognition, object segmentation, speech recognition, natural language procession tasks, drug discovery, and genomics domain [12]. The most common deep learning architectures, such as the convolutional neural networks (CNNs) [13] and recurrent neural networks (RNNs) [14], have been used in APMs prediction [15]-[17]. However, these architectures do not require a lot of feature engineering; it takes the original sequence as input and automatically extracts the hidden features and information through several neural network layers, then predicts the output labels. Many

deep learning approaches have proven their superior performance in the peptide's discovery and design area, like VAEs, Encoder-decoder [18], RNN [19], long-shot-term-memory (LSTM) [20], bi-directional-long-shot-term-memory (Bi-LSTM) [17], and CNN's [21]. However, in other areas, combining two architectures, CNN-RNN, and self-attention mechanism, helped to enhance the accuracy of several classification tasks [22]-[26].

Motivated by the above discussion, this paper proposed a combined DL sequence-based model to predict the potential short AMPs against specific strains of common bacteria that need urgent treatments (E. coli, E. coli ATCC 25922, Pseudomonas aeruginosa, and Pseudomonas aeruginosa ATCC 27853) to help the laboratories in discovering AMPs that have potential to pass the laboratory and clinical trials. Therefore, based on several AMPs databases (DBAASP v.3) [27], DRAMP [28], etc. an appropriate AMPs activity dataset for every target strain was prepared, which consists of AMP sequence structure as text features like "DFKLVRFW" in which each character represents specific amino acid, and the peptide Minimum Inhibitory Concentration (MIC) value as a float number. Moreover, the proposed model takes the advantages of CNN, Bi-LSTM architectures, and self-attention mechanism to improve their performance in discovery the active AMPs against the targeted strains.

Finally, the AMPs activity prediction model versions were evaluated by calculating the curves' plotted rate (ROC), the area under the curve (AUCs), confusion matrices, and Mathew's correlation coefficient (MCC), which AMPs activity prediction models against specific target strains achieved accuracy between 97% to 98% and MCC of 93.13% to 96.06%. As a result, despite the proposed activity prediction model versions proving their performance in discovering the active AMPs against the targeted strains, they are sensitive to the length of the peptides and their amino acid compositions.

The remaining sections of this paper has been ordered as follow: II. Related Works; III. Materials and Methods; IV. Experiments and Results and finally the V. Conclusion section.

II RELATED WORKS

The state-of-the-art works on the computational peptide discovery area were explored and classified into two main categories based on used technology (ML, DL) [4], [7], [8], [29].

A AMPs Predictors based on ML Algorithms

In recent decades, several machine learning approaches have been widely used in drug discovery and design, especially in the AMPs identification field [4]-[8], using several feature selection methods, ML algorisms, and evaluation methodologies. The researchers have used different algorithms like Hidden Markov Model (HMM) [30], artificial neural network (ANN) [31], fuzzy k-nearest neighbor (fuzzy k-NN) [32], logistic regression (LR) [10], [33], support vector machine (SVM) [34], Random Forest (RF) [35], decision tree (DT) [36], Bayesian network (BN) [37], discriminant analysis (DA) [38], and AdaBoost (ADA) [39], which the SVM and RF are the most commonly used [8].

B Discovery AMPs based on Deep Learning Approaches

Veltri and others suggested an improved deep neural network (DNN) model for AMP prediction (AMP Scanner v.2) [20], which they combined between the convolutional layer and the long short-term memory (LSTM) architecture [40] for AMPs. The proposed model consists of an embedding layer responsible for taking the constructed to zero-padded numerical vectors from the peptide sequence and converting them to a fixed-size representation vector. These vectors are passed to the convolutional and LSTM layers to extract their features. Finally, using the fully connected layer and sigmoid function, the probability of the expected label was predicted. However, this work focused on applying DL to distinguish between active and non-active AMPs generally. Also, it focused on physicochemical properties to encode the sequences as numerical vectors where it outperformed the others, similar state-of-the-art models, by achieving 96.48% AUC.

Su and others proposed a new deep learning model for AMP's classification (APIN) [16]. They used the embedding layer to convert numerical fixed-length vectors and a multi-scale convolutional network consisting of several convolutional layers with different filter lengths (max-pooling) to extract the potential hidden features in the peptides sequence. Also, they used DPC and AAC sequence features to improve the model performance. Finally, the fully connected layers and sigmoid function were used to identify the peptide. This work also just focused on AMPs activity prediction generally, where it approves its performance by achieving 97.3% AUC.

Additionally, based on CNN architecture J. Yan et al. [21] introduced another deep learning model for short peptides prediction (Deep-AmPEP30); this model consists of two convolutional layers, max-pooling layers for extracting the implicit features from peptides sequences, fully connected layers with ReLU activation function, and the sigmoid function as output layer. Also, they used AAC and composition-transition-distribution (CTD), pseudo-K-tuple reduced amino acids composition (PseKRAAC) [41], and PseAAC as support features to improve the prediction performance. It is worth mentioning they have been available this model as a public web service. However, this model achieved an AUC of 0.8533 and an accuracy of 77.13% using a benchmark dataset. Finally, though the CCN model accuracy bets the rest model, they found the CNN and RF results were very close together.

Dua and others have investigated several deep neural network architectures for enhancing the AMPs prediction task [19]. The several approaches for Low-level Representation to convert the string sequence to numerical input were explored, like Embedding layer, one-hot-encoding, integer representation, binary representation, and K-mer Count Representation. However, they recommended DNN and DNN-One-Hot ap-

proaches to convert the sequences to input. Also, they explored several DNN architectures like CNN, LSTM, Bi-LSTM, GRU, and simple-RNN. They found that the convolutional layer played an important role in feature extraction to improve the model accuracy, which achieved 93.8% accuracy and 0.972 AUC. In addition, they investigated the attention mechanism in enhancing the model prediction accuracy.

C. Li and others introduced the first deep learning model based using the attention mechanism to identify the AMPs (AMPlify) [17]. Which is based on RNN architecture and consists of bidirectional long short-term memory (Bi-LSTM) layers that enable the model to obtain both forward and backward information about sequence at each time step, and two different types of attention mechanism, context attention (CA), and a multi-head scaled dot-product attention (MHSDPA) to keep the model extract the most important features to help it to be more accurate, which achieved accuracy with 92.79% and AUC with 0.9744.

Jang et al. [42] proposed an attention-based Bi-LSTM+CNN hybrid model for text classification, leveraging CNN for feature extraction, Bi-LSTM for sequential dependencies, and attention for refining relevant features. Their model outperformed standalone CNN, LSTM, and MLP models in sentiment classification tasks, achieving improved precision, recall, and F1 scores.

Wang et al. [43] introduced a regional CNN-LSTM model for sentiment classification, demonstrating that hybrid architectures improve contextual learning and emotional text processing. This study highlighted the effectiveness of integrating CNN and Bi-LSTM for structured sentiment analysis.

Salur et al. [44] further explored hybrid architectures, integrating CNN, Bi-LSTM, and GRU with attention mechanisms for text classification tasks. Their findings confirmed that attention-enhanced models provide better accuracy compared to architectures lacking attention-based refinement.

These studies collectively highlight the superior performance of CNN, Bi-LSTM, and self-attention in classification tasks, reinforcing the rationale for adopting this hybrid combination in AMP activity prediction models.

Finally, based on the above literature reviews, most of the efforts on this area have used machine learning algorithms. In addition, thousands of irrelevant or less important features for peptide sequence negatively impact classification model performance. So, most of the previous proposed efforts competed in using diverse machine learning algorithms and feature selection methods (physicochemical, AAC, etc.) using common published tools to improve the performance of prediction AMP activity and functionality and the distinguish between AMPs and non-AMPs in general, which the feature engineering step needs a lot of time, effort, and a high level of knowledge in the AMP field. Therefore, the most recent research in this field has tended to take advantage of deep learning architectures, especially in feature learning to work forward to improve the performance of these models. However, until now, due to data availability, only some works have focused on preparing DL models able to predict the AMP activity against specific pathogen strains. In this study, we are aim to prepare new AMP activity datasets related to specific pathogen strains. Furthermore, we leverage the combined strengths of CNN, Bi-LSTM, and self-attention mechanisms to design deep learning models capable of accurately identifying active AMPs against the targeted strains [22]-[26], [42]-[44].

III MATERIALS AND METHODS

This paper proposes sequence-based DL models for predicting AMP activity against specific pathogens strains. Figure 1 provides an overview of the entire process, starting from the data collection stage, describing how it is collected from many popular and different AMPs databases, passing through the data pre-processing and processing stages to prepare and adapt the collected data for training the various versions of activity prediction models. Furthermore, it explains the architecture of the proposed model versions and their training process. Also, it shows evaluation methods of AMP's activity prediction models using the curves were plotted (ROC), the area under the curve (AUC), Mathew's correlation coefficient (MCC), and confusion matrix metrics.

