Enhancement Antimicrobial Activity Predictors based on Machine Learning Approaches

Salah G. Abdelkhabir[1], Seham S. Ezz-eldeen[1], Ahmed Ebrahim M. Gabr[1], Ahmed M. Eldakrory[1], Ahmed M. Ali[1], Omnia K. Elkhameesy[1], H. Arafat Ali[1,2], Sarah M. Ayyad[2,], Zainab H. Ali[3,4]*

[1] Faculty Artificial Intelligence, Mansoura University Delta University for Science and Technology, Eldakahlia, 33511, Egypt;

[2,*] Computers and Systems Department, Faculty of Engineering, Mansoura University, Mansoura, 35516, Egypt;

[3] Department of Embedded Network Systems and Technology, Faculty of Artificial Intelligence, Kafrelsheikh University, El-Geish st, Kafrelsheikh, 33516, Egypt;

[4] Department of Electronics and Computer Engineering, School of Engineering and Applied Sciences at Nile University, Giza, Egypt.

Abstract

Recently, the prediction tools of antimicrobial activity revealed a promising avenue for novel antimicrobial peptide (AMP) sequence determination and discovery. Machine Learning (ML) approaches can be utilized to offer the prediction of AMP sequence with great success, which explores alternative strategies to combat antimicrobial resistance and develop effective treatments for infections. The main objective of this chapter is to study and evaluate the predictive ability of modern ML methods to accurately identify the activities of antimicrobial sequences that have been previously described at the protein level through in vitro studies. Two cloud servers and one standalone software are employed to evaluate 20 sequences generated by 26 ML models. The experimental results proved that there are algorithms LightGBM (LGBM), Extra Trees, Random Forest (RF), and SVC have gained the highest rate in the performance metrics of accuracy, recall, Matthew's Correlation Coefficient (MCC), and F1-score. To formally confirm whether the utilized ML approaches have a significant enhancement, we used a dataset with size 6623 instances for both AMP and non-AMP classes. To avoid overfitting, the threshold value is adjusted to 0.9. The best performance was LGBM with an accuracy of 0.92%, MCC of 0.83, recall of 90%, Area Under the Curve (AUC) of 0.97%, precision of 0.91%, and F1-score of 0.92%. Regarding these results, LGBM was demonstrated as highly beneficial for evaluating AMP sequence with antimicrobial potential before proceeding to experimental testing.

Keywords: Antimicrobial peptides; Antimicrobial activity; prediction models; Amino acid composition; Machine learning algorithms.

1. Introduction

Antimicrobial resistance has become a critical issue worldwide, with profound implications for human and animal health, as well as the environment. It involves the ability of microorganisms, including bacteria, viruses, parasites, and fungi, to develop resistance against drugs intended to combat their growth or kill them. The rise of antimicrobial peptides (AMPs) can be attributed to various factors, including the inappropriate and excessive use of antimicrobial medications in human and veterinary medicine, as well as in agriculture [1][2].

One of the key consequences of AMPs is the limited effectiveness of existing antimicrobial treatments, leading to increased morbidity, mortality, and healthcare costs. Infections caused by resistant microorganisms are more difficult to treat, often requiring more expensive and toxic drugs or prolonged hospital stays [3]. This not only affects individual patients but also strains healthcare systems and hampers the ability to effectively control infectious diseases. The development of AMPs is a natural evolutionary process, but human activities have accelerated its occurrence and spread. Inappropriate prescribing practices, self-medication, and the use of antibiotics as growth promoters in livestock contribute to the selection and dissemination of resistant strains. Furthermore, the global interconnectedness of modern society facilitates the rapid global spread of resistant microorganisms, making antimicrobial resistance (AMR) a global health security issue [4].

Addressing AMR requires a multifaceted approach involving prediction methods, such as Machine Learning (ML), Deep Learning (DL), and Reinforcement Learning (RL) [5]-[7]. ML models for AMPs prediction typically utilize a dataset consisting of known AMPs and non-AMPs, along with their corresponding features. These features can include physicochemical properties, amino acid composition, and structural characteristics. By training the model on this dataset, it learns to recognize patterns and relationships between the peptide features and their antimicrobial activity. One commonly used ML approach for AMPs prediction is the use of supervised learning algorithms, such as Naive Bayes (NB), Support vector machines (SVM), K-nearest neighbor (KNN), and Logistic Regression (LR). These algorithms are trained on labeled datasets, where each peptide is assigned, a binary label indicating its antimicrobial activity (active or inactive). The model learns to differentiate between active and inactive peptides based on the provided features and generalizes this knowledge to predict the activity of new, unseen peptides [8]-[10].

The use of mathematical modeling to predict the antimicrobial activity of peptides has gained significant interest in addressing human health and agribusiness challenges. Consequently, several databases and algorithms have been proposed to facilitate the prediction of peptide activity without the need for extensive biochemical experiments. Some widely utilized databases for antimicrobial peptides include DBAASP [11], CS-AMPPred [12], CAMPR3 [13], and BACTIBASE [14]. Most algorithms aimed at building predictive models for peptide activity rely on physicochemical and biochemical parameters of amino acids [15]. These parameters serve as descriptors for developing models, and they typically include factors such as polarity, electrostatic charges, 3D geometry, and

quantum mechanical descriptors. Various machine learning algorithms utilize these descriptors, including partial least squares (PLS) [16], artificial neural networks (ANN) [17], SVM [18], and incremental feature selection [19], among others. By leveraging these algorithms and incorporating physicochemical data, researchers can develop models to predict the antimicrobial activity of peptides. This approach offers a valuable tool for screening and prioritizing potential antimicrobial peptides, facilitating the discovery of effective candidates for further experimental validation.

The main objective of this study is to evaluate the predictive ability of such ML approaches in determining the antimicrobial sequence activities that were previously characterized at the protein level through in vitro studies. Two cloud servers and one standalone software are employed to evaluate 20 sequences generated by 26 ML models. The experimental results proved that there are algorithms LightGBM (LGBM), ExtraTrees, Random Forest (RF), and SVC have gained the highest rate in the performance metrics of accuracy, recall, Matthew's Correlation Coefficient (MCC), and F1-score. To formally confirm whether the utilized ML approaches have a significant enhancement, we used a dataset with size 6623 instances for both AMP and non-AMP classes. To avoid overfitting, the threshold value is adjusted to 0.9. The best performance was LGBM with an accuracy of 0.92%, MCC of 0.83, recall of 90%, Area Under Curve (AUC) of 0.97%, precision of 0.91%, and F1-score of 0.92%. Regarding these results, LGBM was demonstrated as highly beneficial for evaluating AMP sequence with antimicrobial potential before proceeding to experimental testing.

