Ümmü Gülsüm SÖYLEMEZ

ANTIMICROBIAL PEPTIDE ACTIVITY PREDICTION USING MACHINE LEARNING METHODS

A THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By
Ümmü Gülsüm Söylemez
May 2023

A Ph.D. Thesis

AGU 2023

ANTIMICROBIAL PEPTIDE ACTIVITY PREDICTION USING MACHINE LEARNING METHODS

A THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By
Ümmü Gülsüm Söylemez
May 2023

SCIENTIFIC ETHICS COMPLIANCE

I hereby declare that all information in this document has been obtained in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name-Surname:Ümmü Gülsüm Söylemez

Signature :

REGULATORY COMPLIANCE

Ph.D. thesis titled “Antimicrobial Peptide Activity Prediction Using Machine Learning Methods” has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, Graduate School of Engineering & Science.

Prepared By	Advisor	Co-Advisor
Ümmü Gülsüm Söylemez	Assist. Prof. Burcu Bakır Güngör	Prof. Dr. Malik Yousef

Head of the Electrical and Computer Engineering Program

Assoc. Prof. Zafer AYDIN

ACCEPTANCE AND APPROVAL

Ph.D. thesis titled “Antimicrobial Peptide Activity Prediction Using Machine Learning Methods” and prepared by Ümmü Gülsüm SÖYLEMEZ has been accepted by the jury in the Electrical and Computer Engineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

05/05/2023

(Thesis Defense Exam Date)

JURY:

Advisor : Assist. Prof. Burcu BAKIR GÜNGÖR

Member : Assist. Prof. Gökhan BAKAL

Member : Prof. Dr. Zülal KESMEN

Member : Assoc. Prof. Mete ÇELİK

Member : Assist. Prof. Bekir Hakan AKSEBZECİ

APPROVAL:

The acceptance of this Ph.D. thesis has been approved by the decision of the Abdullah Gül University, Graduate School of Engineering & Science, Executive Board dated ….. /….. / 2023 and numbered ……..…… .

…….. /…….. / ……..

(Date)

Graduate School Dean
Prof. Dr. İrfan ALAN

ABSTRACT

ANTIMICROBIAL PEPTIDE ACTIVITY PREDICTION USING MACHINE LEARNING METHODS

Ümmü Gülsüm SÖYLEMEZ
Ph.D. in Electrical and Computer Engineering
Advisor: Assist. Prof. Burcu BAKIR GUNGOR
Co-Advisor: Prof. Malik YOUSEF

May 2023

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 thesis, using the multiple properties of the peptides we aimed to develop machine learning approaches that can predict the antimicrobial activities of the peptides. We have created two datasets for the peptides showing antimicrobial activity against Gram-negative and against Gram-positive bacteria separately. In our first study, ten different physico-chemical properties of the peptides were calculated, and used as features of the peptides. Following the data exploration and data preprocessing steps, a variety of classification models were build with 100-fold Monte Carlo Cross-Validation; and the performance of these models were evaluated. In the second study, we proposed a novel method called AMP-GSM. The method was tested for three datasets, and the prediction performance of AMP-GSM was comparatively evaluated with several feature selection methods and several classifiers. In the last study, using motif matching score with the models of activity against Gram-positive and Gram-negative bacteria, we created novel peptides and predicted the target selectivity of these peptides. The studies presented in this thesis advance the field of computational research by making it easier to predict the possible antimicrobial effects of peptides and to design AMPs in wet laboratory environments.

Keywords: Antibiotic Resistance, Antimicrobial Peptide (AMP) Prediction, Machine Learning, Physico-chemical Properties, Quantitative Structure–activity Relationship.

ÖZET

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE ANTİMİKROBİYAL PEPTİT AKTİVİTE TAHMİNİ

Ümmü Gülsüm SÖYLEMEZ
Elektrik ve Bilgisayar Mühendisliği Anabilim Dalı Doktora
Tez Yöneticisi: Dr. Öğr. Üyesi Burcu Bakır GÜNGÖR
Eş-Danışman: Prof. Dr. Malik YOUSEF
Mayıs-2023

Antimikrobiyal peptitler (AMP'ler), artan antibiyotik direnç sorununun üstesinden gelmek için geleneksel antibiyotiklerin yerine kullanılabilecek umut verici alternatifler olarak kabul edilmektedir. Hesaplamalı tahmin yaklaşımları, in vitro testlerden önce en iyi aday AMP'leri belirlemek ve tasarlamak için artan bir ilgi görmektedir. Bu tezde, peptitlerin birden çok özelliklerini kullanarak, peptitlerin antimikrobiyal aktivitelerini tahmin edebilen makine öğrenimi yaklaşımları geliştirmeyi amaçladık. Gram negatif ve Gram pozitif bakterilere karşı antimikrobiyal aktivite gösteren peptidler için iki ayrı veri seti oluşturduk. İlk çalışmamızda peptitlerin on farklı fiziko-kimyasal özelliği hesaplanmış ve peptidlerin özniteliği olarak kullanılmıştır. Veri keşfi ve veri ön işleme adımlarının ardından, 100-katlı Monte Carlo Çapraz Doğrulama ile çeşitli sınıflandırma modelleri oluşturuldu; ve bu modellerin performansı değerlendirildi. İkinci çalışmada, AMP-GSM adlı yeni bir yöntem önerdik. Yöntem, üç ayrı veri seti için test edildi ve AMP-GSM'nin tahmin performansı, çeşitli öznitelik seçim yöntemleri ve çeşitli sınıflandırıcılar ile karşılaştırmalı olarak değerlendirildi. Son çalışmada, Gram pozitif ve Gram negatif bakterilere karşı aktivite modelleriyle motif eşleştirme skorunu kullanarak yeni peptidler yarattık ve bu peptidlerin hedef seçiciliklerini tahmin ettik. Bu tezde sunulan çalışmalar, peptidlerin olası antimikrobiyal etkilerini tahminlemeyi ve ıslak laboratuvar ortamlarında AMPleri dizayn etmeyi kolaylaştırarak hesaplamalı araştırma alanını ilerletmektedir.

Anahtar kelimeler: Antibiyotik Direnç, Antimikrobiyal Peptit (AMP) Tahmini, Makine Öğrenimi, Fiziko-kimyasal Özellikler, Nicel yapı-aktivite ilişkisi

Acknowledgements

I would like to thank my advisor, Assist.Prof.Dr. Burcu Bakır Güngör, who supported me in every aspect during my doctoral education, guided me at every stage of my thesis work, and never spared her trust and support while carrying out my studies, and whom I see as an idol for my success.

I would like to special thanks to my second advisor, Prof. Dr. Malik Yousef, for his patience, faith and trust in me, and valuable ideas and suggestions. It has been a great honor for me to work with him.

I would like to thank Prof. Dr. Zülal Kesmen, who provides us to study this subject, for guidance through the biological perspective.

I am especially grateful to my husband, Ertürk Söylemez, who made the most beautiful touches to my life and supported me in every way.

I would like to thank my family especially my father Zülküf UZUT and my mother Nisbet UZUT for being in favor of me in any case and believing in me.

I am so grateful to my dear friends Özlem ŞEKER and Zeynep KARAGÖZLÜ who always support me and share our wonderful times.

Finally, special thanks to my little baby girl on the way, for being a good listener and never made me sad while writing this dissertation.

This work is supported by the TUBITAK 1001 program (Project No: 120Z565) to support scientific and technological research projects.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABPs	Antibacterial Peptides
ACPs	Anticancer Peptides
AFPs	Antifungal Peptides
AIPs	Anti-inflammatory Peptides
AMPs	Antimicrobial Peptides
APPs	Antiparasitic Peptides
AUC	Area Under Curve
AVPs	Antiviral Peptides
CMIM	Conditional Mutual Information Maximization
CTD	Composition, Transition, Distribution
DBAASP	Database of Antimicrobial Activity and Structure of Peptides
DNN	Deep Neural Network
DT	Decision Tree
FN	False Negative
FP	False Positive
GSM	Grouping-Scoring-Modeling
IG	Information Gain
KNIME	Konstanz Information Miner
k-NN	k-Nearest Neighbour
LCAMP	Linear Cationic Antimicrobial Peptides
MCCV	Monte Carlo Cross Validation
MICs	Minimum Inhibitory Concentrations
ML	Machine Learning
MRMR	Maximum Relevance Minimum Redundancy
Non-AMPs	Non-Antimicrobial Peptides
PCA	Principal Component Analysis
RF	Random Forest
SVM	Support Vector Machine
TN	True Negative

TP	True Positive
WHO	World Health Organization
XGB	Extreme Gradient Boosting

To my family

Chapter 1

1.Introduction

1.1 The Antibiotic Resistance Problem

One of the biggest concerns to public health is antibiotic resistance, which is getting worse as more and more bacteria develop drug resistance. Microorganisms that became resistant to many types of antibiotics are linked to a related and even more dangerous issue of multidrug-resistant infections. Antibiotic resistance is quickly spreading around the world, endangering the therapeutic efficacy of these drugs that have revolutionized contemporary medicine and saved millions of lives.

Misuse and overuse of antibiotics, as well as a dearth of new drug research by the pharmaceutical sector, have all contributed to the dilemma of antibiotic resistance. Due to this issue, only a small number of medications are successful in treating some opportunistic infections. Unfortunately, some of these pharmaceuticals, like amphotericin B, have the drawback of toxicity, which may prevent patients from receiving additional treatments that involve hazardous chemicals. It is urgently necessary to coordinate efforts to revitalize medical research, deploy innovative medication development techniques, and pursue crisis management measures.

The discovery of more antimicrobial peptides (AMPs) or the creation of peptides from scratch (de novo) have emerged as promising interest areas in antibiotic research in response to the current bacterial resistance crisis and the spread of infectious diseases, both of which pose potential threats to humans. For this reason, AMPs demonstrate the potential for usage as bactericidal and antifungal medications as well as their efficacy in combating bacteria that are multidrug resistant.

For researchers trying to create novel anti-pathogenic medications, the emergence of microbial drug resistance is a difficult problem. This class of medication is found in practically all living things as a component of their innate, non-specific immune systems; AMPs are highly prized as potential building blocks for the creation of human treatments to halt the spread of antibiotic resistance. Because of their distinct qualities as medications—low toxicity, high biological activity, and specificity—AMPs

are desirable therapeutic agents. In order to classify and forecast these naturally occurring AMPs, this dissertation uses computational analysis. The classification and prediction of these naturally occurring AMPs utilizing computational analysis is a key component of this dissertation's assistance for these initiatives. The study presented in this thesis has the potential to aid in the development of new therapies as well as the creation or modification of AMPs used in wet lab studies against multidrug-resistant bacteria.

1.2 Antimicrobial Peptides (AMPs)

An amino acid short chain is known as a peptide. There are 20 amino acids that are found in nature, and they can be combined to form a huge diversity of different compounds. Peptide bonds join the amino acids together in a certain order.

Gram-positive and gram-negative bacteria, viruses, and fungi are just a few of the microbes that AMPs have been shown to effectively destroy. These AMPs act as a first-line defensive mechanism that is already present but may become more active in response to injury and inflammation. Moreover, AMPs have important interactions with host adaptive immune responses and repair, acting outside the first line of defense.

A collection of chemicals known as AMPs contribute significantly to the innate immune system. AMPs are tiny oligopeptides that can permeate lipid-rich membranes and are soluble in aqueous conditions. They can be cationic or anionic, amphipathic molecules with varied amino acid content. All types of life, including bacteria, fungi, plants, vertebrates, and invertebrates, include AMPs, which range in length from five to more than one hundred amino acid residues. These peptides target a wide range of species, including viruses and parasites. Using alternating variation selection operators and a machine learning model that directs the design of sequence space and includes residue sequences with a higher biological activity prediction, new synthetic peptides are synthesized in silico.

1.3 Types of AMPs

1. Antibacterial Peptides (ABPs): ABPs are the AMPs that have received the most attention to date, and the majority of them are cationic. They target bacterial cell membranes and disrupt the lipid bilayer structure, or they block important cellular

processes like protein synthesis and DNA replication. The majority of these AMPs have hydrophobicity, flexibility, and net charge.

2. Antiviral Peptides (AVPs): Viruses seriously endanger human life and have a significant financial impact on animal husbandry. Antiviral peptides are peptides that have the ability to prevent the spread of viruses. The antiviral activity neutralizes viruses by fusing with the viral envelope, weakening membranes so that they cannot infect the host cell, and decreasing the adherence of viruses to membranes [1].

3. Antifungal Peptides (AFPs): The number of AFPs that have been discovered has increased in recent years. Neutral and polar amino acids are commonly found in peptides with primary antifungal action, including those that have been isolated from plants. Fungi can be killed by AFPs by either attacking the cell walls or intracellular components. This binding capacity makes it easier for AFPs to effectively target fungal cells. Cell wall targeting-antifungals kill the target cells by rupturing the fungal membranes, increasing plasma membrane permeability, or by directly generating pores.

4. Antiparasitic Peptides (APPs): Compared to the previous three groups, APPs are a more compact category. By directly interacting with the cell membrane, APPs cause cell death.

