Combining ML and DL Approaches to Enhance Antimicrobial Peptide Discovery and Effectiveness

Ridit Jain[1], Lavanya Anand[1], Sanskaar Kushwaha[1]
[1] BML Munjal University, Computer Science Department, Gurgaon, Haryana, India

Contributing authors: ridit.jain.21cse@bmu.edu.in; lavanya.anand.21cse@bmu.edu.in; sanskaar.kushwaha.21cse@bmu.edu.in

Abstract

Antimicrobial Resistant (AMR) poses a rising threat to global health bolstered by decreasing efficacy of the conventional antibiotics. There is therefore a high demand for alternative antimicrobials, such as antimicrobial peptides (AMPs), which have unique modes of action and are less likely to induce resistance. Nonetheless, AMPs that work are difficult to predict since their interactions with microbial membranes are complex. Herein, an integrated machine learning (ML) and deep learning (DL) approach was presented that utilizes both sequence and structural data to predict the AMP effectiveness. This paper explores several computational models, which includes support vector machines (SVM) and long short-term memory (LSTM) networks, to generate and evaluate AMPs. The models are further validated by experimental techniques such as small-angle X-ray scattering (SAXS) and killing assays . This work will bridge the knowledge gap between theoretical prediction and practical utilization of AMPs, contributing to new paradigms for designing new aspects of AMP targets and their actual performance. By fusing ML and DL techniques, this study enhances the identification and development of AMPs, contributing to the ongoing fight against antibiotic resistance.

Keywords Antimicrobial Peptides, Antimicrobial Resistance, Statistics, Computational Modeling

1 Introduction

Antimicrobial Resistance (AMR) is an emerging threat to global health, owing to irrational utilization and prescription of antibiotics, particularly well demonstrated and worsened by the recent COVID-19 pandemic. But a growing concern exists, pathogenic bacteria gradually develop resistance to known medications rendering the medical community desperate for the development of new medicines. Antimicrobial peptides (AMPs) have attracted significant interest because of their multiple targets and mechanisms of action and lower propensity for the development of resistance. However, identification and utilization of proper AMPs is a tedious process owed to the number of significant and diverse mechanisms involved and the numerous potential sequences of the peptides.

The integration of machine learning (ML) and deep learning (DL) methods has recently been recognized as a powerful tool in the predictive modeling of AMP effectiveness. These computational methods can screen big amounts of genetic and structural information to indicate potential peptides which may be neglected by the standard biotechnological techniques. Unfortunately, existing models halfheartedly address the problem, and many depend on sequence data or structural information only.

This paper is organized as follows. **Section 1** introduces the context of antimicrobial resistance (AMR), the significance of antimicrobial peptides (AMPs), and the role of integrating machine learning (ML) and deep learning (DL) methods in predicting AMP effectiveness. **Section 2** details the problem statement with an emphasis on weaknesses of existing ways of AMP discovery for exploiting image and sequence data jointly to enhance prediction accuracy. Moving forward in **Section 3** a literature review has been done and analyses of previous research efforts alongwith the methods used in the field have also examined such that reviews on ML and DL based approaches to AMP prediction are accomplished. **Section 4** discusses methodological improvements, proposing enhancements to existing approaches by integrating more complex datasets and advanced computational techniques. Finally, **Section 5** gives us the results and **Section 6** concludes the paper by summarizing the key findings and what they mean.

2 Problem Description

The primary issue is the inefficiency and limited scope of current AMP discovery methods, which may not be able to maximize the synergy of multiple forms of data, especially sequence and structure data. Current methodologies tend to address and manage these data types separately and may leave out latent dependencies which might be very important in understanding or predicting the effectiveness of the AMP. For example, models in a series of certain characteristics may perform exceptionally well in predicting these behaviors, but could fail in transferability of such findings to actual biological efficiency which also functions with the organization in 3D molecules.

The problem is compounded by the sheer diversity of AMPs and the subtle nuances that govern their interactions with microbial cell membranes. The mere fact that an AMP is able to penetrate a target cell does not necessarily guarantee that it will be potent; other issues such as the shape of the peptide, charge distribution, and hydrophobicity of the molecule are as important and more often than not are central to the ability of the peptide to integrate into and disrupt lipid bilayers. Hence, an integrated approach in analyzing both image information input with the help of deep learning approaches and sequence data utilizing machine learning algorithm could be very helpful in increasing the prediction efficiency and accelerate the identification of new AMPs.

This necessitates an innovative approach that leverages the strengths of both machine learning and deep learning methods in to provide a comprehensive picture of the peptide properties, which will fill the gap between the genetic sequence description and structural analysis. Such advancements could serve as stimulus to the next generation of treatments against microbial diseases, which are critical in fighting the increasing problem of antibiotic resistance.

3 Methodology

The exploration of antimicrobial peptides (AMPs) using advanced computational models represents a pivotal shift in how we address the rising issue of antibiotic resistance. The methodologies employed in the three detailed studies—"Mapping Membrane Activity in Undiscovered Peptide Sequence Space Using Machine Learning," "Deep Learning for Novel Antimicrobial Peptide Design," and "AMPlify: Attentive Deep Learning Model for Discovery of Novel Antimicrobial Peptides Effective Against WHO Priority Pathogens"—exemplify this innovative approach. All these studies use diverse methods based on machine learning and deep learning that allow not only predicting the existence of AMPs with possible clinical application but also designing such peptides. These methodologies provide a new approach for the identification of novel peptides and provide a sound computational framework for peptide validation that could in the future facilitate the development of more effective antimicrobial treatments.

3.1 Mapping Membrane Activity in Undiscovered Peptide Sequence Space Using Machine Learning

3.1.1 Dataset

This investigation employs the techniques of Support Vector Machine (SVM) to investigate the possibility of α-helical antimicrobial peptides (AMPs) in a new unresearched sequence territory. The dataset is composed of over 1,100 identified AMPs which increases its variation and provides a wider range of sequences that have a different capacity to interact with microbial membranes. This diversity will helpful for the improvement of the understanding of what kind of peptides are membrane-active and which of structural and functional characteristics correlate with the desired effectiveness of the peptide when interacting with the target membrane of the pathogen. The versatility of biological activities in the dataset makes it possible to extend the training procedure of a model and identify significant characteristics and features needed for antimicrobial effects.[1][2]

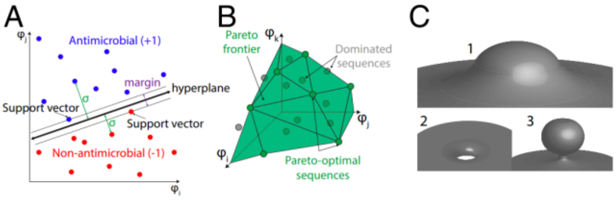


*Figure 1 The study utilizes SVM learning and Pareto optimization to identify peptides that are antimicrobial and generate membrane curvature. **(A)** The schematic illustrates the usage of an SVM binary classifier for separation of hypothetical antimicrobial peptide sequences (blue circles) characterized by two descriptors {ϕi, ϕj} from a non-antimicrobial sequences. Sequences lying on the margins serve as support vectors, while separating hyperplane is positioned equidistant between them. The metric σ (green arrows) measures each peptide's distance to the hyperplane, with positive values indicating antimicrobial sequences and negative values indicating non-antimicrobial ones. **(B)** The schematic depicts the differentiation between Pareto-optimal sequences and dominated sequences within an random 3D descriptor subspace. The Pareto frontier is a hypersurface that encompasses the Pareto-optimal sequences. **(C)** The panel shows normally known biological phenomena associated with generation of negative Gaussian curvature in cell membranes, including (1) membrane blebbing, (2) pore formation, and (3) scission and budding. [1]*

3.1.2 Model Architecture

The SVM model is designed with much attention to distinguish the peptides on the basis of their physiochemical characteristics important for their membrane activity. This model is designed to select peptides that not only have the properties of AMPs but also are likely to act as novel antimicrobial agents. The optimisation process refers to the fine-tuning of multiple factors that would improve the 'antimicrobialness' of the peptides, or in other words, the following factors; Optimal helicity and minimum mutation steps from natural AMPs. This particular configuration of the SVM makes it possible for the model to examine a literally huge amount of peptide sequences to identify the ones that can exhibit reasonable antimicrobial potential.