The rest of this paper is organized as follows: Section 2 describes datasets and predictive model. Section 3 measures and discusses the performance evaluation results. Section 4 wraps up the paper with significant contribution points.

2. Material

2.1. Dataset Description

AMPs relies fundamentally on the quality of the dataset. A high-quality dataset is essential to ensure the accuracy and reliability of models, especially when dealing with sensitive data like AMPs. This work figures out the significance of dataset quality, examining specific factors that influence it, including database sources, annotation quality, and the impact of data variability on model performance.

The dataset forms the foundation upon which predictive models are built, and its ability to encompass diverse peptide sequences directly influences a model's generalizability to unseen data. An imbalanced dataset may result in models that perform well within the training set but struggle when faced with new, unseen AMPs. Therefore, a dataset that captures the broad spectrum of AMPs characteristics is crucial. However, the sources of the database from which the dataset is derived play a crucial role in determining data quality. Databases with rigorous quality control measures contribute to the reliability of the dataset. In contrast, datasets sourced from less reliable or poorly curated databases may introduce noise and inconsistencies, hindering the model's ability

to discern patterns accurately. This work aims to reduce challenges in datasets involving efficiently collecting data, rigorous pre-processing strategies, and ensemble classifier approaches.

2.2 Dataset Collection

The dataset AMPs was gathered from four benchmark datasets as outlined below: i) The antimicrobial peptide database (APD3)[1], encompassing a total of 2619 AMPs, including 261 bacteriocins from bacteria, 4 AMPs from archaea, 7 from protists, 13 from fungi, 321 from plants, and 1972 animal host defense peptides. Within APD3, there are 2169 antibacterial, 172 antiviral, 105 anti-HIV, 959 antifungal, 80 antiparasitic, and 185 anticancer peptides; ii) A Database Linking Antimicrobial Peptides (LAMP)[2], which currently houses 5,547 AMPs sequences, consisting of 3,904 natural AMPs and 1,643 synthetic peptides; iii) Collection of Anti-Microbial Peptides (CAMPR3)[3], which currently contains 10,247 sequences, 757 structures, and 114 family-specific signatures of AMPs; and iv) Data Repository of Antimicrobial Peptides (DRAMP)[4], which encompasses 22,528 entries, including 6,105 general AMPs (comprising natural and synthetic AMP), 16,110 patent AMPs, and 96 AMPs in various stages of drug development (preclinical or clinical stage). In the most recent update, an additional 217 stapled antimicrobial peptides were included, belonging to specific AMPs. This work retrieved all antibacterial AMPs data from these databases, excluding AMPs with a sequence length shorter than 10 amino acids and those containing unusual amino acids such as B, Z, U, X, J, O, I, n, and "-". After destroying duplicate records, we acquired a final set of 6,623 sequences for the AMPs dataset, and many AMPs are less than 50 amino acids in length.

Otherwise, the non-AMP dataset comprises a dual composition, integrating real-world peptides and artificially generated sequences. Authentic peptides were meticulously sourced from UniProt, adhering to specific criteria: i) sequences ranging in length from 10 to 50 amino acids, and ii) devoid of AMP-associated keywords in annotations, including 'Antimicrobial,' 'Antibiotic,' 'Amphibian defense peptide,' and 'Antiviral protein.' This careful selection process ensures the exclusion of peptides explicitly linked to antimicrobial properties, defining a clear distinction for non-antimicrobial peptides within the dataset.
In tandem with authentic peptides, the dataset includes artificially generated sequences crafted by randomly assembling the 20 essential amino acids. Crucially, these artificially generated sequences share an equivalent length distribution with the AMP dataset, providing a controlled framework for representation across various peptide lengths. This deliberate design choice imparts diversity to the non-AMP dataset, facilitating a nuanced exploration of peptide characteristics beyond naturally occurring sequences.
Our curation resulted in two datasets: one containing 6623 sequences for AMP, and another, a unique combination of real-world and artificially generated peptides, establishing a diverse non-AMP dataset [20].

2.3 Customizing Dataset for the Proposed Model Prediction

In this work, customizing the data for the proposed model is crucial as it plays a vital role in improving the model's ability to identify meaningful patterns, resulting in enhanced accuracy and reliable predictions across diverse datasets. To achieve this, we carefully selected a dataset-splitting ratio of 80% for training and 20% for testing. This allocation ensures a balanced

[1] APD3: http://aps.unmc.edu/AP/
[2] LAMP: http://biotechlab.fudan.edu.cn/database/lamp
[3] CAMPR3: www.camp3.bicnirrh.res.in
[4] DRAMP: http://dramp.cpu-bioinfor.org/

distribution, allowing the model to learn effectively during training and enabling robust evaluation during testing.
In addition, the CD-Hit clustering method with a stringent threshold of 99% was employed. This decision was made to enhance the uniqueness of sequences within both the AMPs and non-AMP classes. By applying the 99% threshold individually to each class, we tailored the clustering process to capture the specific characteristics of AMPs and non-AMPs. This approach effectively groups similar sequences within each class, reducing redundancy and enriching the diversity of our dataset. The emphasis on the 99% threshold for each class demonstrates our commitment to meticulous data customization, reducing noise, and strengthening the model's resistance to overfitting. This combined strategy of dataset preparation and splitting ratio enhances the model's generalization capabilities, enabling it to accurately predict new AMP sequences during the two-phase training and testing for each class.

Table 1 presents 20 randomly generated sequences, evenly split between the positive and negative classes. Each sequence is accompanied by its length, showcasing the diversity within the dataset. Notably, the table highlights the most influential amino acid in each sequence. This curated collection serves as a valuable resource for studying sequence-activity relationships, providing insights into the lengths and key amino acids that contribute significantly to the positive or negative classification. The data facilitates a comprehensive understanding of the structural characteristics influencing the antimicrobial peptide classification model.

Table 1: The most influential amino acid and their length.