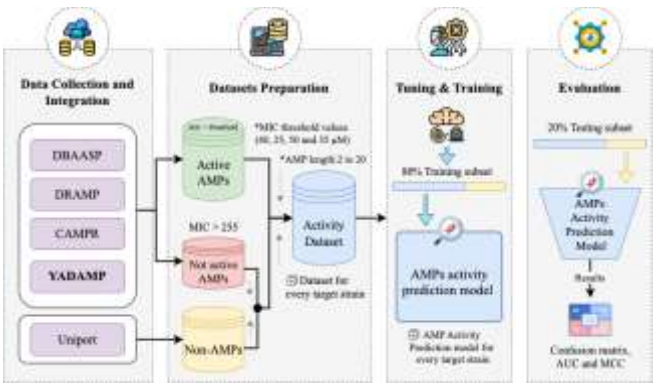

Figure 1 An overview of the four main stages of the proposed work. We began with data collection and integration (step 1), then used this data in preparing the datasets (step 2), tuned and trained the models using these datasets (step 3), and evaluated the models using the testing subsets (step 4).

A Data Collection

The literature of this work shows that many relative researchers have integrated AMPs databases to obtain enough samples to prepare their datasets and train their models using it. So, many of the up-to-date public AMP databases that contain variant information about AMPs DBAASP [27], DRAMP [28], CAMPR [45], Uniport [46], and YADAMP [47] were explored, handled, and integrated. This helped us to collect as many appropriate AMPs as possible to prepare proper datasets for training and evaluating the AMPs activity prediction models and achieve this study's objectives. In this stage, DBAASP (v3) has been chosen as the primary database due to its recency, quality, and data coverage. The others are complementary databases to collect all possible unique samples from each database separately. The sequence, C-

terminus, N-terminus, MIC for the common bacteria, and hemolytic information were collected.

1 DBAASP Database (v3):

The Database of Antimicrobial Activity and Structure of Peptides (DBAASP) has been originally prepared by [27]. The last update of DBAASP, in 2021, contains much information about more than 18,000 peptides like amino acid sequences, toxicity and bioactivity profiles, chemical modification, and 3D structures. It provides two REST endpoints. The first endpoint collects all AMP's primary data, like the peptide name, sequence, C-Terminus, and N-Terminus information. Approximately 15,000 natural AMPs were identified based on their sequence information, which should consist of only natural amino acids. Then using the second endpoint (peptide card), all peptide cards were collected and integrated with their own AMP based on their IDs to create an offline version of this database to give us the ability to extract and explore data more quickly. The peptide cards contain more advanced information about the AMPs, like target group, activities, species, and hemolytic information. More than 12,000 natural AMPs that inhibit the gram-negative bacteria were labeled using the target-Group property. According to the available information, the activity information for the most available samples for gram-negative bacteria strains (Escherichia coli, Escherichia coli ATCC 25922, Pseudomonas aeruginosa, and Pseudomonas aeruginosa ATCC 27853) were considered. Then, using peptide card information, the activity data containing the MIC value and its unit in (μM and μg/ml) for target species from the target-Activities property were extracted. Consequently, the core of our activity dataset is established, containing the AMP's sequences, C-Terminus, N -Terminus, and non-normalized activity information for each target strain. Around 5700 AMPs have activity against the targeted E. coli strains, and 4400 AMPs against the targeted Pseudomonas strains were obtained. Also, the experimentally validated AMPs with low activity profiles against each target strain where their MIC value is more than 256 μg/ml were extracted as negative samples.

2 DRAMP Database:

At 2021, Shi and others proposed a new version of the Data Repository of Antimicrobial Peptides (DRAMP) that includes structures, sequences, activities, hemolytic, patent, physicochemical, and reference information about the AMPs [28]. Overall, it contains more than 22,000 samples, and provides multiple MS Excel files based on the targeted groups, subclasses, and usage purposes. As needed, the special edition of the Gram-negative AMPs database was downloaded, which contains over 2,400 samples. The AMPs activity and hemolytic information are available through a non-standard text in target-organism and hemolytic-activity properties. However, the MIC values and their units from the texts for the four targeted bacteria strains were extracted.

3 CAMPR Database:

Waghu and others introduced an updated version database for the CAMP database [45], consisting of the structures, sequences, activity, hemolytic, and family-specific signatures of prokaryotic and eukaryotic information about more than 10,000 AMPs. Around 2,000 of them are short peptides with an experimentally validated activity against gram-negative bacteria, which are provided as public web pages with the ability to download each page's actual content separately as a flat text file. However, these text files were downloaded and integrated into a single dataset. Then based on the Camp-ID property, the activity information were scraped from web pages and added to the dataset. Next, around 900 new AMPs that did not appear on the primary dataset were identified, and the MIC values and their unit were extracted for each target species. Finally, approximately 75 out of 900 unique AMPs with activity information against target strains were obtained.

3 YADAMP Database:

Piotto and others prepared another antimicrobial peptide database [47] based on its extensive literature search, which includes more than 2,000 AMP with detailed information about AMP's structure, activity, Etc., which made it appropriate for building activity models, but the hemolytic information is absent. However, the last update on this database was in October 2018. This database provided the AMPs information as web pages, so the AMPs sequence and the activity information were scraped, especially the MIC values in the μM unit for all available target species, and saved in an offline data file. Then this data was restructured and used to prepare a dataset for integration prepose by considering the available information for the four target strains mentioned before, in which around 160 unique AMPs sequences that did not appear on the primary dataset were obtained.

5 Uniport Database:

Bateman and others proposed a public database (Uniport) containing millions of protein sequences with their amino acids, which were collected during the past years [46]. The last release of this database on April 2021, and provided as FASTA format and XML files. However, it used to collect the non-AMPs sequences for balancing the activity dataset by increasing the number of negative samples as needed. Following C. Wang's and others approaches [20], [48], [49] we searched for non-AMPs that have lengths less than 20 amino acids and the keywords property not contains these terms: antibiotic, antimicrobial, antifungal, antiviral, fungicide, secretory, secreted, defensin, effector, and excreted. Finally, from the search results we only considered the natural sequences.

B Data Pre-Processing

The main objective of this step is to clean, normalize and integrate the collected data to establish several standard datasets for training and validation of the AMPs activity prediction modes. Several standard of activity datasets were prepared using the collected data from the DBAASP, DRAMP, CAMPR, and YADAMP databases. The required data for the target species were extracted from the DBAASP, DRAMP, and CAMPR databases and put in the same dataset structure, which contains of the AMPs sequence, and the activity information as a tuples data structure containing the

MIC value and its unit in (µM and µg/ml). Afterward, the collected data from the databases mentioned above were integrated into the same dataset. Regarding the DRAMP collected data, the MIC units were rewritten in a standard format. Hence, around 350 new AMPs against E. coli strains and 190 against Pseudomonas strains were obtained and integrated into the primary dataset. Moreover, more than 40 new AMPs against E. coli and Pseudomonas strains were obtained from CAMPR database and integrated with previous data. Nevertheless, the MIC still has different measurement units (µg/ml, µM). Furthermore, YADAMP collected data has just one MIC unit. Therefore, the MIC units were converted from mass concentration (µg/ml) to molar concentration (µM) using the following equation [50] before continuing the integration:

$$(\mu M) = \frac{\left(\frac{\mu g}{ml}\right)}{MW\ in\ KDa} \quad (1)$$

- Micromole. A micromole (μM) is one millionth of a mole (10−6 mol).
- ($\mu g/ml$) is the concentration of one gram of a substance per unit volume of the mixture equal to one cubic meter.
- *MW* is Molecular Weight.
- Dalton (*Da*) is an alternate name for the atomic mass unit, and kilodalton (*KDa*) is 1,000 Daltons.

Then, using the Rdkit package [51] the molecular weight (MW) was calculated by converting the AMP sequence to a simplified molecular-input line-entry system (SMILES), as shown in Figure 2 and calculating the exact molar weight for the AMP.

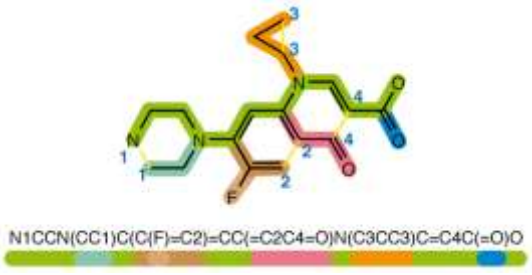

Figure 2 The chemical formula (SMILES) of AMP [52]

Now, all AMPs MIC values have been measured by µM unit. Finally, the collected data from the YADAMP database were integrated with our datasets, which obtained more than 300 new AMPs containing activity information against targeted E. coli and Pseudomonas strains. Finally, the integrated activity dataset's structure consists of features like the AMP sequence, and MIC values in the µM unit for E. coli and Pseudomonas strains.