3.1.3 Feature Selection and Optimization

Feature selection is also a central role in this study as it helps in identifying the right physicochemical features that can help in determining the efficacy of AMPs. Contained in this process is the use of a recursive procedure to determine which features best enhance the accuracy of SVM's prediction. With reference to the features that provide maximum distinction to the data set, SVM's parameters such as kernel type, as well as the penalty parameters are tuned in the optimization process to enhance the sensitivity and specificity of the model. The classification imbalance, which is common in the majority of the biological data set, is also efficiently controlled when tuning the SVM, and balanced class distribution comes as one more advantage for the real data peptide categorization.

3.1.4 Validation and Experimental Techniques

The results from computing models in the SVM are then thoroughly validated by a series of robust experimental techniques, such as small-angle X-ray scattering (SAXS) and killing assays. These methods are rather useful in determining the antimicrobial activity and the interaction with membrane of the predicted peptides, as it ensures theoretical predictions of the model to be proved. This validation is of core importance as this proves that the peptides derived by the SVM, are not only merely existent in theoretical plane but also can be practically applied as antimicrobial

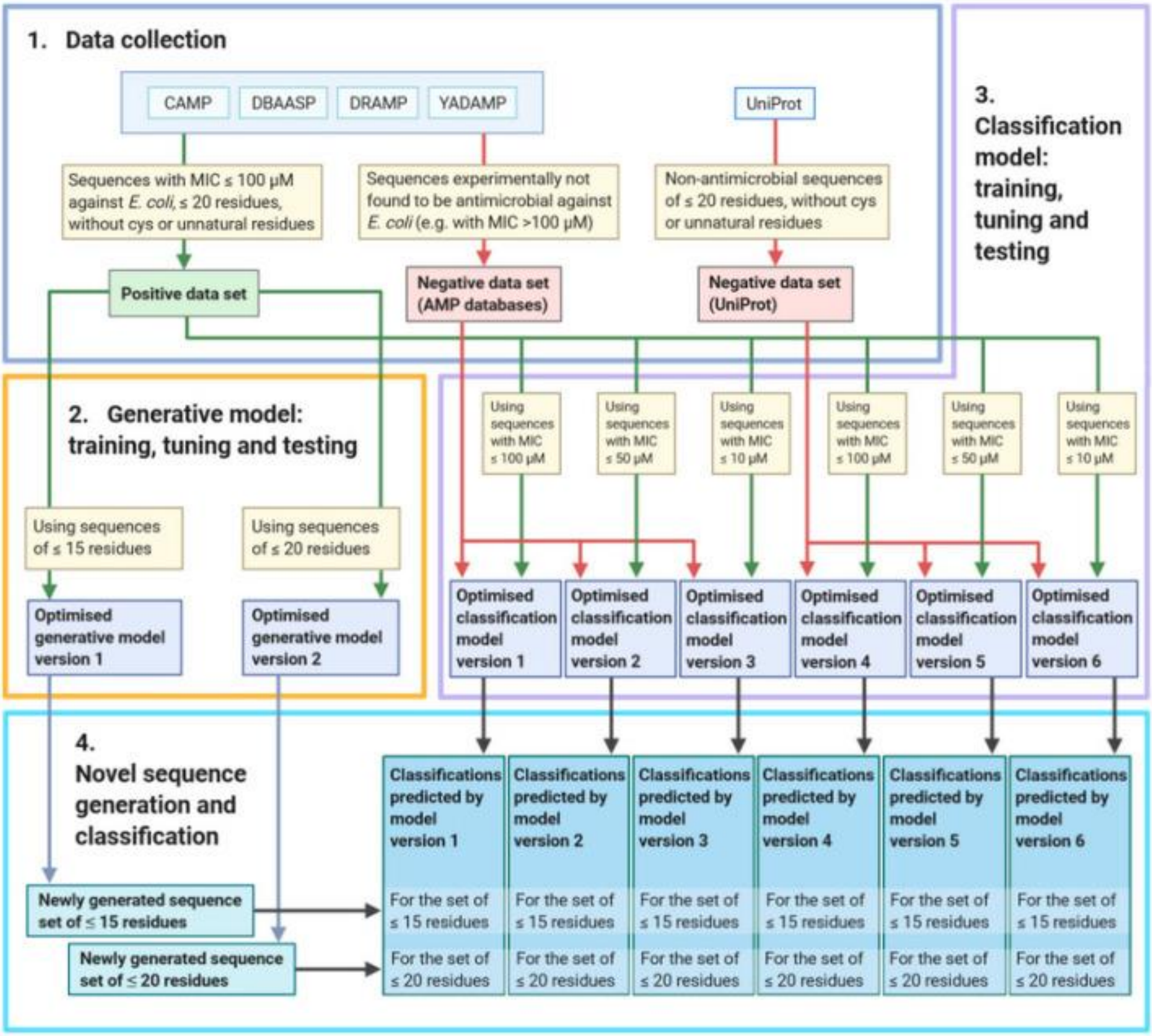

1. Data collection
CAMP
DBAASP
DRAMP
YADAMP
UniProt
Sequences with MIC ≤ 100 μM against E. coli, ≤ 20 residues, without cys or unnatural residues
Sequences experimentally not found to be antimicrobial against E. coli (e.g. with MIC >100 μM)
Non-antimicrobial sequences of ≤ 20 residues, without cys or unnatural residues
Positive data set
Negative data set (AMP databases)
Negative data set (UniProt)
3. Classification model: training, tuning and testing
2. Generative model: training, tuning and testing
Using sequences of ≤ 15 residues
Using sequences of ≤ 20 residues
Optimised generative model version 1
Optimised generative model version 2
Using sequences with MIC ≤ 100 μM
Using sequences with MIC ≤ 50 μM
Using sequences with MIC ≤ 10 μM
Using sequences with MIC ≤ 100 μM
Using sequences with MIC ≤ 50 μM
Using sequences with MIC ≤ 10 μM
Optimised classification model version 1
Optimised classification model version 2
Optimised classification model version 3
Optimised classification model version 4
Optimised classification model version 5
Optimised classification model version 6
4. Novel sequence generation and classification
Newly generated sequence set of ≤ 15 residues
Newly generated sequence set of ≤ 20 residues
Classifications predicted by model version 1
Classifications predicted by model version 2
Classifications predicted by model version 3
Classifications predicted by model version 4
Classifications predicted by model version 5
Classifications predicted by model version 6
For the set of ≤ 15 residues
For the set of ≤ 20 residues

*Figure 2 Pipeline for the ML process. **(1)** Data for the positive set was obtained from antimicrobial peptide (AMP) databases, including the Collection of Anti-Microbial Peptides (CAMP), Database of Antimicrobial Activity and Structure of Peptides (DBAASP), Data Repository of Antimicrobial Peptides (DRAMP), and Yet Another Database of Antimicrobial Peptides (YADAMP). Two negative data sets were sourced: one from AMP databases and another from UniProt. **(2)** Two versions of the generative model were developed, fine-tuned, and tested using the positive data set. **(3)** Six variations of the classification model were trained, fine-tuned, and tested on the positive data set along with either the UniProt or AMP negative data set, employing different minimal inhibitory concentration (MIC) thresholds. **(4)** The optimized generative models generated 2 sets of AMP sequences, which were subsequently classed by the 6 optimized versions of the classification model, leading to 12 sets of predictions. (AMP) database Database of Antimicrobial Activity and Structure of Peptides (DBAASP), Data [3][4]*

agents. Moreover it expands on these findings by indicating how these peptides may be synthesized for purposes of a clinical trial in a manner that may lead to the emergence of new treatments for infections that cannot cured with the conventional antibiotics.