Amino acid	*Sequence*	*Length*
Isoleucine	*QCUQYZXILUCVRTJTRETOMMKIPB*	*26*
Alanine	*IBKVHUEFHZCRIIISLAQOVUMOD*	*25*
Isoleucine	*RGYRGFYKRUZZOAXILGKLQKLLNTUR*	*28*
Cysteine	*BQFCPDJCMQXNUJKKZRTPBAVJTHUIUY*	*30*
Lysine	*ALJBXJVRQFLFVZBILZRKZR*	*22*
Asparagine	*SZRWBHERTTNIUXSKASYB*	*20*
Lysine	*UIXFZOSAZCQOHSUPSHQQ*	*20*
Tyrosine	*USKREMSAMLDTNDTBENSNNVBC*	*24*
Isoleucine	*ZOCXACFEHPZHEUZASXJPGYBEHWWH*	*28*
Isoleucine	*OOVIJIPUYWQGAKLYUSGJRYYMYU*	*26*
Alanine	*QEVOYGVMNSOMEDCVSFGKQFMDIB*	*26*

Amino acid	*Sequence*	*Length*
Aspartic acid	*WXBVRJNQVBGMTUUOFAHVWFCG*	*24*
Proline	*TNUUBBFDJZBQMPNIUAYQ*	*20*
Histidine	*JGMPNBKAYGQSYOWLDEUALRQ*	*23*
Valine	*YUYOJXODNOOSMIYOIISYWEVE*	*24*
Alanine	*JSHQQMRIUHNUNITZVSZYPSVT*	*24*
Leucine	*HUVYZPNYZLICLIQSWYDWJUHWUKHIEC*	*30*
Isoleucine	*FYKUHNWVMAFDYXESVLPO*	*20*
Serine	*DFCEBZGGATWKWFHQDTRFLX*	*22*
Proline	*EVIVVNPWNAAVXWKWMPTGFZRRQ*	*25*

3. Proposed Methodology

Figure 1. depicts the core workflow to build the proposed predictive methodology applied in the study. Our comprehensive analysis revolves around exploring the Amino Acid Composition (AAC) within peptide sequences, with a particular focus on extracting information related to the presence of 20 specific amino acids. These amino acids include 'Alanine', 'Cysteine', 'Aspartic acid', 'Glutamic acid', 'Phenylalanine', 'Glycine', 'Histidine', 'Isoleucine', 'Lysine', 'Leucine', 'Methionine', 'Asparagine', 'Proline', 'Glutamine', 'Arginine', 'Serine', 'Threonine', 'Valine', 'Tryptophan', and 'Tyrosine'. By leveraging this amino acid-centric approach, we aim to gain insights into the compositional characteristics of both AMP and non-AMP sequences.

To conduct a comprehensive examination, we amalgamated the features extracted from both AMP and non-AMP sequences into a unified data frame. The resultant data frame encapsulates valuable information about the frequency and distribution of the specified amino acids across the peptide sequences. As part of our analysis, we sorted the amino acids based on their appearance in descending order. The order of appearance, along with the corresponding frequencies, is as follows: Cysteine, Lysine, Arginine, Glycine, Leucine, Glutamic acid, Tryptophan, Alanine, Histidine, Proline, Phenylalanine, Serine, Threonine, Tyrosine, Aspartic acid, Leucine, Methionine, Isoleucine, Valine, and Glutamine.

Furthermore, to visualize the significance of each amino acid in distinguishing between AMP and non-AMP sequences, we present the results through Figure 2. This figure serves as a graphical representation of the feature importance, shedding light on the amino acids that play a pivotal role in the differentiation of these two categories.

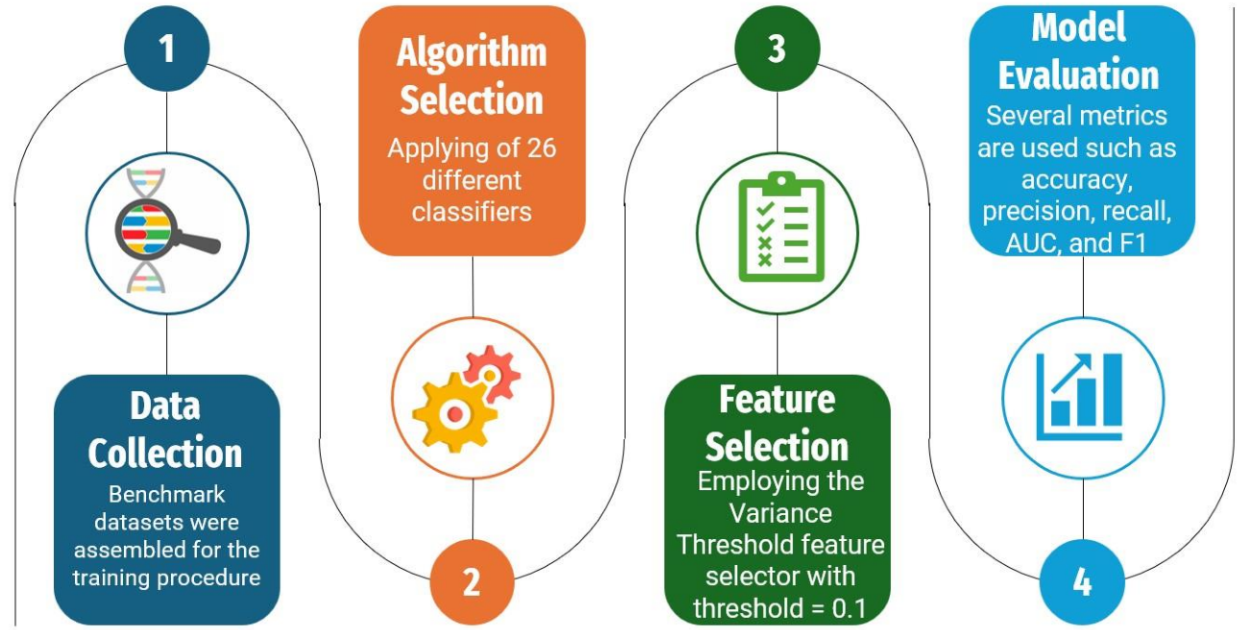


Figure 1: The core workflow to build the proposed predictive methodology.

3.1. Algorithm Selection

We examined several different classifiers to determine which approach best fits our prediction goal as shown in Figure 3. Consequently, the 26 ML algorithms' performance. (LGBM Classifier, Extra Trees Classifier, RF Classifier, KNN Classifier, SVC, Label Spreading, Label Propagation, Bagging Classifier, NuSVC, AdaBoost Classifier, Decision Tree Classifier, Quadratic Discriminant Analysis, Extra Tree Classifier, Calibrated Classifier CV, Linear SVC, Logistic Regression, Linear Discriminant Analysis, Ridge Classifier, Ridge Classifier CV, SGD Classifier, Nearest Centroid, Gaussian NB, Bernoulli NB, Passive Aggressive Classifier, Perceptron, Dummy Classifier) It was evident for both AMP, non-AMP classes that ensemble LGBM outperformed all other ML model types.