Unfortunately, the activity datasets still contain negative MIC values, which are inconsequent values. So, all these values were replaced with Nan. All samples with Nan values that did not have activity information about the targeted strains and the completely similar sequences that were identified using the Cluster Database at High Identity with Tolerance (CD-HIT) webserver [53] were dropped. Around 5000 of 6300 AMPs with activity information against targeted E. coli strains and 2800 of 4600 against Pseudomonas strains are still in the activity datasets. Due to the short peptides having potent activity against target species [54] and the structure of short peptides with residues length less than or equal to 20, they are less complex and form a simple spiral pattern [55]. So, only AMPs with a length less than or equal to 20 were considered. Moreover, the activity feature is added to datasets as a Boolean label. According to the collected data exploration analysis, the MIC threshold values that determine the peptides as active or not were considered based on the Boxplot for the targeted strains E. coli, E. coli ATCC 25922, Pseudomonas, and Pseudomonas ATCC 27853, with MIC values in order (40, 25, 50 and 35 µM) were used to label the positive samples, and to avoid all samples that may be an out-layer and negatively affect the performance of prediction models. The peptide sequences that might not have activity profiles against targeted strains and their MIC value larger than 255 µg/ml were extracted from the DBAASP v3 database and labeled as negative samples. Then, flowing C. Wang's and others, many non-AMP sequences extracted from the Uniport database were integrated with datasets as negative samples to balance positive and negative samples. However, the final output of this step is four different activity datasets were constructed for the target strains E. coli, E. coli ATCC 25922, Pseudomonas, and Pseudomonas ATCC 27853, were included around in order (572, 1533, 450, and 1006) short AMPs with length less than 20 residues and activity profile against targeted strains. Finally, the Figure 3 shows the activity dataset's structure consists of the peptide sequence as a feature and its Boolean label that determine its activity against a specific target strain.

Figure 3 Samples of the integrated activity dataset with feature and its label

C Data Processing

Each dataset still needs more work to be appropriate for training and validating the AMPs prediction model versions against each target strain. This step consists of multiple stages to prepare each dataset and its samples to work fine with the proposed DL models.

Datasets Splitting: The AMPs sequence as (sequence of letters) and their Boolean labeled activity against target species were extracted from the datasets for activity purpose, then each dataset was split into training and testing sets, in which 80% of the primary dataset for training purposes to obtain the optimal weights from the training process, and 20% for testing

purpose to check how the trained model handles with the new AMPs sequence that has not previously been seen [56].

Sequences Padding: The deep learning architectures, like the CNN, Bi-LSTM, etc., take input with the same length, but the AMP sequences in our datasets have a different length (from 2 to 20) residue. Even the padding process enhances the accuracy and performance when fed these architectures by the fixed-length input size [57]. So, all sequences input were padded with zeros to be the same length as the max sequence length in the dataset as illustrated Figure 4.

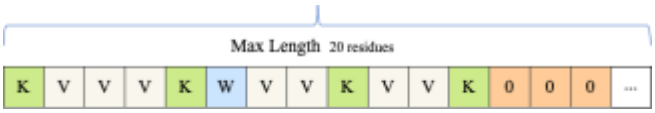

Figure 4 Example of padded sequence with zeros to match the longest sequence in dataset.

Sequences Encoding: Our AMP activity prediction models feds by a sequence of letters representing the amino acids of AMPs sequences, but the deep learning architectures only work with numeric input. Therefore, it is necessary to convert the string sequences to numerical input. According to the literature review, many techniques have been used and evaluated for this purpose [19], [58]. One hot encoding is widely used in deep learning solutions as a way for categorical data encoding, and its recommended compared to integer representation, binary representation, etc., due to its performance, simplicity, and use to encode the input when the solid ordinary relationship between features are absenting [59]. So, the peptide sequences in training and testing subsets were encoded using one-hot-encoding to be the input of activity prediction model versions. Then, as shown Figure 5 each sequence is represented as a 2D matrix of shape [L, 20], in which L represents the padded sequence length and 20 represents the number of natural amino acids added to the padding value (zero). Then these encoded samples were converted to 3D tensors [N, L, 20], in which N represents the number of AMP sequences in the dataset to be the input for our models.

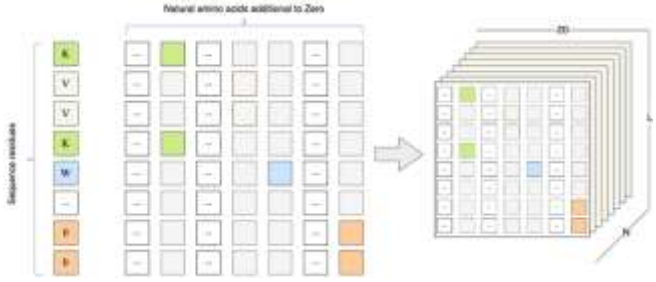

Figure 5 Example of encoding AMP sequence using one-hot-encoding and converting all dataset to 3D tensor.

D Models Topology

The AMP activity prediction problem has been defined as a binary classification task in which the main objective of these activity prediction models is to ascertain whether the suggested peptide sequences act as active AMPs against specifically targeted strains. After conducting multiple experiments with various deep learning model topologies including standalone CNN, LSTM, Bi-LSTM, and their different combinations the selected architecture of CNN, Bi-LSTM, and self-attention achieved the highest accuracy, proving to be the optimal choice for AMP activity prediction. This model topology consists of nine layers, as figured out in Figure 6.

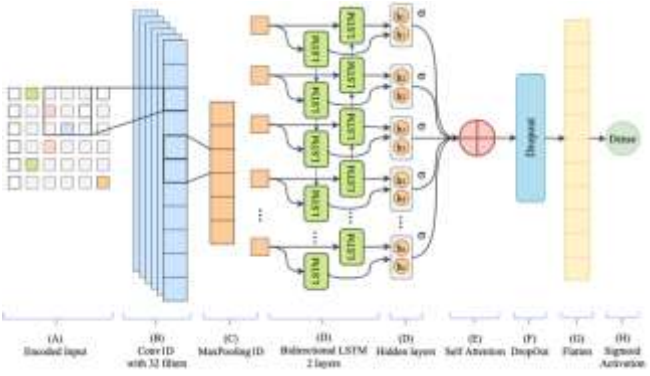

Figure 6 The architecture of the AMPs activity prediction model represented as a nine-layer deep learning layers. (A) The encoded input using one-hot-encoding with max length 20 residues per sequence, (B) 1D conventional layer with Relu activation function, (C) 1D max-pooling layer, (D) two layers of Bi-LSTM, (E) Kears self-attention layer with SoftMax activation function, (F) dropout layer, (G) flatten layer, and (H) dense layer with sigmoid activation function for output prediction.

The workflow begins with (A) encoded input using one-hot encoding with a maximum sequence length of 20 residues. Subsequent layers include (B) a 1D convolutional layer with ReLU activation function to extract sequence patterns, (C) a 1D max-pooling layer to reduce dimensionality and enhance feature detection, (D) two layers of Bi-LSTM to process sequence information bidirectionally for enhanced contextual understanding, and (E) a Keras self-attention layer with SoftMax activation function to prioritize important sequence features. The output is refined through (F) a dropout layer for overfitting prevention, (G) a flatten layer to convert multidimensional data to a 1D vector, and (H) a dense layer with a sigmoid activation function to produce the binary classification result.

Based on prepared datasets, the input is represented as (AMP sequences) and converted to numerical input using the one-hot-encoding technique. The combining 1D CNN and Bi-LSTM deep learning architectures were used to design the topology of activity prediction models to take advantage of both to learn and extract the features from the AMP sequences. The encoded input with shape [L, 20, 20] was fed to a 1D convolutional layer with 32 filters and a kernel size of 3 to identify the sequence patterns by producing the feature maps. The ReLU as a non-linear activation function was applied to the output to increase the non-linearity in the input. Then the output of the 1D convolutional layer was passed to the 1D max-pooling layer to compute the maximum output on each feature map independently. This enhances the model performance and computational cost by reducing the number of parameters; moreover, it controls the over-fitting. After that,

the output from 1D CNN was fed to the two Bi-LSTM layers to learn and extract the features from both sequence directions, from past to future and from future to pass for each time step and saved this information in two hidden states, which helps to improve the context available to its algorithm and increase the available information to the network. Next, the outputs from the Bi-LSTM layers were fed to the self-attention layer [60]. This allows the Bi-LSTM outputs to interact with each other to determine what a sequence residue (input) should pay more attention to by giving them an attention score and producing the outputs from these interactions and attention scores. In addition, the self-attention layer output passed to the dropout layer to reduce the overfitting issue. Finally, a flatten layer was added to reduce the output dimensionality by flattening it to a 1D vector. Due to the final output of this model topology being a binary classification [0,1], the vector was fed to the Dense layer with a sigmoid activation function to predict the final output.

The previous model topology was compiled to be ready for the training process with the following three parameters in order; the first one is the binary_crossentropy was used as a loss function due to the AMP activity perdition model being a binary classification model. Also, the second parameter is the optimizer function which the RMSprop was used, and the third parameter was accuracy as metrics. Moreover, the Keras early stopping callback function was used to control the training process when the learning process stopped improving the accuracy.

