Divergent Approaches to Antimicrobial Peptide Classification: A Comparative Evaluation of AI Techniques

1st Bharathi Mohan G
Department of Computer Science and Engineering,
Amrita School of Computing, Amrita Vishwa Vidyapeetham,
Chennai, India
g_bharathimohan@ch.amrita.edu

2nd Sushma Rajagopal
Department of Computer Science and Engineering,
Amrita School of Computing, Amrita Vishwa Vidyapeetham,
Chennai, India
sushma242003@gmail.com

3rd Srinath Doss
Faculty of Engineering and Technology
Botho university
Botswana
srinath.doss@bothouniversity.ac.bw

***Abstract*—Given the global rise in antibiotic resistance, antimicrobial peptides (AMPs) provide a viable substitute for traditional antibiotics. Through a comparison of transformer-based, deep learning (DL), and machine learning (ML) techniques, this work seeks to classify AMPs. AMP sequences from curated databases were compared with non-AMP sequences obtained from UniProtKB/Swiss-Prot. The study compares performance across many assessment variables to assess how well various algorithms separate AMPs from non-AMPs. Significantly, RoBERTa outperformed DistilBERT (84.43%) in accuracy (86.53%). To provide a thorough knowledge of the algorithms' performance, the research also included a variety of other assessment criteria, including area under the receiver operating characteristic curve (AUC-ROC), recall, F1 score, and accuracy. The performance of the LSTM and GRU models was notable, despite their significantly lower accuracies. Furthermore, conventional machine learning methods such as Naive Bayes and Logistic Regression demonstrated competitive accuracy. This thorough analysis clarifies the benefits and drawbacks of various AMP classification techniques, providing insightful information for next studies on antimicrobial peptides and their potential to fight antibiotic resistance.**

***Index Terms*—Bioactive sequences, peptide classification, Machine learning, Deep learning, LSTM, GRU**

I. Introduction

AMPs are main components of innate immunity, playing a critical role in protecting living things from pathogens. They have strong antibacterial qualities and adaptability in biotechnology, medicine, and agriculture. Correctly classifying peptides as AMPs or non-AMPs is critical for understanding host-pathogen relationships, directing antibiotic treatments, and advancing drug discovery. Many transformer-based, ML and DL models specifically designed for peptide classification have been developed as a result of advances in artificial intelligence (AI) [1]. With the introduction of transformer-based, DL, and ML models in recent years, artificial intelligence (AI) has completely changed the process of categorising peptides. Still, a thorough comparison of different techniques is hard, even with the widespread use of such methodologies. To fill this research vacuum, this work sets out to thoroughly assess transformer, ML, and DL models for AMP categorization. By comparing and evaluating several algorithms' results on various datasets, this work aims to determine the best technique for reliably differentiating between AMPs and non-AMPs. In light of AutoML's growing significance in the field of machine learning, a work aims to close the evaluation gap between different approaches for antimicrobial peptide (AMP) categorization [2]. Understanding the biological pathways and entities that particular medications target is a crucial component of drug development as it facilitates the identification of possible treatment techniques and the mechanisms driving disease processes. The creation of AMPs, which are a potential treatment for infections, may benefit greatly from the efficiency and speed with which this computational approach—which is frequently made possible by text mining techniques—identifies drug-gene interactions [3]. Machine learning-based computational techniques are becoming more and more important in predicting human-virus protein-protein interactions (PPIs) to support experimental efforts. Moreover, by applying transfer learning to small target datasets and tasks, previous knowledge from big source datasets and tasks may be leveraged to improve prediction performance [4] greatly. To develop new drugs, treat infections with antibiotics, and

comprehend host-pathogen interactions, peptide classification into AMPs and non-AMPs is essential. Recent years have seen the development of many transformer-based, ML and DL models for peptide classification as a result of advances in AI. However, a comprehensive evaluation of different approaches for AMP classification is lacking. This paper aims to bridge this gap by evaluating and comparing transformer, ML, and DL models for AMP classification. We compare the performance of many algorithms across multiple datasets to identify the optimal technique for accurately distinguishing between AMPs and non-AMPs.

II. LITERATURE SURVEY

The potential of antimicrobial peptides (AMPs) as alternative treatment agents against drug-resistant infections has attracted a lot of interest in recent years. Due to the growing problem of antimicrobial resistance, scientists are working harder than ever to come up with new methods for determining and creating potent AMPs. Using cutting-edge computational techniques, such deep learning and machine learning models, to forecast the properties and roles of these peptides is one viable strategy.

Researchers have looked at a number of approaches, such as deep learning models, to predict the features and operations of AMPs. In one research, the effectiveness of deep learning models such as BERT transformer, Multi-Layer Perceptron (MLP), and Logistic Regression (LR) for AMP prediction was compared with that of conventional amino acid composition techniques [5]. The outcomes showed how deep learning methods might raise the precision of AMP prediction models.