LGBM and XGBoost algorithms employ different tree growing methods. LGBM is designed to optimize training speed and memory usage while maintaining high accuracy. This is achieved through histogram-based leaf node separation technology, which improves storage efficiency. Leaf-wise tree evolution in LGBM can lead to significant accuracy gains in each iteration, but also carries the risk of overfitting. To improve the efficiency and reliability of the LGBM algorithm, two unique methods are used: gradient-based one-sided sampling (GOSS) and exclusive feature bundling (EFB).

The LGBM classifier is the algorithm of choice due to its excellent performance on key metrics such as accuracy, Matthew correlation coefficient, and confusion matrix. Its integrated approach effectively captures complex patterns in data, resulting in superior predictive capabilities. In

addition, the minimum prediction time of the LGBM classifier is only 0.50 seconds, demonstrating unparalleled efficiency and being very practical for scenarios that require fast decision-making. While alternative models such as extra tree classifier and RF classifier provide competitive accuracy, the LGBM classifier strikes a balance between accuracy and computational efficiency, making it the best choice for a given prediction task. Its reliability and fast prediction are critical for use in a variety of environments.

3.2. Feature Selection

Feature selection is a crucial step in refining ML models, enhancing their interpretability and efficiency. Methods of feature selection have been utilized to eliminate non-essential features prior to the implementation of ML algorithms. In our research, we employed the Variance Threshold method as part of this process. By setting a threshold value, in our case, 0.1, this method systematically identified and excluded features with low variance from the original dataset. The underlying principle is rooted in the notion that features with minimal variance contribute less to the classification task and can be safely omitted. This strategic elimination of less informative features not only reduces dimensionality but also aids in mitigating the risk of overfitting. The practical implementation involved using the scikit-learn library in Python, where the Variance Threshold method was applied to the feature matrix (X). The resulting refined feature set, denoted as X2, was then utilized for training the subsequent classification model.

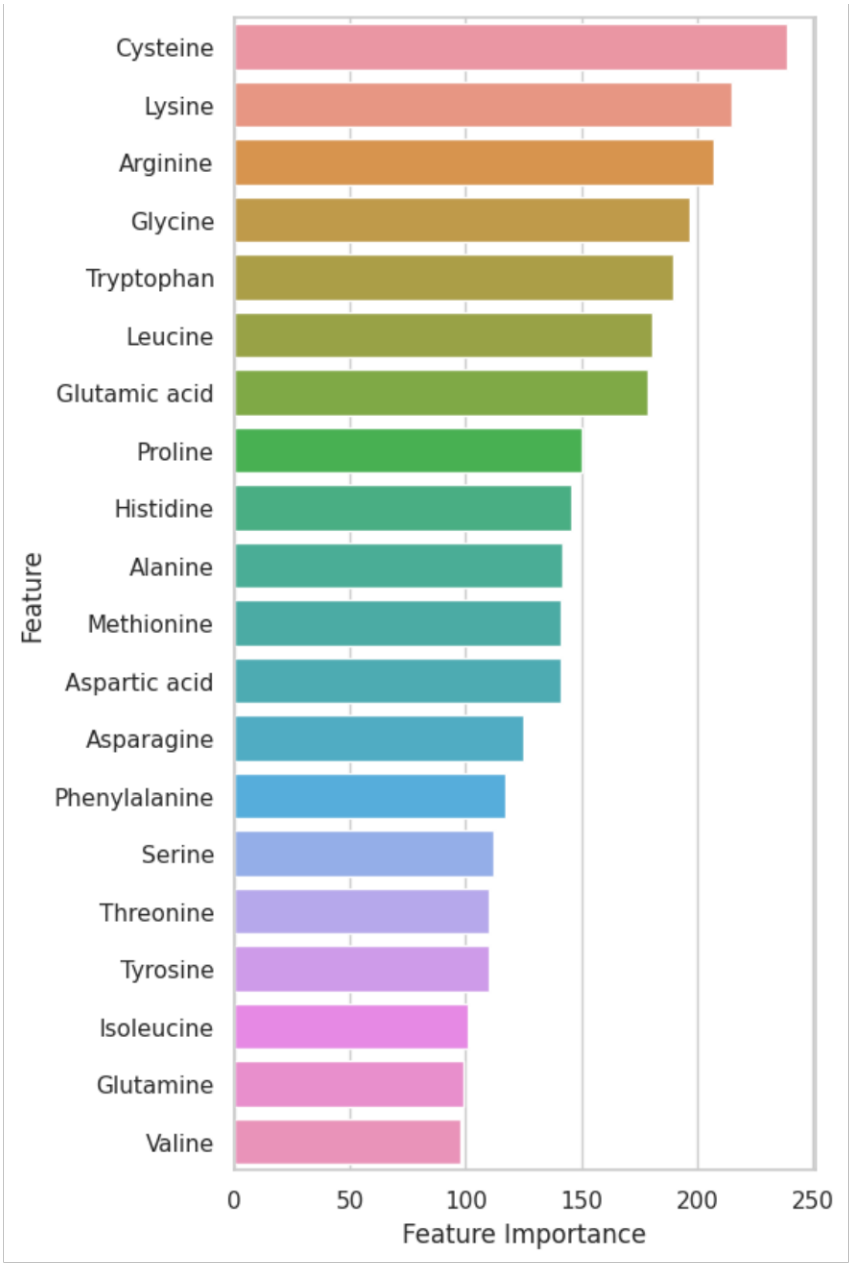


Figure 2: The determination of the significance of each amino acid in distinguishing between AMP and non-AMP sequences

3.3. Evaluation Matrix

In this study several performance metrics are used such as accuracy, recall, area under curve, and F1. Mathematically dataset can be described as: $M = (\rho_i, q_i)$, $i = \{1,2,3, \dots, k\}$ where k denotes total samples of datasets, ρ_i denotes test group in dataset, and q_i denotes the label of dataset. the performance metrics for model evaluation as follows:

(i) Accuracy: the parameter indicates the rate at which the model detects peptide, and it can be described as follows [21, 22]:

$Accuracy = \sum_{i=1}^{n} TP_i/k$, where TP is the true positive.

(ii) Recall: It is useful when false negatives are minimized. The equation of recall is as follows [23, 24]:

$recall = \frac{TP_i}{TP_i + FN_i}$, where FN is the false negative value.