E Models Tuning

Each version of AMPs activity prediction models needs particular optimization to increase their accuracy. Therefore, the Bayesian hyperparameter tuning with k-fold cross-validation was utilized using the Keras tuner for AMPs activity prediction models. In which several hyperparameters were applied and explored like Bi-LSTM units between 32 and 512 with a step value of multiple of 32, dropout rates between 0.2 to 0.7, and different learning rates (0.01, 0.02, 0.03, 0.001, 0.002, 0.003). K-fold cross-validation was applied by splitting the original dataset into 80% for training and 20% for testing peruses, and then the training subset was divided into five folds with k equal to five iterations. During each iteration, four folds were used for training and one for validation that was switched every iteration, as shown in Figure 7, and the accuracy and the average loss were calculated for AMPs activity prediction models. Finally, after ending the hyperparameters tuning process, the best parameters that achieved the best overall performance model were selected to train each version of AMPs activity models using them on the entire training set and to evaluate it using the testing set.

Figure 7 Example of tuning process using 5-fold cross-validation.

F Models Training

Using Google Colab cloud as a storage and computing platform with a GPU unit was used to train four different versions of AMPs activity prediction models based on the best hyperparameters that were determined in the tuning stage, using the training subset of their AMPs activity datasets, with one model for each targeted strain. Then these trained models were saved as h5 files to use later in the evaluation stage.

G Models Evaluation

In order to evaluate the models, two type to tests were used: Statistical tests and Evaluation Metrics. The statistical tests play a crucial role in validating the effectiveness and reliability of predictive models, ensuring that observed performance differences are not due to chance by validating: the dataset distribution - should not follow normal distribution-, the observed differences in performance metrics across the proposed models must be statistically significant, and for the result dependency on the dataset.

For assessing whether AMP sequence lengths follow a normal distribution, Kolmogorov-Smirnov (KS) tests [61] are used for checking the distribution of AMP sequence lengths in the dataset. Ensuring that the data distribution is appropriately structured enhances model generalization and prevents biases in classification. When the dataset deviates from normality, additional balancing techniques such as data augmentation or normalization may be required.

One of the primary statistical methods for showing the model significance is Analysis of Variance (ANOVA) [62], which determines whether there are significant differences in model performance across multiple architectures. ANOVA helps assess variations in key metrics such as accuracy, area under the curve (AUC), and Matthews correlation coefficient (MCC), providing insights into how different models behave across datasets. If significant differences are detected, post-hoc tests like Tukey's HSD [63] can be applied to pinpoint which specific models perform better.

The four optimized prediction versions of AMPs activity prediction models were evaluated by calculating the ROC and the AUCs, which the ROC represents the true positive rate (TRP) versus the false positive rate (FPR) for every decision boundary. In which the Youden index (Yi) was calculated to determine the optimal decision boundary form every ROCs.

$$Yi = specificity + sensitivity - 1 \quad (2)$$

Moreover, the confusion matrix (CM) was generated for every optimized prediction model based on the optimal decision boundary determined in the previous step, which the CM contains the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) that helped us to calculate the specificity, sensitivity, precision, recall, accuracy, and the MCC using the following equations:

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

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (4)$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (5)$$

$$Acuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (6)$$

$$MCC = \frac{(\text{TP} * \text{TN}) - (\text{FP} * \text{FN})}{\sqrt{(\text{TP} + \text{FP}) * (\text{TP} + \text{FN}) * (\text{TN} + \text{FP}) * (\text{TN} + \text{FN})}} \quad (7)$$

Another important technique is the Chi-square Test [64] for Independence, which evaluates whether categorical classifications (AMP vs. non-AMP) are independent of the model or dataset used. By analysing confusion matrices, this test determines whether classification outputs are meaningfully associated with specific bacterial strains. If dependencies are observed, this implies the model effectively captures strain-specific features in AMP prediction.

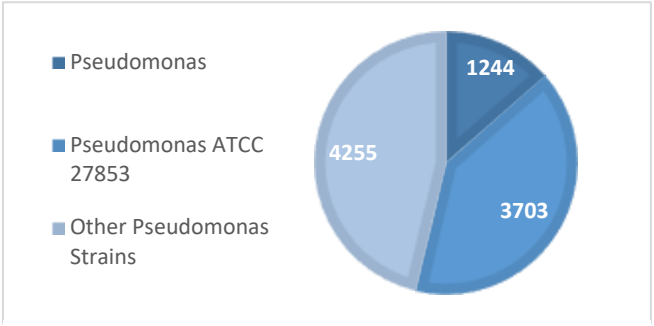

Figure 9 The number of active AMPs against Pseudo. strains in DBAASP

Additionally,

IV Experiments and Results

This section presents and discusses the main results of this paper by showing the collected data from the different AMPs databases and the obtained datasets that are used to train and evaluate the AMPs activity deep learning model versions. Also, it presents and discusses the main results of the AMPs activity prediction model versions for very target strains.

A Data Collection

The activity datasets were collected from multiple AMP databases. The primary database (DBAASP v3) contains more than 12,700 and 9,200 AMPs that have activity information against E. coli and Pseudomonas strains, which these two bacteria have the most available AMPs that have activity information against the other types of bacteria, as shown in Figure 8.

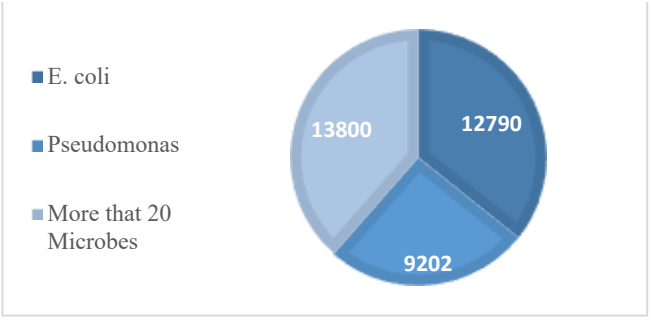

Figure 8 The number of active AMPs against the types of bacteria in DBAASP database

Moreover, the AMPs that have activity information against the top two gram-negative bacteria types (E. coli and Pseudomonas strains) were considered and explored more deeply. The DBAASP contains more than 500 strains of each E. coli and Pseudomonas. The highest numbers of AMPs with activity information were available for E. coli, E. coli ATCC 25922, Pseudomonas, and Pseudomonas ATCC 27853 strains, which contain around 1500, 5600, 1200, and 3,700 AMPs, as shown in Figure 10 and Figure 9. Also, according to our exploration, most pathogens that urgently need treatments and their strains did not have sufficient numbers of AMPs data (positive samples) to train deep learning models that can predict if peptide sequence potentially acts as AMP against it.

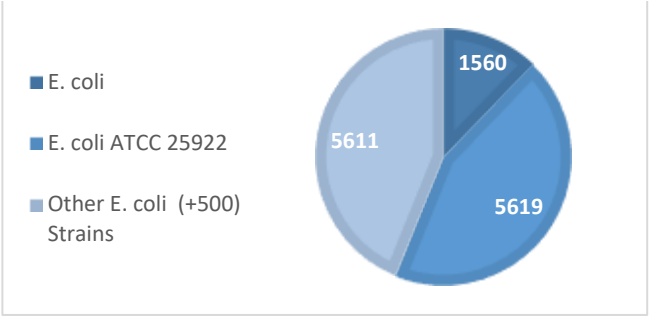

Figure 10 The number of active AMPs against E. coli strains in DBAASP

After considering only The AMPs that have natural residues, the number of available AMPs decreased. So, The DRAMP, CAMPR, and YADAMP were used as support databases to increase the number of available AMPs that have an activity profile against the targeted strains. Only the unique AMPs in support databases and absent from the primary database were considered. Based on databases exploration, TABLE 1 shows the number of obtained unique natural AMPs from the support databases compared to the primary database. The DBAASP v3 contains and covers the most available

information about E. coli and Pseudomonas strains, and the DRAMP and YDAMP still have hundreds of unique AMPs that have activity information against the main strains but are still suffering from containing AMPs against the E. coli ATCC 25922 and Pseudomonas ATCC 27853 strains. Finally, compared to the primary database, the CAMPR did not have enough numbers of unique AMPs against any targeted bacteria.

TABLE 1. THE NUMBER OF UNIQUE NATURAL AMPS IN DATABASES

Target Strain	AMP Databases			
	DBAASP	*DRAMP*	*CAMPR*	*YADAMP*
E. coli	1368	324	19	159
E. coli ATCC 25922	4668	27	44	0
Pseudomonas	971	179	5	42
Pseudomonas ATCC 27853	2935	15	6	0

B Datasets

After the data integration and pre-processing step, all AMPs with MIC values less than or equal to zero were dropped. Also, the natural AMP sequences with a length less than or equal to 20 residues and their MIC value less than or equal to pre-determined MIC threshold value (40, 25, 50 and 35 μM) for targeted strains were only considered. The positive samples in datasets for the (E. Coli, E. Coli ATCC 25922, Pseudomonas, and Pseudomonas ATCC 27853) were around (572, 1533, 450, and 1006). However, in this paper the C and N-terminus AMPs features were not considered due to the data availability. Moreover, until now, the AMP databases did not contain significant numbers of AMPs with activity profiles against other gram-negative bacteria and their strains that need urgent treatments to train deep learning models. On the other hand, a lack of databases containing an experimentally validated non-antimicrobial [49]. Also, the number of extracted peptides with high MIC values that possibly did not have activity profiles against specific bacterial species was very small and insufficient to balance the activity datasets to prevent bias of activity classification models towards the positive samples [65]. So, to avoid this issue, following C. Wang's and others' alternative negative dataset approach, the peptide sequences that did not act as antimicrobials were extracted from the Uniport database and used to balance each activity dataset, which was about (571, 1258, 477, and 1006) non-antimicrobial sequences for targeted strains TABLE 2.