5. Anticancer Peptides (ACPs): Anticancer peptides (ACPs) are a class of bioactive peptides that have the potential to be employed as a novel anticancer treatment. Compared to chemically based drugs, ACPs have a number of benefits, including high specificity, significant tumor penetrating power, and low toxicity to normal cells.

1.4 Role of Computation in AMP Prediction

The study of bioinformatics has significantly improved our understanding of biological processes and their mechanisms. To carry out biological research, bioinformatics creates and employs data, computational tools, and algorithms. In this dissertation, antimicrobial peptides have been classified and predicted using a bioinformatics technique (AMPs).

As more and more diseases become resistant to previously successful antimicrobial medications, the number of available medical treatments declines, potentially leading to a global health security problem. Antibiotic-resistant diseases are evolving, particularly in environments with excessive or improper use of these treatments, unsanitary living quarters, inadequate infection control, or improper food

handling. Antimicrobial-resistant illnesses prolong infection and increase the risk of death when they are unable to fight off therapies using previously successful drugs.

1.5 Literature Review of AMP Prediction

Antimicrobial peptides (AMPs) are part of innate immunity and are natural antibiotics encoded by specific genes [2]. 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 [3]. 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 mechanisms emerge. Therefore, we could face up with an era when infections can no longer be treated with antibiotics [4]. 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 [5]. Studying the properties of antimicrobial peptides in detail is a very important topic for drug design [6]. 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 [4], [5]. Hence, discovering or designing novel antimicrobial peptides became a major field of interest.

The increasing interest in AMPs has recently increased the efforts to discover new peptides with antimicrobial activities. Prior to the time-consuming, costly and difficult production processes, the accurate prediction of the activity of candidate peptides is very important. Along this line, several computational approaches such as de novo computational design [7]–[10], linguistic model [11], [12], pattern insertion algorithm [13]–[16], evolutionary-genetic algorithms [17]–[20] 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 [21], the use of machine learning methods became popular [22]–[25]. Machine learning is a computational technique, where the generated models can make predictions via learning the data [26]. 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 [27]. Lata *et al.* used the Support Vector Machine (SVM) method for prediction and classification of peptides on data which was collected from Antimicrobial Peptides Database [25]. Their model is based on aminoacid composition; and using five-fold cross validation they obtained 92.14% accuracy [25]. Burdukiewicz *et al.* attempted to identify essential AMP potential regions via applying Random Forest (RF) as a classification algorithm [28]. Chung *et al.* makes predictions for antimicrobial peptides on different organisms including amphibians, humans, fish, insects, plants, bacteria,and mammals [29]. Amino acid (aa) compositions, amino acid pairs, and the physicochemical properties are used as features. They performed feature selection, and applied RF, SVM, k-Nearest Neighbor (kNN) algorithms. They reported that RF generated the best result, which was over 92% accuracy on all tested organisms [29]. Bhadra *et al.* also utilized a RF algorithm for AMP prediction using physicochemical properties as features [24]. 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 96% accuracy with 10 fold cross validation technique and reduced feature sets [24]. Wang *et al.* combined sequence alignment with feature selection methods for classification of AMPs [30]. 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 [22]. There are many computational tools to predict AMPs based on machine learning approaches [18], [31]–[35]. Also, deep learning methods have been started to apply to antimicrobial peptides prediction problems. Bhadra *et al.* presents a method called deepAMP for sequences shorter than 30 aa. In their method they use an optimal feature set of reduced amino acid composition with convolutional neural network and obtain 77% accuracy. They also compare their results with RF and SVM algorithms. While the RF model gives close accuracy (75%) to CNN, the model used for SVM has a lower accuracy (72%) [36]. Su *et al.* designed a deep neural network which consists of an embedding layer and multi-convolutional layers for AMP identification. Compared with the existing models, their model achieved a higher accuracy score (92%) [37]. Schneider *et al.* used self organizing maps as input layers for their feedforward neural

network on AMP data and obtained 92% reclassification accuracy with balanced prediction on samples [38]. Witten *et al.* reported a convolutional neural network model for the classification and regression of AMPs [39]. They used Minimum Inhibition Concentration (MIC) values for regression and compared with ridge regression and kNN neighbors algorithms. They showed that CNN has better root mean squared error value (0.501) than others. Also for the classification part, when their CNN model is compared with other state-of-art methods, they have shown that higher prediction performance (97%) is obtained. Beltran *et al.* proposed a new feature selection approach to concentrate on molecular descriptors [40]. Their approach is applied on six benchmark datasets for evaluation. Also, they compared their results with state-of-the-art prediction tools and showed that their model outperforms these tools for prediction of antimicrobial and antibacterial peptides. In addition to the above-mentioned research efforts, some recent studies also used deep neural networks for the prediction of antimicrobial peptides[41]–[44]. However, there is no standardization in terms of the use of machine learning methods for the AMP prediction.

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 [45], [46]. However, in these databases, there is no standardization in terms of the experimental methods that are used to measure the activity of the AMPs *in vitro*. On the other hand, the antimicrobial activity of several AMPs have been predicted *in silico.* However, these algorithms do not take into account the physico-chemical and structural properties of the peptides and the mechanism of antimicrobial action against specific target microorganisms. Therefore, there is a need for new approaches based on the structure-activity relationship to accurately predict the antimicrobial activity of candidate peptides before synthesis.

In the last decade, a vast number of studies focused on the development of computational methods for determining the antimicrobial activity of natural or synthetic AMPs. However, the vast majority of these methods do not take into account the specific properties of bacterial targets. However, an AMP can exhibit different mechanisms of action against different target microorganisms. AMPs firstly interact with the bacterial cell wall and hence it is considered that the cell wall composition greatly affects the antimicrobial activity of AMPs [47]. It is also well known that Gram-positive and Gram-negative bacteria have different cell-surface architectures. For

example, Gram-negative bacteria have a thin peptidoglycan cell wall, surrounded by an outer membrane mainly containing lipopolysaccharide. Gram-positive bacteria lack an outer membrane but the cell wall contains thicker peptidoglycan layer and teichoic acids. Cell surface envelopes play a crucial role in the penetration and initial interaction of AMP. Therefore, the prediction of the antimicrobial activity of AMPs need be considered separately for these two different bacterial groups. For this reason, in this thesis we aimed to develop different machine learning approaches based on physico-chemical and structural properties of peptides and to predict their activities against Gram-positive and Gram-negative bacteria, separately. For this purpose, two different data sets were created in this study by selecting the peptides that are active against i) *E. coli ATCC 25922, P. aeruginosa ATCC 27853* and *A. baumannii* species for Gram-negative bacteria; and ii) *S. aureus ATCC 25923, L. monocytogenes ATTC 7644* and *B. cereus ATCC 11778* species for Gram-positive bacteria.

Chapter 2

2.Materials and Methods

2.1 Datasets and Data Preprocessing

2.1.1 Dataset 1: Linear Cationic Antimicrobial Peptide Dataset from DBAASP Database

In this thesis, as a data resource, several AMP databases were investigated. Database of Antimicrobial Activity and Structure of Peptides (DBAASP v.2. http://dbaasp.org, accessed on 10 August 2021) [48] was chosen due to the following reasons: (i) DBAASP is one of the most comprehensive AMP databases and it is widely used in literature. (ii) This database provides users with detailed information about the activity of thousands of peptides, where the antimicrobial activity has been tested experimentally or in silico against more than 4200 different organisms (bacteria, fungi, some parasites, viruses, and cancer cells). (iii) DBAASP has an application programmable interface (API). (iv)While most other databases were outdated, DBAASP is being updated frequently. Therefore, in this study, we have compiled our dataset from DBAASP.

In Figure 2.1, we illustrate our data preprocessing steps. In terms of synthesis type, ribosomally synthesized peptides, non-ribosomally synthesized peptides, and synthetic peptides were included in our datasets (Figure 2.1, Step 1). In terms of peptide complexity, we focused on monomers since 90% of the peptides in databases are monomeric peptides which consist of only one polypeptide chain (Figure 2.1, Step 2). Most of the property calculation algorithms recognize natural amino acids. Hence, the peptides which contain non-standard amino acids, or which have N and C terminal modifications were removed from the datasets (Figure 2.1, Step 3 and 4).

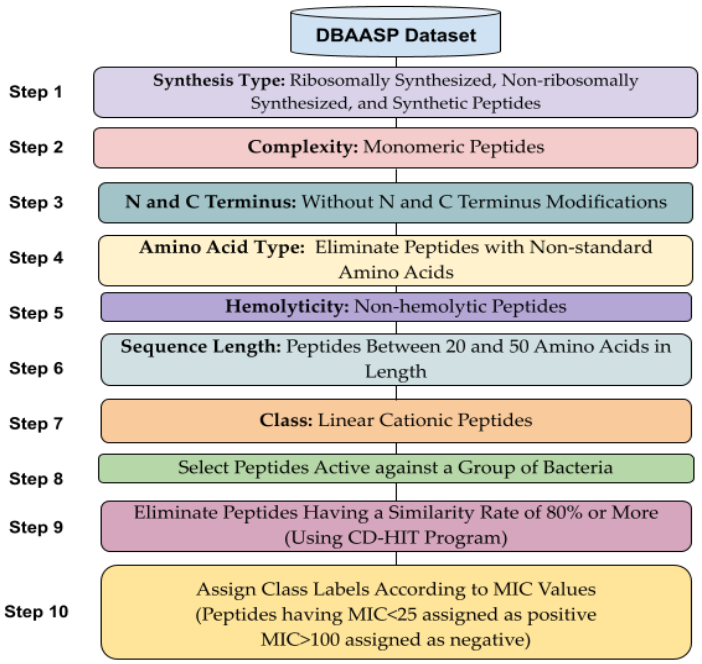


Figure 2.1 Workflow of Data Preprocessing

In this thesis, we plan to perform de novo antimicrobial peptide design by using the dataset that we have compiled. 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 [49]. Hence, here we focused on non-hemolytic peptides, and the peptides having hemolytic activity against human erythrocytes were removed from the datasets (Figure 2.1, Step 5).

AMPs exhibit their antimicrobial effects mainly through two different mechanisms. The membrane-targeting AMPs disrupt cell membrane integrity and lead to cytoplasmic leakage while the AMPs that use non-membrane targeting mechanisms mainly inhibit essential intracellular functions by interfering with DNA, RNA or proteins. AMPs shorter than 20 aa usually exert their antimicrobial effect by using non-membrane target mechanisms and they are defined as cell-penetrating antimicrobial peptides [50,51]. However, in this study, we focused on membrane-active peptides which are generally longer than 20 aa. Among the peptides longer than 20 aa in DBAASP, most of the peptide entries are shorter than 50 aa, hence we have selected the peptides with lengths ranging from 20 to 50 aa (Figure 2.1, Step 6).

Linear cationic antimicrobial peptides (LCAMPs) are the largest class of AMPs and they are widely found in different organisms [50]. Therefore, LCAMPs which have antimicrobial activity against Gram-negative bacteria including *Escherichia coli ATCC*

25922, *Pseudomonas aeruginosa ATCC 27853*, *Acinetobacter baumannii* species, and Gram-positive bacteria including *Staphylococcus aureus ATCC 25923*, *Listeria monocytogenes ATTC 7644*, *Bacillus cereus ATCC 11778* species are selected from the DBAASP (Figure 2.1, Step 7 and 8).

The CD-HIT program was used to eliminate the sequences that have more than 80% identity (Figure 2.1, Step 9). The CD-HIT program is widely used in the AMP prediction problem for removing highly similar sequences[52-60].

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 [51]. Hence, we utilized MIC values provided in DBAASP for each protein against different target organisms. All concentration units were converted to μg/mL using the molecular weights of the peptides. While the peptides having MIC value < 25 μg/mL against one of our target organisms are assigned as positive (antimicrobial), the peptides having MIC >100 μg/mL are assigned as negative (non-antimicrobial) (Figure 2.1, Step 10). This procedure is repeated separately for our Gram-negative and Gram-positive datasets. Hence, we assigned a class label to each peptide in our dataset.

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.

2.1.2 Dataset 2: Antimicrobial Peptide Dataset from APD Database

Veltri et al. provided a dataset containing 1778 AMPs and 1778 non-AMPs, which are available in the APD vr.3 database [56]. AMP peptides are active against Gram-negative and/or Gram-positive bacteria. These AMPs are filtered by removing sequences that are less than 10 amino acids long, and those that share 90% sequence identity using the CD-HIT program [34]. Additionally, non-AMPs are filtered by removing sequences less than 10 amino acids in length and those that share 40% sequence identity with the CD-HIT program [34]. The features are represented with a sequence-to-vector conversion, in which peptide sequences are encoded into uniform numerical vectors of length 200. Further details can be found in [56].

2.1.3 Dataset 3: Anti-Inflammatory Peptide Dataset

Manavalan et al. provided another dataset in [61], which is slightly different from other antimicrobial peptide datasets, since it includes anti-inflammatory peptides (AIPs). Using the IEDB (The Immune Epitope Database), they extracted positive and negative linear peptides that passed experimental validation [62,63]. A positive label was assigned to a peptide if it caused any of the anti-inflammatory cytokines to be produced in mouse and human T-cell experiments. Anti-inflammatory cytokines testing negative for linear peptides were regarded as negative. This dataset included 1258 AIPs and 1887 non-AIPs.