(iii) Precision: It measures the percentage of accurately detected samples in each class. It expresses as follows [25, 26]:

$Precision_i = \frac{TP_i}{TP_i + FP_i}$, where FP is the false positive value.

(iv) F1-score: the harmonic means of Precision and Recall. A value of 1 indicates that the model is performing flawlessly. F1 can be expressed as [27, 28]:

$$F1 = \frac{2(TP_i)}{2(TP_i + FP_i + FN_i)} = \frac{2.sensitivity \times Precision_i}{2(senstivity + Precision_i)}$$

(v) MCC: the model's ability to discriminate between positive and negative instances. It can be expressed as follows [29]:

$$MCC = \frac{TP_i.TN_i - FP_i.FN_i}{\sqrt{TP_i.FP_i \times TN_i.FN_i}}.$$

4. Results and Discussion

4.1. Implementation

In our experiment, the model was designed based on python3, and the Keras2.7 library was utilized to train our model. For programming purposes, the Google Colab online editor was employed. We utilized the Gradio API to determine whether a given sequence is an AMP or a non-AMP in our research. This integration provides a straightforward user interface where users can input peptide sequences effortlessly.

4.2. Overall Performance

We can see from Figure 3 and Table 2 the result of the examination of the proposed model based on LGBM against other ML algorithms. Moreover, table 3 depicts the values of the control parameters specific to each classifier. These parameters value will directly affect the overall performance.

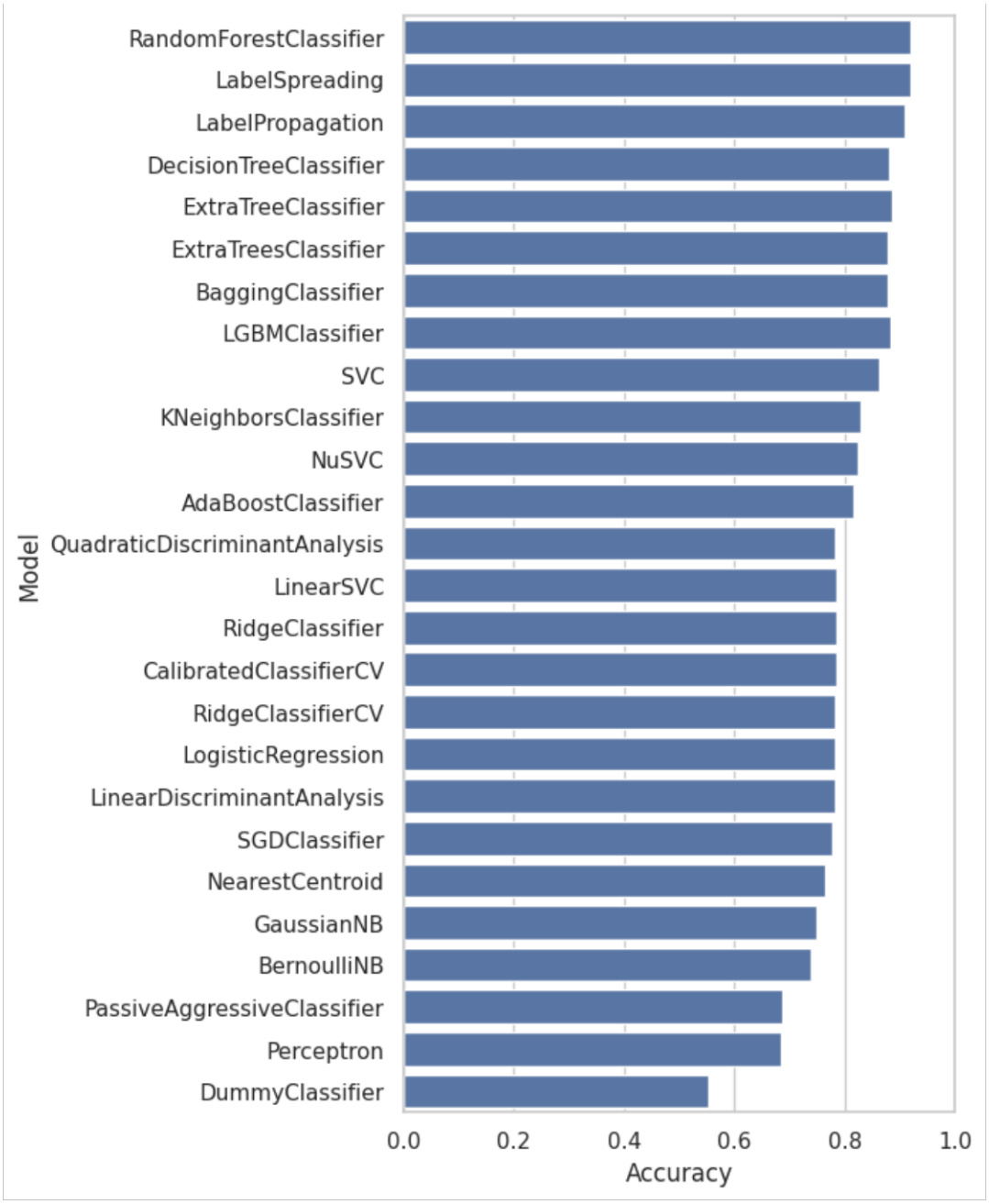


Figure 3: Select the best fit algorithm.

Table 2: Comparative analysis of proposed model based on LGBM classifier vs. other ML algorithms.

Model	Accuracy	Balanced	F1	MCC	Time
LGBM	0.92	0.92	0.92	0.83	0.51
Extra Trees	0.92	0.91	0.92	0.84	1.67
RF	0.91	0.90	0.91	0.81	2.43
KNN	0.88	0.89	0.88	0.77	0.28
SVC	0.89	0.89	0.89	0.77	3.65
Label Spreading	0.88	0.88	0.88	0.76	8.40
Label Propagation	0.88	0.88	0.88	0.76	5.19
Bagging	0.88	0.88	0.88	0.76	0.99
NuSVC	0.86	0.86	0.86	0.72	5.33
AdaBoost	0.83	0.83	0.83	0.65	1.23
Decision Tree	0.82	0.82	0.82	0.64	0.29
QDA	0.82	0.81	0.81	0.62	0.12
Extra Tree	0.78	0.78	0.78	0.56	0.12
Calibrated Classifier CV	0.79	0.78	0.79	0.57	4.61
Linear SVC	0.79	0.78	0.78	0.56	1.20
LR	0.78	0.78	0.78	0.56	0.14
Linear Discriminant	0.78	0.78	0.78	0.56	0.20
Ridge	0.78	0.78	0.78	0.56	0.14
Ridge Classifier CV	0.78	0.78	0.78	0.56	0.26
SGD	0.78	0.77	0.78	0.55	0.25
Nearest Centroid	0.76	0.76	0.76	0.52	0.07
Gaussian NB	0.75	0.74	0.75	0.49	0.06
Bernoulli NB	0.74	0.74	0.74	0.47	0.11
Passive Aggressive	0.69	0.69	0.69	0.39	0.10
Perceptron	0.68	0.69	0.69	0.38	0.11
Dummy	0.55	0.50	0.39	0.00	0.09

Table 3: values of the parameters specific to each classifier.