In a different investigation, the minimum inhibitory concentration (MIC) of AMPs against the bacteria Escherichia coli was precisely predicted by the MBC-Attention model, which performed better than traditional machine learning approaches [6]. The minimum inhibitory concentration (MIC) is a critical metric that establishes the minimum amount of an antimicrobial agent needed to stop the development of a certain bacterium. Researchers can minimise possible negative effects and resistance development by optimising the dose and efficacy of AMPs by precise prediction of the MIC.

By employing transfer learning from pretrained protein models to predict antifungal peptide activity, another research project showed performance that was equivalent to that of current methods [7]. Transfer learning is a method that makes use of the information gleaned from one task to improve a model's performance on a related task. Here, the scientists employed pretrained protein models, which were optimised for the particular purpose of antifungal peptide activity after being trained on sizable protein sequence databases. This method made use of the extensive representational information that the pretrained models had amassed, which might enhance the antifungal peptide prediction model's capacity for generalisation and overall effectiveness.

A further research that highlighted the shortcomings of existing predictors in accurately predicting the functional properties of AMPs [8] used the iAMPCN deep-learning approach to identify AMPs and their functional activities. The iAMPCN model demonstrated the ability of deep learning approaches to capture the complex sequence patterns and structural elements that dictate the functional properties of AMPs. It did this by utilising a convolutional neural network architecture specifically designed for the job of AMP prediction.

Researchers also investigated latent spaces for AMP design, looking into five deep learning models, including Transformer, in an effort to address the rising problem of bacterial resistance [9]. This work aimed to use deep learning models' generative skills to produce new AMP sequences with enhanced antibacterial properties. The goal of the study was to find new peptide sequences that could be able to go around established resistance mechanisms by investigating the latent regions that these models learnt. Researchers compared the effectiveness of balanced and unbalanced training sets for antifungal peptide prediction in a different investigation using protein pretrained models [10]. Machine learning algorithms may encounter difficulties and provide biassed predictions when dealing with imbalanced datasets, which include a large underrepresentation of one class (such as active or inactive peptides). The researchers gained insights into the resilience and generalisation capabilities of these models while achieving performance equivalent to state-of-the-art antifungal peptide prediction algorithms by methodically assessing the influence of training set imbalance.

Furthermore, AMPTrans-lstm, a deep generative model, was introduced to facilitate logical AMP design [11]. The goal of this model was to produce a wide range of useful peptides with antibacterial activity against different diseases. Researchers want to speed up the identification and development of new antimicrobial agents (AMPs) and maybe produce more potent therapies for drug-resistant illnesses by utilising the capabilities of deep learning and generative modelling approaches.

TransImbAMP, a different deep learning system, combined transformer architecture with imbalanced multi-label learning in an effort to increase the precision of antimicrobial peptide prediction [12]. By using specific training methods designed for multi-label classification issues, this method addressed the issue of unbalanced datasets, where specific antimicrobial activity or targets may be underrepresented. TransImbAMP aimed to deliver more accurate and dependable forecasts of antimicrobial peptide activity by combining the benefits of transformer models with strategies for managing class imbalance. In a different paper, a novel approach to AMP prediction based on sequence multidimensional representation was devised [13], going beyond earlier methods. This new method made use of sophisticated representation techniques to capture the complex connections between the antibacterial properties of amino acid sequences. The goal of the study was to improve the AMP prediction models' interpretability and predictive capability by integrating multidimensional representations of peptide sequences. This might lead to the discovery of new information about the structural and functional factors that influence antimicrobial activity.

The AMPDeep model [14] tackled the issue of sparse

data, which might impede the efficacy of machine learning models. This approach overcomes the restrictions of limited or unbalanced datasets by utilising transfer learning techniques to harness information from similar tasks or domains. Through the application of transfer learning, AMPDeep was able to surpass earlier studies in the prediction of AMP hemolytic activity—a critical characteristic that establishes the peptides' potential toxicity towards human red blood cells.

Moreover, machine learning approaches were applied, accounting for unique characteristics of the target microbe, to improve predictions of AMP activity against certain bacteria [15]. This method acknowledged that in order to properly forecast the efficacy of AMPs, it is critical to take into account the distinct properties of the pathogenic bacteria, including cell wall composition, virulence factors, and resistance mechanisms. In order to create more specialised and focused treatment approaches, researchers incorporated information unique to bacteria into the prediction models. When HydrAMP beat earlier techniques in peptide synthesis challenges, it proved the potential of this novel generative model for antimicrobial peptide discovery [16]. HydrAMP created new AMP sequences with desired characteristics, such increased stability, decreased toxicity, and better antibacterial efficacy, by using cutting-edge generative modelling approaches. HydrAMP sought to speed up the identification and refinement of AMPs by utilising generative models, which might result in more powerful and focused antimicrobial treatments.

