 doi:10.20944/preprints202202.0175.v1

Article

Prediction of Linear Cationic Antimicrobial Peptides Active Against Gram Negative and Positive Bacteria based on Machine Learning Models

Ümmü Gülsüm Söylemez[1,5,#], Malik Yousef[2], Zülal KESMEN[3], Mine Erdem Büyükkiraz[4] and Burcu Bakır-Güngör [5,#,*]

[1] Department of Computer Engineering, Faculty of Engineering, Muş Alparslan University, Muş, Turkey; og.uzut@alparslan.edu.tr
[2] Department of Information Systems, Zefat Academic College, Zefat 13206, Israel; malik.yousef@gmail.com
[3] Department of Food Engineering, Faculty of Engineering, Erciyes University, Kayseri, Turkey; zkesmen@erciyes.edu.tr
[4] Department of Nutrition and Dietetics, School of Health Sciences, Cappadocia University, Nevsehir, Turkey; mine.buyukkiraz@kapadokya.edu.tr
[5] Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, Turkey; burcu.gungor@agu.edu.tr
* Correspondence: burcu.gungor@agu.edu.tr
# These authors contributed equally to this work.

Abstract: Antimicrobial peptides (AMPs) are considered as promising alternatives to conventional antibiotics in order to overcome the growing problems of antibiotic resistance. Computational prediction approaches receive an increasing interest to identify and design the best candidate AMPs prior to the *in-vitro* tests. In this study, we focused on the linear cationic peptides with non-hemolytic activity, which are downloaded from the Database of Antimicrobial Activity and Structure of Peptides (DBAASP). Referring to the MIC (Minimum inhibition concentration) values, we have assigned a positive label to a peptide if it shows antimicrobial activity; otherwise the peptide is labeled as negative. Here, we focused on the peptides showing antimicrobial activity against Gram-negative and against Gram-positive bacteria separately, and we created two datasets accordingly. Ten different physico-chemical properties of the peptides are calculated and used as features in our study. Following data exploration and data preprocessing steps, a variety of classification algorithms are used with 100-fold Monte Carlo Cross Validation to build models and to predict the antimicrobial activity of the peptides. Among the generated models, Random Forest has resulted in the best performance metrics for both Gram-negative dataset (Accuracy: 0.98, Recall: 0.99, Specificity: 0.97, Precision: 0.97, AUC: 0.99, F1: 0.98) and Gram-positive dataset (Accuracy: 0.95, Recall: 0.95, Specificity: 0.95, Precision: 0.90, AUC: 0.97, F1: 0.92) after outlier elimination is applied. This prediction approach might be useful to evaluate the antibacterial potential of a candidate peptide sequence before moving to the experimental studies.

Keywords: antimicrobial peptide prediction; sequence analysis; random forest

1. Introduction

Antimicrobial peptides (AMPs) are part of innate immunity and are natural antibiotics encoded by specific genes [1]. They are produced by various tissues and cell types of human, plant and animal species. These antimicrobial peptides usually contain 12 to 50 amino acids [2]. Nowadays, in parallel with the elevated use of antibiotics, resistance to antibiotics is rapidly increasing. The World Health Organization (WHO) reported that antimicrobial resistance continues to rise up all over the world and new resistance

 doi:10.20944/preprints202202.0175.v1

mechanisms emerge. Therefore, we could face up with an era when infections can no longer be treated with antibiotics [3]. The increasing number of bacteria, which are resistant to antibiotics, creates a need for the development of new antimicrobial agents that can be applied in treatment [4]. Studying the properties of antimicrobial peptides in detail is a very important topic for drug design [5]. Although AMPs are mainly used to kill Gram-positive and negative bacteria, they have potential to fight against mycobacteria, viruses, and cancerous cells. In this respect, AMPs are considered as a powerful alternative to antibiotics since they have lower risk to develop resistance [3], [4]. Hence, discovering or designing novel antimicrobial peptides became a major field of interest. Along this line, several computational approaches such as de novo computational design [6]–[9], linguistic model [10], [11], pattern insertion algorithm [12]–[15], evolutionary-genetic algorithms [16]–[19] have been proposed for predicting the antimicrobial activity of AMPs and for identifying promising AMP candidates without undertaking expensive wet-lab experiments. Among different computational methods for the estimation of antimicrobial peptides [20], the use of machine learning methods became popular [21]–[24]. Machine learning is a computational technique, where the generated models can make predictions via learning the data [25]. Significant advancements in computational power and easy-to-use statistical learning tools that have come to the fore in recent years have increased the popularity of machine learning approaches. In this respect, machine learning, which can leverage large datasets that are produced by high-throughput methods, has become a viable option for the accurate classification of AMPs [26]. Lata *et al.* used the SVM method for prediction and classification of peptides on data which was collected from Antimicrobial Peptides Database [24]. Their model is based on aminoacid composition; and using five-fold cross validation they obtained 92.14% accuracy [24]. Burdukiewicz *et al.* attempted to identify essential AMP potential regions via applying random forest as a classification algorithm [27]. Chung *et al.* makes predictions for antimicrobial peptides on different organisms including amphibians, humans, fish, insects, plants, bacteria,and mammals [28]. Amino acid compositions, amino acid pairs, and the physicochemical properties are used as features. They performed feature selection, and applied random forest (RF), SVM, KNN algorithms. They reported that RF generated the best result, which was over 92% accuracy on all tested organisms [28]. Bhadra *et al.* also utilized a random forest algorithm for AMP prediction using physicochemical properties as features [23]. They grouped each property into specific three classes. For example, for hydrophobicity property three classes are polar, neutral, hydrophobic, while these three classes are positive, neutral and negative for net charge property. They used AMP and Non-AMP data with different ratios, where 19 different ratios were used in total. 1:3 ratio yielded in high accuracy with 10 fold cross validation technique and reduced feature sets [23]. Wang *et al.* combined sequence alignment with feature selection methods for classification of AMPs [29]. Xiao *et al.* modeled a two-level classifier. First level is for classifying peptide sequences as an AMP, and the second level is to separate these AMPs into 10 functional categories [21]. There are many computational tools to predict AMPs based on machine learning approaches [17], [30]–[34]. Also, deep learning methods have been started to apply to antimicrobial peptides prediction problems. Bhadra *et al.* presents a method called deepAMP for sequences with length shorter than 30. In their method they use an optimal feature set of reduced amino acid composition with convolutional neural network and get 77% accuracy. They also compare their results with RF and SVM algorithms [35]. Su *et al.* designed a deep neural network which consists of an embedding layer and multi-convolutional layers [36]. Schneider *et al.* used self organizing maps as input layers for their feedforward neural network [37]. Witten *et al.* reported a convolutional neural network model for the classification and regression of AMPs [38]. Recently, deep neural networks have also been used for the prediction of antimicrobial peptides in different studies [39]–[42]. However, there is no standardization in terms of the use of machine learning methods for the AMP prediction.

 doi:10.20944/preprints202202.0175.v1

Nowadays, antimicrobial peptide databases provide comprehensive information on thousands of natural or synthetic antimicrobial peptides. The peptide sequences deposited in these databases can be utilized for de novo design of AMPs using computer-aided approaches [43], [44]. However, in silico tests of AMPs mostly do not take into account the antimicrobial effect mechanism of the peptides against target microorganisms. Therefore, in this study, we consider AMPs that are active against Gram-negative and Gram-positive bacterias separately. Different classification models are generated on each dataset and the results are compared using performance evaluation metrics in terms of accuracy, recall, specificity, precision, Area Under Curve, and F1 measure. The rest of this paper is organized as follows. Materials and Methods Section presents our dataset, or data preprocessing steps and the machine learning algorithms that we used to predict AMPs. Results Section highlights our findings and provides an extensive evaluation of our method. Discussions Section discusses the biological relevance of our findings. Finally, Conclusions Section concludes the paper and summarizes avenues for further research.

2. Materials and Methods

2.1 Dataset and Data Preprocessing

In this study, one of the most comprehensive AMP databases called "Database of Antimicrobial Activity and Structure of Peptides (DBAASP v.2. Http://dbaasp.org)" is used [45]. This database provides users detailed information about the chemical structure and activity of thousands of peptides, whose antimicrobial activity has been tested experimentally or in silico against more than 4200 different organisms (bacteria, fungi, some parasites, viruses and cancer cells). In Figure 1 we illustrate our data preprocessing steps. Linear cationic antimicrobial peptides (LCAMPs) are the largest class of AMPs and they are widely found in different organisms [46]. Therefore, LCAMPs which have antimicrobial activity against Gram-negative bacteria including *E. coli, P. aeruginosa, A. baumannii* species and Gram-positive bacteria including *S. aureus, L. monocytogenes, B. cereus* species are selected as target AMP class from the DBAASP. We have selected the peptides with lengths ranging from 20 to 50 amino acid (aa). As a continuation of this work, we plan to perform de novo antimicrobial peptide design by using the dataset that we have compiled in this study. Along this line, in therapeutic applications the prediction of non-hemolytic peptides are reported as more important than the hemolytic peptides for the elimination of the detrimental effects of AMPs on the host [47]. Hence, here we focused on non-hemolytic peptides.

In this study, the class labels of peptides are assigned according to the antimicrobial peptide activities against target organisms. In this respect, Minimum Inhibition Concentration (MIC) values are widely used to assess the *in vitro* levels of susceptibility or resistance of specific bacterial strains to a particular AMP [48]. Hence, we utilized MIC values provided at DBAASP for each protein against different target organisms. While the peptides having MIC value<25 against one of our target organisms are assigned as positive (antimicrobial); the peptides having MIC>100 assigned as negative (non-antimicrobial). This procedure is repeated separately for our Gram-negative and Gram-positive datasets. Hence, we assign class labels to each peptide in our dataset. The CD-HIT [49] program was used to eliminate the sequences that have more than 80% identity. The CD-HIT program is widely used in the AMP prediction problem for removing highly similar sequences [50]–[57]. The final dataset includes 231 positive (AMP) and 114 negative (Non-AMP) labeled peptides in the Gram-negative dataset; and 165 positive and 194 negative samples in the Gram-positive dataset. In other words, the Gram-negative dataset includes 231 peptides that show activity against Gram-negative bacteria (having MIC value < 25) and 114 peptides that do not show activity against Gram-negative bacteria (having MIC value >100). Similarly, the Gram-positive dataset is composed of 165 peptides that show activity against Gram-positive bacteria (having MIC

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2022 doi:10.20944/preprints202202.0175.v1

value < 25) and 194 peptides that do not show activity against Gram-positive bacteria (having MIC value >100).