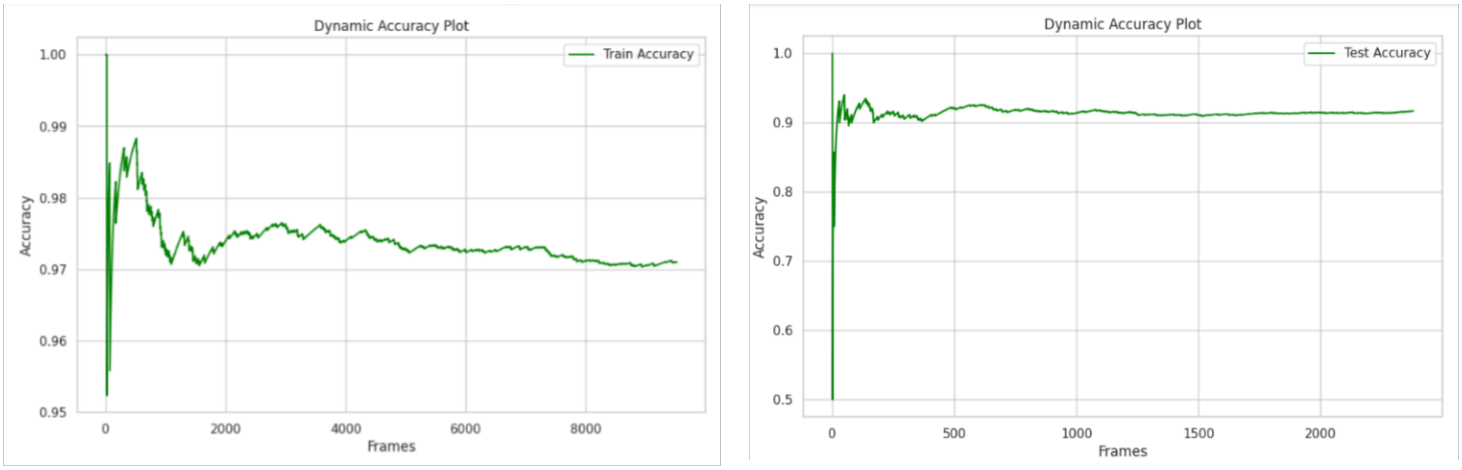
Classifier	Parameters
LGBM	boosting_type = gbdt, num_leaves=31, learning_rate=0.1

	min_child_samples=20 min_child_weight=0.001
Extra Trees	n_estimators=100, max_depth=None, min_samples_split=2
RF	n_estimators=100, max_depth=None, min_samples_split=2
KNN	n_neighbors=5, weights=uniform, algorithm=auto
SVC	C=1.0, kernel=rbf, gamma=scale
Label Spreading	kernel=rbf, alpha=1.0
Label Propagation	kernel=rbf, gamma=scale, alpha=1.0
Bagging Classifier	n_estimators=10, max_samples=1.0, max_features=1.0
NuSVC	nu=0.5, kernel = rbf, gamma=scale
AdaBoost	n_estimators =50, learning_rate =1.0
Decision Tree	criterion=gini, splitter=best, max_depth=None
Quadratic Discriminant Analysis	priors=None, reg_param=0.0, store_covariance=False
Extra Tree	criterion=gini, splitter=random, max_depth=None
Calibrated Classifier CV	base_estimator=None, method=sigmoid, cv=3
Linear SVC	C=1.0, penalty=l2, loss=squared_hinge
Logistic Regression	penalty=l2, C=1.0, solver=lbfgs
Linear Discriminant Analysis	solver=svd, shrinkage=None, priors=None

Ridge	alpha=1.0, solver=auto, max_iter=None
Ridge Classifier CV	Alphas = (0.1, 1.0, 10.0), store_cv_values=False
SGD	loss=hinge, penalty=l2, alpha=0.0001
Nearest Centroid	metric=euclidean, shrink_threshold=None
Bernoulli NB	alpha=1.0, binarize=0.0, fit_prior=True
Passive Aggressive	C=1.0, fit_intercept=True, max_iter=1000, tol=0.001
Perceptron	Penalty =None, alpha=0.0001, fit_intercept =True, max_iter =1000
Dummy	Strategy =stratified

The selection of the LGBM for training is a crucial step in developing a reliable prediction system. LGBM provides efficiency and effectiveness, particularly in handling large datasets and providing high-performance results. Once the model is trained, evaluating its performance using various metrics becomes essential for a comprehensive understanding of its capabilities.

During the evaluation phase, key metrics such as accuracy, confusion metrics, and the Matthews correlation coefficient (MCC) are utilized. Accuracy is a fundamental measure that indicates the proportion of correct predictions. In this case, the training set exhibits an impressive accuracy of approximately 97%, as shown in Figure 4a, demonstrating the model's ability to make accurate predictions on the data it was trained on. Additionally, the testing set maintains a commendable accuracy of around 91%, as depicted in Figure 4b, indicating that the model generalizes well to new, unseen data.



(a) Accuracy on training (b) Accuracy on testing

Figure 4: Performance comparison of accuracy in training and testing operations

Delving further into the evaluation metrics, MCC offers a nuanced perspective by considering the entire confusion matrix. For the training set, the MCC reaches approximately 94% see Figure 4, indicating a robust performance in terms of both true positive and true negative predictions. However, on the testing set, the MCC is slightly lower, at around 83%, see Figure 6.