Furthermore, current developments in AMP design and discovery that make use of both traditional machine learning and deep learning methods were described, acknowledging the limitations and challenges that persist [17]. This thorough analysis emphasised the benefits and drawbacks of several computational techniques, highlighting the necessity of ongoing innovation and development to address the difficulties pertaining to AMP prediction and design.

In contrast, a different study found that machine learning models such as CatBoost outperformed deep learning techniques for AMP prediction [18]. This research cast doubt on the widely held belief that deep learning methods are always better and made clear how crucial it is to carefully consider which modelling techniques are most appropriate for a given job or set of datasets.

The suggestion of an AI-based model for predicting the minimum inhibitory concentration (MIC) of antibacterial peptides against ESKAPEE pathogens [19] offered additional proof of the validity of prediction models. Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, and Enterobacter species are among the drug-resistant bacterial species referred to by the abbreviation ESKAPEE. The suggested AI model sought to aid in the development of efficient antimicrobial therapeutics against drug-resistant illnesses by precisely predicting the minimum inhibitory concentration (MIC) of antibacterial peptides against these difficult pathogens.

Classifying AMPs with sequence similarity methods yielded high area under the curve (AUC) values, a statistic frequently used to assess the effectiveness of binary classification models [20]. In order to capture the complex links between peptide sequences and their antimicrobial properties, our strategy combines the benefits of sequence similarity approaches with Support Vector Machines (SVMs), a potent family of machine learning algorithms. Through the achievement of high AUC values for both microbe-specific and general antibacterial action, this work illustrated the potential benefits of combining domain-specific information with sophisticated machine learning approaches to enhance AMP prediction.

As demonstrated by a study that found mistakes in these models for AMP prediction, deep learning models do not always perform better than shallow models [21]. This resulted in a study that highlighted the significance of closely assessing the resilience and performance of deep learning models, in addition to investigating alternate modelling methods, such shallow machine learning techniques, which could provide comparable or even better performance under certain conditions. A deep multi-task learning model was developed to predict the antibacterial activity of spider venom peptides in order to tackle the problem of species-specific antimicrobial activity [22]. There is evidence that a wide variety of peptides with possible antibacterial qualities can be found in spider venom. Nevertheless, these peptides' activities can range greatly between various microbial species. The suggested model sought to concurrently predict the antibacterial activity of spider venom peptides against many target species by utilising a multi-task learning technique. Through the use of shared representations acquired from many tasks (such as activity against various microbial species), this method enhanced the model's overall prediction performance and generalisation abilities. Collaborative filtering and link prediction algorithms were also looked at in order to forecast antimicrobial efficacy for untested peptide-drug combos, in an effort to get over the restrictions of constrained AMP datasets [23]. Utilising patterns of similarity between users and objects, collaborative filtering is a popular approach in recommender systems that generates personalised suggestions. Collaborative filtering was used in the AMP prediction context to find possibly useful peptide-drug combos based on the reported antibacterial activity of comparable peptides or medications. In contrast, link prediction sought to deduce, from the observable links in the data, the presence of unknown linkages (such antimicrobial activity) between entities (pathogens and peptides). Combining these methods allowed researchers to possibly speed up the identification and development of new antimicrobial medicines by overcoming the sparsity of existing AMP data and enabling more precise forecasts of antimicrobial effectiveness.

In a research that compared computer prediction models for linear cationic antimicrobial peptides against both Gram-positive and Gram-negative bacteria, the Random Forest model performed better [24]. Since Random Forest is resilient, interpretable, and can handle complicated, non-linear interactions, it is a commonly utilised ensemble learning approach across a wide range of fields. Random Forest mixes many decision trees. This work showed how well the Random Forest model

predicted the antimicrobial activity of linear cationic peptides, a type of AMPs distinguished by positively charged amino acid residues and a linear shape. The Random Forest technique demonstrated potential for the effective screening and creation of linear cationic AMPs with strong activity against a broad spectrum of bacterial pathogens by surpassing existing computer models.

The creation of innovative and efficient antimicrobial treatments has emerged as a critical global health goal as the threat posed by antibiotic resistance keeps growing. Antimicrobial peptides (AMPs) have become attractive options because of their distinct modes of action, broad-spectrum efficacy, and perhaps decreased susceptibility to the development of resistance. However, because of the intricate interactions between peptide sequence, structure, and antimicrobial activity, designing and optimising AMPs continues to be difficult. Computational approaches, including as machine learning and deep learning techniques, have been crucial in improving our comprehension and capacity for prediction of AMP functions and properties. By using the abundance of information on peptide sequences, antimicrobial activity, and structural characteristics, these methods have created prediction models that may direct the logical creation and identification of new AMPs.