2.2 Feature Generation

Machine learning algorithms paved the way for the discovery of novel AMPs. Since ML models require numerical or categorical data (features) as an input, an informative encoding of proteins is crucial. Unfortunately, the development of appropriate encodings for proteins is a major challenge, and hence the feature generation problem for peptides has not been entirely solved so far. Therefore, the development of novel amino acid encodings is an active stand-alone research branch. A recent review paper [64] discussed state-of-the-art encodings of amino acids as well as their properties in sequence-based and structure-based aggregation.

2.2.1 Generation of Physico-Chemical Features (Descriptors)

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 physico-chemical 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 [10, 47,64-67]. As Spanig et al. noted in their recent review paper, the physico-chemical property encoding is also utilized by several web servers such as AVPpred [68] and DBAASP [48] in order to perform database queries, classify, and retrieve peptides. Moreover, physico-chemical properties have been employed in different studies to predict the antimicrobial effects of synthetic peptides [69] or to find substructures with antimicrobial potency in larger proteins [70].

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 [47]. 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 [**6**]. 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 2.1 presents example sequences that are included in our Gram-negative dataset. As shown in Table 2.1, along with 10 physico-chemical properties, each peptide has a class label as 0 or 1, where 0 implies that the peptide is not active against Gram-negative bacteria, and 1 implies that the peptide is active against these bacteria.

Table 2.1 An example of AMP and non-AMP peptides included in our Gram-negative dataset and their physico-chemical properties, excerpted from DBAASP [48].

Name of Seq.	Seq.	SL	NHM	NH	NC	IP	PD	TA	DCP	LM	PA	Mean MIC	CLASS
XPF-B2	GWASKIGTQLGKMAKVGLKEFVQS	24	1,11	-0,25	3	10,7	15	76	0,09	0,16	0	256,81	0
Ovalbumin (271-290)	SNVMEERKIKVYLPRMKMEE	20	0,13	-0,28	1	9,38	30	67	-0,11	0,29	0	800	0
MBI 29 A1	KWKSFIKKLTSVLKKVVTTALPALIS	26	1,03	-0,54	6	11,37	12	106	0,16	0,27	3,4	9,33	1
Cyanophlycti	FLNALK	21	1,69	-0,24	5	11,74	15	88	-0,03	0,25	0	12	1

n	NFA KTA GKR LKS LLN												
	...												

*Seq.: Sequence, SL: Sequence Length, NHM: Normalized Hydrophobic Moment, NM: Normalized Hydrphobicity, NC: Net Charge, IP: Isoelectric Point, PD: Penetration Depth, TA: Tilt Angle, DCP: Disordered Conformation Propensity, LM: Linear Moment, PA: Propensity in vitro Aggregation, MIC: Minimum Inhibiton Concentration

2.2.2 Generation of Sequence-based, Structure-based, Linguistic-based Features

Several studies have provided web servers or standalone programs to calculate features from peptide sequences [71-73]. These tools are reviewed in detail in [64]. Propy tool, which is developed by Cao et al. provides five feature groups with 13 subfeatures from proteins or peptide sequences [74]. Chen et al. developed iFeature tool, which calculates 18 feature groups and also provides clustering and feature selection on protein and peptide sequences [75]. PyBioMed is another Python package that computes features not only from protein, DNA sequences but also from chemical structures [76]. It is a frequently used tool in this field due to its wide scope in attribute definition [77-79]. The PyProtein [76] is a module of PyBioMed for calculating the structural and physico-chemical features of proteins and peptides. It computes five feature groups including physico-chemical, amino acid composition, pseudo-amino acid composition (PseAAC), Composition, Transition and Distribution (CTD) of physico-chemical properties, autocorrelation, sequence order, and conjoint triad. These features are also known as different Chou's PseAAC modes [74]. For our Gram-positive and Gram-negative datasets, 1497 features including amino acid composition (20), dipeptide composition (400), CTD composition (21), CTD transition (21), CTD distribution (105), Moran autocorrelation (240), Geary autocorrelation (240), Moreau–Broto autocorrelation (240), Quasi-sequence-order descriptors (100), Sequence order coupling number (60), Pseudo Amino Acid Composition (50) are calculated via freely available PyProtein module in PyBioMed python package [76]. These features are also used in other studies for AMP prediction using machine learning [65, 80].

2.3 Data Exploration

2.3.1 Principal Component Analysis for Outlier Detection and Elimination

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 [81]. 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 [82-85].

2.4 Feature Selection Techniques

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 [86]. 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 recent tools. Such an approach was adopted by different tools such as SVM RCE, SVM-RCE-R [87-89], maTE [90], CogNet [91], miRcorrNet [92], miRModuleNet [93], and Integrating Gene Ontology Based Grouping and Ranking [94]. Recently, these tools and their competitors were reviewed in [95].

2.4.1 Maximum Relevance — Minimum Redundancy (mRMR)

mRMR is a filtering method that tries to select the most relevant features with the class labels, while simultaneously trying to minimize the redundancy between the selected features. This algorithm starts with an empty set, uses mutual information to balance the features, and then combines sequential search with forward selection to identify the best subset of attributes [96].

2.4.2 Conditional Mutual Information Maximization (CMIM)

Conditional Mutual Information Maximization (CMIM) strikes a balance between the candidate feature's ability to forecast the future and its independence from other characteristics that have already been chosen by using conditional mutual information to calculate distance [97].

2.4.3 Extreme Gradient Boosting (XGB)

Extreme Gradient Boosting (XGB) is another mostly used feature selection method. Importance in XGB assigns a score based on the usefulness or value of each feature in building the boosted decision trees within the model. An attribute's relative relevance increases as more and more decision trees use it to make important decisions [98].

2.4.4 Information Gain (IG)

Information gain (IG) is utilized for feature selection by assessing each variable's gain in relation to the target variable. For each of the independent features, we determine the information gain. The traits would then be ranked according to their individual information gains in descending order. We would choose a cutoff point and incorporate all features above the cutoff point into the machine learning algorithms [99].

2.5 Machine Learning Classifiers

The modeling of systems that make predictions by making inferences on the data along with statistical and mathematical operations with computers is called machine learning. Machine learning creates a model, analyzes data and produces a result. This model learns using data and constantly updates itself to make accurate predictions. Machine learning algorithms make decisions by learning from datasets, rather than acting according to a set of predefined rules. Machine learning algorithms can be used in many fields such as data mining, natural language processing, image processing, robotics and bioinformatics.

The increasing interest in AMPs has recently increased the efforts to discover new peptides with antimicrobial activities. Prior to the time-consuming, costly and difficult production processes, the accurate prediction of the activity of candidate peptides is very important. Along this line, several computational approaches 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, the use of machine learning methods became popular. Based on this purpose, we used traditional machine learning classifiers in this thesis.

2.5.1 Random Forest (RF)

Random Forests (RF) 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. Basically, the algorithm creates a decision tree for each sample, and the estimated value result of each decision tree is formed. Voting is performed for each value formed as a result of the prediction. Finally, the algorithm generates the result by choosing the most voted value for the final guess [100].

Random forest parameters either help the model be more predictive or make the model's training process simpler. There are many parameters that are used and optimized for increasing performance of the algorithm. "Max_features" parameter is the maximum number of features Random Forest is allowed to try in individual tree. "n_estimator" parameter refers the number of trees you want to build before taking the maximum voting of predictions. The longest path from the root node to the leaf node is referred to as a tree's "max_depth" parameter in Random Forest. We used the default parameter values for all parameters obtained by scikit-learn library.

2.5.2 Support Vector Machines (SVM)

One of the discriminative classifiers used in machine learning is the support vector machine. Finding a hyperplane that best discriminates between two or more classes is the main goal of SVM [101]. Both linear and nonlinear datasets can be classified using SVM. Data from several classes can be linearly separated from one another in a variety of ways when using linear separation. As a linear decision boundary cannot divide the data in nonlinear separation, a non-linear mapping is used instead. A linear hyperplane is discovered that divides the data samples in the new space after the data samples in the input feature space are mapped to a higher dimensional space. The optimization problem can be resolved without explicitly shifting the data points to the new space by formulating it in dual space and utilizing the kernel method.

There are two most important hyperparameter that is SVM used. The penalty for the classifier is determined by the C parameter. The margin will be minimal if C is very large since there will be a high penalty for misclassification training. If the C is low, there will be a low penalty and high margin. A single training point's radius of influence is controlled by the gamma parameter. Low gamma values suggest a wide similarity radius, which causes more points to be grouped together. To be included in the same group (or class) with high gamma values, the points must be quite close to one another. We used the default parameter values for C and gamma parameters obtained by scikit-learn library.

2.5.3 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 method, a weak learner is used to train the training set at first. After the training phase, wrongly predicted samples are crucial for this algorithm. The erroneously learnt training data from the first iteration is retrained by giving it more priority in the following training phase [102].

Different hyperparameters are used for AdaBoost Algorithm. The "base_estimator" parameter is used to indicate the kind of weak learner or base learner that can be employed. The "n_estimators" parameter refers the number of base estimators or weak learners we want to use in our dataset. The "learning_rate" parameter is provided to shrink the contribution of each classifier.

2.5.4 LogitBoost

LogitBoost is a boosting classification algorithm which has been developed to provide solutions to the overfitting problem experienced in AdaBoost. This algorithm linearly reduces the errors in the training. As both conduct an additive logistic regression, LogitBoost and AdaBoost are similar to one another. AdaBoost reduces the exponential loss, whereas LogitBoost reduces the logistic loss [103].

The LogitBoost Algorithm has the same hyperparameters with AdaBoost Algorithm. These parameters are explained in section 2.5.3.

2.5.5 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 [104]. Decision trees are a type of machine learning algorithm used for classification and regression tasks. They are a graphical representation of a set of decisions and the possible consequences of those decisions, and are constructed starting at the root node and working their way through the tree, making decisions based on the values of the properties at each node. Decision trees are useful in data science as they are easy to understand and interpret and can handle both continuous and categorical data [104].

The “criterion” parameters refers to how to measure the quality of a split in a decision tree. The “Max_Depth” parameter is used for identifying the maximum depth of the tree. The “Min_Samples_Split” parameter is the minimum samples required to split an internal node. The “Min_Samples_Leaf” parameter is the minimum samples required to be at a leaf node. The “Max_Features” parameter identifies the number of features to consider when looking for the best split. All parameters are used with their default parameters.

2.5.6 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. This method classifies data by taking into account the majority of votes among the "k" points that are closest to the unlabeled data point. It operates on unobserved data and will look for the k-most comparable cases in the training dataset. Different metrics are used to determine the distance between two points such as Euclidean distance, Hamming distance [105].

The “n_neighbors” parameter is used for the number of neighbours.The “metric” parameter used to decide which distance metric to be used will calculating the similarity.

2.5.7 Stacking

Stacking, one of the most common ensemble machine learning techniques, is used to estimate many nodes in order to create a new model and enhance model

performance. Using stacking, multiple models can be trained to handle related issues and then create a new, more effective model based on the combined results.

In the thesis for first study, two stacking ensembles were built by combining various classifiers. The first ensemble technique combines the support vector machine with k-Nearest Neighbor and uses Logistic Regression as the meta-learner. The second ensemble technique combines LogitBoost with k-Nearest Neighbor and uses Random Forest as the meta-learner.

2.6 Performance Metrics

The following formulas were used to calculate a number of quantitative metrics, including Accuracy, Sensitivity, Specificity, Precision, F1-measure and Balanced Accuracy in order to assess the performance of the model:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{2.1}$$

$$\text{Recall (Sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.4}$$

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{2.5}$$

where TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative. Furthermore, we used Area Under the Curve (AUC) for performance evaluation. AUC is one of the most crucial evaluation criteria for assessing the effectiveness of any classification model. The level or measurement of separability is represented by AUC. It reveals how well the model can differentiate across classes. According to our study, the higher the AUC, the better the model is at distinguishing between samples with negative (non-AMP) and positive (AMP).

Chapter 3

3.Prediction of Linear Cationic Antimicrobial Peptides Active against Gram-Negative and Gram-Positive Bacteria Based on Machine Learning Models

3.1 Motivation

In this study, we aimed to develop a machine learning approach based on physico-chemical and structural properties of peptides and to predict their activities against Gram-positive and Gram-negative bacteria, separately. For this purpose, two different data sets were created in this study by selecting the peptides that are active against (i) *E. coli ATCC 25922*, *P. Aeruginosa ATCC 27853*, and *A. baumannii* species for Gram-negative bacteria, and (ii) *S. Aureus ATCC 25923*, *L. Monocytogenes ATTC 7644*, and *B. cereus ATCC 11778* species for Gram-positive bacteria. 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 (AUC), F1 measure.

3.2 Model Construction

As illustrated in Figure 3.1, 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). 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 [106]. 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%.