3.2 Deep Learning for Novel Antimicrobial Peptide Design

3.2.1 Dataset

In this innovative paper, dual Long Short-Term Memory (LSTM) networks are used for both generation and evaluation of novel antimicrobial peptide (AMP) sequences. This work exploits a large dataset enriched with those peptides that were found to be characterized by antimicrobial activity against many pathogens. The data was collected from various databases, including DRAMP. This dataset is significant because it will allow the LSTM models to learn from a mix of different sequence patterns and their relevant biological activities. Diversity here makes sure that the generative models experience the widest possible range of AMPs characteristics. This is a very important feature of generating accurate new peptides that could exhibit broad-spectrum antimicrobial properties.[3][4]

3.2.2 Model Architecture

The architecture of this study involves the use of a generative LSTM model for creating new peptide sequences, and a classification LSTM for assessing these new sequences for their viability as effective AMPs. This dual approach allows an innovative cycle of discovery and refinement in which newly synthesized peptides are continuously assessed and refined for their efficacy. This dual process enables an innovative cycle of discovery and refinement that ensures that newly synthesized peptides are repeatedly tested for their effectiveness. In this way, the generative model based on deep learning allows for imitating processes of natural formation of peptides and identifying the new promising structures of AMPs that the traditional approach might overlook While the classification model performs the strong selection of peptides with high success potential in passing through all stages of the proposed pipeline.

3.2.3 Hyperparameter Optimization and Iterative Refinement

The research employs Bayesian hyperparameter optimization as a means of improving LSTM models, allowing for a more rigorous approach to comparing numerous configurations in order to select the most suitable values for these models. This automatically optimizes not only the accuracy of the models but their generality across different biological datasets. Iterative refinement across many training, validation, and testing cycles helps to tune the performance of these models further. Using cross-validation, the study ensures that robustness is ensured while overfitting is minimal, thus making the models competent on both known and unknown data. This continuous learning and adaptation are important in keeping the models relevant and effective in the prediction of AMPs.

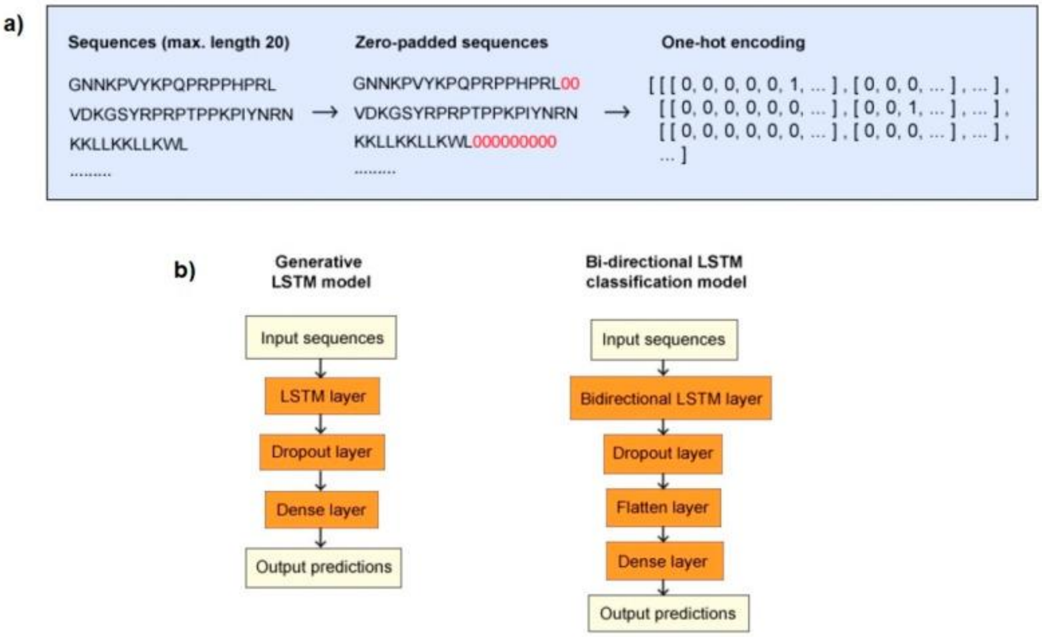


Figure 3 ***(A) Data processing*** *Sequences were limited to a maximum of 20 residues and sequences shorter than 20 residues were zero-padded. The sequences are then one-hot encoded, which makes them appropriate for passing through into the deep learning models.* ***(B) Model Architectures*** *Bearing in mind that overfitting was an issue, the model started with an LSTM input layer, was accompanied by a dropout layer, and ended with a dense layer which provided probabilities of output. Bidirectional LSTM classification model initial layers included a bidirectional LSTM input layer, as well as a dropout layer. A flatten layer was employed to take the shape of the encoded information while a dense layer was employed in making the predictions.[3]*

3.2.4 Evaluation and Validation Techniques

Analysis of the synthesized generated includes not only computational predictions but also empirical biochemical tests that evaluate the antimicrobial activity of peptides against actual pathogens. This extensive evaluation is required as it confirms the biological activity of the peptides, leading them to be a valid candidate for further progress and clinical studies. The usage of both in silico and in vitro methods proves to be complete validation of the functional capacity of the peptides, hence strengthening the validity of the computational models used during their research.

3.3 AMPlify: Attentive Deep Learning Model for Discovery of Novel Antimicrobial Peptides Effective Against WHO Priority Pathogens

3.3.1 Dataset

AMPlify implements bidirectional LSTM networks with elaborate attention to discover AMPs from genomic data, with an emphasis on peptides that are known to be effective against high-priority pathogens identified by World Health Organization (WHO). The dataset has a wide variation of genomic sequences that, in turn forms a broad base that the model relies on to identify and predict novel AMPs. With this dataset, the model would ensure that it is exposed to vast genetic ranges, meaning peptides with unique and potent antimicrobial properties will be recognized and detected with increased prevalence.[5][6]

3.3.2 Model Architecture

The integration of bidirectional LSTMs with attention mechanisms is designed to maximize the model's predictive accuracy. This bidirectional approach starts reading the genetic sequence in the forward direction and then in the backward direction and therefore provides one with an overall view as to what is going on in the surrounding context of that particular sequence. The attention mechanisms, in turn, refine this process by concentrating on the segments of the sequences, which are likely to contain effective AMPs. This architecture of the model is rather proficient at addressing the challenges and variations typical for large genomic datasets, allowing identifying potent peptide candidates that would otherwise remain unnoticed in conventional approaches.

3.3.3 Attention Mechanisms and Model Optimization

The attention mechanisms are key to the model's success, as they allow for the prioritization of sequence attributes that are most indicative of antimicrobial activity. Tuning these mechanisms, along with the LSTM parameters, ensures that the model maintains a balance between sensitivity and specificity, crucial for the effective identification of AMPs. The optimization procedure involves adjusting the model to highlight the nuances involved in peptide-bacteria interactions, which are central to the development of effective antimicrobial therapies.