The capacity of machine learning and deep learning models to identify complex patterns and correlations in data that may not be immediately obvious using more conventional analytical techniques is one of its main advantages. Through the use of extensive datasets including peptide sequences and their corresponding antimicrobial activity, these models are capable of revealing the fundamental sequence-structure-function correlations that control the effectiveness and selectivity of AMPs. Furthermore, other sequence-based tasks, such as protein structure prediction and functional annotation, have shown impressive results for deep learning approaches like transformer models and convolutional neural networks. Their suitability for the task of AMP prediction and design stems from their capacity to autonomously develop hierarchical representations and grasp long-range relationships within sequences.

In addition to predictive modelling, the de novo creation of AMPs has demonstrated the potential of generative deep learning techniques. Through the utilisation of generative models, scientists may investigate the extensive sequence space and produce unique peptide sequences possessing advantageous antibacterial characteristics. This strategy might expedite the identification of novel AMPs and get beyond the drawbacks of conventional optimisation and screening techniques. Nonetheless, there are several difficulties in applying deep learning and machine learning methods to the field of AMP design and prediction. The scarcity of well-curated, high-quality datasets is a major barrier. Data shortage or imbalance problems may arise from the time- and resource-intensive nature of AMP data production and experimental validation. In order to overcome this difficulty, scholars have investigated a number of approaches, including collaborative filtering, data augmentation, and transfer learning, which make use of expertise from adjacent fields to fill in the gaps in data.

The robustness and interpretability of these models provide another difficulty. Despite the impressive predictive accuracy that deep learning models have shown, their internal workings are frequently opaque and challenging to understand. The process of converting model predictions into useful information for AMP design and optimisation may be hampered by this lack of interpretability. Concerns concerning deep learning models' dependability and resilience in practical applications are also raised by the possibility that they are vulnerable to adversarial assaults or that they do not generalise effectively to new or out-of-distribution data.

Researchers have looked at alternate modelling strategies, such as ensemble methods, kernel-based methods, and hybrid models that combine the best features of many approaches, in an effort to lessen these difficulties. In order to increase the models' interpretability and generalisation abilities, further efforts have been made to include domain-specific knowledge, such as structural details and physicochemical parameters, into the modelling process. Looking ahead, the synergistic combination of computational techniques, experimental validation, and domain expertise is expected to propel future developments in the field of AMP prediction and design. Machine learning and deep learning models are expected to be more and more important in hastening the discovery and development of novel antimicrobial therapies as the amount of high-quality data becomes more readily available and our comprehension of the underlying mechanisms governing antimicrobial activity grows.

Transforming the findings from these computational tools into practical therapeutic treatments would need interdisciplinary cooperation among microbiologists, pharmacological experts, and computer researchers. Furthermore, by incorporating these predictive models into experimental workflows and drug discovery pipelines, AMP development may be streamlined and the time and resources needed for screening and optimisation can be decreased.

In conclusion, there is great potential for addressing the worldwide challenge of antimicrobial resistance through the use of machine learning and deep learning approaches to the field of antimicrobial peptide prediction and design. Through the utilisation of data-driven strategies and the most recent developments in computational techniques, scientists can open up new directions in the logical development and identification of effective and targeted antimicrobial treatments. To ensure that computational approaches play a key role in the development of successful antimicrobial therapies in the future, however, continued efforts are required to address the issues of data scarcity, model interpretability, and translational hurdles.

III. METHODOLOGY

This study's technique was painstakingly designed to tackle the categorization of AMPs using a multimodal approach that included model selection, training, assessment, and data preparation. Firstly, the dataset consisting of AMPs sourced

from the Antimicrobial Peptide Database (APD3) and the Database of Anuran Defense Peptides (DADP), as well as non-AMPs sampled from the UniProtKB/Swiss-Prot database, was collected. These sequences were then preprocessed to standardize their format for model training.