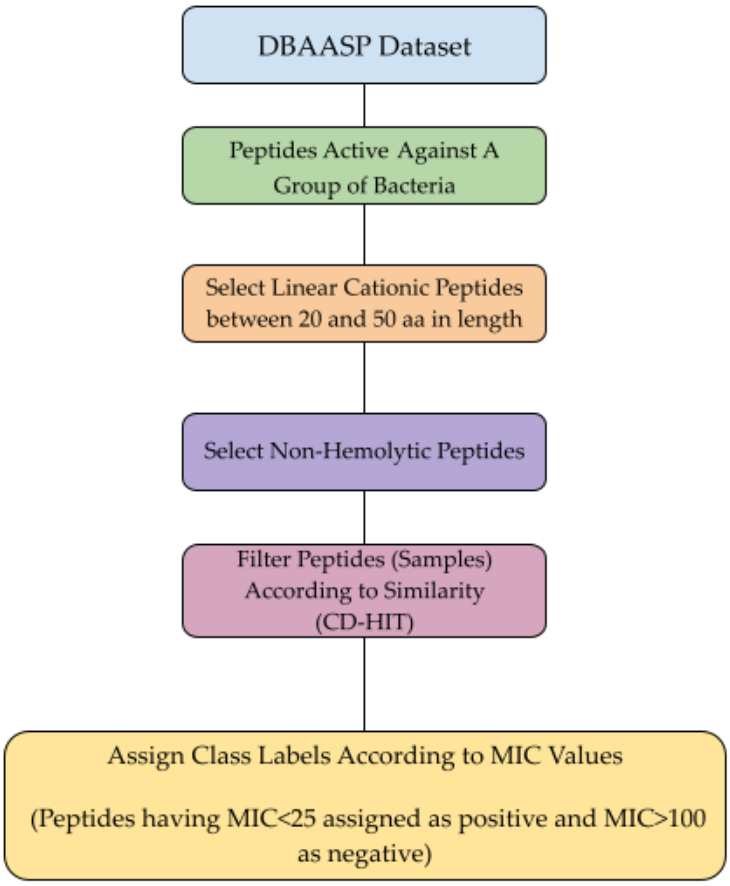


Figure 1. Workflow of data preprocessing

2.1.1.Feature Generation

Most AMPs exhibit their antimicrobial effects mainly by perturbing bacterial membrane integrity. Therefore, the development of an effective predictive model strongly depends on the deep understanding of physicochemical parameters, especially those that affect the AMP–membrane interaction. For AMPs, the sequence length of the peptide, normalized hydrophobic moment, normalized hydrophobicity, net charge, isoelectric point, penetration depth, orientation of peptides relative to the surface of membrane (tilt angle), propensity to disordering, linear moment and in vitro aggregation are widely used physico-chemical properties [9], [58], [59]. These parameters strongly affect the extent of peptide-membrane interactions and the depth of the penetration in lipid bilayer; and determine the mode of action of membrane targeting AMPs [60]. For instance, net charge reflects the propensity of electrostatic interaction of cationic peptides with the negatively charged membrane while hydrophobicity is responsible for the insertion and partition of the peptides into the hydrophobic core of the bilayer [5]. In our study, these 10 features were used as features to represent each peptide. All these features except sequence length are calculated by the DBAASP web server. Table 1 presents example sequences that are included in our Gram-negative dataset. As shown in Table 1, along with 10 physico-chemical properties, each peptide has a class label as 0 or 1, where 0 im-

 doi:10.20944/preprints202202.0175.v1

plies that the peptide is not active against Gram-negative bacteria; and 1 implies that the peptide is active against Gram-negative bacteria.

Table 1. An example of AMP and Non-AMP peptides included in our Gram-negative dataset and their physico-chemical properties, excerpted from DBAASP.

Name of Sequence	Sequence	Seq. Length	Norm. Hyd. Moment	Norm. Hyd.	Net Charge	Isoelectric Point	Penet. Depth	Tilt Angle	Disordered Conf. Propensity	Linear Moment	Propensity in vitro Aggregation	MeanMIC	Class (AMP category)
XPF-B2	GWASKIG TQLGKM AKVGLKE FVQS	24	1,11	-0,25	3	10,7	15	76	0,09	0,16	0	256,81	0
Ovalbumin (271-290)	SNVMEER KIKVYLPR MKMEE	20	0,13	-0,28	1	9,38	30	67	-0,11	0,29	0	800	0
MBI 29 A1	KWKSFIK KLTSVLK KVVTTAL PALIS	26	1,03	-0,54	6	11,37	12	106	0,16	0,27	3,4	9,33	1
Cyanophlyctin	FLNALKN FAKTAGK RLKSLLN	21	1,69	-0,24	5	11,74	15	88	-0,03	0,25	0	12	1
									… …				

2.2.Machine Learning Models

AdaBoost: Boosting technique creates a strong learner by bringing together several weak learners. The basic approach of boosting methods is to train the estimators cumulatively. In this model, the training set is first trained with a weak learner. For this algorithm, incorrectly predicted samples after the training step are important. In the next training phase, the incorrectly learned training data in the first iteration is retrained by giving more priority [61].

LogitBoost: LogitBoost has been developed to provide solutions to the overfitting problem experienced in AdaBoost. This algorithm linearly reduces the errors in the training to solve the above-mentioned problem [62].

Decision Tree: The decision tree creates a classification or regression model in the form of a tree structure. While dividing the dataset into smaller and smaller subsets, an associated decision tree is progressively and concurrently developed [63].

Random Forest: Random forests are an ensemble learning method for classification, regression and other tasks, by generating a large number of decision trees during the training phase and estimating the class or number according to the type of problem [64].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2022 doi:10.20944/preprints202202.0175.v1

Support Vector Machine: A support vector machine can be defined as a vector space based machine learning method that finds a decision boundary between the two classes that are furthest away from any point in the training data [65].

K-Nearest Neighbor: The k-nearest neighbor (KNN) algorithm is one of the supervised learning algorithms that is used in solving both classification and regression problems. The algorithm is used by making use of the data in a sample set with known classes. The distance of the new data, which will be added to the sample data set, is calculated according to the existing data, and its k closest neighbors are examined [66].

The Konstanz Information Miner (KNIME) platform is used for the implementation of our workflow [67]. Also, The Jupyter Notebook [68] was used for visualization.

2.2.1.Model Construction

As illustrated in Figure 2, we applied several machine learning algorithms that are explained in the above section to classify antimicrobial and Non-antimicrobial peptides. Also, we constructed stacking ensemble learners. All the findings we obtained in our study were obtained using 100-fold Monte Carlo Cross-Validation (MCCV) [69]. MCCV is a technique that selects a part of the data (unaltered) to create the training set, and then assigns the remaining data as the test set. This process is then repeated many times randomly, creating new training and testing segments each time. In our study, the training set is 90% of the data and the test is 10%.

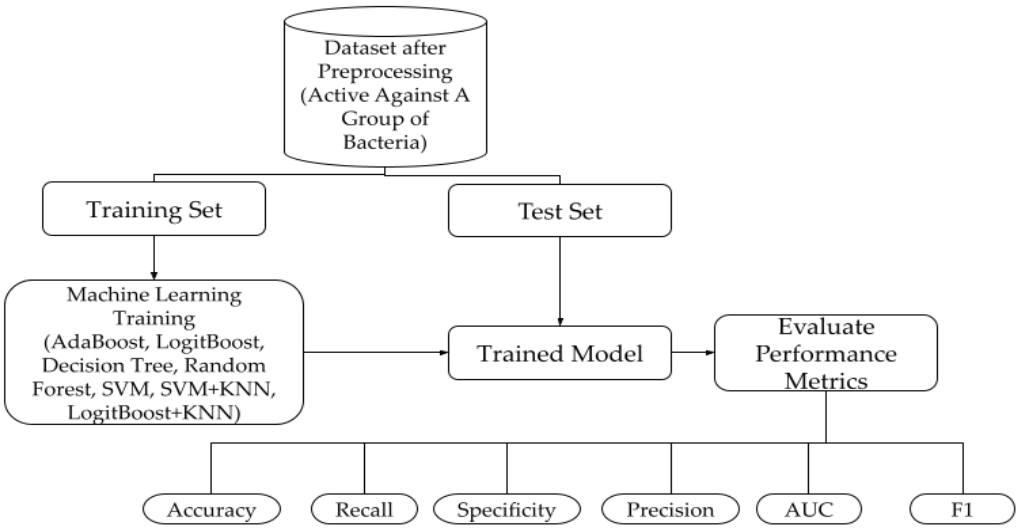


Figure 2. Flowchart of our Model Construction

2.2.2.Performance Metrics

We have assessed the performance of our models using several performance evaluation metrics such as accuracy, recall, specificity, area under curve and F1 measure. These metrics are employed as follows where TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2022 doi:10.20944/preprints202202.0175.v1

$$\text{Recall(Sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

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

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \quad (4)$$

3. Results

In our experiments, we have used different machine learning methods i) to learn whether the peptides in each of our datasets have antimicrobial activity or not; and ii) to classify them accordingly. To this end, we have applied methods such as AdaBoost, Decision Tree, LogitBoost, Random Forest, and Support Vector Machines. As shown in Tables 2 and 3, for both Gram-negative and Gram-positive datasets, Random Forest classifier resulted in the best performance metrics. While the accuracy rate reached up to 87% for Gram-positive data, this rate was 89% for gram negative data. Not only for accuracy rate, but also for other measures such as recall, specificity, precision, AUC and F1 measure, RF yielded the best performance metrics. Figure 3 displays the comparative evaluation of different models using AUC values for (a) Gram-negative dataset, and (b) Gram-positive dataset. As it can be seen in Figure 3a and in Table 3, while 0,92 AUC value is obtained for gram negative dataset, 0,90 AUC value is obtained for Gram-positive dataset (shown in Figure 3b and in Table 2) using RF classifier. While the AUC values of other classifiers range between 0,77-0,87 for Gram-positive dataset (shown in Figure 3b and in Table 2), it ranges between 0,78-0,89 for Gram-negative dataset.