TABLE 2. THE NUMBER OF AMPS THAT HAVE ACTIVITY PROFILE AGAINST TARGETED STRAINS IN DATASETS

Target Strain	Dataset Samples	
	Positive Samples	*Negative Samples*
E. coli	572	571
E. coli ATCC 25922	1533	1258
Pseudomonas	450	447
Pseudomonas ATCC 27853	1006	1006

The result of KS shows that the sequence lengths do not follow a normal distribution since that the sequence lengths do not follow a normal distribution since KS Statistic D = 0.0833 and p-value = 1.03×10^{-09}.

C Models Topology

In this work, four DL activity prediction models were built, each designed to predict AMP activity against a specific target strain. These models were built using various deep learning architectures, including CNN, LSTM, Bi-LSTM, and different combinations of these layers. Through extensive experimentation with multiple model topologies, we assessed the effectiveness of each configuration. The selected combination integrating CNN, Bi-LSTM, and a self-attention mechanism achieved the best predictive accuracy, demonstrating its strength in feature extraction and contextual sequence understanding. This combination was inspired by its proven success in several fields, such as text classification, biomedical sequence analysis, and natural language processing, where hybrid models have significantly improved classification accuracy and pattern recognition [22]-[26], [42]-[44]. By adapting this approach to AMP activity prediction, we leveraged its ability to enhance feature learning and optimize model performance for identifying active peptides against specific pathogen strains.

D Models Tunning

The choice of hyperparameter combinations is one of the essential factors affecting the performance of deep learning models [66]. The traditional or manual way of tuning the DL models wastes time and effort, and very hard to try every possible value of hyperparameters. So, the performance of the proposed models was automated tune using Bayesian hyperparameter optimization due to its capability to take the performance from prior hyperparameter combinations into account when determining the new hyperparameter to select the best parameters that allow the model to achieve the best overall result. This makes it one of the most effective methods in this area [67]. Using automated Bayesian hyperparameter optimization, the best hyperparameters obtained based on the validation accuracy for the AMPs activity prediction models include 32 filters, hidden units ranging from 192 to 512, dropout rates between 0.5 and 0.7, and a consistent learning rate of 0.003, as detailed in TABLE 3.

TABLE 3. HYPERPARAMETERS OPTIMIZATION FOR POTENTIAL AMPS PREDICTION MODELS

Activity Model	Hyperparameters				*Validation Acc.*
	Filters	*Hidden Units*	*Dropout rate*	*Learning rate*	
E. coli	32	192	0.5	0.003	96.53
E. coli ATCC 25922	32	512	0.7	0.003	96.55
Pseudomonas	32	352	0.5	0.003	94.97
Pseudomonas ATCC 27853	32	352	0.5	0.003	97.26

E Models Training History

Four AMPs activity prediction models were trained using the targeted strain's activity datasets based on the best parameters determined by the reported Bayesian hyperparameter optimization TABLE 3. The activity models against (E. coli, E. coli ATCC 25922, Pseudomonas, and Pseudomonas ATCC 27853) achieved training accuracy (100, 99.37, 100, and 100) % in (26, 39, 27, 27) epochs respectively, as shown Figure 11. Based on our experiments, it has been observed that larger datasets containing consistent samples make the training process smoother.

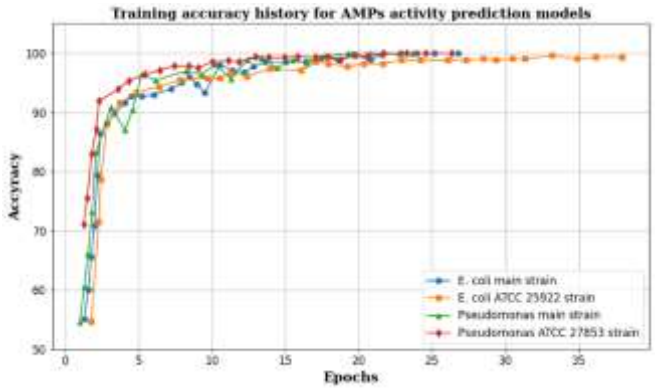

Figure 11 Training accuracy history for AMPs activity prediction models for every target strain.

F Models Evaluation

The four AMPs activity prediction models were thoroughly evaluated using testing subsets specific to each bacterial strain, comprising 229 samples for E. coli, 558 samples for E. coli ATCC 25922, 179 samples for Pseudomonas aeruginosa, and 402 samples for Pseudomonas aeruginosa ATCC 27853. The evaluation was conducted by calculating the ROC curve and its associated AUC values, which quantify the model's ability to accurately distinguish between active AMPs sequences and non-antimicrobial sequences. Additionally, the Yi index and MCC were computed to provide a more holistic performance assessment. A higher AUC value reflects superior model performance and predictive capability. The models achieved exceptional results, with AUC values of 0.983 for E. coli with an optimal decision boundary of 0.67379 Figure 12, 0.995 for E. coli ATCC 25922 with an optimal decision boundary of 0.97447 Figure 13, 0.994 for Pseudomonas aeruginosa with an optimal decision boundary of 0.93355 Figure 14, and 0.991 for Pseudomonas aeruginosa ATCC 27853 with an optimal decision boundary of 0.96656 Figure 15. These outcomes underscore the reliability and efficiency of the AMPs prediction models in identifying active peptides against specific pathogen strain.

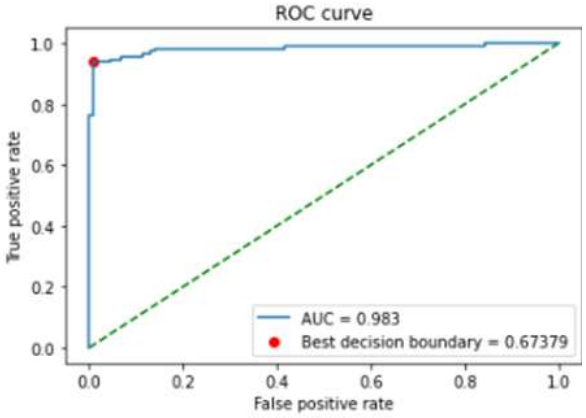

Figure 12 The ROC and AUC of AMPs activity prediction model against E. coli main strain.

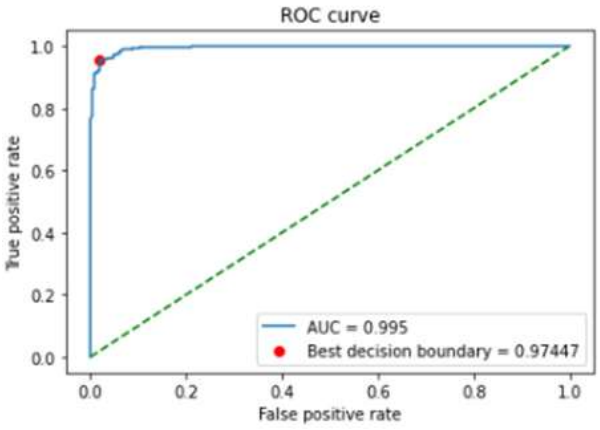

Figure 13 The ROC and AUC of AMPs activity prediction model against E. coli ATCC 25922 strain.

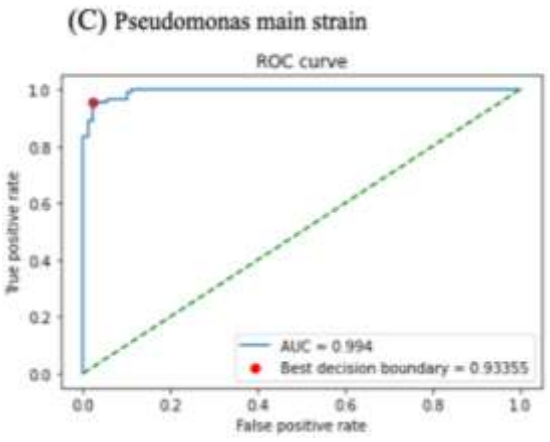

Figure 14 The ROC and AUC of AMPs activity prediction model against Pseudomonas aeruginosa strain.