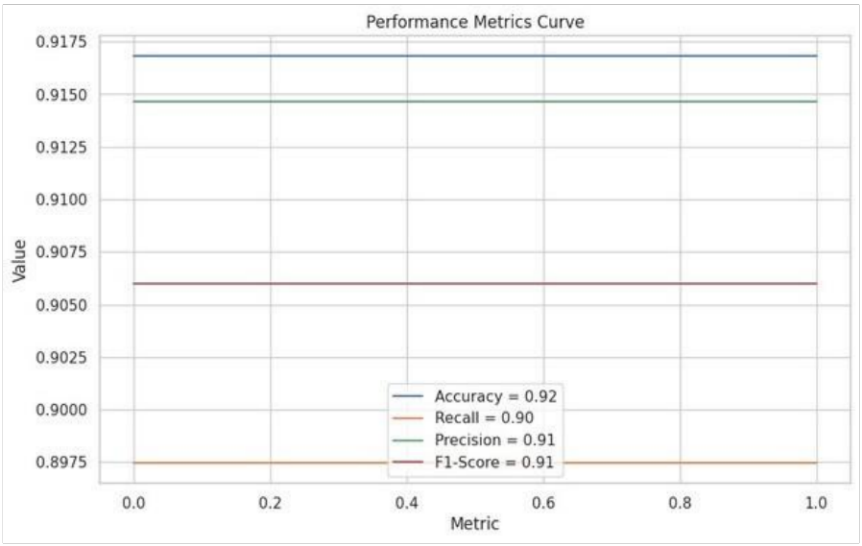
4.3. Classifier comparative analysis based on the optimal features.

The performance of the proposed model is evaluated in this section to enhance the classification between AMP and non-AMP sequences. The experiment was conducted between different ML classifier algorithms as shown in Table 4. The validation of high-rate accuracy was LGBM and Extra Trees with 0.92%, then MCC with 0.84% for Extra Trees and 0.83% for LGBM. Recall is another performance metric to evaluate our proposed model. The highest rate was KNN with 0.92%, the second highest rate was LGBM with a slight difference 90%. LGBM, Extra Trees, and RF were gain the highest rate in measuring AUC with 0.97%. the first rank in the precision was Extra Trees with 0.93%, the second one RF 0.92%, and LGBM comes in the third with 0.91%. In the F1-score, the LGBM and Extra Trees have the highest rate among other compared algorithms where both have the same value of 0.92%.

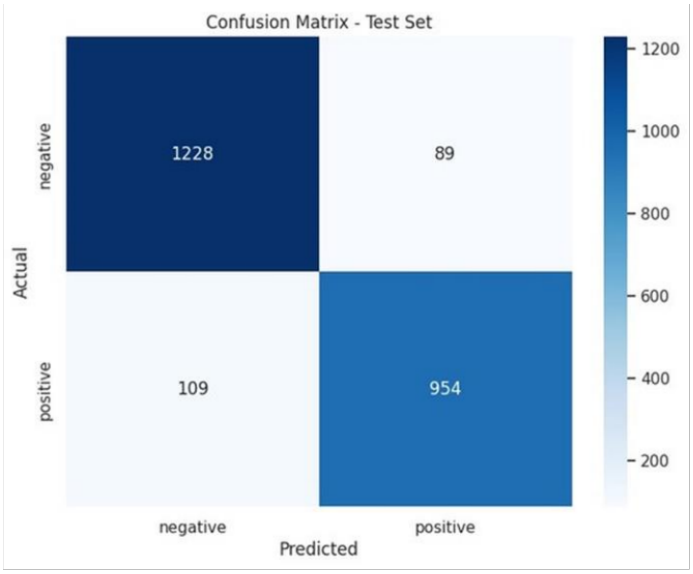
Table 4: The performance metrics of the ML approaches.

Classifier	Accuracy	Balanced Accuracy	MCC	Recall	AUC	Precision	F1-score
LGBM	**0.92**	**0.92**	0.83	**90%**	**0.97%**	0.91%	**0.92**
Extra Trees	**0.92**	0.91	**0.84**	0.88%	**0.97%**	**0.93%**	**0.92**
RF	0.91	0.90	0.81	0.87%	**0.97%**	0.92%	0.91
KNN	0.88	0.89	0.77	0.92%	0.94%	0.81%	0.88
SVC	0.89	0.89	0.77	0.87%	0.95%	0.87%	0.89

Figure 5 (a) presents underscores the significance of incorporating additional key metrics, namely F1 score, recall, and precision, in the assessment of a classification model. It emphasizes that a holistic and thorough evaluation of the model's performance extends beyond a singular metric, necessitating a nuanced consideration of multiple factors to gauge its effectiveness accurately. This visual representation serves as a visual testament to the importance of a comprehensive approach, highlighting the interconnected nature of these metrics in providing a more nuanced and insightful evaluation of classification model performance. Figure 5 (b) provides the confusion matrix with TP, FP, TN, and FN of the new sample prediction.

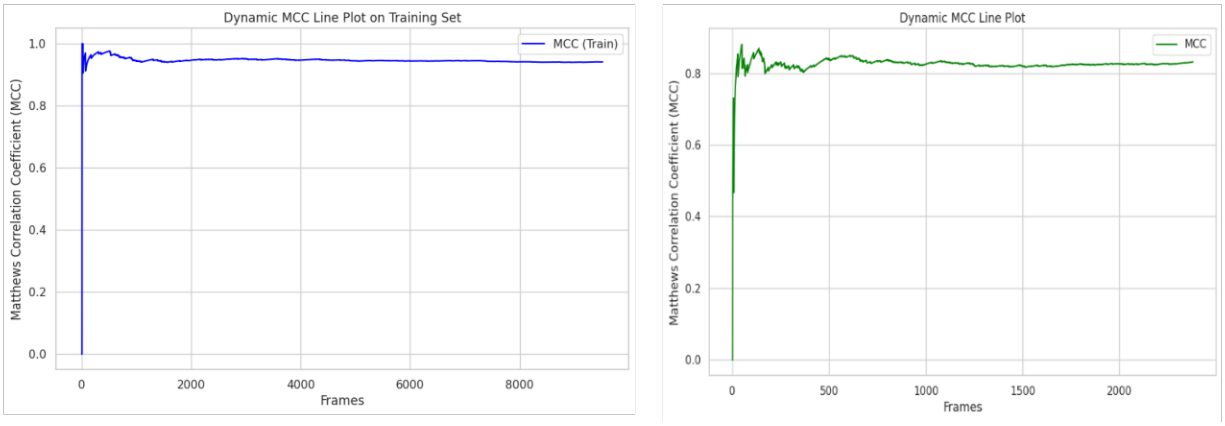


(a)



(b)

Figure 5: (a) shows different performance metrics including accuracy, recall, precision, and F1-score. (b) shows Confusion matrix with TP, FP, TN, and FN.