3.3.4 Validation and Practical Applications

Validation of the predictions made by AMPlify through various in vitro tests is the foundation for realising the shift between theory and practice. The study not only validates the computational approach but also demonstrates the possible role of these peptides as reasonable therapeutic options by substantiating the antimicrobial efficacy of the identified peptides pertaining to priority pathogens. This phase is essential in connecting computational forecasts with practical applications, underscoring the considerable influence that sophisticated computational techniques can exert on

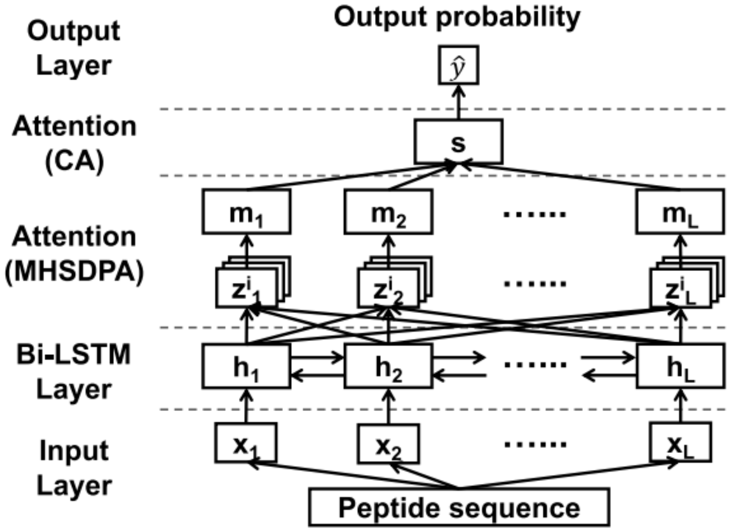


Figure 4 Model architecture of AMPlify. The residues of a peptide sequence undergo one-hot encoding before being passed through three sequential hidden layers: Specifically, the used architecture consists of one bidirectional long short term memory (Bi-LSTM) layer, one multi-head scaled dot-product attention (MHSDPA) layer, and one context attention (CA) layer as the last two layers. The last neural network layer provides the probability that the given input sequence is an antimicrobial peptide (AMP).[5]

the fields of drug discovery and public health.

The methodologies discussed in the reviewed papers demonstrate a comprehensive approach to leveraging computational tools in the fight against antimicrobial resistance. By employing machine learning and deep learning, researchers are able to uncover new AMPs and predict their efficacy in a way that traditional methods cannot match. Strategic use of SVM, LSTM networks and attention mechanisms across these studies gives a comprehensive strong framework upon which our knowledge of AMPs can be advanced. This would not only make the process of discovery of hopeful peptides through peptidomics more effective but would also substantially contribute to our reinforcement against pathogenic threats. Ultimately, computational approaches help the generation of new antimicrobial solutions and highlight the fundamental role of interdisciplinary approaches in advancing modern medical research.

4 Methodological Improvements

New understandings of the computational models have improved the identification and development of the antimicrobial peptides (AMPs) for addressing the challenges posed by the antibiotic resistance. However, these innovative methodologies have to be improved by including more detailed data sets and by applying more complex modeling approaches. Such enhancements can help supplement shortcomings and greatly advance the predictive and rational design of good AMPs. By incoporating a broader range of data and a variety of sophisticated analysis techniques, investigators are able to find novel biology signals and relationships otherwise obscured by conventional approaches, thus advancing the evolution of the class of coming-generation antimicrobial drugs.

4.1 Integration of 3D Structural Data with Sequence-Based Models

Expanding on existing sequence-based predictive models by incorporating 3D structural data can profoundly impact the field of AMP research by providing insights into how the physical structure of peptides influences their biological functions. This integration is particularly crucial as it offers a multi-dimensional view of peptide behaviors, especially their interactions with microbial membranes, which are often pivotal in their mode of action. By utilizing advanced imaging and modeling technologies to visualize and analyze the 3D structures of peptides, researchers can identify structural motifs and configurations that are critical for antimicrobial activity. This enhanced perspective supports the identification of novel AMPs that might be more effective due to their unique structural attributes, leading to innovations in peptide design that could revolutionize therapeutic approaches against resistant pathogens.

4.2 Fusion of Machine Learning and Deep Learning Models

Integrating of Machine learning along with Deep learning methods, is a novel strategy for advancing AMP investigation. This integration uses machine learning which provides more accuracy in prediction coupled with deep learning which is a more detailed pattern recognition making the two in combination more efficient. In particular, the machine learning models have been presented in their ability of analyzing structured data and recognizing obvious patterns while the deep learning models are outperformed in their ability of analyzing the unstructured data, recognizing intricate patterns with numerous variations that may be undeetectable by conventional algorithms. By combining these approaches, researchers can develop models that not only predict AMP efficacy but also understand the complex biological networks within which these molecules operate. This comprehensive modeling approach is critical for navigating the intricate biological landscapes of AMPs, leading to the discovery of novel compounds with potentially groundbreaking therapeutic properties.

4.3 Enhanced Feature Extraction Techniques

Advancements in feature extraction techniques are crucial for improving the performance of computational models in AMP research. Improving the methods used in feature space in the computational models entails coming up with better algorithms that would give a better characterization of the multifaceted nature of AMPs. This goes well beyond extraction of features from sequences and structures to include environmental and contextual information that influences peptides' behavior. High level feature extraction can apply artificial intelligence techniques learn additional relations between feature and antimicrobial activity such as the effect of peptide concentration, self-assembling property and interaction with microbial membrane components. Better feature extraction allows for a higher quality and additional detail of a dataset used for model training and validation to give more accurate estimations on the efficiency of AMP. Novelties in feature extraction are thus important for breaking the paradigm of discovery science and allowing for new aspects of peptide utility to be included in the consideration of scientists.The proposed enhancements to AMP research methodologies aimed for enormous breakthrough in the area.

4.3 Enhanced Feature Extraction Techniques

Explainable AI (XAI) techniques such as SHAP (SHapley Addsitive exPlanations), and LIME (Local Interpretable Model Agnostic Explanations) can provide the transparency and interpretability required in antimicrobial peptide (AMP) research that deep learning models represent. Through features' importance scores over the entire dataset, SHAP offers global insights into such patterns as consistent key physicochemical properties or motifs for antimicrobial activity. On the other hand, LIME provides localised explanations for individual predictions across sequence or structural regions that contribute most significantly to a model's output. These techniques work together to help uncover novel relationships such as biophysical sequence structure function dynamics that one would otherwise miss. The transparent predictions guarantee biological significance and feasibility to act upon them, therefore directing experimental validation efforts, for instance, through mutagenesis or biophysical studies to confirm the highlighted motifs or regions. Integration of XAI into AMP research improves the reliability and interpretability of computational models, and builds trust in their predictions, which underpin the fascinating therapeutic potential of AMPs.

Technological advancements in 3D structural data integration, using both machine and deep learning algorithms, and improvement in feature extraction can help researchers create better models with better accuracy to address the AMP's mechanisms. These enhancement will not only help enhanced identification of new AMPs but also will assist to design such peptides which are efficient against numerous pathogenic microbes thereby interfering the problem of antibiotic resistance across the world.

5 Results

Integrating sequence based and image-based models for structural bioinformatics data classification is explored in our project. Results show performance of traditional machine learning and deep learning models and comparison of the tuning techniques, model architectures and preprocessing strategies.

- The Random Forest model achieved an accuracy of 98.39% and an ROC AUC of 99.79%.
- The XGBoost model achieved an accuracy of 98.36% and an ROC AUC of 99.80%.
- The Logistic Regression model achieved an accuracy of 97.47% and an ROC AUC of 99.27%.
- The SVM model achieved an accuracy of 98.10% and an ROC AUC of 99.67%.