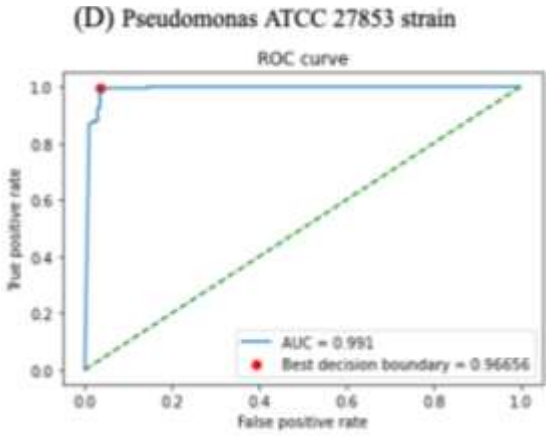

Figure 15 The ROC and AUC of AMPs activity prediction model against Pseudomonas ATCC 27853 strain.

Furthermore, the confusion matrices were constructed based on the calculated optimal decision boundaries for each model, as illustrated in Figure 16. These matrices provided detailed insights into the performance metrics of the AMPs activity prediction models for the target bacterial strains: E. coli, E. coli ATCC 25922, Pseudomonas aeruginosa, and Pseudomonas aeruginosa ATCC 27853. The following results were obtained: sensitivity values of 93.85%, 95.43%, 95.55%, and 99.50%; specificity values of 99.13%, 98%, 97.75%, and 96.51%; Youden Index (Yi) scores of 0.9299, 0.9344, 0.9330, and 0.9601; precision values of 99.07%, 98.32%, 97.72%, and 96.61%; MCC scores of 99.13%, 93.19%, 93.32%, and 96.06%; and accuracy levels of 97%, 97%, 97%, and 98%, respectively.

These metrics highlight the robustness and reliability of the proposed models in accurately classifying AMPs against the corresponding pathogen strains. The high sensitivity scores demonstrate the models' ability to detect active AMP sequences effectively, while the specificity values validate their capacity to distinguish non-antimicrobial sequences accurately. Similarly, the Yi index and MCC scores reflect the balanced and consistent performance of the models across different datasets. Overall, the calculated metrics underline the efficiency of the hybrid deep learning approach in addressing complex AMP classification challenges.

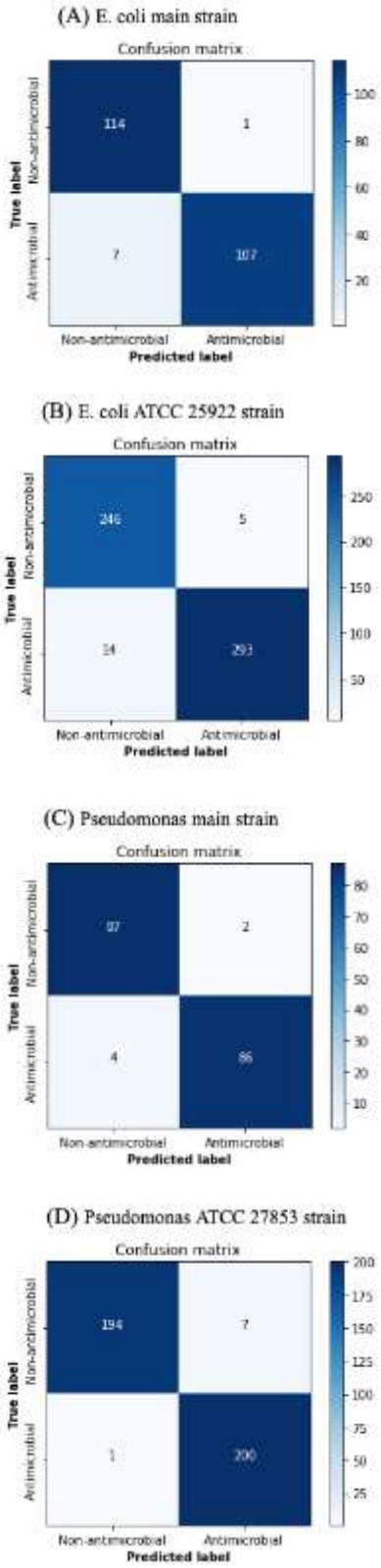

Figure 16 The Confusion matrices for the four AMPs activity prediction model. (A) The conducted confusion matrix for E. coli main strain model, (B) The conducted confusion matrix for E. coli ATCC 25922 strain model, (C) The conducted confusion matrix for Pseudomonas

main strain model, (D) The conducted confusion matrix for Pseudomonas ATCC 27853 strain model.

TABLE 4 presents the evaluation results of the proposed AMPs activity prediction models for four target bacterial strains (*E. coli*, *E. coli ATCC 25922*, *Pseudomonas aeruginosa*, and *Pseudomonas aeruginosa ATCC 27853*). The evaluation metrics include the Area Under the Curve (AUC), Matthews Correlation Coefficient (MCC), and overall accuracy (ACC), collectively demonstrating the models' ability to distinguish active antimicrobial peptides from non-antimicrobial sequences. The results highlight the robust performance of the prediction models, achieving consistently high AUC values ranging from 0.983 to 0.995, MCC percentages between 93.13% and 96.06%, and accuracy levels of 97–98%.

These findings confirm the effectiveness and reliability of the proposed models in accurately predicting AMP activity across varied dataset sizes, effectively differentiating between short AMPs and non-antimicrobial peptides. It is worth noting that the accuracy of the models is closely aligned, which is attributed to the similarity observed in the negative samples used during training. This emphasizes the efficiency of the proposed hybrid topology in distinguishing active AMPs from non-active sequences, showcasing its strength in enhancing predictive capabilities.

TABLE 4. THE EVALUATION RESULTS FOR THE AMPS ACTIVITY PREDICTION MODELS

Activity Prediction Model	Evaluation Metrics		
	AUC	*MCC (%)*	*ACC (%)*
AMPs against E. coli	0.983	93.13	97
AMPs against E. coli ATCC 25922	0.995	93.19	97
AMPs against Pseudomonas	0.994	93.32	97
AMPs against Pseudomonas ATCC 27853	0.991	96.06	98

[a.] AUC = Area Under the Curve, MCC = Mathews correlation coefficient, ACC = Accuracy.

TABLE 5 provides detailed evaluation metrics for the AMPs activity prediction models against the four target bacterial strains (E. coli, E. coli ATCC 25922, Pseudomonas aeruginosa, and Pseudomonas aeruginosa ATCC 27853). The metrics include specificity (SPEC), sensitivity (SENS), Youden Index (Yi), and the optimal decision threshold values for classification. Specificity reflects the models' ability to correctly identify non-antimicrobial peptides, while sensitivity highlights their effectiveness in accurately detecting active AMPs. The Youden Index (Yi) serves as a measure of model performance by balancing specificity and sensitivity, with higher values indicating better predictive capability. Lastly, the optimal threshold values were calculated for each model, representing the decision boundary that maximizes the Yi metric.

The data in TABLE 5 demonstrates the strong performance of the proposed models, with specificity ranging from 96.51% to 99.13% and sensitivity from 93.85% to 99.50%. These results highlight the robustness and accuracy of the models in distinguishing active AMPs from non-active peptides for each target strain.

TABLE 5. More detailed information about evaluation results

Activity Prediction Model	Evaluation Metrics			
	SPEC (%)	*SENS (%)*	*Yi*	*Best threshold*
AMPs against E. coli	99.13	93.85	0.9299	0.67379
AMPs against E. coli ATCC 25922	98	95.43	0.9344	0.97447
AMPs against Pseudomonas	97.75	95.55	0.93.30	0.93355
AMPs against Pseudomonas ATCC 27853	96.51	99.50	0.9601	0.96656

[a.] SPEC = Specificity, SENS = Sensitivity, Yi = Youden index.

However, direct comparison between the proposed AMPs prediction models and other state-of-the-art models is disproportionate, subjective, and difficult due to several factors, including the lack of standardized AMPs activity datasets and the purpose of the prediction models, which it trained to predict activity profiles for specific strains or trained as general prediction models. Regardless of the difference between the related approaches, methodologies, methods, and models used in this task, TABLE 6 shows performance comparisons between our models and other state-of-the-art works.

TABLE 6. EVALUATION AND PERFORMANCE COMPARISON BETWEEN THE STATE-OF-THE-ART ACTIVITY MODELS AND OUR ACTIVITY MODELS.

Model	Evaluation Metrics			*Ref.*
	AUC	*ACC (%)*	*Target*	
AMP Scanner v.2	0.9648	-	General	[20]
APIN	0.973	92.55	General	[16]
Deep-AmPEP30	0.8533	77.13	General	[21]
DNN-Conv2	0.972	93.8	General	[19]
AMPlify	0.9744	92.79	General	[68]
SP tool	-	81.0	E. coli ATCC 25922 Pseudomonas ATCC 27853	[69]
Model V.6	0.981	95.7	E. coli	[70]
Our activity prediction models	**0.983**	**97**	**E. coli**	-
	0.995	**97**	**E. coli ATCC 25922**	-
	0.994	**97**	**Pseudomonas**	-
	0.991	**98**	**Pseudomonas ATCC 27853**	-

[a.] AUC = Area Under the Curve, ACC = Accuracy, Ref = Reference.