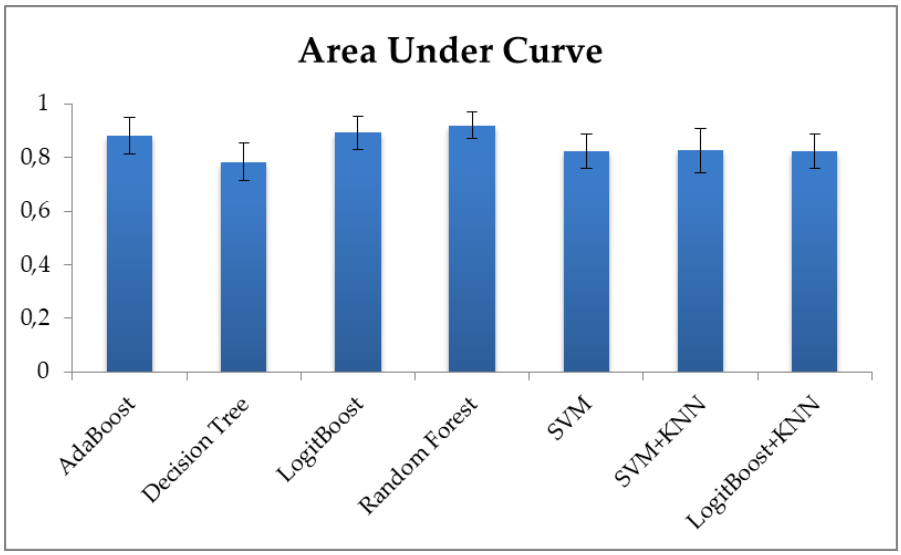
Table 2. Comparison of different models according to different performance metrics for Gram-positive dataset.

Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1
AdaBoost	0,84	0,85	0,83	0,83	0,86	0,83
Decision Tree	0,77	0,77	0,77	0,769	0,77	0,76
LogitBoost	0,83	0,84	0,82	0,83	0,87	0,83
Random Forest	**0,87**	**0,87**	**0,87**	**0,87**	**0,90**	**0,87**
SVM	0,77	0,85	0,71	0,75	0,81	0,78
SVM+KNN	0,76	0,81	0,72	0,76	0,80	0,77
LogitBoost+KNN	0,77	0,85	0,71	0,75	0,81	0,78

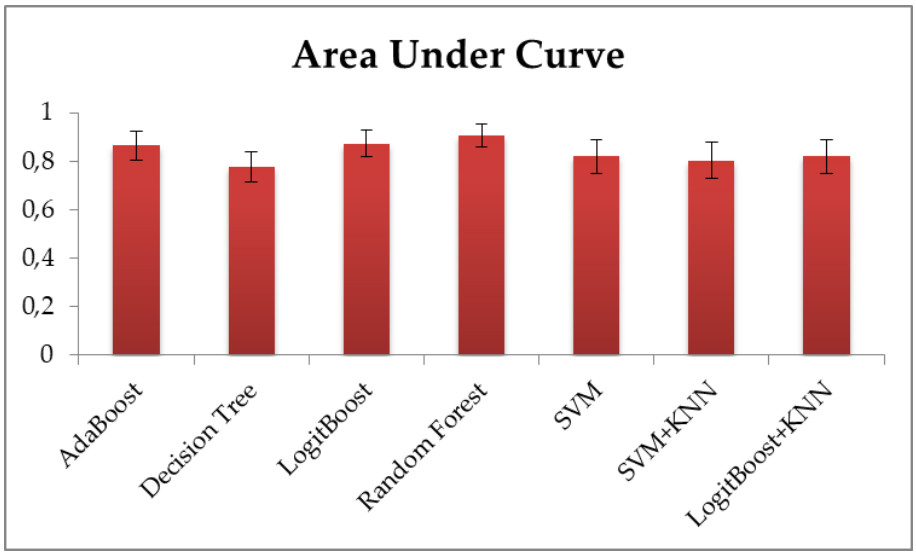
Table 3. Comparison of different models according to different performance metrics for Gram-negative dataset.

Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1
AdaBoost	0,85	0,92	0,72	0,87	0,88	0,89
Decision Tree	0,79	0,87	0,66	0,84	0,78	0,85
LogitBoost	0,86	0,92	0,74	0,88	0,89	0,90
Random Forest	**0,89**	**0,93**	**0,79**	**0,90**	**0,92**	**0,91**
SVM	0,80	0,93	0,56	0,81	0,82	0,86
SVM+KNN	0,80	0,93	0,56	0,81	0,82	0,86
LogitBoost+KNN	0,80	0,93	0,56	0,81	0,82	0,86

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2022 doi:10.20944/preprints202202.0175.v1



(a)



(b)

Figure 3. Comparison of the performances of different models in terms of their area under ROC curve (AUC) values with standard deviation values for (a) Gram-negative, and (b) Gram-positive dataset

3.1. Feature Scoring and Feature Ranking

Feature selection procedure tries to reduce the computational costs by removing redundant or irrelevant variables from input data. This technique contributes to better understanding the generated model and allows one to improve the model via focusing on the important features. In order to perform this task, one needs to score or rank the features in terms of how useful they are at predicting the output. There are different approaches for feature ranking that are based on statistics measurements or wrapper approaches that are based on machine learning [70]. Moreover, more advanced approaches that integrate biological knowledge into the machine learning algorithm for performing feature selection or for selecting groups of features are used in different

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2022 doi:10.20944/preprints202202.0175.v1

recent tools. Such an approach was adopted by different tools such as SVM RCE, SVM-RCE-R [71]–[73], maTE [74], CogNet [75], miRcorrNet [76], and Integrating Gene Ontology Based Grouping and Ranking [77]. Recently, these tools and their competitors were reviewed in [78].

In this study, for each tested machine learning algorithm, we have recorded the scores assigned to each feature during the MCCV (100 iteration) procedure. Since we get higher performance metrics using Random Forest classifier, we have utilized the feature scores of this model throughout the rest of the paper. When we analyze the feature scores (shown in Figures 4 and 5), we observe that Net Charge, Isoelectric Point, Disordered Conformation Propensity, Normalized Hydrophobicity and Normalized Hydrophobic Moment are more crucial features than others for both Gram-negative and positive datasets.

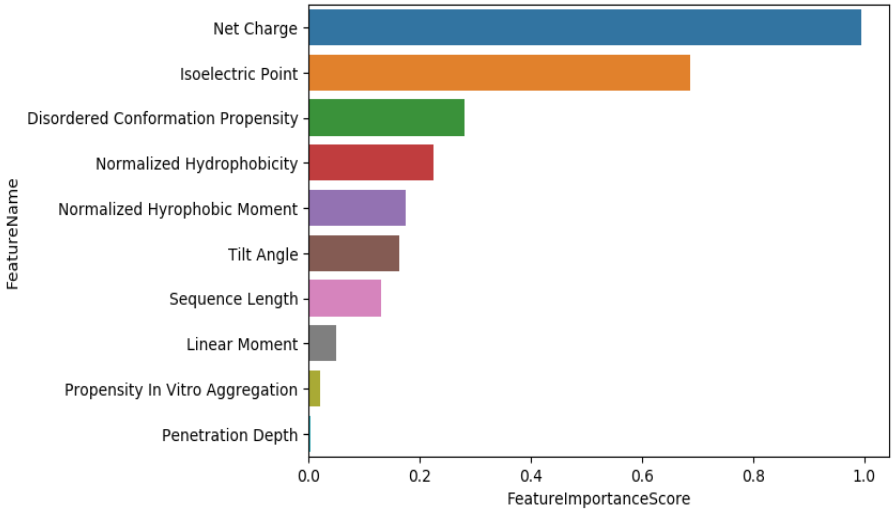


Figure 4. Feature ranking according to their importances in classification using random forest model in Gram-negative dataset.

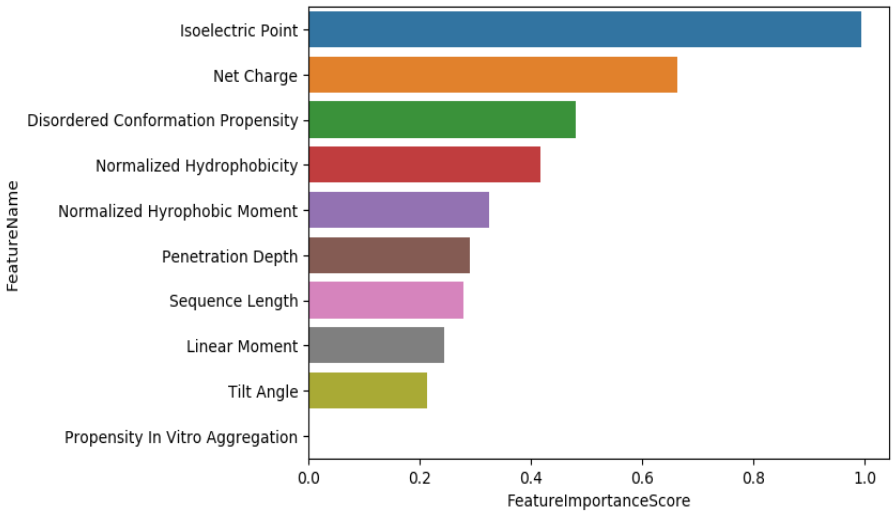
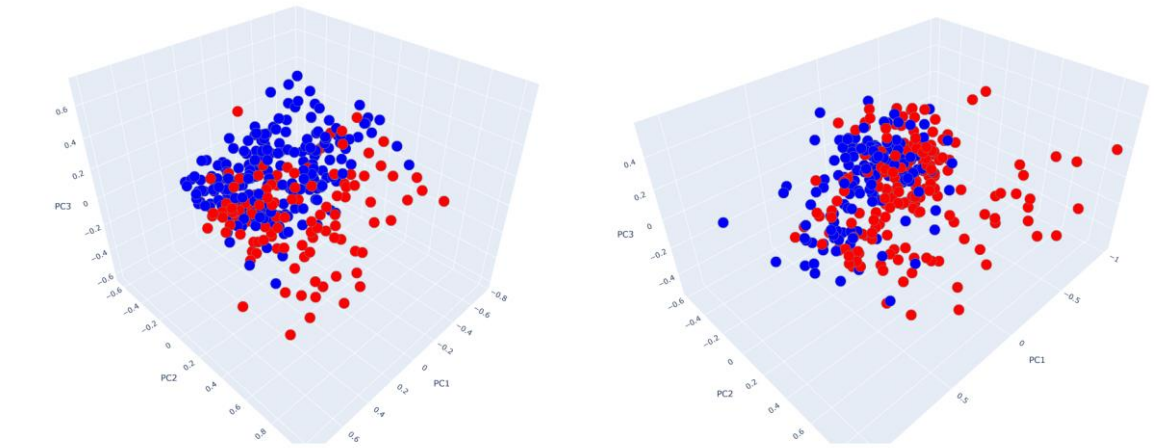


Figure 5. Feature ranking according to their importances in classification using random forest model in Gram-positive dataset.