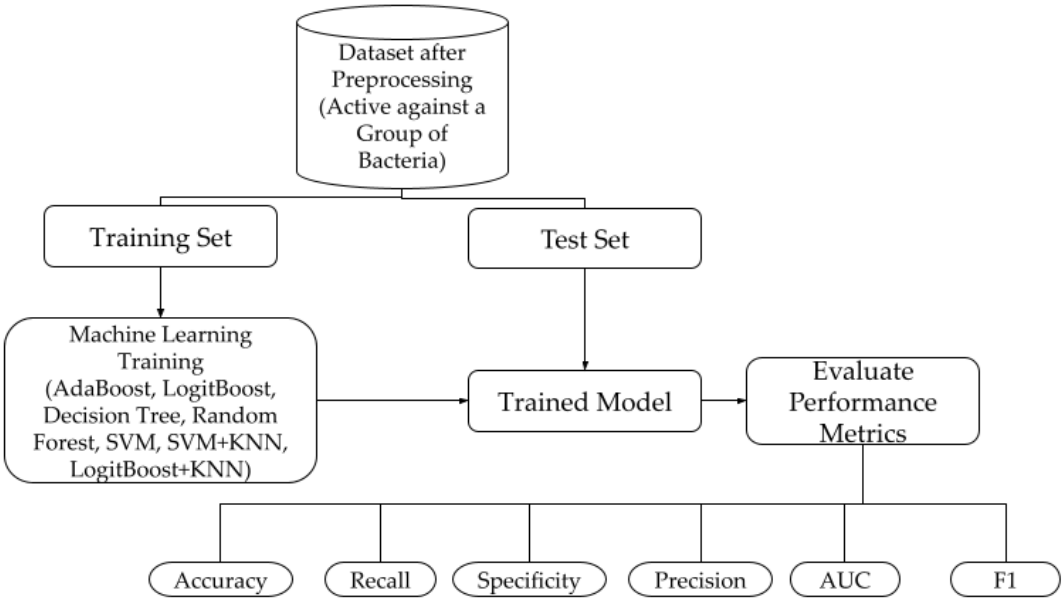


Figure 3.1 Flowchart of Model Construction

3.3 Results

3.3.1 Training Models Using Physico-Chemical Features

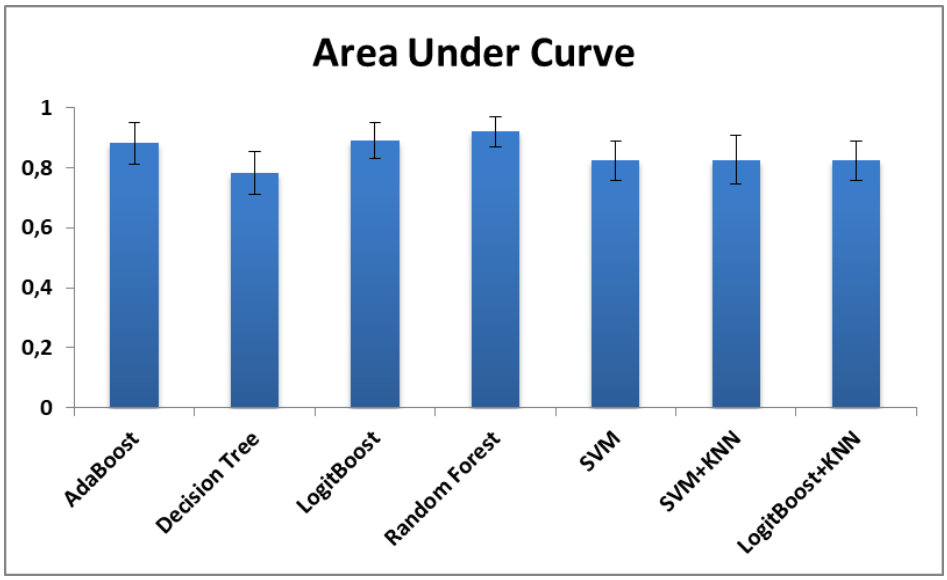
In our experiments, firstly we have used the above-mentioned ten physico-chemical features and different machine learning methods i) to learn whether the peptides in each of our datasets have antimicrobial activity or not; and ii) to classify them accordingly. To this end, we have applied methods such as AdaBoost, Decision Tree, LogitBoost, RF, and SVM. As shown in Tables 3.1 and 3.2, for both Gram-negative and Gram-positive datasets, RF classifier resulted in the best performance metrics. While the AUC rate reached up to 90% for Gram-positive data, this rate was 92% for Gram-negative data. Not only for AUC rate, but also for other measures such as accuracy, recall, specificity, precision and F1 measure, RF yielded the best performance metrics. Figure 3.2 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 3.2(a) and in Table 3.1, while 92% AUC value is obtained for gram negative dataset, 90% AUC value is obtained for Gram-positive dataset (shown in Figure 3.2(b) and in Table 3.2) using RF classifier. While the AUC values of other classifiers range between 0,77-0,87 for Gram-positive dataset (shown in Figure 3.2(b) and in Table 3.2), it ranges between 0,78-0,89 for Gram-negative dataset.

Table 3.1 Comparison of different models according to different performance metrics for Gram-negative dataset, using physico-chemical features.

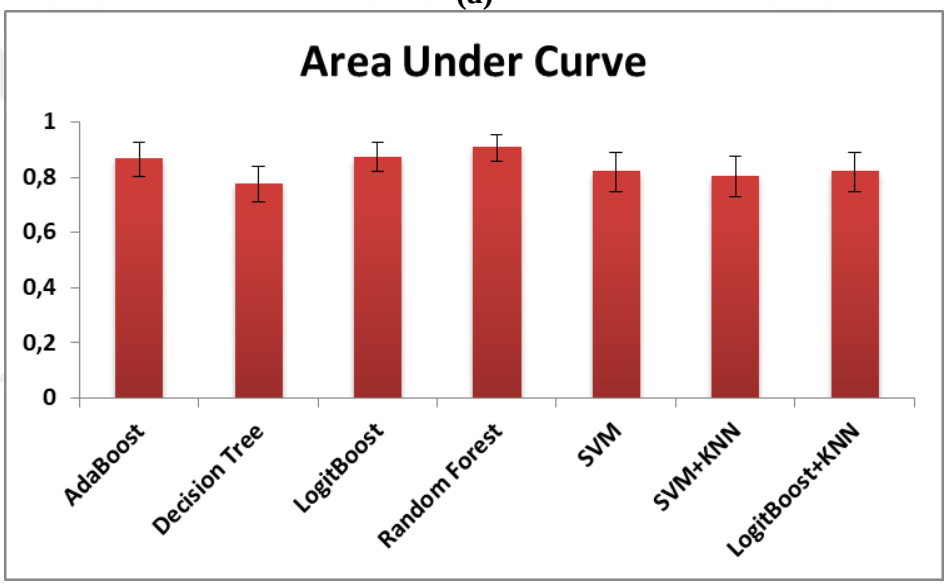
Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1
AdaBoost	0,85±0,06	0,92±0,06	0,72±0,20	0,87±0,07	0,88±0,06	0,89±0,04
Decision Tree	0,79±0,06	0,87±0,07	0,66±0,24	0,84±0,07	0,78±0,07	0,85±0,04
LogitBoost	0,86±0,05	0,92±0,06	0,74±0,16	0,88±0,06	0,89±0,06	0,90±0,03
RF	**0,89±0,05**	**0,93±0,04**	**0,79±0,16**	**0,90±0,06**	**0,92±0,05**	**0,91±0,03**
SVM	0,80±0,05	0,93±0,06	0,56±0,21	0,81±0,07	0,82±0,06	0,86±0,03
SVM+kNN	0,80±0,07	0,93±0,05	0,56±0,25	0,81±0,08	0,82±0,08	0,86±0,04
LogitBoost+kNN	0,80±0,05	0,93±0,06	0,56±0,21	0,81±0,07	0,82±0,06	0,86±0,03

Table 3.2 Comparison of different models according to different performance metrics for Gram-positive dataset, using physico-chemical features.

Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1
AdaBoost	0,84±0,06	0,85±0,08	0,83±0,14	0,83±0,10	0,86±0,06	0,83±0,05
Decision Tree	0,77±0,07	0,77±0,10	0,77±0,16	0,769±0,09	0,77±0,06	0,76±0,05
LogitBoost	0,83±0,06	0,84±0,09	0,82±0,15	0,83±0,10	0,87±0,05	0,83±0,05
RF	0,87±0,04	0,87±0,07	0,87±0,08	0,87±0,07	0,90±0,04	0,87±0,04
SVM	0,77±0,07	0,85±0,11	0,71±0,19	0,75±0,12	0,81±0,06	0,78±0,05
SVM+kNN	0,76±0,08	0,81±0,11	0,72±0,21	0,76±0,13	0,80±0,07	0,77±0,05
LogitBoost+kNN	**0,77±0,07**	**0,85±0,11**	**0,71±0,19**	**0,75±0,12**	**0,81±0,06**	**0,78±0,05**



(a)



(b)

Figure 3.2 Comparison of the performances of different models in terms of their AUC values with standard deviation values for (a) Gram-negative, and (b) Gram-positive datasets, using physico-chemical features.

3.3.2 Results for Feature Scoring and Feature Ranking

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 RF classifier, we have utilized the feature scores of this model throughout the rest of the thesis. The weighted decrease in node impurity divided by the likelihood of reaching that node is used to determine a feature's importance for RF classifier. The node probability can be computed by dividing the

total number of samples by the number of samples that reach the node. The higher the score the more important the feature. When we analyze the feature scores (shown in Figures 3.3 and 3.4), 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.

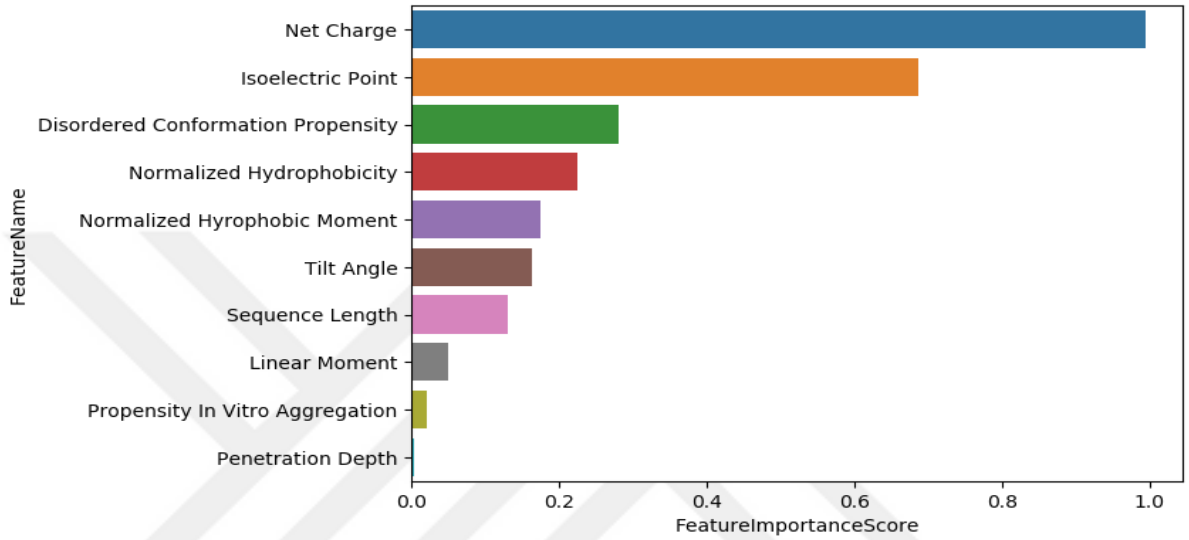


Figure 3.3 Feature ranking according to their importances in classification using random forest model in Gram-negative dataset.

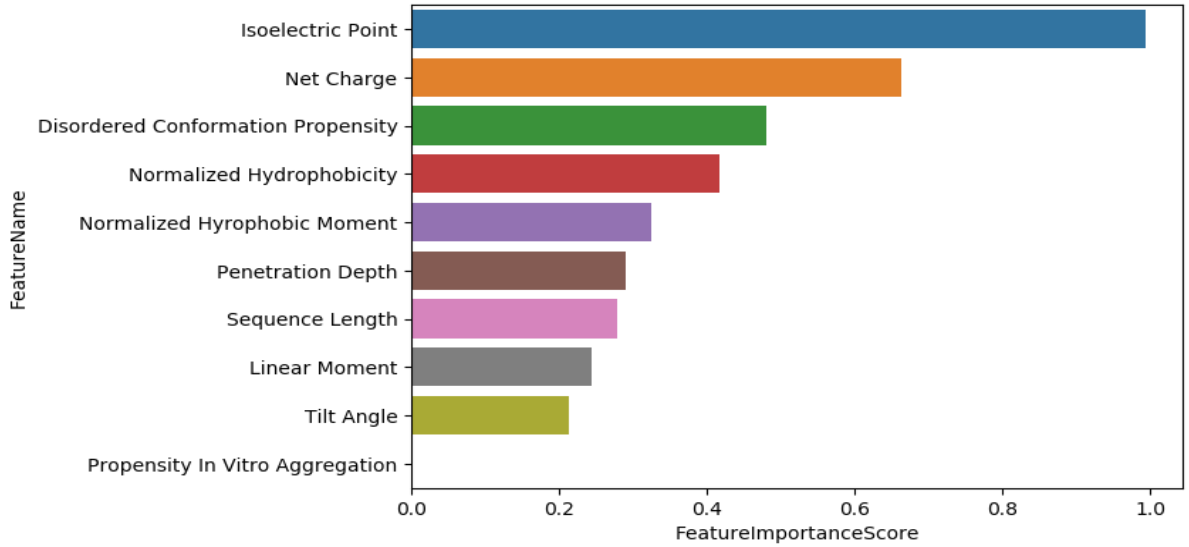


Figure 3.4 Feature ranking according to their importances in classification using RF model in Gram-positive dataset.