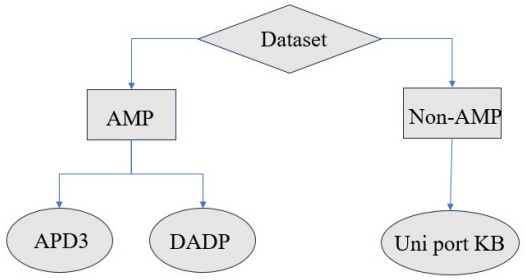

Fig. 1. Dataset

Fig 1 represents the split of dataset into antimicrobial and non-antimicrobial peptide, The preprocessing steps included normalization, where all sequences were changed to uppercase letters to ensure uniformity. Additionally, every peptide sequence was truncated to a unchanging length of 22 characters and padded with special characters (#) to match the desired length.

$$\text{One-hot encoding: Each amino acid} = [0, 0, ..., 1, ..., 0] \tag{1}$$

Equation 1 suggests that a one-hot encoding scheme was applied to represent each amino acid in the peptide sequences, using the above mathematical equation. Once the dataset was preprocessed, various ML, DL, and transformer-based models were trained for AMP classification. Traditional ML models were trained using TF-IDF vector representations of the peptide sequences. Deep learning models, including LSTM and GRU networks, were utilized for learning sequential patterns from the peptide sequences.

$$\text{Backpropagation: } \frac{\partial L}{\partial w} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial h} \cdot \frac{\partial h}{\partial w} \tag{2}$$

$$\text{Gradient Descent: } w_{t+1} = w_t - \alpha \cdot \frac{\partial L}{\partial w} \tag{3}$$

The equation 2 and 3 suggests that the models used, were trained using RNNs and optimized using backpropagation and gradient descent algorithms. Additionally, transformer-based models such as DistilBERT and RoBERTa were trained using pre-trained transformer architectures specifically designed for sequence classification tasks.

Attention Mechanism

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

Equation 4 suggests that these models leverage attention mechanisms to capture long-range dependencies in the peptide sequences. Following training, conventional assessment measures including as accuracy, precision, recall, and F1-score were used to assess the models' performance. The training dataset was divided into testing and training sets, and the performance of the trained models' classification was evaluated on the testing dataset. The model predictions were compared to the ground truth labels in the testing dataset to determine the assessment measures.

The performance of the ML, DL, and transformer-based models was compared based on their accuracy scores and other relevant evaluation metrics. Additionally, considerations such as computational efficiency and scalability were taken into account to provide insights into the practical feasibility of each approach.

Fig. 2. Methodology

In Fig 2 the methodology employed a systematic approach to data preprocessing, model training, and evaluation, utilizing a diverse range of models to perform AMP classification. The use of mathematical equations illustrated the underlying principles of model architectures and optimization algorithms, enhancing the understanding of the methodology employed in this study.

IV. RESULTS AND DISCUSSION

In this study, we employ a diverse set of evaluation metrics to comprehensively assess the performance of our classification models in the domain of biological sequence analysis. Each evaluation metric offers unique insights into different aspects of model performance, providing valuable information for refining and optimizing our classification approach.

TABLE I
MODEL EVALUATION METRICS

Model	Accuracy	Precision	Recall	F1-Score	AUC
DistilBERT	0.844	0.856	0.847	0.851	0.916
RoBERTa	0.889	0.876	0.920	0.898	0.940
Naive Bayes	0.865	0.902	0.835	0.867	0.867
Logistic Regression	0.847	0.893	0.807	0.848	0.850
Random Forest	0.847	0.908	0.790	0.845	0.851
LSTM	0.757	0.819	0.693	0.751	0.761
GRU	0.746	0.858	0.619	0.719	0.753

Table 1 includes information on how well each model performs in terms of recall, accuracy, precision, F1 score, and area under the ROC curve. Transformer models—in particular,

RoBERTa—clearly outperform other models in the majority of measures, demonstrating their efficacy in your classification task. While deep learning models—particularly the GRU model—perform comparably to other models, machine learning algorithms also demonstrate competitive performance.

As a basic indicator of the overall performance of the model, accuracy shows the percentage of examples correctly identified in each class. High accuracy indicates that our algorithms correctly forecast across several categories and efficiently capture the underlying patterns in biological sequences. For our study, precision is very important since it gauges how reliable positive forecasts are. When it comes to situations where false positives might have serious repercussions, like misdiagnosis or false alarms, a high precision means that most positive forecasts are true. The model's recall measures how well it can recognise every positive case among all the real positive examples. This parameter is critical to ensure that our algorithms reliably identify significant biological sequences, especially in situations when the absence of positive occurrences might have dire consequences. The F1 Score provides a fair evaluation of our models' performance by integrating recall and accuracy into a single statistic. It offers a thorough assessment that takes into account both false positives and false negatives, and it is appropriate for situations in which striking a balance between recall and precision is required. Accurate categorization of biological sequences depends on the model's ability to differentiate between distinct classes, which is shown by a higher AUC-ROC. Together, these many assessment indicators provide us a thorough grasp of our model's performance, allowing us to pinpoint areas for development and boost its efficiency in biological sequence categorization tasks.