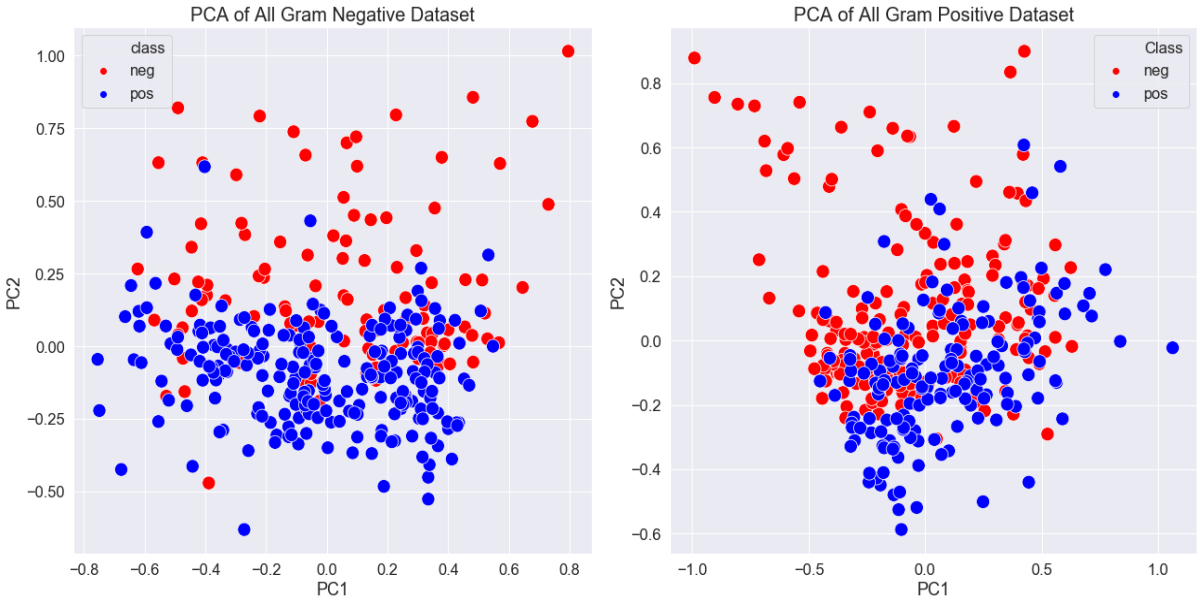
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2022 doi:10.20944/preprints202202.0175.v1

3.2.Data Exploration

In order to obtain the underlying structure of the data, we apply Principal Component Analysis (PCA) on Gram-negative and Gram-positive datasets separately. PCA is a dimensionality reduction technique that maps the data in high dimensional space (here each dimension corresponds to a physico-chemical property of a peptide) to a lower dimensional space (usually 2D or 3D) preserving the original structure of the data [79]. This technique is commonly used to highlight variation in a dataset and to capture strong patterns. Hence, PCA helps to visualize the data and the outliers. PCA has been applied to antimicrobial peptide data in several studies for data exploration and outlier detection purposes [80]–[83]. In our study, we also applied PCA to our dataset for visualizing the AMP and Non-AMP samples. In Figure 6, we present PCA results of the Gram-negative dataset (Figure 6a, 6c), and of the Gram-positive dataset (Figure 6b, 6d). While Figures 6a, 6b refer to the PCA results in 3D, Figures 6c, 6d refer to the PCA results in 2D. Interactive 3D plots are provided as supplementary material. We observe in Figure 6 that there are some outlier samples (peptides) in both Gram-negative and positive datasets.



(a) (b)



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2022 doi:10.20944/preprints202202.0175.v1

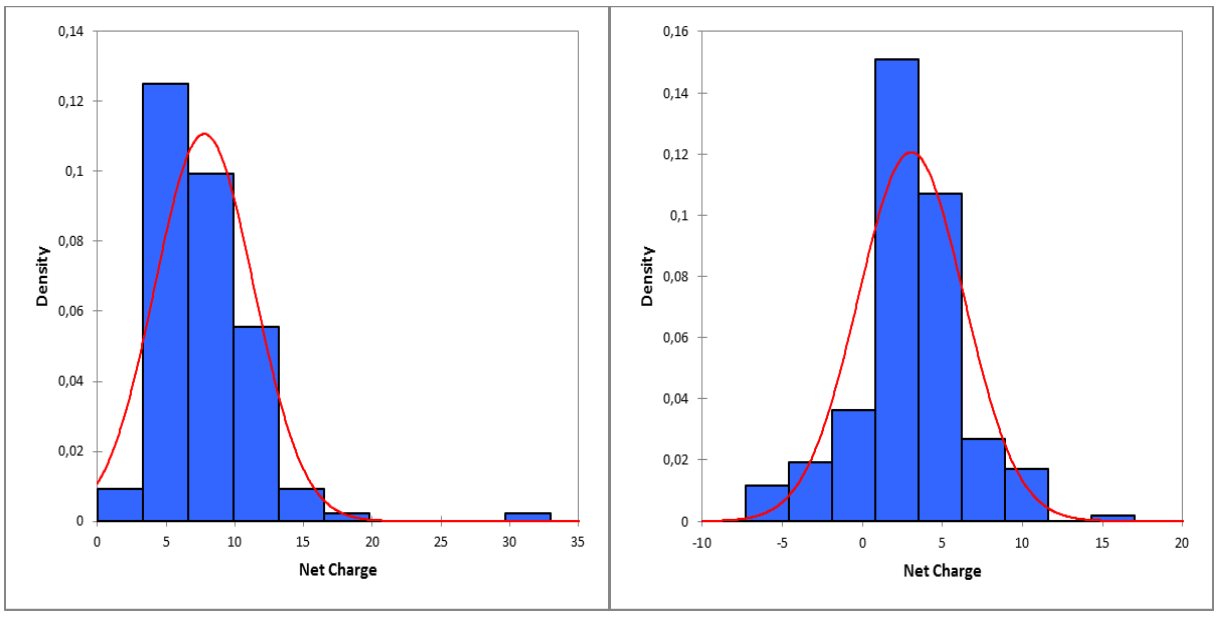
(c) (d)

Figure 6. Principal component analysis results for Gram-negative dataset are shown in (a) and (c); for Gram-positive dataset are shown in (b) and (d). While 3D plots are presented in (a) and (b), 2D plots are presented in (c) and (d).

3.3.Outlier Detection and Elimination

The presence of outliers can result in a poor fit and lower predictive modeling performance in classification or regression problems. For most machine learning datasets, due to the large number of input variables, the identification and removal of outliers is challenging by only using simple statistical methods. There are different computational approaches for outlier detection. One of those approaches depends on novelty detection based on machine learning [84], more specifically on one-class approaches [85], [86].

In this study, in order to have a more homogenous group of peptides having antimicrobial activities, we wanted to eliminate outlier samples (peptides) if one of their physico-chemical features acts as an outlier. To see the distribution of the attributes in positive class (AMP) and negative class (Non-AMP), we plotted the histograms for each feature. Figure 7 presents two histograms drawn for the Net Charge feature of the Gram-positive dataset for a) AMP class, b) Non-AMP class. It can be observed from Figure 7 that while the net charge values are in the range of [0 , 31] for AMP class, it is in the range of [-6, 16] for the negative class. Based on our analysis using such histograms, we define a certain range of values for each feature for the positive class (AMP, the peptides having antimicrobial activity). We perform this analysis separately for the Gram-positive dataset and the Gram-negative dataset and we eliminate the peptides in the positive class if their physico-chemical properties are outside of this predefined range. The range for each attribute is shown in Table 4.



(a) (b)

Figure 7. Graphical representation of Net Charge feature of the Gram-positive dataset. Histogram of (a) AMP class, (b) Non-AMP class.

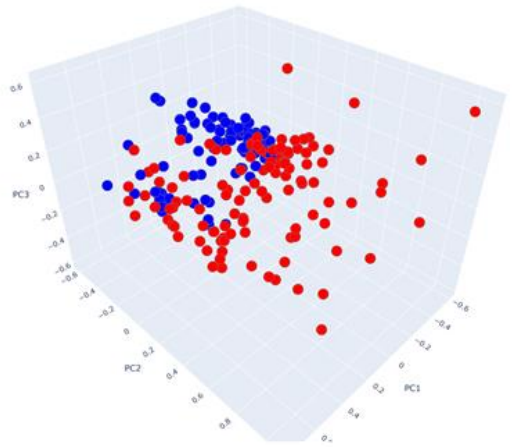
Table 4. Minimum and maximum values of each feature that are used in outlier elimination

Features	Gram-negative Dataset	Gram-positive Dataset

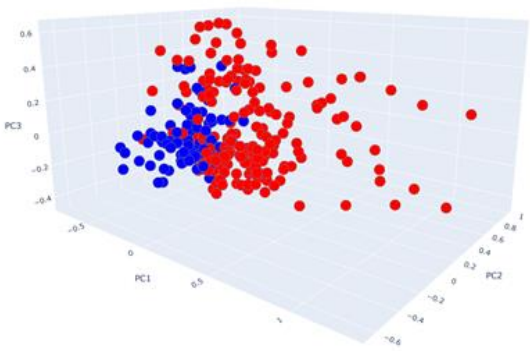
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2022 doi:10.20944/preprints202202.0175.v1

	Minimum threshold	Maximum threshold	Minimum threshold	Maximum threshold
Hydrophobic Moment	0.4	2	0.1	1.7
Normalized Hydrophobicity	-0.9	0.55	-0.8	1
Net Charge	5	13	4	13
Isoelectric Point	10.5	13	10	13
Penetration Depth	13	30	12	30
Tilt Angle	40	150	30	152
Linear Moment	0.1	0.4	0.15	0.32
Propensity in vitro Aggregation	0	250	0	87
Disordered Conformation Propensity	-0.5	0.08	-0.85	0.15

At the end of the outlier elimination step, we get 194 Non-AMPs and 88 AMPs for the Gram-positive dataset; 114 Non-AMPs and 90 AMPs for the Gram-negative dataset. In Figure 8, we present PCA results of the Gram-negative dataset (shown in a, c); and of the Gram-positive dataset (shown in b,d) after outlier detection and elimination. While PCA plots are presented in 3D in (Figure 8a, b), they are presented in 2D in (Figure 8c, d). While the red colors refer to Non-AMPs, blue colors indicate AMPs. Compared with Figure 6, Figure 8 implies that the positive class members are better separated from negative class members for both datasets after outliers are eliminated.