3.3.3 Results for Outlier Detection and Elimination

In our study, we applied PCA to our dataset for visualizing the AMP and Non-AMP samples. In Figure 3.5, we present PCA results of the Gram-negative dataset (Figure 3.5(A), 3.5(C)), and of the Gram-positive dataset (Figure 3.5(B), 3.5(D)). While Figures 3.5(A), 3.5(B) refer to the PCA results in 3D, Figures 3.5(C), 3.5(D) refer to the PCA results in 2D. We observe in Figure 3.5 that there are some outlier samples (peptides) in both Gram-negative and positive datasets.

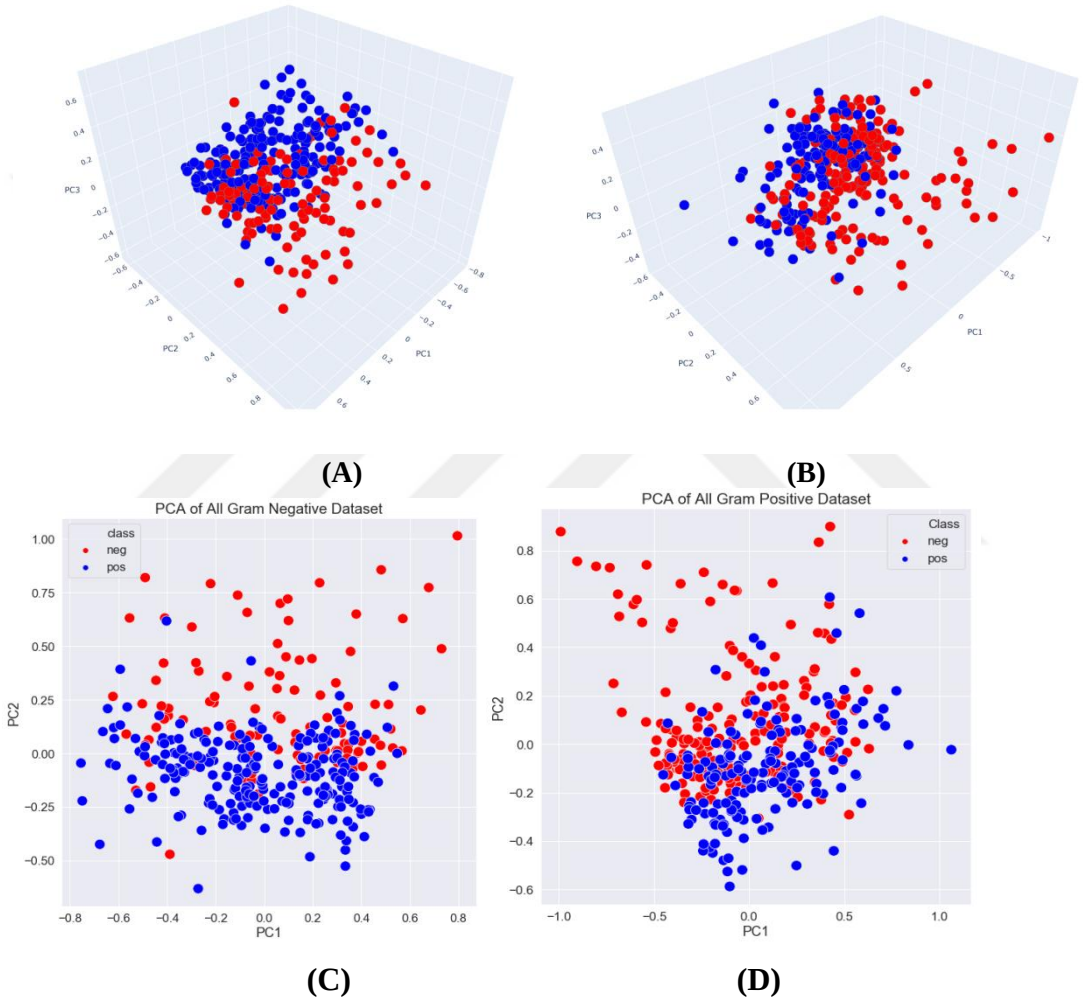
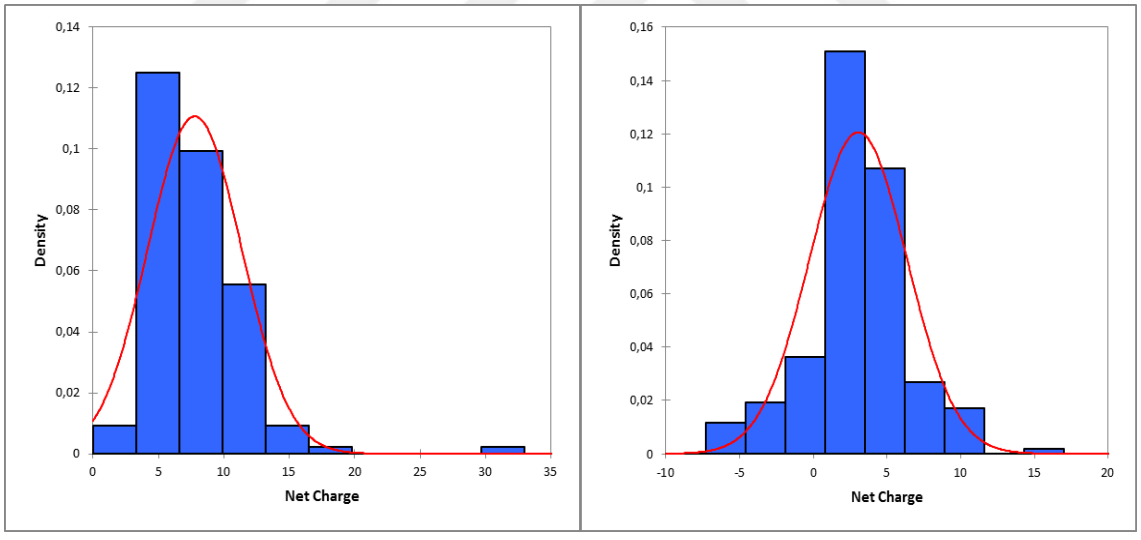


Figure 3.5 Principal component analysis results for Gram-negative dataset are shown in (A) and (C); for Gram-positive dataset are shown in (B) and (D). While 3D plots are presented in (A) and (B), 2D plots are presented in (C) and (D).

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 [107], more specifically on one-class approaches [108-112].

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 3.6 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 3.6 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 3.3.



(A) **(B)**

Figure 3.6 Graphical representation of Net Charge feature of the Gram-positive dataset. Histogram of (A) AMP class, (B) Non-AMP class.

Table 3.3 Minimum and maximum values of each feature that are used in outlier elimination.

Features	Gram-negative Dataset	Gram-positive Dataset

	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 3.7, we present PCA results of the Gram-negative dataset (shown in A, C); and of the Gram-positive dataset (shown in Figure B,D) after outlier detection and elimination. While PCA plots are presented in 3D in (Figure 3.7(A), (B)), they are presented in 2D in (Figure 3.7(C), (D)). While the red colors refer to Non-AMPs, blue colors indicate AMPs. Compared with Figure 3.5, Figure 3.7 implies that the positive class members are better separated from negative class members for both datasets after outliers are eliminated.

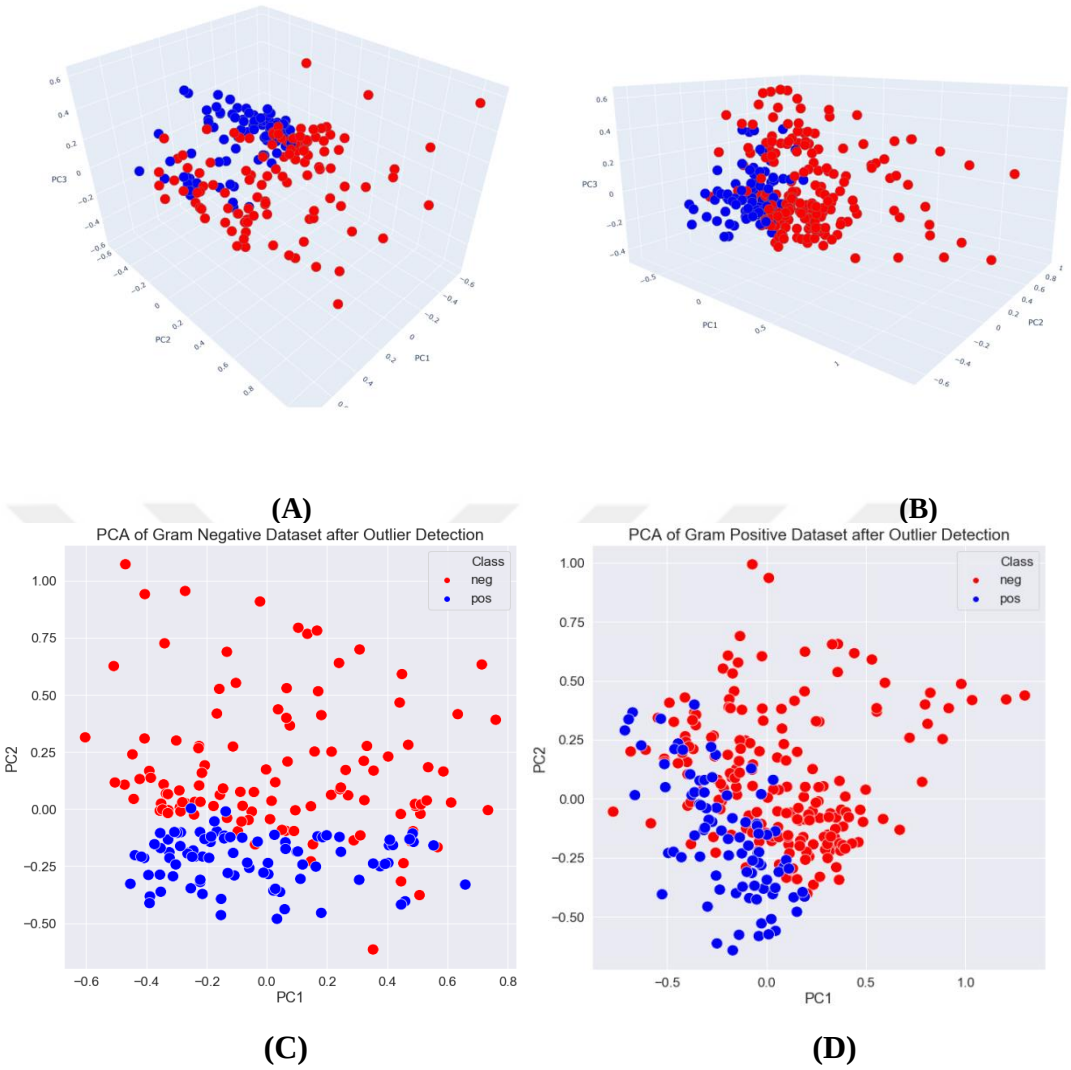


Figure 3.7 Principal component analysis of Gram-negative dataset (shown in A,C) and of Gram-positive dataset (shown in B,D) after outlier detection and elimination, shown in 3D in (A, B) and in 2D in (C, D).

Using two of the datasets after outlier elimination, we repeated our classification experiment as explained in the methods section. As shown in Tables 3.4 and 3.5, when outlier removal is applied, we have obtained higher performance metrics. As presented in Tables 3.4 and 3.5, the AUC rate increased by 7% and reached 99% AUC for the Gram-negative dataset, while this score is obtained as 97% for the Gram-positive dataset.