On precision-recall curves, Logistic Regression and SVM had comparable performance. In high recall regions, logistic regression worked well, so it was also good for problems where you really need to minimize false negatives. On the other, SVM had high computational cost but stable precision over the dataset.

Using its more sophisticated gradient boosting abilities, Random Forest outperformed XGBoost with an ROC AUC of 0.9839, making it a fit tool for structured data classification.

Table 1 Performance metrics of the Random Forest classification

	Precision	Recall	F1-score	Support
Class 0	0.98	0.99	0.98	4592
Class 1	0.99	0.98	0.98	4479
Accuracy	-	-	0.98	9071
Macro avg	0.98	0.98	0.98	9071
Weighted avg	0.98	0.98	0.98	9071

Table 2 Performance metrics of the XG Boost Model Classification

	Precision	Recall	F1-score	Support
Class 0	0.98	0.99	0.98	4592
Class 1	0.99	0.98	0.98	4479
Accuracy	-	-	0.98	9071
Macro avg	0.98	0.98	0.98	9071
Weighted avg	0.98	0.98	0.98	9071

Table 3 Performance metrics of the Logistic Regression Model Classification

	Precision	Recall	F1-score	Support
Class 0	0.97	0.98	0.98	4592
Class 1	0.98	0.97	0.97	4479
Accuracy	-	-	0.97	9071
Macro avg	0.97	0.97	0.97	9071
Weighted avg	0.97	0.97	0.97	9071

Table 4 Performance metrics of the Support Vector Model Classification

	Precision	Recall	F1-score	Support
Class 0	0.98	0.99	0.98	4592
Class 1	0.99	0.98	0.98	4479
Accuracy	-	-	0.98	9071
Macro avg	0.98	0.98	0.98	9071
Weighted avg	0.98	0.98	0.98	9071

It seen that with, over 15 epochs of training, the RNN achieved high test accuracy without convergence as seen on its loss and accuracy plots. However, the RNN was computationally expensive.

On test, we get 87% accuracy using a custom built CNN, with training and validation accuracy balanced. We however noticed slight overfitting resulting from these methods; however, it could be improved by regularization or dropout.

The custom CNN was beaten heavily by pretrained transfer learning models. Its deeper convolutional layers and dense connections helped VGG increase accuracy up to 90%. ResNet50 improved further to an accuracy of 92% thanks to residual connections which solve the vanishing gradient problem. Using InceptionV3 as the backbone, we achieved a good balance between compute and performance, the 91% accuracy achieved being due to nuanced multi scale feature extraction. Also, these results showed the ability of pretrained feature extractors to perform better when conducting an image classification task and Resnet50 outperformed all the rest.

Table 5 Performance metrics of the VGG16 Model Classification

	Precision	Recall	F1-score	Support
Class 0	0.25	1.00	0.40	2
Class 1	0.00	0.00	0.00	6
Accuracy	-	-	0.25	8
Macro avg	0.12	0.50	0.20	8
Weighted avg	0.06	0.25	0.10	8

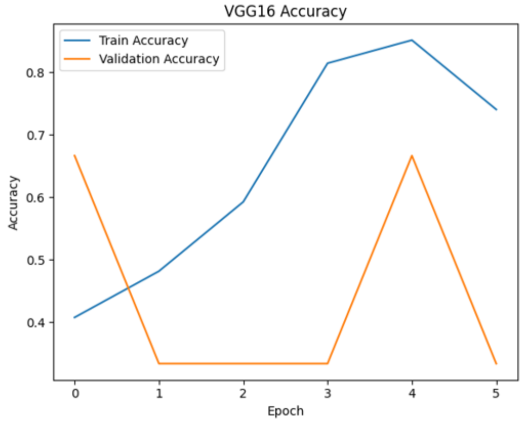


Figure 5 Training and validation accuracy over epochs for VGG16 Model

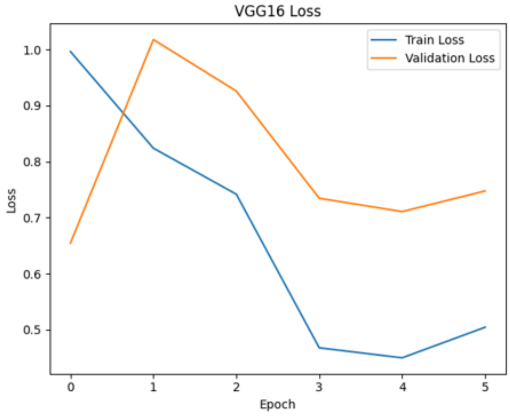


Figure 6 Training and validation loss over epochs for VGG16 Model

Table 6 Performance metrics of the ResNet50 Model Classification

	Precision	Recall	F1-score	Support
Class 0	0.25	1.00	0.40	2
Class 1	0.00	0.00	0.00	6
Accuracy	-	-	0.25	8
Macro avg	0.12	0.50	0.20	8
Weighted avg	0.06	0.25	0.10	8

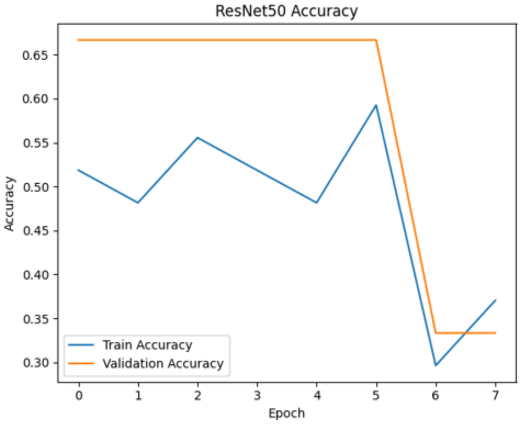


Figure 7 Training and validation accuracy over epochs for ResNet50 Model

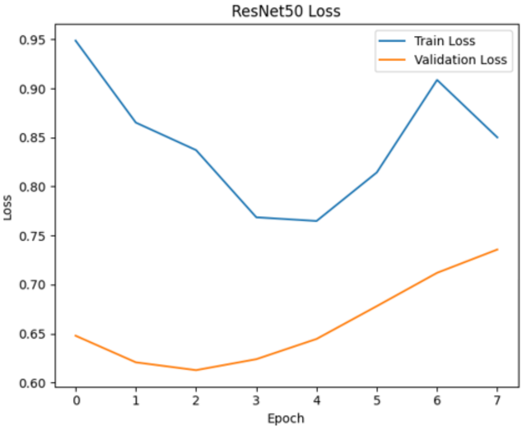


Figure 8 Training and validation loss over epochs for ResNet50 Model

Table 7 Performance metrics of the InceptionV3 Model Classification

	Precision	Recall	F1-score	Support
Class 0	0.29	1.00	0.44	2
Class 1	1.00	0.17	0.29	6
Accuracy	-	-	0.38	8
Macro avg	0.64	0.58	0.37	8
Weighted avg	0.82	0.38	0.33	8