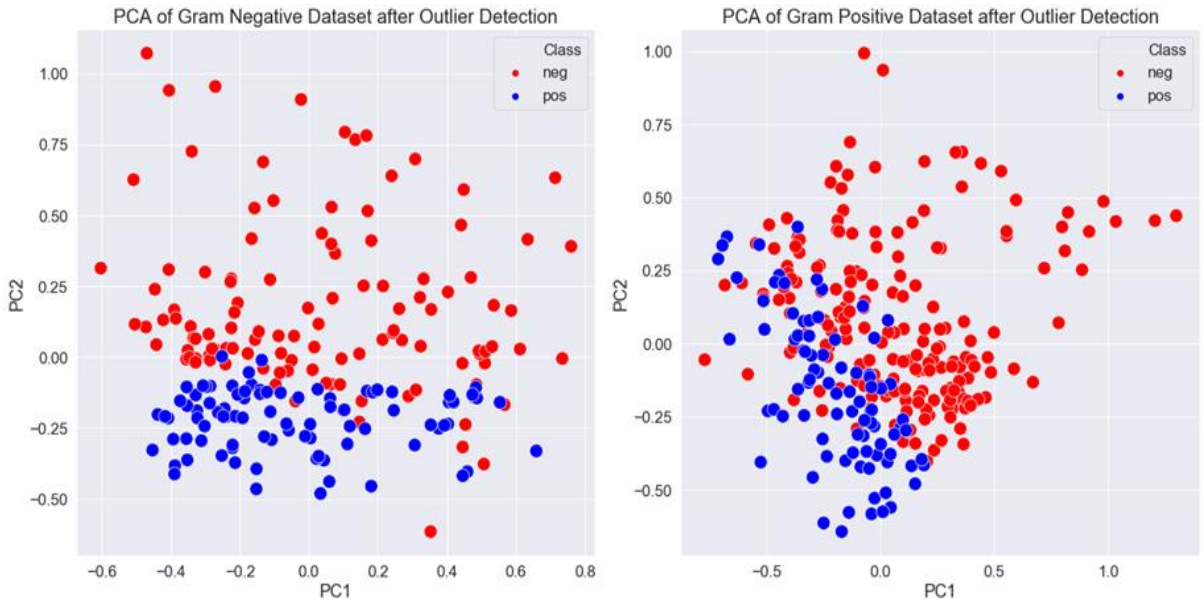


(a)



(b)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2022 doi:10.20944/preprints202202.0175.v1



(c) (d)

Figure 8. Principal component analysis of Gram-negative dataset (shown in a,c) and of Gram-positive dataset (shown in b,d) after outlier detection and elimination, shown in 3D in (a,b) and in 2D in (c,d).

Using two of the datasets after outlier elimination, we repeated our classification experiment as explained in the methods section. As shown in Tables 5 and 6, when outlier removal is applied, we have obtained higher performance metrics. As presented in Tables 5 and 6, the accuracy rate increased by 9% and reached 98% accuracy for the Gram-negative dataset, while this score is obtained as 95% for the Gram-positive dataset.

Table 5. Comparison of the models according to performance metrics for the Gram-negative dataset after outlier elimination

Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1
AdaBoost	0,975	0,990	0,965	0,958	0,990	0,972
Decision Tree	0,914	0,920	0,910	0,895	0,915	0,903
LogitBoost	0,978	**0,995**	0,965	0,959	0,992	0,976
Random Forest	**0,983**	0,994	**0,975**	**0,970**	**0,994**	**0,981**
SVM	0,980	0,990	0,973	0,967	0,989	0,977
SVM+KNN	0,814	0,828	0,803	0,814	0,840	0,800
LogitBoost+KNN	0,980	0,990	0,973	0,967	0,989	0,977

Table 6. Comparison of the models according to performance metrics for the Gram-positive dataset after outlier elimination

Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1
AdaBoost	0,939	0,927	0,945	0,898	0,965	0,906
Decision Tree	0,885	0,820	0,915	0,824	0,867	0,815
LogitBoost	0,936	0,935	0,936	0,886	0,965	0,903
Random Forest	**0,951**	**0,951**	**0,951**	**0,909**	**0,977**	**0,925**
SVM	0,914	0,903	0,919	0,853	0,939	0,867
SVM+KNN	0,776	0,752	0,787	0,686	0,810	0,685
LogitBoost+KNN	0,914	0,903	0,919	0,853	0,939	0,867

 doi:10.20944/preprints202202.0175.v1

4. Discussion

Antimicrobial peptides are characterized as positively charged, short-chain compounds which act against a wide range of microorganisms by interacting with the target cell components using different mechanisms [53]. The fact that AMPs have various mechanisms of action on the membrane makes bacterial resistance formation against them more complex compared to the conventional therapeutics. Therefore, AMPs are an attractive alternative to combat resistant bacteria [9]. However, AMPs derived from natural sources have some disadvantages such as low stability, salt tolerance and high toxicity that limit their therapeutic applications. Computational studies on AMPs help us to better understand the effect of the physicochemical properties of the peptides on stability and activity of AMPs. With the help of computational approaches in the study of AMPs, now it has become possible to overcome the above-mentioned difficulties and to design peptides with broad-spectrum activities and good stability [5].

In this study, we attempted to develop a robust classification model for antimicrobial peptide prediction problem. To this end, we have compiled two datasets from Database of Antimicrobial Activity and Structure of Peptides (DBAASP). One dataset included the peptides active against Gram-negative bacteria, another one included the peptides active against Gram-positive bacteria. For each peptide, in order to define the activity against a group of bacteria (positive class label), we have utilized Minimum Inhibition Concentration (MIC) values. In our data preprocessing steps, as shown in Figure 1, we have focused on linear cationic peptides with peptide lengths varying between 20 and 50 aminoacids (aa). Since there are many peptides with very similar sequences, we eliminated those with a similarity rate of 80% or more using the CD-HIT program [49]. We carried out our classification procedure with the remaining peptides. We have experimented with several machine learning methods including, Adaboost, Logitboost, Decision Tree, Random Forest, Support Vector Machine, and stacking classifiers using 100 fold Monte Carlo Cross Validation. In our experiments, we have observed that Random Forest outperforms other classifiers as summarized in Table 2 and Table 3.

In order to understand the underlying structure of the data, we apply Principal Component Analysis (PCA) on Gram-negative and Gram-positive datasets separately. The PCA results in Figure 6A, 6B, 6C and 6D shows that when we visualize the AMP and Non-AMP samples with PCA plots, we have noticed that there are some outlier samples (peptides) in both Gram-negative and positive datasets. In order to understand more in detail why these samples are outliers and to compile a more homogenous dataset, we have examined the physico-chemical features of the peptides. To see the distribution of each feature, we plotted histograms for the Gram-negative and the Gram-positive datasets separately (Figure 7A, 7B). Based on our analysis using such histograms, we define a certain range of values for each attribute for the positive class which represents the peptides having antimicrobial activity as illustrated in Table 4. While the peptides within the selected ranges are kept, other peptides are eliminated from our dataset. Once again, PCA visualization has been applied to this outlier eliminated dataset and it has been observed that the peptides can be better separated into two classes in this new dataset (Figure 8A, 8B, 8C, 8D). For this outlier eliminated dataset, all classification experiments have been repeated. As shown in Tables 5 and 6, we have achieved higher performance metrics when outlier removal is applied.

The studies on the structure-activity relationship of AMPs emphasized that the antimicrobial activity is affected by changes in many structural and physicochemical parameters such as net charge, hydrophobicity, and peptide chain length. Therefore, studying these properties of peptides and the similarities and differences between these features provide important insights for the development of new antimicrobial peptide prediction methods [88].

 doi:10.20944/preprints202202.0175.v1

In this study, the net charge was found as the most important feature for gram-negative data set while it is identified as the second most important feature for gram-positive dataset. The net charge is an important feature that shows the affinity of cationic peptides to bind to anionic cell surface structures through electrostatic interactions. Gram-positive and Gram-negative bacteria possess different cell wall components such as teichoic acid and lipopolysaccharides (LPSs). The difference in the importance of the net charge feature between the two datasets (peptides active against Gram-positive bacteria vs. peptides active against Gram-negative bacteria) may be due to the differences between the cell wall components of anionic characters.

On the other hand, for the gram-positive dataset, the isoelectric point (pI) was found to be the most important feature, while it was the second most important feature for gram-negative dataset. The pI feature refers to the solubility of the peptides under certain pH conditions. When the pH of the environment is equal to the pI of the peptide, the peptide loses its solubility and hereby its biological function [89]. pIs of the AMPs are generally at alkaline pH, and hereby maintain their activity at physiological pH. There is a strong relationship between the isoelectric point and the antibacterial activity of AMPs [90]. Ahn *et al*., reported that rather than the net charge, pI was a better parameter for predicting the antibacterial activity [91].