Table 3.4 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**	**BalancedAcc.**
AdaBoost	0,97±0,03	0,99±0,03	0,96±0,04	0,95±0,05	0,99±0,01	0,97±0,03	0,97±0,04
Decision Tree	0,91±0,06	0,92±0,08	0,91±0,08	0,89±0,09	0,91±0,06	0,90±0,06	0,91±0,08
LogitBoost	0,97±0,03	**0,99±0,02**	0,96±0,05	0,95±0,05	0,99±0,01	0,97±0,03	0,98±0,03
RF	**0,98±0,02**	0,99±0,02	**0,97±0,04**	**0,97±0,05**	**0,99±0,01**	**0,98±0,03**	**0,98±0,03**
SVM	0,98±0,02	0,99±0,03	0,97±0,04	0,96±0,04	0,98±0,01	0,97±0,03	0,98±0,03
SVM+kNN	0,81±0,11	0,82±0,14	0,80±0,24	0,81±0,16	0,84±0,10	0,80±0,09	0,81±0,19
LogitBoost+kNN	0,98±0,02	0,99±0,03	0,97±0,04	0,96±0,04	0,98±0,01	0,97±0,03	0,98±0,03

Table 3.5 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	Bal.Acc.
AdaBoost	0,93±0,04	0,92±0,08	0,94±0,06	0,89±0,09	0,96±0,03	0,90±0,05	0,93±0,07
Decision Tree	0,88±0,05	0,82±0,12	0,91±0,06	0,82±0,11	0,86±0,07	0,81±0,09	0,86±0,09
LogitBoost	0,93±0,05	0,93±0,09	0,93±0,07	0,88±0,11	0,96±0,03	0,90±0,07	0,93±0,08
RF	0,95±0,03	0,95±0,07	0,95±0,05	0,90±0,09	0,97±0,02	0,92±0,05	0,95±0,06
SVM	0,91±0,04	0,90±0,11	0,91±0,06	0,85±0,11	0,93±0,04	0,86±0,06	091±0,09
SVM+kNN	0,77±0,10	0,75±0,16	0,78±0,20	0,68±0,17	0,81±0,08	0,68±0,08	0,76±0,18
LogitBoost+kNN	0,91±0,04	0,90±0,11	0,91±0,06	0,85±0,11	0,93±0,04	0,86±0,04	0,91±0,09

3.3.4 Training models Using an Extended Set of Features

In addition to the physico-chemical features, structural properties, sequence order, compositional features, the pattern of terminal residues, amino acid composition, dipeptide composition, autocorrelation, pseudo-amino acid composition and sequence order properties have been suggested as additional features for representing amino acid sequences[64-68]. Hence, in our experiments we have also tested the effect of different features, in addition to the ten physico-chemical features. As explained in Methods Section, amino acid composition, pseudo amino acid composition, autocorrelation and sequence order properties are calculated for the peptides included in our dataset. These 1497 additional features were added to the initially calculated 10 physico-chemical features, and our final dataset included 1507 features in total. Using the datasets including the extended set of features, we have repeated our classification experiment as explained in the methods section. For both Gram-negative and Gram-positive datasets, when an extended set of features are utilized, the obtained performance metrics (as shown in Tables 3.6 and 3.7) were slightly lower than the performance metrics obtained using only ten physico-chemical features (as shown in Tables 3.4 and 3.5). For the Gram-negative dataset, while the extended set of features yielded 98% AUC with LogitBoost, physico-chemical features yielded 99% AUC with RF. For Gram-positive dataset, while the model using an extended set of features achieved 95% AUC with RF, the generated model using only ten physico-chemical features achieved 97% AUC.

Table 3.6 Comparison of the models according to performance metrics for the Gram-negative dataset with 1507 features.

Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1	Bal.Acc.
Adaboost	0.96±0.03	0.98±0.04	0.95±0.05	0.94±0.06	0.98±0.02	0.96±0.03	0.96±0.05
DT	0.90±0.06	0.90±0.09	0.90±0.08	0.88±0.09	0.90±0.06	0.88±0.07	0.90±0.08
LogitBoost	**0.97±0.03**	**0.98±0.03**	**0.95±0.06**	**0.95±0.06**	**0.98±0.01**	**0.96±0.03**	**0.97±0.04**
RF	0.95±0.04	0.98±0.04	0.94±0.06	0.93±0.07	0.98±0.02	0.95±0.04	0.96±0.05

Table 3.7 Comparison of the models according to performance metrics for the Gram-positive dataset with 1507 features.

Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1	Bal.Acc.
Adaboost	0.89±0.05	0.88±0.10	0.90±0.08	0.82±0.11	0.93±0.04	0.84±0.07	0.89±0.09
DT	0.82±0.10	0.74±0.14	0.86±0.16	0.75±0.13	0.80±0.07	0.73±0.10	0.80±0.15
LogitBoost	0.90±0.05	0.89±0.09	0.91±0.07	0.84±0.11	0.94±0.03	0.85±0.06	0.90±0.08
RF	**0.92±0.04**	**0.91±0.09**	**0.92±0.06**	**0.86±0.10**	**0.95±0.03**	**0.88±0.06**	**0.92±0.08**

3.3.5 Training models Using an Extended Set of Features and Applying Feature Selection

There are a high number of features (1507) in the extended feature set. To remove redundant features and select informative ones, we repeated our experiments with different feature selection methods including Information Gain (IG), Maximum Relevance-Minimum Redundancy (MRMR), Conditional Mutual Information Maximization (CMIM), XGBoost (XGB). We have focused on the top 3 scoring features in both Gram-negative and Gram-positive datasets. The performance metrics obtained after feature selection are presented in Tables 3.8 and 3.9 for Gram-negative and Gram-positive datasets, respectively. For the Gram-negative dataset, the generated LogitBoost model with the three selected features by XGBoost resulted in the best performance metric (96% AUC) among all other tested classifiers, all other tested feature selection methods. The top 3 selected features on the Gram-negative dataset are GearyAuto_Steric14 from Geary Autocorrelation set, PAAC42 from pseudo-aminoacid composition, and PolarityT13 from composition, transition and distribution of physico-chemical properties. On the Gram-negative dataset, the performance of the physico-chemical feature set (99% AUC with RF with 10 features) was still higher than the performance of the extended feature set (98% AUC with LogitBoost with 1507 features); and than the performance of the extended feature set after feature selection (96% AUC with LogitBoost with 3 features).

For the Gram-positive dataset, the generated RF model with the three selected features by Information Gain resulted in the best performance metric (94% AUC) among all other tested classifiers, all other tested feature selection methods. On the Gram-positive dataset, the performance of the physico-chemical feature set (97% AUC with RF with 10 features) was still higher than the performance of the extended feature set (95% AUC with RF with 1507 features); and than the performance of the extended feature set after feature selection (94% AUC with RF with 3 features). It is interesting to note that on the Gram-positive dataset, the top 3 scoring features of the extended descriptors are isoelectric point, net charge, disordered conformation propensity, which all belong to our initial 10 physico-chemical features.

When we compare the performance metrics before and after feature selection is applied on the extended set of features, we observed that for Gram-positive and for Gram-negative datasets, the AUC performance metrics only decreased by 1% and 2%, respectively, when three selected features are used to generate the model (as compared with the 1507 features included in the extended set of features). That is to say that using only 3 features yields satisfactory performance results (96% and 94% AUC) for Gram-negative and Gram-positive datasets, respectively.

Table 3.8 Comparison of the models according to performance metrics for the Gram-negative dataset after feature selection (XGBoost).

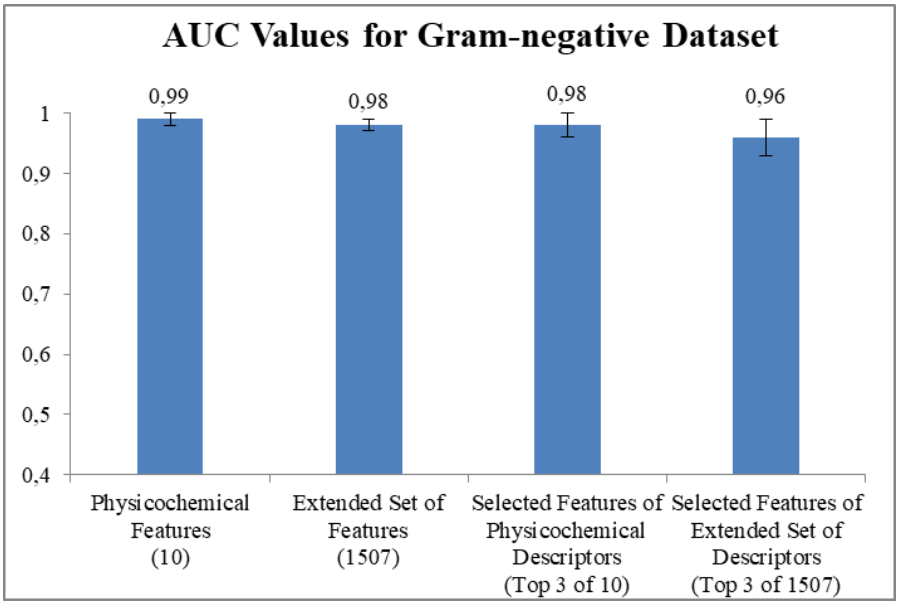
Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1
Adaboost	0.94±0.05	0.97±0.05	0.91±0.09	0.91±0.09	0.95±0.06	0.93±0.07
DT	0.90±0.07	0.90±0.11	0.89±0.09	0.87±0.10	0.90±0.08	0.88±0.10
LogitBoost	0.94±0.05	0.98±0.04	0.91±0.09	0.90±0.09	0.96±0.06	0.94±0.07
RF	0.94±0.05	0.97±0.06	0.92±0.08	0.91±0.08	0.96±0.05	0.93±0.07

Table 3.9 Comparison of the models according to performance metrics for the Gram-positive dataset after feature selection (Information gain).

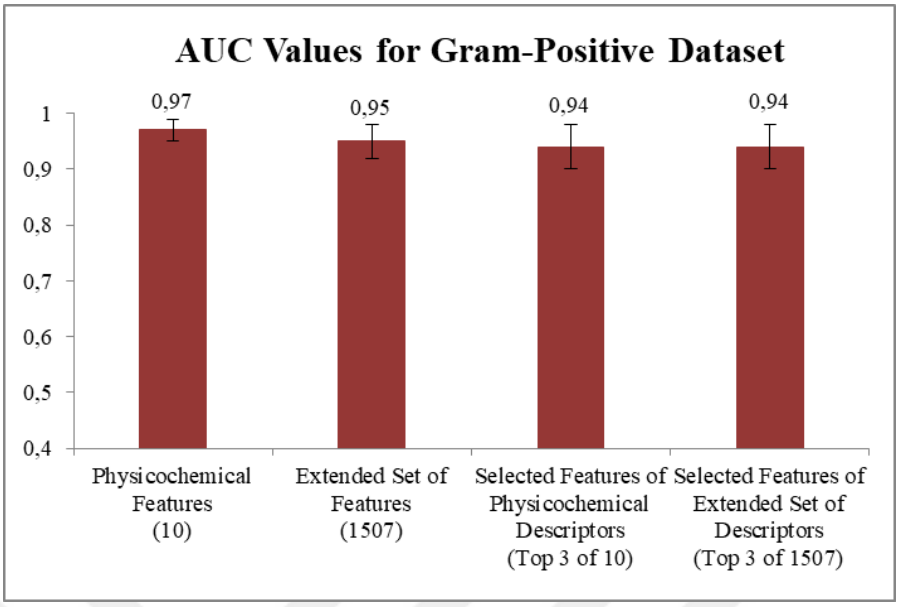
Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1
Adaboost	0.86±0.06	0.91±0.10	0.83±0.10	0.74±0.12	0.90±0.05	0.80±0.07

DT	0.83±0.10	0.77±0.12	0.86±0.16	0.76±0.14	0.82±0.07	0.75±0.10
LogitBoost	0.87±0.05	0.90±0.10	0.86±0.08	0.77±0.11	0.91±0.04	0.82±0.06
RF	0.90±0.04	0.89±0.10	0.91±0.07	0.84±0.11	0.94±0.04	0.86±0.06

Similarly, to compare the performance metrics of the models which use physico-chemical features with the models which use the extended set of features, we reduced the number of features in the original dataset to the same number of features (top 3 scoring features). For this purpose, we applied the same feature selection strategy on our original dataset which includes only physico-chemical features. We wanted to test whether a certain number of attributes will be sufficient for prediction. In Figure 3.8 we present the AUC values obtained using i) 10 physico-chemical features, ii) extended set of features (1507 features), iii) top 3 scoring features of physico-chemical descriptors, iv) top 3 scoring features of extended descriptors. As illustrated in Figure 3.8(A), for the Gram-negative dataset, the models which use the physico-chemical features yield the best AUC score (99%). For this dataset, the extended features and the top 3 scoring features (normalized hydrophobicity, normalized hydrophobic moment, net charge) of the physico-chemical features generate the same AUC values (98%). It can be observed from Figure 3.8(B) that on the Gram-positive dataset, the model which uses physico-chemical features achieves 97% AUC and hence obtains better performance metrics than the extended dataset and than the models using top 3 scoring features. On the Gram-positive dataset, the top 3 scoring features of physico-chemical descriptors are net charge, isoelectric point, disordered conformation propensity.



(A)



(B)

Figure 3.8 Comparison of the AUC results before and after feature selection is applied on physico-chemical features and extended set of features for (A) Gram-negative, and (B) Gram-positive dataset.

3.4 Discussions

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 [54]. 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. 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[6].

In this study, a machine learning-based approach was developed for the first time to separately predict the peptides active against Gram-positive and Gram-negative bacteria. It is well known that Gram-positive and Gram-negative bacteria have different

cell-surface architectures. For example, Gram-negative bacteria have a thin peptidoglycan cell wall, surrounded by an outer membrane mainly containing lipopolysaccharide. Gram-positive bacteria lack an outer membrane but the cell wall contains thicker peptidoglycan layer and teichoic acids. Cell surface envelopes play a crucial role in the penetration and initial interaction of AMP. Therefore, the prediction of the antimicrobial activity of AMPs need be considered separately for these two different bacterial groups. For this purpose, in this study two different data sets were created by selecting peptides that are active against i) *E. coli ATCC 25922, P. aeruginosa ATCC 27853* and *A. baumannii* species for Gram-negative bacteria; and ii) *S. aureus ATCC 25923, L. monocytogenes ATTC 7644* and *B. cereus ATCC 11778* species for Gram-positive bacteria.