For the model significance, the one-way ANOVA test shows that the last model is significantly differs from the others in terms of predictive performance (e.g., MCC), since the test values came as: F-statistic: 6.21 and p-value: 0.0021 for the null hypothesis "the performance of all the models is equal".

Tukey's Honestly Significant Difference (HSD) test was conducted to identify which model pairs showed statistically significant differences. The test revealed that the model targeting P. aeruginosa ATCC 27853 significantly outperformed the others in terms of MCC and AUC. This aligns with empirical evaluation results, where this model achieved the highest AUC (0.991) and MCC (96.06%).

Finally, Chi Square test for independence result, Chi-square statistic: 21.45 and p-value: 0.0006, confirms that the models effectively learned strain-specific features, confirming the models' capacity for targeted AMP classification.

V CONCLUSIONS

The peptide-based drugs usage has increased in treating diseases caused by antibiotic-resistant pathogens due to their unique properties compared to antibiotics. However, the discovery of this type of treatment are expensive and take a lot of time and effort. In the last decade, several machine learning methodologies and algorithms have been used to speed up the validation and selection of the most potential peptide sequences to nominate them for clinical trials. Nevertheless, it still required a lot of feature engineering, vast experience, and knowledge in this domain. Recently, deep learning has proven its performance in many fields, especially in the drug discovery and design domain, and has inspired many researchers to take advantage of this domain due to its performance and automatic feature extraction advantages, and high learnability. Unfortunately, most research has tended to distinguish generally between AMPs and non-AMPs. A few of them touched on proposing models that differentiate between AMPs with activity against a specific target bacteria strain and non-active AMPs.

This work proposed a complete sequence-based AMPs activity prediction model using CNN, Bi-LSTM, and the self-attention mechanism that can determine the AMPs have activity against E. coli, E. coli ATCC 25922, Pseudomonas, and Pseudomonas ATCC 27853 strains, by collecting the required AMPs information from different up-to-data antimicrobial activity databases and integrating them to establish up-to-date activity datasets for every target strain. The DBAASP v.3 databases were the top rich, quality ones and had the most available data about the targeted strains. However, until now, the explored databases contained neither experimentally validated negative samples in non-antimicrobial peptides nor peptides lacking activity profiles against the target strains. So, this point needs more work from the laboratory side by reporting the results of negative experiments and publishing them in AMPs databases because the accuracy of deep learning models is affected by negative data as well as positive sets.

The AMPs activity prediction models demonstrated impressive results, achieving accuracies between 97–98% and MCC scores ranging from 0.93 to 0.96. Nonetheless, the models' performance is sensitive to the diversity of training datasets, particularly in terms of sequence length variability and amino acid order differences among peptide samples. To enhance model robustness, future training datasets should prioritize broader coverage of peptide sequences, including outlier samples and sequences with varying residue orders.

Looking ahead, the findings demonstrate that combining CNN and Bi-LSTM deep learning architectures with a self-attention mechanism is highly effective in distinguishing AMPs active against specific pathogen strains from non-active ones. As a future direction, we propose exploring the potential of training large-scale deep learning models on generalized AMP datasets and fine-tuning them to predict activity for specific target strains. This approach could address the challenge posed by the limited availability of AMPs for certain bacteria while maintaining the precision required for strain-specific activity prediction. Moreover, leveraging advanced techniques such as transfer learning and attention-based architectures offers opportunities to enhance model performance and optimize computational efficiency. These methods can expand the scope and adaptability of AMP discovery by enabling scalable solutions that adapt to diverse datasets and pathogen-specific requirements.

By incorporating these strategies, future research can advance peptide-based drug design, paving the way for more accurate, impactful, and innovative solutions in combating antibiotic-resistant pathogens.

REFERENCES

[1] M. S. Zharkova *et al.*, "Application of antimicrobial peptides of the innate immune system in combination with conventional antibiotics-a novel way to combat antibiotic resistance?," *Front Cell Infect Microbiol*, vol. 9, no. APR, 2019.

[2] H. Jenssen, P. Hamill, and R. E. W. Hancock, "Peptide antimicrobial agents," *Clin Microbiol Rev*, vol. 19, no. 3, pp. 491–511, 2006.

[3] B. Mishra and G. Wang, "The importance of amino acid composition in natural amps: An evolutional, structural, and functional perspective," *Front Immunol*, vol. 3, no. JUL, pp. 2010–2013, 2012.

[4] Q. Wu, H. Ke, D. Li, Q. Wang, J. Fang, and J. Zhou, "Recent Progress in Machine Learning-based Prediction of Peptide Activity for Drug Discovery," *Curr Top Med Chem*, vol. 19, no. 1, pp. 4–16, 2019, doi: 10.2174/1568026619666190122151634.

[5] S. Basith, B. Manavalan, T. Hwan Shin, and G. Lee, "Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening," *Med Res Rev*, vol. 40, no. 4, pp. 1276–1314, 2020.

[6] M. S. M. Serafim *et al.*, "The application of machine learning techniques to innovative antibacterial discovery and development," *Expert Opin Drug Discov*, vol. 15, no. 10, pp. 1165–1180, 2020.

[7] N. Stephenson *et al.*, "Survey of Machine Learning Techniques in Drug Discovery," *Curr Drug Metab*, vol. 20, no. 3, pp. 185–193, 2018.

[8] J. Xu *et al.*, "Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides," *Brief Bioinform*, Mar. 2021.

[9] X. Xiao and Z. B. You, "Predicting minimum inhibitory concentration of antimicrobial peptides by the pseudo-amino acid composition and Gaussian kernel regression," *Proceedings - 2015 8th International Conference on BioMedical Engineering and Informatics, BMEI 2015*, pp. 301–305, Feb. 2016.

[10] D. Veltri, U. Kamath, and A. Shehu, "Improving Recognition of Antimicrobial Peptides and Target Selectivity through Machine Learning and Genetic Programming," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 14, no. 2, pp. 300–313, Mar. 2017.

[11] S. Gull and F. Minhas, "AMP0: Species-Specific Prediction of Anti-microbial Peptides Using Zero and Few Shot Learning," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 19, no. 1, pp. 275–283, 2022.

[12] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature Publishing Group*, 2015.

[13] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans Neural Netw*, vol. 8, no. 1, pp. 98–113, 1997.

[14] A. M. and G. H. Alex Graves, "Speech Recognition with Deep Recurrent Neural Networks , Department of Computer Science, University of Toronto," *Department of Computer Science, University of Toronto*, vol. 3, no. 3, pp. 45–49, 2013.

[15] P. Pandey, M. A. Bender, R. Johnson, and R. Patro, "Supplementary Material for deBGR: An Efficient and Near-Exact Representation of the Weighted de Bruijn Graph 1 An Algorithm for Computing Abundance Corrections," *Bioinformatics*, vol. 323, no. 1, pp. 46–1, 2017.

[16] X. Su, J. Xu, Y. Yin, X. Quan, and H. Zhang, "Antimicrobial peptide identification using multi-scale convolutional network," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–10, 2019.

[17] C. Li *et al.*, "AMPlify: Attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens," *bioRxiv*, pp. 1–37, 2020.

[18] S. N. Dean and S. A. Walper, "Variational autoencoder for generation of antimicrobial peptides," *ACS Omega*, vol. 5, no. 33, pp. 20746–20754, 2020, doi: 10.1021/acsomega.0c00442.

[19] M. Dua, D. Barbará, and A. Shehu, "Exploring deep neural network architectures: A case study on improving antimicrobial peptide recognition," *EPiC Series in Computing*, vol. 70, pp. 182–191, 2020.

[20] D. Veltri, U. Kamath, and A. Shehu, "Deep learning improves antimicrobial peptide recognition," *Bioinformatics*, vol. 34, no. 16, pp. 2740–2747, 2018.

[21] J. Yan *et al.*, "Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning," *Mol Ther Nucleic Acids*, vol. 20, no. June, pp. 882–894, 2020.

[22] Y. Xu, T. Li, and H. Fang, "An Antibiotic Terahertz Spectrum Recognition Method Based on CNN and Attention-BiLSTM," in *2022 8th International Conference on Control, Automation and Robotics (ICCAR)*, IEEE, Apr. 2022, pp. 197–202.

[23] L. Xiao, L. Zhang, F. Niu, X. Su, and W. Song, "RETRACTED: Remaining useful life prediction of wind turbine generator based on 1D-CNN and Bi-LSTM," *Int J Fatigue*, vol. 163, p. 107051, Oct. 2022.

[24] Y. Li, X. Li, Y. Liu, Y. Yao, and G. Huang, "MPMABP: A CNN and Bi-LSTM-Based Method for Predicting Multi-Activities of Bioactive Peptides," *Pharmaceuticals*, vol. 15, no. 6, p. 707, Jun. 2022.

[25] A. Singh and D. Vij, "CNN-LSTM based Social Media Post Caption Generator," in *2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, IEEE, Feb. 2022, pp. 205–209.