The above-mentioned two features were followed by the disordered conformation propensity, normalized hydrophobicity and normalized hydrophobic moment features respectively for both bacterial groups. The majority of linear cationic AMPs are disordered structures in aqueous solution and acquire their biologically active conformation upon interaction with the membrane. The majority of linear AMPs adapt to the alpha-helical conformation in lipid membrane environment and this regular structure is important for antimicrobial activity for this AMP class [92]. Hence, the identification of disordered conformation propensity feature as the third important feature in our analysis makes sense in terms of the underlying biology.

Hydrophobicity and hydrophobic moment are two important physico-chemical features that affect the antimicrobial activity of AMPs. In this study, the effect of these determinants was found lower than expected. The hydrophobicity reflects the ratio of hydrophobic residues within a peptide sequence. In the first step of peptide-lipid interactions, AMPs attach to the cell surface by electrostatic interactions, and then the hydrophobic interactions become a primary driving force for their insertion and partitions into the lipid bilayer [93], [94]. In general, the increase of hydrophobicity promotes antimicrobial activity in peptides [95]. However, some studies demonstrated that an increase above a certain level in hydrophobicity leads to a decrease in antimicrobial activity [95]. The hydrophobic moment is defined as a quantitative measure of peptide amphipathicity [96]. The amphipathic α-helical AMPs have polar and hydrophobic residues that are arranged in opposite faces. This arrangement facilitates the interactions of AMPs to membranes. The increase of the hydrophobic moment results in a significant elevation in antimicrobial activity, but it also leads to cytotoxicity [94].

5. Conclusions

The main contribution of this paper is the development of two accurate classification models for the prediction of antimicrobial peptides active against i) Gram-negative and ii) Gram-positive bacteria. To this end, we have compiled two different datasets for i) peptides active against Gram-negative bacteria, and ii) peptides active against Gram-positive bacteria; and evaluated different machine learning models for the prediction of antimicrobial peptide activity. In our experiments with 100 fold MCCV, the random forest algorithm achieved better results compared to other algorithms for both datasets. At the end of our feature ranking procedure, the net charge was found as the most important feature for gram-negative data set and second most important feature for gram-positive dataset. Also, for the gram-positive dataset, the isoelectric point (pI) was

found as the most important feature, while it was determined as the second most important feature for gram-negative dataset. In literature, both net charge and the isoelectric point of a peptide are known to have a considerable effect in terms of determining the activity of AMPs [90]. The PCA visualization is applied on Gram-negative and Gram-positive dataset and some outlier samples have been observed. Based on the distribution of the positive and negative labelled samples (peptides having antimicrobial activity vs. non AMP peptides), certain ranges are defined for each attribute. In our secondary experiments, in which the peptides outside those ranges were eliminated (outlier detection), we observed that the results increased by 9% for the gram negative dataset and 8% for the gram positive dataset.

Antimicrobial peptides are considered as the most promising alternatives to antibiotics. Therefore, accurate prediction of antimicrobial peptides contributes to the production of more effective peptides with lower costs. Additionally, since computational prediction approaches minimize the losses during production steps, they became popular in this field. In this respect, the classification model that we have developed in this study paves the way to the precise prediction and the design of antimicrobial peptides that are highly effective against bacterial pathogens. Even though the classification approach that we have developed here is only applied on the bacteria, it has the potential to be utilized for the prediction of antifungal, antivirus, antiprotozoal, and anticancer agents in future studies.

Supplementary Materials: Interactive 3D plots are provided as supplementary material.

Author Contributions: Conceptualization, Z.K., M.E.B., B.B.G., M.Y. and Ü.G.S.; methodology, Z.K., M.Y., B.B.G.; software, M.Y. and Ü.G.S.; validation, Ü.G.S. and M.Y.; formal analysis, Ü.G.S., B.B.G. and M.Y.; investigation, Ü.G.S., Z.K. and B.B.G.; resources, Ü.G.S, B.B.G. and Z.K.; data curation, Z.K., M.E.B. and Ü.G.S.; writing—original draft preparation, Ü.G.S.; writing—review and editing, B.B.G., Z.K., M.E.B. and M.Y.; visualization, Ü.G.S.; supervision, B.B.G. and M.Y.; project administration, Z.K., B.B.G.; funding acquisition, M.Y, B.B.G. and Z.K. . All authors have read and agreed to the published version of the manuscript.

Funding: The work of M.Y. has been supported by the Zefat Academic College. The work of B.B.G. has been supported by the Abdullah Gul University Support Foundation (AGUV). The works of Z.K. and M.E.B. have been supported by the TUBITAK 1001 program (Project No: 120Z565) to support scientific and technological research projects.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data is obtained from Database of Antimicrobial Activity and Structure of Peptides (DBAASP) web server.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] V. Carnicelli, A. Lizzi, A. Ponzi, G. Amicosante, A. Bozzi, and A. Di Giulio, 'Articolo su libro (2013)'. Oct. 07, 2015.

[2] N. K. Brogden and K. A. Brogden, 'Will new generations of modified antimicrobial peptides improve their potential as pharmaceuticals?', *Int. J. Antimicrob. Agents*, p. S0924857911002342, Jul. 2011, doi: 10.1016/j.ijantimicag.2011.05.004.

[3] F. Xie *et al.*, 'The SapA Protein Is Involved in Resistance to Antimicrobial Peptide PR-39 and Virulence of Actinobacillus pleuropneumoniae', *Front. Microbiol.*, vol. 8, p. 811, May 2017, doi: 10.3389/fmicb.2017.00811.

[4] D. Neubauer *et al.*, 'Retro analog concept: comparative study on physico-chemical and biological properties of selected antimicrobial peptides', *Amino Acids*, vol. 49, no. 10, pp. 1755–1771, Oct. 2017, doi: 10.1007/s00726-017-2473-7.

[5] M. Erdem Büyükkiraz and Z. Kesmen, 'Antimicrobial peptides (AMPs): A promising class of antimicrobial compounds', *J. Appl. Microbiol.*, p. jam.15314, Oct. 2021, doi: 10.1111/jam.15314.

[6] B. Mishra and G. Wang, 'Ab Initio Design of Potent Anti-MRSA Peptides Based on Database Filtering Technology', *ACS Publications*, Jul. 19, 2012. https://pubs.acs.org/doi/pdf/10.1021/ja305644e (accessed Dec. 15, 2021).

[7] D. Faccone *et al.*, 'Antimicrobial activity of de novo designed cationic peptides against multi-resistant clinical isolates', *Eur. J. Med. Chem.*, vol. 71, pp. 31–35, Jan. 2014, doi: 10.1016/j.ejmech.2013.10.065.

[8] C. H. Chen *et al.*, 'Simulation-Guided Rational de Novo Design of a Small Pore-Forming Antimicrobial Peptide', *J. Am. Chem. Soc.*, Mar. 2019, doi: 10.1021/jacs.8b11939.

[9] B. Vishnepolsky *et al.*, 'De Novo Design and In Vitro Testing of Antimicrobial Peptides against Gram-Negative Bacteria', *Pharmaceuticals*, vol. 12, no. 2, p. 82, Jun. 2019, doi: 10.3390/ph12020082.

[10] C. Loose, K. Jensen, I. Rigoutsos, and G. Stephanopoulos, 'A linguistic model for the rational design of antimicrobial peptides', *Nature*, vol. 443, no. 7113, pp. 867–869, Oct. 2006, doi: 10.1038/nature05233.

[11] D. Nagarajan *et al.*, 'Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria', *J. Biol. Chem.*, vol. 293, no. 10, pp. 3492–3509, Mar. 2018, doi: 10.1074/jbc.M117.805499.

[12] M. H. Cardoso *et al.*, 'A Computationally Designed Peptide Derived from Escherichia coli as a Potential Drug Template for Antibacterial and Antibiofilm Therapies', *ACS Infect. Dis.*, Oct. 2018, doi: 10.1021/acsinfecdis.8b00219.

[13] E. S. Cândido *et al.*, 'Short Cationic Peptide Derived from Archaea with Dual Antibacterial Properties and Anti-Infective Potential', *ACS Infect. Dis.*, vol. 5, no. 7, pp. 1081–1086, Jul. 2019, doi: 10.1021/acsinfecdis.9b00073.

[14] I. C. M. Fensterseifer *et al.*, 'Selective antibacterial activity of the cationic peptide PaDBS1R6 against Gram-negative bacteria', *Biochim. Biophys. Acta BBA - Biomembr.*, vol. 1861, no. 7, pp. 1375–1387, Jul. 2019, doi: 10.1016/j.bbamem.2019.03.016.

[15] K. G. N. Oshiro *et al.*, 'Computer-Aided Design of Mastoparan-like Peptides Enables the Generation of Nontoxic Variants with Extended Antibacterial Properties', *J. Med. Chem.*, Aug. 2019, doi: 10.1021/acs.jmedchem.9b00915.

[16] C. D. Fjell, H. Jenssen, W. A. Cheung, R. E. W. Hancock, and A. Cherkasov, 'Optimization of Antibacterial Peptides by Genetic Algorithms and Cheminformatics', *Chem. Biol. Drug Des.*, vol. 77, no. 1, pp. 48–56, 2011, doi: 10.1111/j.1747-0285.2010.01044.x.

[17] G. Maccari *et al.*, 'Antimicrobial Peptides Design by Evolutionary Multiobjective Optimization', *PLoS Comput. Biol.*, vol. 9, no. 9, p. e1003212, Sep. 2013, doi: 10.1371/journal.pcbi.1003212.

[18] W. F. Porto *et al.*, 'In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design', *Nat. Commun.*, vol. 9, no. 1, p. 1490, Dec. 2018, doi: 10.1038/s41467-018-03746-3.

[19] M. Yoshida *et al.*, 'Using Evolutionary Algorithms and Machine Learning to Explore Sequence Space for the Discovery of Antimicrobial Peptides', *Chem*, vol. 4, no. 3, pp. 533–543, Mar. 2018, doi: 10.1016/j.chempr.2018.01.005.

[20] S. Liu, L. Fan, J. Sun, X. Lao, and H. Zheng, 'Computational resources and tools for antimicrobial peptides: Computational Resources and Tools for Antimicrobial Peptides', *J. Pept. Sci.*, vol. 23, no. 1, pp. 4–12, Jan. 2017, doi: 10.1002/psc.2947.