As mentioned above, in this study, we have an important biological question. The whole study aims to answer this biological question via developing a specific classification model for AMP prediction, separately for Gram-positive and Gram-negative datasets. For this reason, we created a new AMP prediction dataset from publicly available DBAASP dataset by filtering for specific values (as shown in Figure 2.1 and as explained in detail in the Data Preprocessing section). In this study, we have only focused on linear cationic antimicrobial peptides. Among these peptides, we selected the peptides having antimicrobial activity against above-mentioned species. For each peptide, in order to define the activity against a group of bacteria (positive class label), we have utilized MIC values. Since we focus on the membrane-targeting AMPs, we have selected the peptides with lengths ranging from 20 to 50 aa. Here we focused on non-hemolytic peptides because in therapeutic applications the prediction of non-hemolytic peptides is reported as more important than the hemolytic peptides for the elimination of the detrimental effects of AMPs on the host. 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[34]. We carried out our classification procedure with the remaining peptides.

The antimicrobial activity of the peptides (AMP or Non-AMP class) was predicted separately for each bacterial group by using different physico-chemical properties. For each bacterial group, different models were developed using different classification algorithms. We have experimented with several machine learning methods including, Adaboost, Logitboost, Decision Tree, RF, SVM, and stacking classifiers using 100 fold MCCV. In our experiments using ten physico-chemical

features, we have observed that RF outperforms other classifiers. As summarized in Tables 3.1 and 3.2, 0.92 and 0.90 AUC values were obtained for Gram-negative and Gram-positive datasets, respectively. Also, in this research effort, for the first time, feature scoring and feature ranking were performed for Gram-positive and Gram-negative datasets separately, and the importance (score) of each feature in these two data sets was compared.

In order to understand the underlying structure of the data, we apply PCA on Gram-negative and Gram-positive datasets separately. The PCA results in Figure 3.5(A), 3.5(B), 3.5(C) and 3.5(D) 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 3.6(A), 3.6(B)). 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 3.3. 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 3.7(A), 3.7(B), 3.7(C), 3.7(D)). For this outlier eliminated dataset, all classification experiments have been repeated. As shown in Tables 3.4 and 3.5, 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 [113].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. In other words, the positive charge of the cationic AMPs

enables an electrostatic interaction with the negatively charged bacterial cell wall components [114]. The outer surface of the Gram-negative bacteria contains lipopolysaccharides (LPS), while Gram-positive bacteria contain acidic polysaccharides (teichoic acids). These structures confer a net negative charge to the surface of both Gram-positive and Gram-negative bacteria. In addition, the inner membrane of Gram-negative bacteria and the single membrane of Gram-positive bacteria are composed of negatively charged phospholipids. The net positive charge is the most conserved property of AMPs, making it possible to bind to the negatively charged outer surface of the bacteria [115]. Therefore, the net charge of AMPs has an essential role in the administration of peptide−membrane interactions resulting in the disruption of the membrane integrity [6]. The consistency of the results obtained with this computational study with the previous experimental results also supports the validity of the computational models created in this study. As mentioned above, 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 is defined as the pH at which the net charge of a protein/peptide is equal to zero. In other words, a protein has zero net charge at its isoelectric point. As the pH of the environment gets closer to the isoelectric point of the peptide, the net charge on the peptide surface gradually decreases and peptide-peptide interaction increases. Proteins have minimum solubility at or near their isoelectric point while protein solubility increases when pH moves away from pI. The pI is a feature that is closely related to the peptide charge and directly affects solubility. When the pH is equal to the pI of the peptide, the peptide loses its solubility and hereby its biological function [116]. Therefore, pI has an important role to exhibit the AMP's antimicrobial activity. pIs of the AMPs are generally at alkaline pH, and hereby maintain their activity at physiological pH. Therefore, the isoelectric point is another important feature that administers the antibacterial activity of AMPs [117-120]. Ahn *et al.*, reported that rather than the net charge, pI was a better parameter for predicting the antibacterial activity [121]. Our results are in accordance with the previous literature, supporting the feature

ranking analysis performed in this study. Along this line, the findings of this study support the idea that isoelectric point and the net charge are two main descriptors of antimicrobial peptides.

In our experiments, 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 LCAMPs 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 [122]. 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 [123,124]. In general, the increase of hydrophobicity promotes antimicrobial activity in peptides [125]. However, some studies demonstrated that an increase above a certain level in hydrophobicity leads to a decrease in antimicrobial activity [125]. The hydrophobic moment is defined as a quantitative measure of peptide amphipathicity [126]. 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 [124].

In addition to the physico-chemical descriptors, we have comparatively evaluated the effect of structure based and sequence based features on the classification performance. To this end, we have computed an extended set of features including amino acid composition, dipeptide composition, pseudo amino acid composition, CTD of physico-chemical properties, different autocorrelations, quasi-sequence-order descriptors, sequence order coupling number, separately for Gram-positive and Gram-negative datasets. We have compared the performances of the models which use only

the physico-chemical features with the models which use an extended set of features, separately for Gram-positive and Gram-negative datasets. As shown in Tables 3.6 and 3.7, the addition of an extended set of features did not improve performance metrics, even lowered the metrics slightly. For the Gram-positive dataset, when we applied feature selection on the extended set of features, we observed that all three selected features (isoelectric point, net charge, disordered conformation propensity) belong to the physico-chemical features category. Among 1507 different descriptors belonging to the structure based, linguistic based, sequence based, physico-chemical based classes in the extended dataset, the identification of the three physico-chemical descriptors as the top three scoring features was noteworthy. These three physico-chemical descriptors are computed from sequence information only. A similar observation is reported for miRNAs in[127-129]. In these studies, it is shown for miRNAs that the use of sequence information only (k-mer representation) is just enough for the prediction, while different studies use structure information, motif representation and k-mer for that purpose. Khabbaz *et al.*[65] imported AMPs with reported quantitative hemolytic activity from DBAASP and extracted 1541 features from physico-chemical, structure, sequence categories. They trained models using SVM classifier with radial basis function (RBF) and Polynomial kernels, Linear Support Vector Classifier (LSVC), RF, Naïve Bayes and kNN. In their experiments, the top three scoring features (aggregation propensity, polarity, charge density) among the 1541 features belong to the physico-chemical category. They have also applied feature selection and reported the performance metrics for 90 selected features among 1541 features. Among the selected 90 features, three features (aggregation propensity in vivo, charge density, isoelectric point) in top ten scoring features belong to the physico-chemical features. In their study, the performance metrics reported after feature selection (including 90 features) were very close to the performance metrics before apöplying feature selection (with 1541 features).

The models developed in this study are mainly based on physico-chemical features because as a continuation of this work, we are working on de-novo antimicrobial peptide design by using the datasets that we have compiled in this study, and by using the classification models that we have developed in this study, separately for Gram-positive and Gram-negative datasets. Before synthesizing de-novo peptides, we would like to computationally evaluate the antimicrobial activity of these candidate peptides using our classification model. During the wet-lab part of our future studies (when we synthesize those peptides), we need to know about those physico-chemical

features. As a future work, once we identify a promising candidate (a de-novo peptide), we plan to continue with the recombinant peptide production steps in wet-lab, and we plan to test the antimicrobial activity of this peptide against Gram positive or Gram negative bacteria in wet-lab.

Chapter 4

4.AMP-GSM: Prediction of Antimicrobial Peptides via a Grouping–Scoring–Modeling Approach

4.1 Motivation

Most of the machine learning models utilized for AMP prediction are based on the physico-chemical properties of antimicrobial peptides, such as net charge, isoelectric point, hydrophobic moment, penetration depth, tilt angle, etc. [130,131]. Apart from physico-chemical properties, there are also approaches that include sequence-based features, including amino acid composition, dipeptide composition, tripeptide composition, etc. [1332-135]. Additionally, there are studies that involve structure-based, linguistic-based features [11,64,65]. In the present study, for the antimicrobial peptide prediction task, we aim to develop a new computational approach that incorporates different types of AMP features and takes advantage of the characteristics of these groups. We attempt to show that one can increase the antimicrobial peptide prediction performance by using the selected groups of features that are identified with the proposed AMP-GSM method.

4.2 Proposed Model

We used three different datasets which is explained in Section 2.1. Also, we used different feature selection techniques that is explained in Section 2.4 for proposed model.

In our earlier studies, in order to improve classification performance, we proposed grouping-based feature elimination techniques, e.g., SVM RCE [87], SVM-RCE-R [88], and SVM-RCE-R-OPT [89]. Recently, we proposed numerous tools which incorporate biological information into the machine learning algorithm to accomplish feature selection or to choose groups of features. maTE [90], CogNet [91], miRcorrNet

[92], miRModuleNet [93], PriPath [129], 3Mint [130], and Integrating Gene Ontology-Based Grouping and Ranking [94] followed this strategy. This technique is known as the GSM approach [95], which is the primary motivation for the development of our proposed approach within this study.The workflow of the proposed approach, AMP-GSM, is presented in Figure 4.1. AMP-GSM includes three main components: grouping, scoring, and modeling.

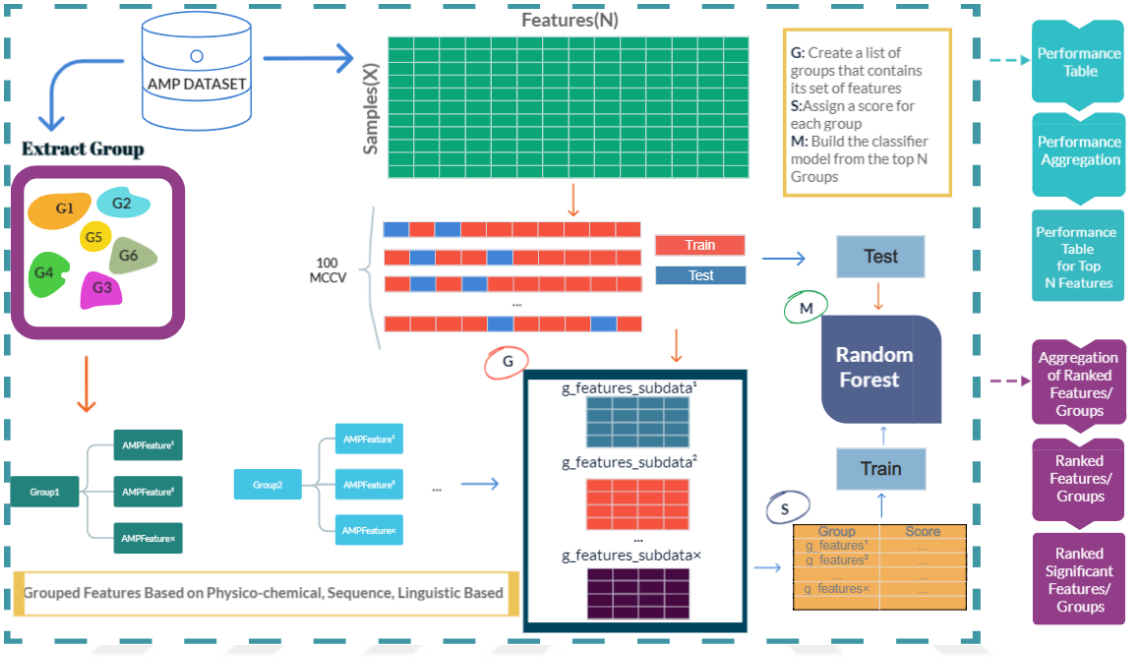


Figure 4.1 AMP-GSM workflow based on grouping, scoring, and modeling.

4.2.1 Grouping Peptides Based on Physico-Chemical, Sequence-Based, Structure-Based, and Linguistic-Based Features

The grouping component generates a list of groups where each group is composed of a feature set. An example output of this component is shown in Table 4.1. In our study, we have 12 groups, including physico-chemical, amino acid composition, dipeptide composition, physico-chemical composition, physico-chemical transition, physico-chemical distribution, normalized Moreau–Broto autocorrelation, Moran autocorrelation, Geary autocorrelation, sequence order coupling number, quasi-sequence order, and pseudo-amino acid composition. Each group has its own feature set. In Figure 4.2, the distribution of the features into different groups are shown.

Table 4.1 A list of feature groups and the features that are associated with them, based on [48,76].

Group	Feature Set	Number of Features

Physico-chemical	Sequence Length, Normalized Hydrophobic Moment, Normalized Hydrophobicity, Net Charge, Isoelectric Point, Penetration Depth...	10
Amino acid composition	A, C, E, D, G, F, I, H, K, M…	20
Dipeptide composition	GW, GV, GT, GS, GR, GQ, ME, MD, MG, MF, MA, MC, MM, ML, MN…	400
Physico-chemical composition	_NormalizedVDWVC2, _PolarizabilityC2, _PolarizabilityC3, _ChargeC1…	21
Physico-chemical transition	_SecondaryStrT13, _SecondaryStrT12, _HydrophobicityT23, _NormalizedVDWVT23, _ChargeT12…	21
Physico-chemical distribution	_NormalizedVDWVD1075, _PolarityD1075, _SecondaryStrD2075, _SolventAccessibilityD1100…	105
Normalized Moreau–Broto autocorrelation	MoreauBrotoAuto_ResidueASA28, MoreauBrotoAuto_ResidueVol30…	240
Moran autocorrelation	MoranAuto_FreeEnergy8, MoranAuto_FreeEnergy9, MoranAuto_Steric8 …	240
Geary autocorrelation	GearyAuto_Mutability23, GearyAuto_Mutability21, GearyAuto_FreeEnergy24, …	240
Sequence order coupling number	QSO26, QSO27, QSO_ex50, QSO_ex24, QSO_ex18, QSO_ex19…	60
Quasi-sequence-order	Taugrant23, taugrant24, tausw8, tausw9, tausw6, tausw7…	100
Pseudo-amino acid composition	PAAC34, PAAC35, APAAC20, PAAC38, PAAC39…	50

The groups created within the Grouping step are utilized to generate sub-datasets from the initial data. Each sub-data is composed of the properties belonging to the features within a particular group, retaining the original class labels of the peptides.