(a) (b)

Figure 6: (a) Shows MCC visualization on training dataset. (b): shows Figure 6: MCC visualization on testing dataset.

5. Conclusion

Studying metalloproteins, metalloproteomes, and their evolutionary tendencies in nature is made possible by the potent tool that bioinformatics provides. In this article, we focus on examining the performance of machine learning to predict the AMP sequence. We have proposed a new bioinformatics tool to evaluate 20 sequences generated. Among the several models created using distinct classifiers like linear discriminant, decision tree, KNN, the LGBM outperformed with the highest performance metrics on a dataset with size 6623 instances. In future endeavors, it's necessary to design a deep learning framework that is more precise and all-encompassing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data available on request due to ethical restrictions.

References

[1] Porto, William F., Állan S. Pires, and Octavio L. Franco. "Antimicrobial activity predictors benchmarking analysis using shuffled and designed synthetic peptides." Journal of Theoretical Biology 426 (2017): 96-103.

[2] Riedling, Olivia, Allison S. Walker, and Antonis Rokas. "Predicting fungal secondary metabolite activity from biosynthetic gene cluster data using machine learning." Microbiology Spectrum (2024): e03400-23.

[3] Han, So-Ra, et al. "Evidential deep learning for trustworthy prediction of enzyme commission number." Briefings in Bioinformatics 25.1 (2024): bbad401.

[4] Kiggundu, Reuben, et al. "A One Health approach to fight antimicrobial resistance in Uganda: Implementation experience, results, and lessons learned." Biosafety and Health (2024).

[5] Toropova, Alla P., et al. "Prediction of antimicrobial activity of large pool of peptides using quasi-SMILES." BioSystems 169 (2018): 5-12.

[6] Bournez, Colin, et al. "CalcAMP: a new machine learning model for the accurate prediction of antimicrobial activity of peptides." Antibiotics 12.4 (2023): 725.

[7] Barcin, Tunga, et al. "Deep learning approach to the discovery of novel bisbenzazole derivatives for antimicrobial effect." Journal of Molecular Structure 1295 (2024): 136668.

[8] Arif, Muhammad, et al. "iMRSAPred: Improved Prediction of Anti-MRSA Peptides Using Physicochemical and Pairwise Contact-Energy Properties of Amino Acids." ACS Omega (2024).

[9] Aguilera-Puga, Mariana D. C., et al. "Accelerating the discovery and design of antimicrobial peptides with artificial intelligence." Computational Drug Discovery and Design. New York, NY: Springer US, 2023. 329-352.

[10] Teimouri, Hamid, Angela Medvedeva, and Anatoly B. Kolomeisky. "Bacteria-Specific Feature Selection for Enhanced Antimicrobial Peptide Activity Predictions Using Machine-Learning Methods." Journal of Chemical Information and Modeling 63.6 (2023): 1723-1733.

[11] Pirtskhalava, Malak, et al. "DBAASP v. 2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides." Nucleic acids research 44.D1 (2016): D1104-D1112.

[12] Porto, William F., Állan S. Pires, and Octavio L. Franco. "CS-AMPPred: an updated SVM model for antimicrobial activity prediction in cysteine-stabilized peptides." PLoS One 7.12 (2012): e51444.

[13] Waghu, Faiza Hanif, et al. "CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides." Nucleic acids research 44.D1 (2016): D1094-D1097.

[14] Hammami, Riadh, et al. "BACTIBASE: a new web-accessible database for bacteriocin characterization." BMC microbiology 7 (2007): 1-6.

[15] Speck-Planche, Alejandro, et al. "First multitarget chemo-bioinformatic model to enable the discovery of antibacterial peptides against multiple Gram-positive pathogens." Journal of chemical information and modeling 56.3 (2016): 588-598.

[16] Jenssen, HÅvard, et al. "QSAR modeling and computer-aided design of antimicrobial peptides." Journal of peptide science: an official publication of the European Peptide Society 14.1 (2008): 110-114.

[17] Torrent, Marc, et al. "Connecting peptide physicochemical and antimicrobial properties by a rational prediction model." PloS one 6.2 (2011): e16968.

[18] Webb-Robertson, Bobbie-Jo M. "Support vector machines for improved peptide identification from tandem mass spectrometry database search." Mass Spectrometry of Proteins and Peptides: Methods and Protocols (2009): 453-460.

[19] Gabere, Musa Nur, and William Stafford Noble. "Empirical comparison of web-based antimicrobial peptide prediction tools." Bioinformatics 33.13 (2017): 1921-1929.

[20] Datasets, https://axp.iis.sinica.edu.tw/AI4AMP/helppage.html , accessed in February 2024

[21] Waghu, Faiza Hanif, and Susan Idicula-Thomas. "Collection of antimicrobial peptides database and its derivatives: Applications and beyond." Protein Science 29.1 (2020): 36-42.

[22] Chicco, Davide, and Giuseppe Jurman. "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification." BioData Mining 16.1 (2023): 1-23.

[23] Hesterkamp, Thomas. "Antibiotics clinical development and pipeline." How to Overcome the Antibiotic Crisis: Facts, Challenges, Technologies and Future Perspectives (2015): 447-474.

[24] Ayyad, Sarah M., et al. "10 A Multimodal MR-Based." *Handbook of Texture Analysis: AI-Based Medical Imaging Applications* (2024): 209.

[25] Ayyad, Sarah M., et al. "Prostate cancer detection using histopathology image analysis." *Artificial Intelligence in Cancer Diagnosis and Prognosis, Volume 3: Brain and prostate cancer*. Bristol, UK: IOP Publishing, 2022. 11-1.

[26] Ayyad, Sarah M., et al. "A new framework for precise identification of prostatic adenocarcinoma." *Sensors* 22.5 (2022): 1848.

[27] Balaha, Hossam Magdy, et al. "Early Diagnosis of Prostate Cancer Using Parametric Estimation of IVIM from DW-MRI." *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023.

[28] Balaha, Hossam Magdy, et al. "Precise Prostate Cancer Assessment Using IVIM-Based Parametric Estimation of Blood Diffusion from DW-MRI." *Bioengineering* 11.6 (2024): 629.

[29] Lee, Hao-Ting, et al. "A large-scale structural classification of antimicrobial peptides." BioMed research international 2015 (2015).