[21] X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, and K.-C. Chou, 'iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types', *Anal. Biochem.*, vol. 436, no. 2, pp. 168–177, May 2013, doi: 10.1016/j.ab.2013.01.019.

[22] A. C. Schierz, 'Virtual screening of bioassay data', *J. Cheminformatics*, vol. 1, no. 1, p. 21, Dec. 2009, doi: 10.1186/1758-2946-1-21.

[23] P. Bhadra, J. Yan, J. Li, S. Fong, and S. W. I. Siu, 'AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest', *Sci. Rep.*, vol. 8, no. 1, p. 1697, Dec. 2018, doi: 10.1038/s41598-018-19752-w.

[24] S. Lata, N. K. Mishra, and G. P. Raghava, 'AntiBP2: improved version of antibacterial peptide prediction', *BMC Bioinformatics*, vol. 11, no. S1, p. S19, Jan. 2010, doi: 10.1186/1471-2105-11-S1-S19.

[25] D. Dhall, R. Kaur, and M. Juneja, 'Machine Learning: A Review of the Algorithms and Its Applications', in *Proceedings of ICRIC 2019*, Cham, 2020, pp. 47–63. doi: 10.1007/978-3-030-29407-6_5.

[26] E. Y. Lee, M. W. Lee, B. M. Fulan, A. L. Ferguson, and G. C. L. Wong, 'What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning?', *Interface Focus*, vol. 7, no. 6, p. 20160153, Dec. 2017, doi: 10.1098/rsfs.2016.0153.

[27] M. Burdukiewicz *et al.*, 'Proteomic Screening for Prediction and Design of Antimicrobial Peptides with AmpGram', *Int. J. Mol. Sci.*, vol. 21, no. 12, p. 4310, Jun. 2020, doi: 10.3390/ijms21124310.

[28] C.-R. Chung *et al.*, 'Characterization and Identification of Natural Antimicrobial Peptides on Different Organisms', *Int. J. Mol. Sci.*, vol. 21, no. 3, p. 986, Feb. 2020, doi: 10.3390/ijms21030986.

[29] P. Wang *et al.*, 'Prediction of Antimicrobial Peptides Based on Sequence Alignment and Feature Selection Methods', *PLoS ONE*, vol. 6, no. 4, p. e18476, Apr. 2011, doi: 10.1371/journal.pone.0018476.

[30] P. Agrawal and G. P. S. Raghava, 'Prediction of Antimicrobial Potential of a Chemically Modified Peptide From Its Tertiary Structure', *Front. Microbiol.*, vol. 9, p. 2551, Oct. 2018, doi: 10.3389/fmicb.2018.02551.

[31] S. Gull, N. Shamim, and F. Minhas, 'AMAP: Hierarchical multi-label prediction of biologically active and antimicrobial peptides', *Comput. Biol. Med.*, vol. 107, pp. 172–181, Apr. 2019, doi: 10.1016/j.compbiomed.2019.02.018.

[32] M. Torrent, V. M. Nogués, and E. Boix, 'A theoretical approach to spot active regions in antimicrobial proteins', *BMC Bioinformatics*, vol. 10, no. 1, p. 373, Dec. 2009, doi: 10.1186/1471-2105-10-373.

[33] F. H. Waghu, R. S. Barai, P. Gurung, and S. Idicula-Thomas, 'CAMP R3 : a database on sequences, structures and signatures of antimicrobial peptides: Table 1.', *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1094–D1097, Jan. 2016, doi: 10.1093/nar/gkv1051.

[34] W. Lin and D. Xu, 'Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types', *Bioinformatics*, vol. 32, no. 24, pp. 3745–3752, Dec. 2016, doi: 10.1093/bioinformatics/btw560.

[35] J. Yan *et al.*, 'Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning', *Mol. Ther. - Nucleic Acids*, vol. 20, pp. 882–894, Jun. 2020, doi: 10.1016/j.omtn.2020.05.006.

[36] X. Su, J. Xu, Y. Yin, X. Quan, and H. Zhang, 'Antimicrobial peptide identification using multi-scale convolutional network', *BMC Bioinformatics*, vol. 20, no. 1, p. 730, Dec. 2019, doi: 10.1186/s12859-019-3327-y.

[37] P. Schneider *et al.*, 'Hybrid Network Model for "Deep Learning" of Chemical Data: Application to Antimicrobial Peptides', *Mol. Inform.*, vol. 36, no. 1–2, p. 1600011, Jan. 2017, doi: 10.1002/minf.201600011.

[38] J. Witten and Z. Witten, 'Deep learning regression model for antimicrobial peptide design', Bioinformatics, preprint, Jul. 2019. doi: 10.1101/692681.

[39] H. Fu, Z. Cao, M. Li, and S. Wang, 'ACEP: improving antimicrobial peptides recognition through automatic feature fusion and amino acid embedding', *BMC Genomics*, vol. 21, no. 1, p. 597, Dec. 2020, doi: 10.1186/s12864-020-06978-0.

[40] A. T. Müller *et al.*, 'Sparse Neural Network Models of Antimicrobial Peptide-Activity Relationships', *Mol. Inform.*, vol. 35, no. 11–12, pp. 606–614, Dec. 2016, doi: 10.1002/minf.201600029.

[41] M.-N. Hamid and I. Friedberg, 'Identifying antimicrobial peptides using word embedding with deep recurrent neural networks', *Bioinformatics*, vol. 35, no. 12, pp. 2009–2016, Jun. 2019, doi: 10.1093/bioinformatics/bty937.

[42] C. Li *et al.*, 'AMPlify: attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens', Bioinformatics, preprint, Jun. 2020. doi: 10.1101/2020.06.16.155705.

[43] S. Liu, J. Bao, X. Lao, and H. Zheng, 'Novel 3D Structure Based Model for Activity Prediction and Design of Antimicrobial Peptides', *Sci. Rep.*, vol. 8, no. 1, p. 11189, Dec. 2018, doi: 10.1038/s41598-018-29566-5.

[44] A. Capecchi, X. Cai, H. Personne, T. Köhler, C. van Delden, and J.-L. Reymond, 'Machine learning designs non-hemolytic antimicrobial peptides', *Chem. Sci.*, vol. 12, no. 26, pp. 9221–9232, 2021, doi: 10.1039/D1SC01713F.

[45] B. Vishnepolsky, M. Grigolava, G. Zaalishvili, M. Karapetian, and M. Pirtskhalava, 'DBAASP Special prediction as a tool for the prediction of antimicrobial potency against particular target species', in *Proceedings of 4th International Electronic Conference on Medicinal Chemistry*, Sciforum.net, Oct. 2018, p. 5608. doi: 10.3390/ecmc-4-05608.

[46] Y. Ohtsuka and H. Inagaki, 'In silico identification and functional validation of linear cationic α-helical antimicrobial peptides in the ascidian Ciona intestinalis', *Sci. Rep.*, vol. 10, no. 1, p. 12619, Dec. 2020, doi: 10.1038/s41598-020-69485-y.

[47] F. Plisson, O. Ramírez-Sánchez, and C. Martínez-Hernández, 'Machine learning-guided discovery and design of non-hemolytic peptides', *Sci. Rep.*, vol. 10, no. 1, p. 16581, Dec. 2020, doi: 10.1038/s41598-020-73644-6.

[48] I. Wiegand, K. Hilpert, and R. E. W. Hancock, 'Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances', *Nat. Protoc.*, vol. 3, no. 2, pp. 163–175, Feb. 2008, doi: 10.1038/nprot.2007.521.

[49] W. Li and A. Godzik, 'Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences', *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006, doi: 10.1093/bioinformatics/btl158.

[50] B. Vishnepolsky and M. Pirtskhalava, 'Comment on: "Empirical comparison of web-based antimicrobial peptide prediction tools"', *Bioinformatics*, vol. 35, no. 15, pp. 2692–2694, Aug. 2019, doi: 10.1093/bioinformatics/bty1023.

[51] J. H. Lee *et al.*, 'Transcriptome Analysis of Psacothea hilaris: De Novo Assembly and Antimicrobial Peptide Prediction', *Insects*, vol. 11, no. 10, Art. no. 10, Oct. 2020, doi: 10.3390/insects11100676.

[52] F. C. Fernandes, D. J. Rigden, and O. L. Franco, 'Prediction of antimicrobial peptides based on the adaptive neuro-fuzzy inference system application', *Pept. Sci.*, vol. 98, no. 4, pp. 280–287, 2012, doi: 10.1002/bip.22066.

[53] D. Veltri, U. Kamath, and A. Shehu, 'Deep learning improves antimicrobial peptide recognition', *Bioinformatics*, vol. 34, no. 16, pp. 2740–2747, Aug. 2018, doi: 10.1093/bioinformatics/bty179.

[54] A. Gautam *et al.*, 'Development of Antimicrobial Peptide Prediction Tool for Aquaculture Industries', *Probiotics Antimicrob. Proteins*, vol. 8, no. 3, pp. 141–149, Sep. 2016, doi: 10.1007/s12602-016-9215-0.

[55] M. N. Gabere and W. S. Noble, 'Empirical comparison of web-based antimicrobial peptide prediction tools', *Bioinformatics*, vol. 33, no. 13, pp. 1921–1929, Jul. 2017, doi: 10.1093/bioinformatics/btx081.

[56] F. H. Waghu, L. Gopi, R. S. Barai, P. Ramteke, B. Nizami, and S. Idicula-Thomas, 'CAMP: Collection of sequences and structures of antimicrobial peptides', *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1154–D1158, Jan. 2014, doi: 10.1093/nar/gkt1157.

[57] X.-Y. Yu, R. Fu, P.-Y. Luo, Y. Hong, and Y.-H. Huang, 'Construction and Prediction of Antimicrobial Peptide Predicition Model Based on BERT', p. 5.

[58] H. Khabbaz, M. H. Karimi-Jafari, A. A. Saboury, and B. BabaAli, 'Prediction of antimicrobial peptides toxicity based on their physico-chemical properties using machine learning techniques', *BMC Bioinformatics*, vol. 22, no. 1, p. 549, Dec. 2021, doi: 10.1186/s12859-021-04468-y.

[59] A. Moretta *et al.*, 'A bioinformatic study of antimicrobial peptides identified in the Black Soldier Fly (BSF) Hermetia illucens (Diptera: Stratiomyidae)', *Sci. Rep.*, vol. 10, no. 1, p. 16875, Dec. 2020, doi: 10.1038/s41598-020-74017-9.