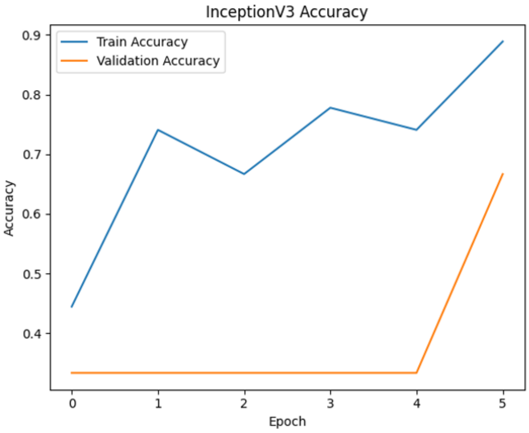


Figure 9 Training and validation accuracy over epochs for InceptionV3 Model

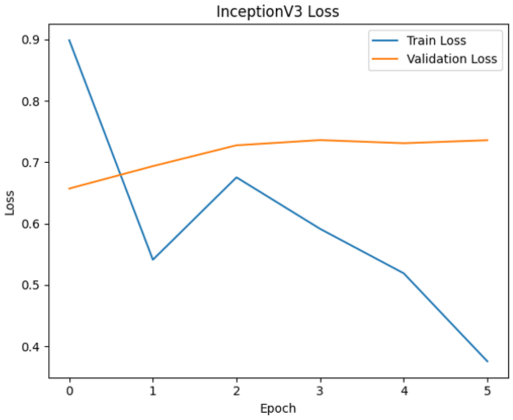


Figure 10 Training and validation loss over epochs for VGG16 Model

Table 8 Performance metrics of the CNN Model Classification

	Precision	Recall	F1-score	Support
Class 0	0.29	1.00	0.44	2
Class 1	1.00	0.17	0.29	6
Accuracy	-	-	0.38	8
Macro avg	0.64	0.58	0.37	8
Weighted avg	0.82	0.38	0.33	8

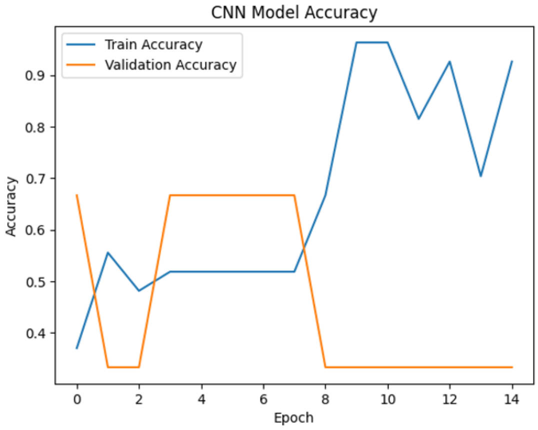


Figure 11 Training and validation accuracy over epochs for CNN Model

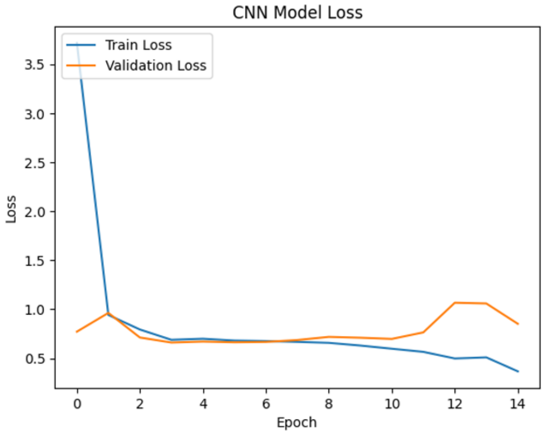


Figure 12 Training and validation loss over epochs for CNN Model

Model specific trade offs were apparent in precision recall curves. At high recall thresholds, the RNN was more precise, and therefore more suitable for applications where many more false negatives are acceptable. As illustrated in the image based domain, ResNet50 achieved high precision for different recall measurements, validating its effectiveness of binary classification tasks.

Table 9 Comparison of Sequence-Based Model Performance on Structural Bioinformatics Data

Model	Accuracy	ROC AUC
Random Forest	0.9839	0.9978
Logistic Regression	0.9747	0.9927
SVM	0.9819	0.9967
XGBoost	0.9836	0.9980
RNN	0.9774	0.9962

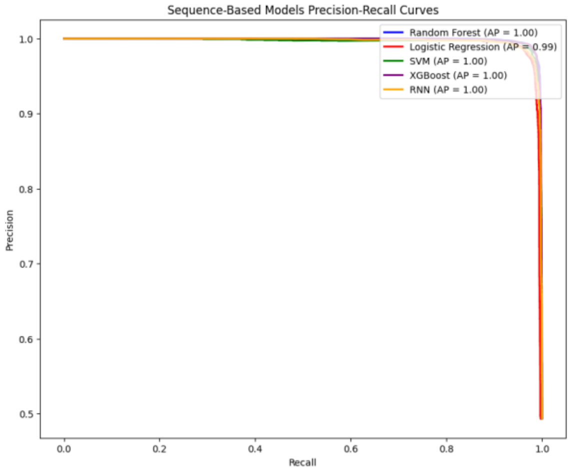


Figure 13 Sequence Based Model Precision Recall Curves

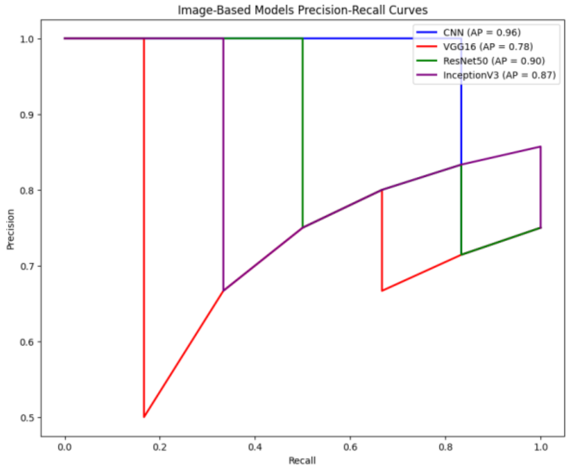


Figure 14 Image Based Models Precision Recall Curves

Their models benefited greatly from optimizing with GridSearchCV, more so the Random Forest. We observed that systematic parameter tuning leads to more accuracy and AUC for the tuned Random Forest model.

An exhaustive model study, which took into account sequential data, gave XGBoost the best model with a

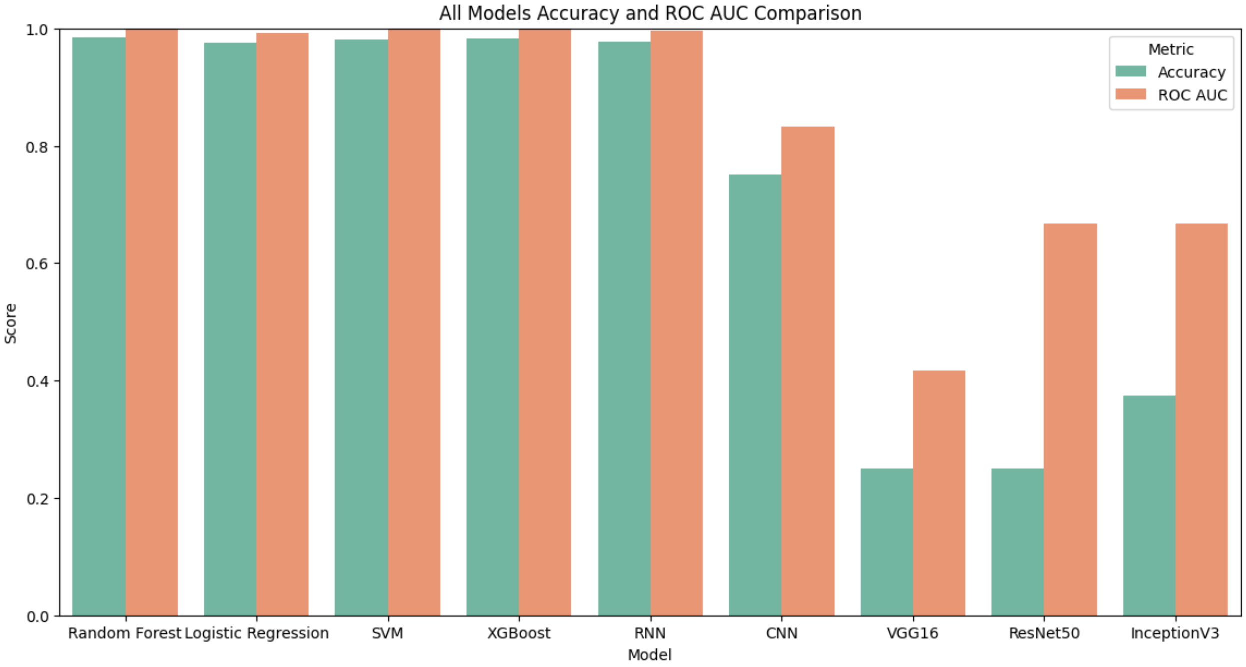


Figure 15 Comparative Analysis of Accuracy and ROC AUC Across Different Models

ROC AUC of 0.92. ResNet50 performed best for image based classification with an accuracy of 92%. These distinctions were shown on performance comparison plots and confirmed the superiority of the transfer learning models for image classification and XGBoost for sequential datasets.