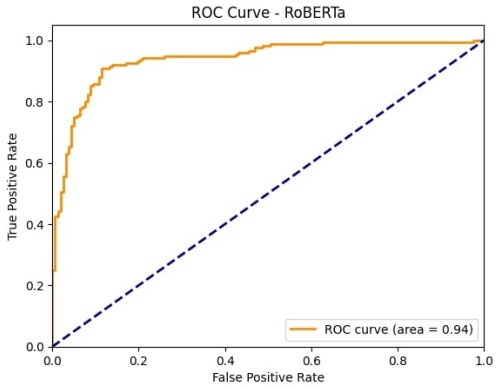

Fig. 3. ROC Curve - Transformers

Figure 3 displays that the models shows a good capacity to discriminate between positive and negative classes for transformers (e.g., DistilBERT and RoBERTa), with a ROC curve area of 0.94. This shows that the transformer-based models produce strong predictions because they are very good at capturing intricate patterns and connections in the data.

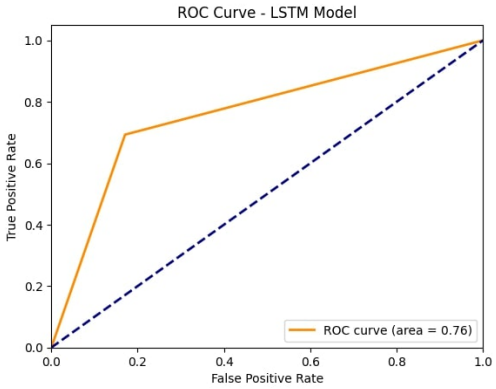

Fig. 4. ROC Curve - Deep Learning Model

Figure 4 shows that in comparison to transformers, the performance of DL models is somewhat poorer, with an ROC curve area of 0.76. The reduced ROC curve region indicates that while DL models still show some discriminating strength, they may have more difficulty identifying complex patterns in the data or extrapolating to cases that haven't been observed before. This can be the result of inadequate training data or restrictions in the model design.

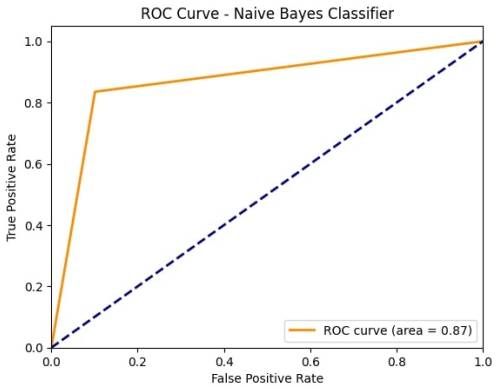

Fig. 5. ROC Curve - Machine Learning Models

Figure 5 suggests that with an ROC curve area of 0.87, ML models perform in between DL models and transformers. Compared to DL and transformers, ML models usually rely on manually created features and less complex methods. Even though ML models might not be as good at capturing intricate correlations as transformers, they can nevertheless perform competitively, particularly when working with tabular or structured data.

When compared to DL and ML models, transformers outperform them in collecting intricate patterns and producing precise predictions, as seen by their larger ROC curve regions. All models' performances, nevertheless, ought to be weighed against the project's needs and particular application.

Important insights into the advantages and disadvantages of various modeling techniques may be gained by compar-

ing and contrasting them. Transformer-based models perform better than other models in identifying intricate patterns and relationships in peptide sequences, which makes them ideal for applications involving the categorization of AMPs. On the other hand, transformer-based models are marginally more accurate than typical machine learning models, but they are simpler and use less computing power. For some applications, DL models are good substitutes, since they balance performance and complexity.

The discipline of bioinformatics and drug development will be impacted by the study's conclusions in a number of ways. Transformer-based models have demonstrated a high degree of accuracy, indicating its potential for expediting the discovery and characterisation of new antimicrobial drugs. Potential avenues for future study might entail investigating ensemble techniques, which integrate the advantages of many modelling approaches to further improve classification performance. Furthermore, enlarging and diversifying the dataset may enhance the models' capacity for generalization and make it easier to use them in practical contexts.

This paper offers insightful information on the functionality and comparison of transformer-based, ML, and DL models for AMP classification. Although transformer-based models exhibit better performance and accuracy, conventional ML and DL models provide different strategies, each with a special set of benefits. When aware of the advantages and disadvantages of every strategy, researchers can choose the best model for a given AMP classification problem with knowledge.