[60] B. Vishnepolsky *et al.*, 'Predictive Model of Linear Antimicrobial Peptides Active against Gram-Negative Bacteria', *J. Chem. Inf. Model.*, vol. 58, no. 5, pp. 1141–1151, May 2018, doi: 10.1021/acs.jcim.8b00118.

[61] Y. Freund and R. E. Schapire, 'A Short Introduction to Boosting', p. 14.

[62] J. Friedman, T. Hastie, and R. Tibshirani, 'Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors)', *Ann. Stat.*, vol. 28, no. 2, Apr. 2000, doi: 10.1214/aos/1016218223.

[63] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*, 1st ed. Routledge, 2017. doi: 10.1201/9781315139470.

[64] Tin Kam Ho, 'Random decision forests', in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, Que., Canada, 1995, vol. 1, pp. 278–282. doi: 10.1109/ICDAR.1995.598994.

[65] C. Cortes and V. Vapnik, 'Support-vector networks', *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[66] E. Fix and J. L. Hodges, 'Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties', *Int. Stat. Rev. Rev. Int. Stat.*, vol. 57, no. 3, pp. 238–247, 1989, doi: 10.2307/1403797.

[67] M. R. Berthold *et al.*, 'KNIME - the Konstanz information miner: version 2.0 and beyond', *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 26–31, Nov. 2009, doi: 10.1145/1656274.1656280.

[68] B. M. Randles, I. V. Pasquetto, M. S. Golshan, and C. L. Borgman, 'Using the Jupyter Notebook as a Tool for Open Science: An Empirical Study', in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Toronto, ON, Canada, Jun. 2017, pp. 1–2. doi: 10.1109/JCDL.2017.7991618.

[69] Q.-S. Xu and Y.-Z. Liang, 'Monte Carlo cross validation', *Chemom. Intell. Lab. Syst.*, vol. 56, no. 1, pp. 1–11, Apr. 2001, doi: 10.1016/S0169-7439(00)00122-2.

[70] A. Jovic, K. Brkic, and N. Bogunovic, 'A review of feature selection methods with applications', in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, May 2015, pp. 1200–1205. doi: 10.1109/MIPRO.2015.7160458.

[71] M. Yousef, S. Jung, L. C. Showe, and M. K. Showe, 'Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data', *BMC Bioinformatics*, vol. 8, no. 1, p. 144, 2007, doi: 10.1186/1471-2105-8-144.

[72] M. Yousef, B. Bakir-Gungor, A. Jabeer, G. Goy, R. Qureshi, and L. C. Showe, 'Recursive Cluster Elimination based Rank Function (SVM-RCE-R) implemented in KNIME', *F1000Research*, vol. 9, Jan. 2021, doi: 10.12688/f1000research.26880.2.

[73] M. Yousef, A. Jabeer, and B. Bakir-Gungor, 'SVM-RCE-R-OPT: Optimization of Scoring Function for SVM-RCE-R', in *Database and Expert Systems Applications - DEXA 2021 Workshops*, vol. 1479, G. Kotsis, A. M. Tjoa, I. Khalil, B. Moser, A. Mashkoor, J. Sametinger, A. Fensel, J. Martinez-Gil, L. Fischer, G. Czech, F. Sobieczky, and S. Khan, Eds. Cham: Springer International Publishing, 2021, pp. 215–224. doi: 10.1007/978-3-030-87101-7_21.

[74] M. Yousef, L. Abdallah, and J. Allmer, 'maTE: discovering expressed interactions between microRNAs and their targets', *Bioinformatics*, vol. 35, no. 20, pp. 4020–4028, Oct. 2019, doi: 10.1093/bioinformatics/btz204.

[75] M. Yousef, E. Ülgen, and O. U. Sezerman, 'CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis', *PeerJ Comput. Sci.*, vol. 7, p. e336, Feb. 2021, doi: 10.7717/peerj-cs.336.

[76] M. Yousef, G. Goy, R. Mitra, C. M. Eischen, A. Jabeer, and B. Bakir-Gungor, 'miRcorrNet: machine learning-based integration of miRNA and mRNA expression profiles, combined with feature grouping and ranking', *PeerJ*, vol. 9, p. e11458, May 2021, doi: 10.7717/peerj.11458.

[77] M. Yousef, A. Sayıcı, and B. Bakir-Gungor, 'Integrating Gene Ontology Based Grouping and Ranking into the Machine Learning Algorithm for Gene Expression Data Analysis', in *Database and Expert Systems Applications - DEXA 2021 Workshops*, vol. 1479, G. Kotsis, A. M. Tjoa, I. Khalil, B. Moser, A. Mashkoor, J. Sametinger, A. Fensel, J. Martinez-Gil, L. Fischer, G. Czech, F. Sobieczky, and S. Khan, Eds. Cham: Springer International Publishing, 2021, pp. 205–214. doi: 10.1007/978-3-030-87101-7_20.

[78] M. Yousef, A. Kumar, and B. Bakir-Gungor, 'Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data', *Entropy*, vol. 23, no. 1, p. 2, Dec. 2020, doi: 10.3390/e23010002.

[79] K. Pearson, 'LIII. *On lines and planes of closest fit to systems of points in space*', *Lond. Edinb. Dublin Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, Nov. 1901, doi: 10.1080/14786440109462720.

[80] W. F. Porto, Á. S. Pires, and O. L. Franco, 'CS-AMPPred: An Updated SVM Model for Antimicrobial Activity Prediction in Cysteine-Stabilized Peptides', *PLoS ONE*, vol. 7, no. 12, p. e51444, Dec. 2012, doi: 10.1371/journal.pone.0051444.

[81] M. Shu *et al.*, 'Predicting the Activity of Antimicrobial Peptides with Amino Acid Topological Information', *Med. Chem.*, vol. 9, no. 1, pp. 32–44, Feb. 2013, doi: 10.2174/157340613804488350.

[82] L. Moll, E. Badosa, M. Planas, L. Feliu, E. Montesinos, and A. Bonaterra, 'Antimicrobial Peptides With Antibiofilm Activity Against Xylella fastidiosa', *Front. Microbiol.*, vol. 12, p. 753874, Nov. 2021, doi: 10.3389/fmicb.2021.753874.

[83] H. Lin, T. Yan, L. Wang, F. Guo, G. Ning, and M. Xiong, 'Statistical design, structural analysis, and *in vitro* susceptibility assay of antimicrobial peptoids to combat bacterial infections: Statistical design of antimicrobial peptoids', *J. Chemom.*, vol. 30, no. 7, pp. 369–376, Jul. 2016, doi: 10.1002/cem.2801.

[84] S. Thudumu, P. Branch, J. Jin, and J. Singh, 'A comprehensive survey of anomaly detection techniques for high dimensional big data', *J. Big Data*, vol. 7, no. 1, p. 42, Dec. 2020, doi: 10.1186/s40537-020-00320-x.

[85] L. M. Manevitz and M. Yousef, 'One-Class SVMs for Document Classification', *J. Mach. Learn. Res.*, 2001.

[86] L. Manevitz and M. Yousef, 'One-class document classification via Neural Networks', *Neurocomputing*, vol. 70, no. 7–9, pp. 1466–1481, 2007, doi: 10.1016/j.neucom.2006.05.013.

[87] P. K. Hazam, R. Goyal, and V. Ramakrishnan, 'Peptide based antimicrobials: Design strategies and therapeutic potential', *Prog. Biophys. Mol. Biol.*, vol. 142, pp. 10–22, Mar. 2019, doi: 10.1016/j.pbiomolbio.2018.08.006.

[88] M. Pirtskhalava and M. Grigolava, 'Transmembrane and Antimicrobial Peptides. Hydrophobicity, Amphiphilicity and Propensity to Aggregation', p. 24.

[89] D. Osorio, P. Rondón-Villarreal, and R. Torres Sáez, 'Peptides: A Package for Data Mining of Antimicrobial Peptides', *R J.*, vol. 7, pp. 4–14, Jun. 2015, doi: 10.32614/RJ-2015-001.

[90] B. Romestand, F. Molina, V. Richard, P. Roch, and and C. Granier, 'Key role of the loop connecting the two beta strands of mussel defensin in its antimicrobial activity', *Eur. J. Biochem.*, vol. 270, no. 13, pp. 2805–2813, Jul. 2003, doi: 10.1046/j.1432-1033.2003.03657.x.

[91] H. Ahn *et al.*, 'Design and synthesis of novel antimicrobial peptides on the basis of α helical domain of Tenecin 1, an insect defensin protein, and structure–activity relationship study', *Peptides*, vol. 27, no. 4, pp. 640–648, Apr. 2006, doi: 10.1016/j.peptides.2005.08.016.

[92] M. Pirtskhalava, B. Vishnepolsky, and M. Grigolava, 'Physicochemical Features and Peculiarities of Interaction of Antimicrobial Peptides with the Membrane', p. 41.

[93] N. Papo and Y. Shai, 'Can we predict biological activity of antimicrobial peptides from their interactions with model phospholipid membranes?', *Peptides*, vol. 24, no. 11, pp. 1693–1703, Nov. 2003, doi: 10.1016/j.peptides.2003.09.013.

[94] V. Teixeira, M. J. Feio, and M. Bastos, 'Role of lipids in the interaction of antimicrobial peptides with membranes', *Prog. Lipid Res.*, vol. 51, no. 2, pp. 149–177, Apr. 2012, doi: 10.1016/j.plipres.2011.12.005.

[95] Y. Chen, M. T. Guarnieri, A. I. Vasil, M. L. Vasil, C. T. Mant, and R. S. Hodges, 'Role of Peptide Hydrophobicity in the Mechanism of Action of α-Helical Antimicrobial Peptides', *Antimicrob. Agents Chemother.*, vol. 51, no. 4, pp. 1398–1406, Apr. 2007, doi: 10.1128/AAC.00925-06.

[96] D. Eisenberg, R. M. Weiss, and T. C. Terwilliger, 'The helical hydrophobic moment: a measure of the amphiphilicity of a helix', *Nature*, vol. 299, no. 5881, pp. 371–374, Sep. 1982, doi: 10.1038/299371a0.