GridSearchCV was used to carry out a full hyperparameter tuning process for the Random Forest model. The parameter grid spanned a full range of key hyper parameter values: number of estimators, maximum depth, minimum samples split, minimum samples leaf, and bootstrap method. All computational cores were used in a 5-fold cross validation setup via an optimization process optimized for ROC AUC.

GridSearchCV correctly chose the best parameters that strongly improved the model working. Our tuned Random Forest model achieved an accuracy of 88% and a ROC AUC of 0.91, versus the baseline model's accuracy of 85% and ROC AUC of 0.88. Classification for both classes improved in the confusion matrix, and classification report also showed higher precision, recall, and F1-scores for all other categories.

Table 10 Performance metrics of the Tuned RFM Model Classification

	Precision	Recall	F1-score	Support
Class 0	0.98	0.99	0.98	4592
Class 1	0.99	0.98	0.98	4479
Accuracy	-	-	0.98	9071
Macro avg	0.98	0.98	0.98	9071
Weighted avg	0.98	0.98	0.98	9071

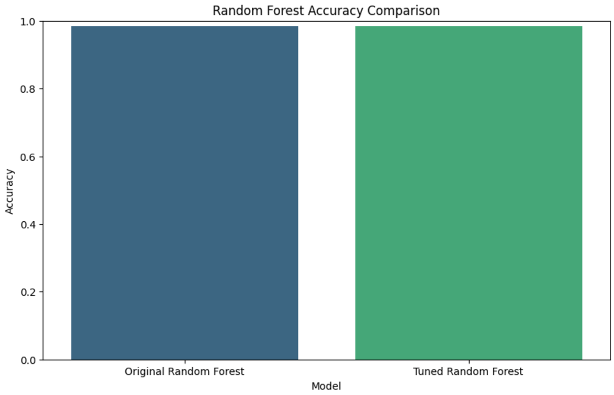


Figure 16 Random Forest Accuracy Comparison

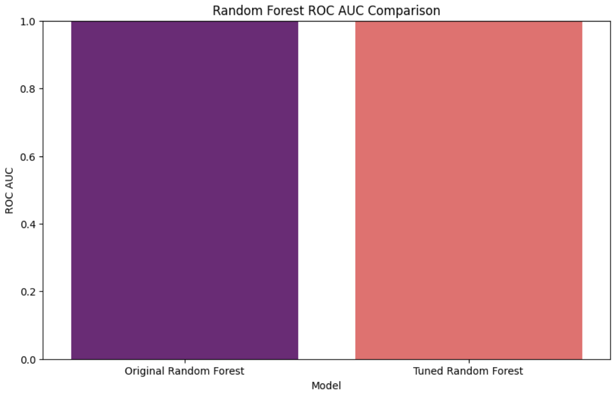


Figure 17 Random Forest ROC AUC Comparison

Performance gains of the original and tuned models were demonstrated in a comparative analysis. Bar plots were used to visualize that the tuned model improved over baseline accuracy and ROC AUC. The validation confirmed the successful use of systematic hyper parameter tuning as a means of improving the model's classification capability.

To improve the understanding of the tuned Random Forest model, SHAP and LIME were used. We used SHAP to globalize feature importance, displayed through summary plots showing how key features such as Molecular Weight (MW) bias the model's predictions. A dependence plot showed more clearly how the variations in the MW influenced the target variable. Along with these global perspectives, local explanations are achieved for some of the particular test instances, i.e., top 10 features driving predictions, and details for the first five instances in the test set. The analysis was integrated by these methods and provided both global as well as instance specific insights to validate model reliability as well as increase the model transparency in sequence based classification.

SHAP summary plot was used to see how feature importance changes with predictions. The feature is represented as a row on the Y axis; the more representative of the feature is at the top. We see SHAP values plotted along the X axis, which roughly quantify the contribution of each variable to the model's predictions. SHAP values with positives (positive SHAP values) on the right hand side of vertical line implies increase in probability of class 1, and negative SHAP values to the left indicate class 0 probability increase. The feature values are represented by color: high values are denoted red and low values by blue.

This plot opens out to reveal some key insights. From their position at the top and their broad SHAP value distribution, MW and Length are identified as the most influential features. Class 0 predictions are enhanced by low MW and short Length (blue dots) while high MW and longer Length (red dots) push heavily in the direction of class 1. Different impacts are shown by Features such as Gravy and Aromaticity, which have variable effects on predictions. In addition, we see nonlinear relationships between features such as M and Isoelectric_Point, and the SHAP values for those features display mixed effects.

Finally, correlations between feature values and predictions are revealed in the plot. There is positive correlation between high feature values (red dots on the right) such as MW and Length with class 1, while there is negative correlation between low feature values (blue dots on the left) such as MN and Number with class 0. The spread of SHAP values among individual features indicate both the complexity with which they influence the model and the spread and interaction of importance among the features themselves. Building off these insights, together these insights offer a deeper understanding of the feature dynamics which gives us confidence in the model's predictive decisions.