V. CONCLUSION AND FUTURE SCOPE

Our thorough examination of many classification strategies for separating AMPs from non-AMPs has produced insightful information on the state of bioinformatics and drug development computational techniques. Of all the approaches evaluated, RoBERTa performed very well, demonstrating its ability to classify sequences with an astounding accuracy rate of 86.53%. This result emphasises how important it is to use pretrained transformer structures, which have proven to be extremely effective at capturing complex patterns and representations seen in biological sequences. Apart from RoBERTa, our research also highlighted the robustness and effectiveness of traditional ML models, such Logistic Regression and Naive Bayes. These models demonstrated competitive performance, with accuracy ratings ranging from 84.43% to 86.53%, despite their ease of use and interpretability. This implies that although cutting-edge DL architectures have attracted a lot of interest, conventional ML techniques are still strong competitors, especially in situations where computing efficiency and transparency are critical. Furthermore, we investigated the capability of DL models, such as GRU and LSTM, in AMP classification tasks. DL models were not as accurate as RoBERTa and ML models, but they were still able to distinguish between AMP and non-AMP sequences with some degree of accuracy because of their sophisticated comprehension of sequential data and temporal connections. Even while DL and ML models perform well, it's important to recognise the accompanying difficulties, such overfitting and disappearing gradients. When faced with small datasets, DL models in particular may have problems that need to be properly mitigated by thorough regularisation approaches and hyperparameter adjustment.

In terms of the future, our work opens up a number of research directions. Increasing the range and variety of datasets may improve the generalisation and resilience of the model in different biological environments. Furthermore, investigating ensemble methods that capitalise on the complementing advantages of several modelling paradigms may result in increased classification resilience and accuracy. Lastly, there is a strong chance to improve model performance and adaptability to domain-specific issues by investigating transfer learning techniques, such as optimising pretrained models on AMP-specific tasks. Large-scale language models' pretrained representations can speed up learning and make it easier to apply information from related domains to the current task. To sum up, our research provides an extensive analysis of several computational methods for classifying AMPs, illuminating their advantages, disadvantages, and potential areas for further investigation. Through the combined use of deep learning architectures, transformer-based models, and conventional ML techniques, we may further progress the creation of efficient antibacterial medications and add to the larger field of computational biology and drug discovery.

REFERENCES

[1] Rangarajan, P.K., Gurusamy, B.M., Rajasekar, E. et al. Retroactive data structure for protein–protein interaction in lung cancer using Dijkstra algorithm. Int. j. inf. tecnol. 16, 1239–1251 (2024). https://doi.org/10.1007/s41870-023-01557-4

[2] C. Spandana, I. V. Srisurya, S. Aasha Nandhini, R. P. Kumar, G. Bharathi Mohan and P. Srinivasan, "An Efficient Genetic Algorithm based Auto ML Approach for Classification and Regression," 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 2023, pp. 371-376, doi: 10.1109/IDCIoT56793.2023.10053442. keywords: Radio frequency;Productivity;Machine learning;Forestry;Network architecture;Iterative methods;Internet of Things;AutoML;Genetic Algorithm;Possum Dataset;Binary Classification;Regression;Image Classification;Fitness Function;Random Forest (RF),

[3] Anand, S., Iyyappan, O.R., Manoharan, S., Anand, D., Jose, M.A., Shanker, R.R. (2022). Text Mining Protocol to Retrieve Significant Drug–Gene Interactions from PubMed Abstracts. In: Raja, K. (eds) Biomedical Text Mining. Methods in Molecular Biology, vol 2496. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-2305-3_2

[4] I. R. Oviya, S. Sravya N and K. Raja, "R2V-PPI: Enhancing Prediction of Protein-Protein Interactions Using

Word2Vec Embeddings and Deep Neural Networks," 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2024, pp. 1-7, doi: 10.1109/ICAECT60202.2024.10469595.

[5] Lobanov, M. Y., Slizen, M. V., Dovidchenko, N. V., Panfilov, A. V., Surin, A. A., Likhachev, I. V., Galzitskaya, O. V. Comparison of deep learning models with simple method to assess the problem of antimicrobial peptides prediction. Molecular Informatics, 2200181. https://doi.org/10.1002/minf.202200181

[6] Yan J, Zhang B, Zhou M, Campbell-Valois F, Siu SWI. 2023. A deep learning method for predicting the minimum inhibitory concentration of antimicrobial peptides against Escherichia coli using Multi-Branch-CNN and Attention. mSystems 8:e00345-23. https://doi.org/10.1128/msystems.00345-23

[7] Lobo, F.; González, M.S.; Boto, A.; Pérez de la Lastra, J.M. Prediction of Antifungal Activity of Antimicrobial Peptides by Transfer Learning from Protein Pretrained Models. Int. J. Mol. Sci. 2023, 24, 10270. https://doi.org/10.3390/ijms241210270

[8] Jing Xu, Fuyi Li, Chen Li, Xudong Guo, Cornelia Landersdorfer, Hsin-Hui Shen, Anton Y Peleg, Jian Li, Seiya Imoto, Jianhua Yao, Tatsuya Akutsu, Jiangning Song, iAMPCN: a deep-learning approach for identifying antimicrobial peptides and their functional activities, Briefings in Bioinformatics, Volume 24, Issue 4, July 2023, bbad240, https://doi.org/10.1093/bib/bbad240

[9] 1. Renaud S, Mansbach R. Latent Spaces for Antimicrobial Peptide Design. ChemRxiv. 2023; doi:10.26434/chemrxiv-2022-m3900-v2 This content is a preprint and has not been peer-reviewed.