[26] R. Mahajan and V. Mansotra, "Predicting Geolocation of Tweets: Using Combination of CNN and BiLSTM," *Data Sci Eng*, vol. 6, no. 4, pp. 402–410, Dec. 2021.

[27] M. Pirtskhalava *et al.*, "DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics," *Nucleic Acids Res*, vol. 49, 2021.

[28] G. Shi *et al.*, "DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides," *Nucleic Acids Res*, no. August, pp. 0–9, 2021.

[29] A. Capecchi, X. Cai, H. Personne, T. K. ¨ Ohler, and J.-L. Reymond, "Machine learning designs non-hemolytic antimicrobial peptides ," 2021.

[30] B.-J. Yoon, "Hidden Markov Models and their Applications in Biological Sequence Analysis," *Curr Genomics*, vol. 10, no. 6, pp. 402–415, 2009.

[31] M. Torrent, D. Andreu, V. M. Nogués, and E. Boix, "Connecting peptide physicochemical and antimicrobial properties by a rational prediction model," *PLoS One*, vol. 6, no. 2, pp. 1–8, 2011.

[32] X. Xiao, P. Wang, W. Z. Lin, J. H. Jia, and K. C. Chou, "IAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Anal Biochem*, vol. 436, no. 2, pp. 168–177, 2013.

[33] E. G. Randou, D. Veltri, and A. Shehu, "Systematic analysis of global features and model building for recognition of antimicrobial peptides," *2013 IEEE 3rd International Conference on Computational Advances in Bio and Medical Sciences, ICCABS 2013*, pp. 1–6, 2013.

[34] E. Y. Lee, B. M. Fulan, G. C. L. Wong, and A. L. Fer-

guson, "Mapping membrane activity in undiscovered peptide sequence space using machine learning," *Proc Natl Acad Sci U S A*, vol. 113, no. 48, pp. 13588–13593, 2016.

[35] P. Bhadra, J. Yan, J. Li, S. Fong, and S. W. I. Siu, "AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest," *Sci Rep*, vol. 8, no. 1, pp. 1–10, 2018.

[36] C. R. Chung, T. R. Kuo, L. C. Wu, T. Y. Lee, and J. T. Horng, "Characterization and identification of antimicrobial peptides with different functional activities," *Brief Bioinform*, vol. 21, no. 3, pp. 1098–1114, 2020.

[37] R. Barrett, S. Jiang, and A. D. White, "Classifying antimicrobial and multifunctional peptides with Bayesian network models," *Peptide Science*, vol. 110, no. 4, pp. 1–9, 2018.

[38] P. Feng, Z. Wang, and X. Yu, "Predicting antimicrobial peptides by using increment of diversity with quadratic discriminant analysis method," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 16, no. 4, pp. 1309–1312, 2019.

[39] H. Fu, Z. Cao, M. Li, X. Xia, and S. Wang, "Prediction of Anuran Antimicrobial Peptides Using AdaBoost and Improved PSSM Profiles," *ACM International Conference Proceeding Series*, 2020.

[40] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," *Physica D*, vol. 404, p. 132306, 2020.

[41] Y. Zuo, Y. Li, Y. Chen, G. Li, Z. Yan, and L. Yang, "PseKRAAC: A flexible web server for generating pseudo K-tuple reduced amino acids composition," *Bioinformatics*, vol. 33, no. 1, pp. 122–124, 2017.

[42] B. Jang, M. Kim, G. Harerimana, S. U. Kang, and J. W. Kim, "Bi-LSTM model to increase accuracy in text classification: Combining word2vec CNN and attention mechanism," *Applied Sciences (Switzerland)*, vol. 10, no. 17, Sep. 2020.

[43] J. Wang, L.-C. Yu, K. R. Lai, and X. Zhang, "Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model."

[44] M. U. Salur and I. Aydin, "A Novel Hybrid Deep Learning Model for Sentiment Classification," *IEEE Access*, vol. 8, pp. 58080–58093, 2020.

[45] F. H. Waghu, R. S. Barai, P. Gurung, and S. Idicula-Thomas, "CAMPR3: A database on sequences, structures and signatures of antimicrobial peptides," *Nucleic Acids Res*, vol. 44, no. D1, pp. D1094–D1097, 2016.

[46] A. Bateman *et al.*, "UniProt: A hub for protein information," *Nucleic Acids Res*, vol. 43, no. D1, pp. D204–D212, 2015.

[47] S. P. Piotto, L. Sessa, S. Concilio, and P. Iannelli, "YADAMP: Yet another database of antimicrobial peptides," *Int J Antimicrob Agents*, vol. 39, no. 4, pp. 346–351, 2012.

[48] J. Witten and Z. Witten, "Deep learning regression model for antimicrobial peptide design," *bioRxiv*, 2019.

[49] C. Wang, S. Garlick, and M. Zloh, "Deep learning for novel antimicrobial peptide design," *Biomolecules*, vol. 11, no. 3, pp. 1–17, 2021.

[50] NovaTeinBio, "Converting Protein Mass Concentration to Molar Concentration, Or Vice Versa," https://www.novusbio.com/resources/calculators.

[51] G. Landrum, "Rdkit documentation," *Release*, vol. 1, pp. 1–79, 2019.

[52] I. W. Hamley, *Introduction to Peptide Science*, 1st ed. Wiley, 2020.

[53] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012.

[54] S. Ramesh, T. Govender, H. G. Kruger, B. G. de la Torre, and F. Albericio, "Short AntiMicrobial Peptides (SAMPs) as a class of extraordinary promising therapeutic agents," *Journal of Peptide Science*, no. February, pp. 438–451, 2016.

[55] D. Nagarajan *et al.*, "Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria," *Journal of Biological Chemistry*, vol. 293, no. 10, pp. 3492–3509, Mar. 2018.

[56] B. Ghojogh and M. Crowley, "The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial," May 2019, [Online]. Available: http://arxiv.org/abs/1905.12787

[57] M. Dwarampudi and N. V. S. Reddy, "Effects of padding on LSTMs and CNNs," Mar. 2019.

[58] F. Taho, "Antimicrobial Peptide Host Toxicity Prediction with Transfer Learning for Proteins," 2020.

[59] G. De Clercq, "Deep Learning for Classification of Dna Functional Sequences," 2018.

[60] "GitHub - CyberZHG/keras-self-attention: Attention mechanism for processing sequential data that considers the context for each timestamp." Accessed: Dec. 12, 2022. [Online]. Available: https://github.com/CyberZHG/keras-self-attention

[61] G. Fasano and A. Franceschini, "A multidimensional version of the Kolmogorov–Smirnov test," *Mon Not R Astron Soc*, vol. 225, no. 1, pp. 155–170, Mar. 1987, doi: 10.1093/MNRAS/225.1.155.

[62] L. Sthle and S. Wold, "Analysis of variance (ANOVA)," *Chemometrics and Intelligent Laboratory Systems*, vol. 6, no. 4, pp. 259–272, Nov. 1989.

[63] H. Abdi and L. J. Williams, "Encyclopedia of Research Design", Accessed: May 08, 2025. [Online].

Available: http://www.utd.edu/~herve
[64] R. J. Tallarida and R. B. Murray, "Chi-Square Test," *Manual of Pharmacologic Calculations*, pp. 140–142, 1987.
[65] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, Nov. 2016, doi: 10.1007/s13748-016-0094-0.
[66] J. Wu, X. Y. Chen, H. Zhang, L. D. Xiong, H. Lei, and S. H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, Mar. 2019.
[67] M. Feurer and F. Hutter, "Hyperparameter Optimization," 2019, pp. 3–33.
[68] C. Li *et al.*, "AMPlify: Attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens," *bioRxiv*, pp. 1–37, 2020.
[69] B. Vishnepolsky *et al.*, "Predictive Model of Linear Antimicrobial Peptides Active against Gram-Negative Bacteria," *J Chem Inf Model*, vol. 58, no. 5, pp. 1141–1151, May 2018.
[70] C. Wang, S. Garlick, and M. Zloh, "Deep learning for novel antimicrobial peptide design," *Biomolecules*, vol. 11, no. 3, pp. 1–17, 2021.

Abdullah Abu Nada is the head of the software and development department at the University of Palestine. He received his B.S. in Software Engineering from the University of Palestine and his MSc in Data Science from the Islamic University of Gaza. He has many achievements in software development and has won many awards in this field. Now, he is looking forward to making his path in scientific research and being part of innovative and challenging research projects. Abu Nada's research contributions include natural language processing, information security, and computational discovery and design of antimicrobial peptides.

Iyad H. Alshami currently is the dean of Faculty of Information Technology, Islamic University of Gaza. He received his B.Sc. degree in Mathematics and computer in 1999, M.Sc. degree in Information technology 2011 from Islamic University of Gaza, Palestine. He got his PhD in Software Engineering in 2016 from University Technology Malaysia (UTM), during his PhD journey he awarded many times as best year student. Iyad's research interests include Deep Learning/Machine Learning with applied orientation in many domains such as Indoor Positioning, Healthcare & diagnosis, Speech recognition and more. He has many valuable published research in these domains.