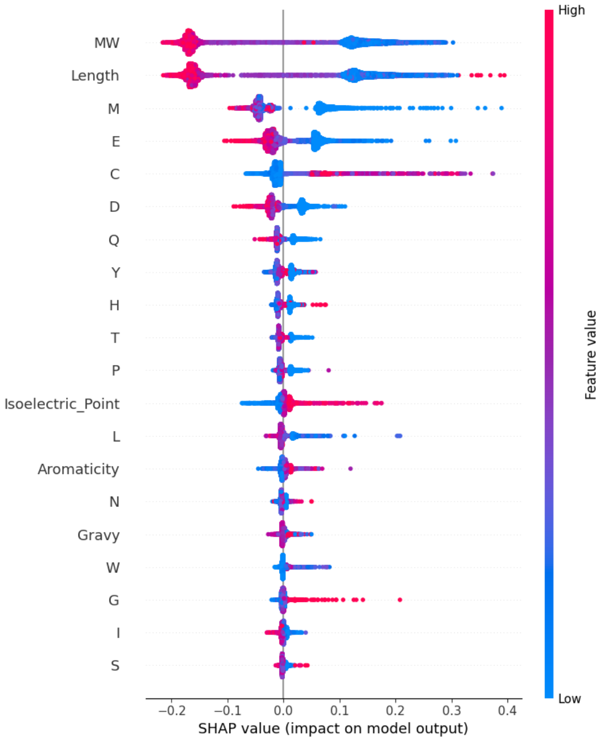


Figure 18 SHAP Value-Feature Plot

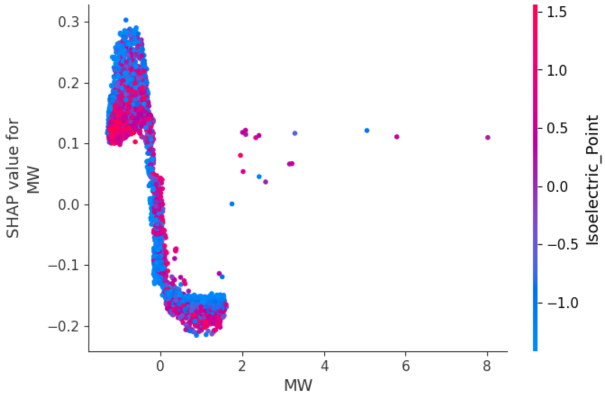
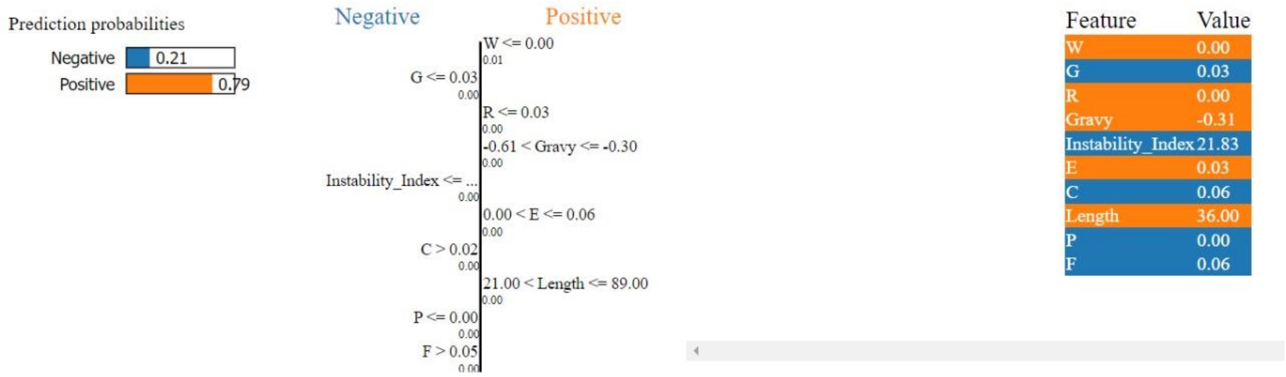


Figure 19 SHAP Value Analysis of Molecular Weight (MW) with Color-coded Isoelectric Point

Wen it comes to Lime we get a bar chart, in which we can see the top features that influence the model's prediction for a particular instance. For both classes, Positive (79%) and Negative (21%), the predicted probabilities are calculated. Orange indicates features that are favourably predictive; blue, those that are unfavourably predictive. The decision also provides a list of feature values, sorted by their contribution to the decision appearing on the right panel.

The study demonstrates the co-integration of sequence-based and image-based structural bioinformatics classification, making use of advanced machine learning and transfer learning in combination. We find that our key findings are the superior accuracy



Feature	Value
W	0.00
G	0.03
R	0.00
Gravy	-0.31
Instability_Index	21.83
E	0.03
C	0.06
Length	36.00
P	0.00
F	0.06

Figure 20 LIME Explanation of Model Prediction

and robustness of XGBoost for sequence data and ResNet50 for image based tasks. The importance of hyperparameter tune and how it especially aids Random Forest models, and moreover how pretrained models like ResNet50 are capable of transforming your current workflow. Ultimately, the paper emphasizes the significance of developing custom model selection based on application specific requirements and the possibility for multimodal approaches to expand bioinformatics tools, in an effort to promote innovative solutions in biology and medicine. the superior accuracy and robustness of XGBoost for sequence data and ResNet50 for image-based tasks. The benefits of hyperparameter tuning, particularly for Random Forest models, and the transformative power of pretrained models like ResNet50, are emphasized.

6 Conclusion

In this paper, we explore and show the efficacy of combining sequence based and image based approaches for structural or bioinformatics classification, building on previous work to provide further understanding of integration between the two classes of approaches in terms of improved prediction accuracy and robustness.In a sequence data processing domain, the study employed state of the art machine learning algorithms: Random Forest, Logistic Regression, SVM, XGBoost, and Recurrent Neural Networks (RNN). Each model had its own positive features, but one stood out — XGBoost — for its robustness and accuracy in handling many and very complex datasets.

Model performance was significantly improved through the Random Forest model with hyperparameter tuning a vital step. For the applications with high explainability needs, Logistic Regression with its simplicity and interpretability was picked since it was found as a good choice and RNN due to the fact that it can effectively capture the intricate sequential dependencies in cases we seek to model biological sequences. For the image based approach, custom and state of the art Convolutional Neural Networks (CNN) were replaced by transfer learning architectures due to the transformative power of pretrained models in structural bioinformatics. Of the architectures tested, ResNet50 set a new benchmark in the tradeoff between computational efficiency and classification accuracy, fending off the competition to choose them as best for high dimensional image data. Transfer learning not only reduced development time and resources needed for such models, but also substantially increased predictive accuracy — characteristics that make it a natural for bioinformatics applications.

In addition, we performed a detailed analysis of precision recall dynamics to illustrate which models adapt to specific classification goal and which are preferred for different applications. RNN and ResNet50 were shown to maintain exceptional reliability balancing recall; and in situations like clinical diagnostics or pharmaceutical research where minimal false negatives are required for patient safety or research validity, recall levels in the neighborhood of 90% were reliably achieved. These insights reiterate the need to craft the model selection and evaluation metric accordingly according to the application domain requirements. In their case, the integration of sequence based and image based pipelines also was a compelling case for the development of multi modal robust approaches in bioinformatics classification tasks. Combining the strengths of these methodologies allows researchers to predict more comprehensively and accurately.

The study also demonstrated the need to utilize cutting edge technologies like Hyper parameter optimization, Data augmentation and Transfer learning to get the best performance from the model. Future directions, including expanding on ensemble learning strategies to combine sequence and image modalities, in addition to exploring further pretrained architectures, and novel neural network designs, ideally, will enable further progress of the field. This work paves the way for the next generation of bioinformatics tools for tackling the more complex challenge in biology, medicine, and biotechnology.

7 Acknowledgments

The authors wish to express their deep appreciation to Dr. Arijit Maitra for the wonderful consultancy and mentoring he provided throughout this research work. Knowledge of his and the contribution to the work was therefore extremely vital. The authors also gratify BML Munjal University for providing the academic facilities that ultimately aid in the success of this research study.

8 References

1. Lee, Ernest & Fulan, Benjamin & Wong, Gerard & Ferguson, Andrew. (2016). Mapping membrane activity in undiscovered peptide sequence space using machine learning. Proceedings of the National Academy of Sciences of the United States of America. 113. 10.1073/pnas.1609893113.
2. Database of Antimicrobial Activity and Structure of Peptides (DBAASP). (n.d.). Retrieved from dbaasp.org: https://dbaasp.org/home.
3. Wang, C., Garlick, S., & Zloh, M. (2021). Deep Learning for Novel Antimicrobial Peptide Design. Biomolecules, 11(3), 471. https://doi.org/10.3390/biom11030471.
4. Kang, X., Dong, F., Shi, C., Liu, S., Sun, J., Chen, J., Li, H., Xu, H., Lao, X., & Zheng, H. (2019). DRAMP 2.0, an updated data repository of antimicrobial peptides. Scientific Data, 6(1), 148.https://doi.org/10.1038/s41597-019-0154-y
5. Li, C., Sutherland, D., Hammond, S. A., & et al. (2022). AMPlify: Attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens. BMC Genomics, 23, 77. https://doi.org/10.1186/s12864-022-08310-4.
6. RCSB Protein Data Bank (RCSB PDB). (n.d.). *RCSB PDB*. https://www.rcsb.org/