[10] Li, C., Warren, R.L. Birol, I. Models and data of AMPlify: a deep learning tool for antimicrobial peptide prediction. BMC Res Notes 16, 11 (2023). https://doi.org/10.1186/s13104-023-06279-1

[11] Mao, Jiashun, Guan, Shenghui, Chen, Yongqing, Zeb, Amir, Sun, Qingxiang, Lu, Ranlan, Dong, Jie, Wang, Jianmin, Cao, Dongsheng. Application of a deep generative model produces novel and diverse functional peptides against microbial resistance. Computational and Structural Biotechnology Journal,2023, https://doi.org/10.1016/j.csbj.2022.12.029

[12] Yuxuan Pang, Lantian Yao, Jingyi Xu, Zhuo Wang, Tzong-Yi Lee, Integrating transformer and imbalanced multi-label learning to identify antimicrobial peptides and their functional activities, Bioinformatics, Volume 38, Issue 24, 15 December 2022, Pages 5368–5374, https://doi.org/10.1093/bioinformatics/btac711

[13] Dong B, Li M, Jiang B, Gao B, Li D and Zhang T (2022) Antimicrobial Peptides Prediction method based on sequence multidimensional feature embedding. Front. Genet. 13:1069558. doi: 10.3389/fgene.2022.1069558

[14] Salem, M., Keshavarzi Arshadi, A. Yuan, J.S. AMPDeep: hemolytic activity prediction of antimicrobial peptides using transfer learning. BMC Bioinformatics 23, 389 (2022). https://doi.org/10.1186/s12859-022-04952-z

[15] Chem. Inf. Model. 2023, 63, 6, 1723–1733 Publication Date:March 13, 2023 https://doi.org/10.1021/acs.jcim.2c01551

[16] Szymczak, P., Możejko, M., Grzegorzek, T. et al. Discovering highly potent antimicrobial peptides with deep generative model HydrAMP. Nat Commun 14, 1453 (2023). https://doi.org/10.1038/s41467-023-36994-z

[17] Yan, J.; Cai, J.; Zhang, B.; Wang, Y.; Wong, D.F.; Siu, S.W.I. Recent Progress in the Discovery and Design of Antimicrobial Peptides Using Traditional Machine Learning and Deep Learning. Antibiotics 2022, 11, 1451. https://doi.org/10.3390/antibiotics11101451

[18] J. -C. Yu, K. Ni and C. -T. Chen, "Sequence-based Prediction of Antimicrobial Peptides with CatBoost Classifier," 2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE), Taichung, Taiwan, 2022, pp. 217-220, doi: 10.1109/BIBE55377.2022.00053.

[19] R. Sharma, S. Shrivastava, S. K. Singh, A. Kumar, A. K. Singh and S. Saxena, "Artificial intelligence-based model for predicting the minimum inhibitory concentration of antibacterial peptides against ESKAPEE pathogens," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2023.3271611.

[20] Redshaw, Joseph, Ting, Darren S. J., Brown, Alex, Hirst, Jonathan D., Gärtner, Thomas. Krein support vector machine classification of antimicrobial peptides, 2023, Digital Discovery, 10.1039/D3DD00004D

[21] César R García-Jacas, Sergio A Pinacho-Castellanos, Luis A García-González, Carlos A Brizuela, Do deep learning models make a difference in the identification of antimicrobial peptides?, Briefings in Bioinformatics, Volume 23, Issue 3, May 2022, bbac094, https://doi.org/10.1093/bib/bbac094

[22] Lee B, Shin MK, Yoo JS, Jang W and Sung J-S (2022) Identifying novel antimicrobial peptides from venom gland of spider Pardosa astrigera by deep multi-task learning. Front. Microbiol. 13:971503. doi: 10.3389/fmicb.2022.971503

[23] J. Chem. Inf. Model. 2023, 63, 12, 3697–3704, Publication Date:June 12, 2023, https://doi.org/10.1021/acs.jcim.3c00137

[24] Söylemez, Ü.G.; Yousef, M.; Kesmen, Z.; Büyükkiraz, M.E.; Bakir-Gungor, B. Prediction of Linear Cationic Antimicrobial Peptides Active against Gram-Negative and Gram-Positive Bacteria Based on Machine Learning Models. Appl. Sci. 2022, 12, 3631. https://doi.org/10.3390/app12073631