Let $Group^n$ represent the n-th group from the Gram-negative or Gram-positive AMP dataset, which includes x different features and is denoted as

$Group^n = \{AMPFeature^1, AMPFeature^3, \ldots, AMPFeature^x, \text{class label}\}$

and

Let $g\_features\_subdata^s$ represent the s-th group created by the grouping part of the AMP-GSM model, which includes a number of groups (i), denoted as

$g\_features\_subdata^s = \{Group^1, Group^3, \ldots, Group^i\}$

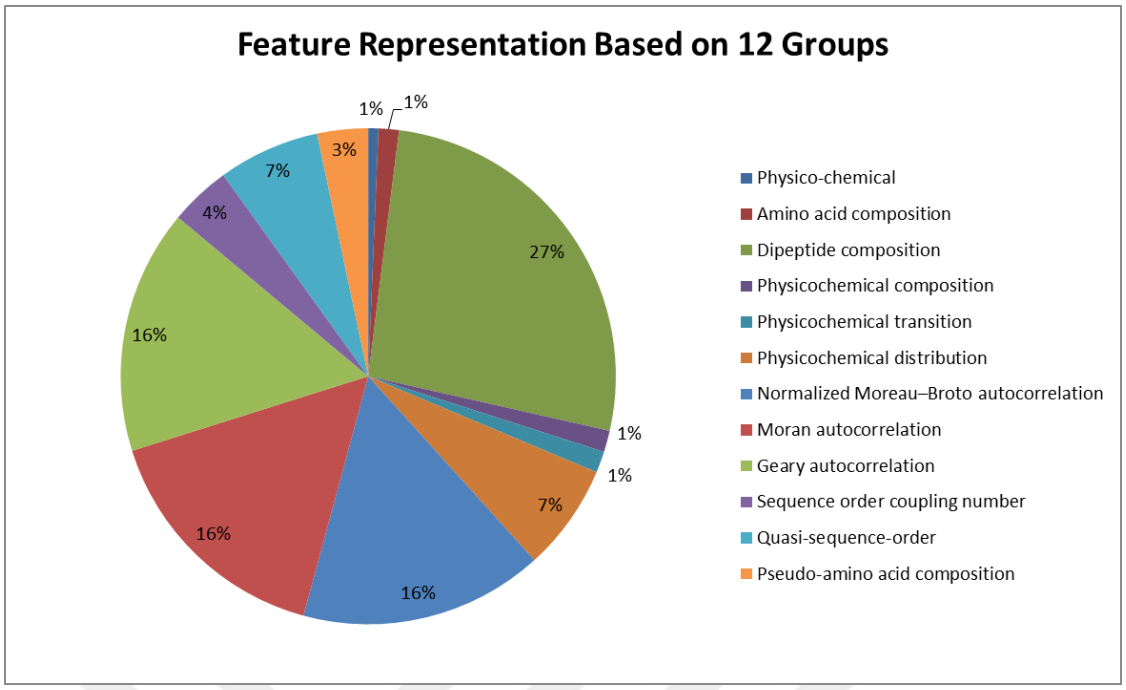


Figure 4.2 Feature representation based on different groups for antimicrobial peptides.

4.2.2 Scoring the Groups

The scoring component gives a score to each group that is created by the grouping component. At the end of this step, each group will have their own score. This score shows how well the group can distinguish between negative and positive classes. In order to determine this score, a 100-fold Monte Carlo cross validation (MCCV) procedure is used [106]. In our experiments, 90% of the data is used as the training set, and 10% is used as the test set.

The scoring component produces lists of AMP groups and the features linked to them that are slightly different after each iteration. Consequently, a prioritization strategy needs to be applied to those lists. We applied rank aggregation techniques similar to those proposed in miRcorrNet [92]. In this regard, the RobustRankAggreg R package [138] was integrated into our workflow. Each element in the aggregated list was given a *p*-value by the RobustRankAggreg, which indicates how well it was ranked relative to the predicted value. The list of the groups that are ordered by scores is the final output of the scoring component.

4.2.3 Modeling Component

After defining the informative groups of features, the model can then be tested on the group with the highest ranking, or cumulatively on the top j groups. In our experiments we decided to use 10 for j. In other words, while maintaining the original labels, we build sub-data using the features related to the top 10 group categories. Applying machine learning to this new subset of data results in the creation of the model, which is then tested using the test set. The model built using the top-ranked groups is evaluated as the final part of our method. We used the Random Forest (RF) Classifier for the modeling part. We split the data into 90% training and 10% testing. We applied 100-fold MCCV for evaluation.

The Konstanz Information Miner (KNIME) platform was used to implement all three components of our method [88].

4.3 Results

We tested AMP-GSM for Dataset 1, which includes a Gram-negative and a Gram-positive dataset, as mentioned above; details are presented in dataset and dataset preprocessing section(Section 2.1). Furthermore, we applied different feature selection methods on those datasets. Additionally, we ran our approach on other existing datasets (explained in Section 2.1) to compare our results with other methods.

4.3.1 Performance Evaluation of AMP-GSM on the Gram-Negative Dataset in Dataset 1

The Gram-negative dataset includes 231 positive (AMP) and 114 negative (non-AMP) samples. Average 100-fold MCCV performance metrics of AMP-GSM for the combined top 10 groups for the Gram-negative dataset are shown in Table 4.2. The first column represents the number of groups, and the second column shows the number of cumulative features. The performance of the top-ranked group is given in the last row, where number of groups = 1. Using 10 features on average, we obtained 95% accuracy and 99% AUC. The features from the initial top-ranked group and the second-highest scoring group are combined. The tenth row of Table 4.2, where number of groups = 2 displays the performance metrics derived for the top two groups cumulatively. AMP-GSM reports the cumulative performance metrics for the top 10 groups.

Table 4.2 Performance results of the AMP-GSM approach on the Gram-negative dataset of Dataset 1 (for 12 groups and 100-fold MCCV).

#Groups	#Features (Mean)	Acc (Mean)	Sn (Mean)	Sp (Mean)	F-Measure (Mean)	AUC (Mean)	Pr (Mean)
10	1039.99	0.92	0.91	0.93	0.91	0.98	0.92
9	807.19	0.93	0.93	0.92	0.92	0.98	0.91
8	714.01	0.94	0.94	0.93	0.93	0.98	**0.92**
7	571.08	0.92	0.93	0.92	0.91	0.98	0.90
6	439.45	0.92	0.93	0.91	0.91	0.98	0.90
5	306.47	0.92	0.94	0.90	0.91	0.98	0.89
4	190.91	0.92	0.94	0.91	0.91	0.98	0.90
3	121.25	0.93	0.95	0.91	0.92	0.98	0.90
2	60.6	0.93	0.95	0.92	0.93	0.99	0.91
1	10	**0.95**	**0.98**	0.93	**0.95**	**0.99**	0.92

* The average number of features is shown in the column #Features. Firstly, the features from the top group are used to create a model, which is then tested using the testing part of the data. Secondly, the top one and two groups are used to create a model, which is subsequently tested. Similarly, the model is created using the features from the top 10 groups, and it is then tested. Acc: Accuracy, Sn: Sensitivity, Sp: Specificity, AUC: Area Under Curve, Pr: Precision.

4.3.2 Performance Evaluation of AMP-GSM on the Gram-Positive Dataset in Dataset 1

The Gram-positive dataset in Dataset 1 includes 165 positively labeled (AMP) and 194 negatively labeled (non-AMP) samples. Average 100-fold MCCV performance metrics of AMP-GSM for the combined top 10 groups for the Gram-positive dataset are shown in Table 4.3. The first column represents the number of groups, and the second column shows the number of cumulative features. The performance of the top-ranked group is given in the last row, where number of groups = 1. Using 10 features on average, we obtained 92% for accuracy and 98% for AUC metric. The features from the initial top-ranked group and the second-highest scoring group are combined. The tenth row of Table 4.3 where number of groups = 2, displays the performance metrics derived for the top 2 groups cumulatively. AMP-GSM reports the cumulative performance metrics for the top 10 groups.

Table 4.3 Performance results of the AMP-GSM approach on the Gram-positive dataset of Dataset 1 (for 12 groups and 100-fold MCCV).

#Groups	#Features (Mean)	Acc (Mean)	Sn (Mean)	Sp (Mean)	F-Measure (Mean)	AUC (Mean)	Pr (Mean)

10	1026.75	0.88	0.69	**0.96**	0.77	0.95	**0.91**
9	795.75	0.88	0.72	0.96	0.79	0.95	0.90
8	657.26	0.87	0.70	0.95	0.77	0.95	0.89
7	526.35	0.88	0.73	0.94	0.78	0.95	0.88
6	351.87	0.88	0.75	0.94	0.80	0.95	0.87
5	226.48	0.89	0.78	0.94	0.82	0.96	0.88
4	160.75	0.89	0.77	0.94	0.81	0.95	0.87
3	103.51	0.90	0.80	0.95	0.83	0.96	0.89
2	44.28	0.91	0.82	0.95	0.85	0.96	0.90
1	10	**0.92**	**0.89**	0.93	**0.87**	**0.98**	0.87

* The average number of features is shown in the column #Features. Firstly, the features from the first top group are used to create a model, which is then tested using the testing part of the data. Secondly, the top one and two groups are used to create a model, which is subsequently tested. Similarly, the model is created using the features from the top 10 groups for j = 10, and is then tested. Acc: Accuracy, Sn: Sensitivity, Sp: Specificity, AUC: Area Under Curve, Pr: Precision.

It is worth noting that both for the Gram-negative and Gram-positive datasets of Dataset 1, all members of the top scoring group originated from the physico-chemical group, and this group showed the highest performance results. It was observed that the scoring made using only physico-chemical features obtained much better results than the groups formed by adding other features.

We re-ran our method by removing this physico-chemical group from the grouping to demonstrate how important physico-chemical properties are in antimicrobial peptide prediction. As seen in Table 4.4, when the physico-chemical properties are extracted, on the Gram-negative dataset, a single group generated by AMP-GSM includes 38.27 features (averaged over 100-fold MCCV iterations), and this group achieves an accuracy of only 87% and an AUC value of 93%. However, when physico-chemical properties are included, this rate was 95% for accuracy and 99% for AUC (as shown in Table 4.2). Likewise, for the Gram-positive dataset of Dataset 1, when the physico-chemical properties are removed, 82% accuracy and 90% AUC were obtained (shown in Table 4.5) by using a single group, including 37.67 features (averaged over 100-fold MCCV iterations), while 92% accuracy and 98% AUC values were obtained when physico-chemical properties were included in the analysis (shown in Table 4.3).

Table 4.4 Performance results of AMP-GSM approach for the Gram-negative dataset of Dataset 1 without using physico-chemical properties (for 11 groups and 100-fold MCCV).

#Groups	#Features (Mean)	Acc (Mean)	Sn (Mean)	Sp (Mean)	F-Measure (Mean)	AUC (Mean)	Pr (Mean)
3	169.73	0.88	0.87	0.89	0.86	0.96	0.87
2	100.65	0.89	**0.89**	0.89	0.87	0.95	0.86
1	38.27	0.87	0.86	0.88	0.85	0.93	0.85

* Acc: Accuracy, Sn: Sensitivity, Sp: Specificity, AUC: Area Under Curve, Pr: Precision.

Table 4.5 Performance results of the AMP-GSM approach for the Gram-positive dataset of Dataset 1 without using physico-chemical properties (for 11 groups and 100-fold MCCV).

#Groups	#Features (Mean)	Acc (Mean)	Sn (Mean)	Sp (Mean)	F-Measure (Mean)	AUC (Mean)	Pr (Mean)
3	135.41	0.85	0.69	0.92	0.74	0.92	0.81
2	85.27	0.84	0.67	0.91	0.72	0.91	0.79
1	37.67	0.82	0.63	0.91	0.68	0.90	0.78

* Acc: Accuracy, Sn: Sensitivity, Sp: Specificity, AUC: Area Under Curve, Pr: Precision.

4.3.3 Ranking of the Groups

We ranked the groups by the RobustRankAggreg method, applied on the Gram-negative and Gram-positive datasets of Dataset 1. The results are presented in Appendix Table A1.

4.3.4 Comparative Evaluation of the Proposed Method with Other Feature Selection Methods and Classifiers

We have a total of 1508 features for the sequences in Dataset 1. Feature selection techniques attempted to eliminate redundant and unimportant features. As explained in the Materials and Methods section, we experimented with the use of mRMR, IG, XGBoost, and CMIM feature selection methods for the AMP prediction problem. Additionally, the effectiveness of different classification methods, such as RF, SVM, LogitBoost, Decision Tree, and AdaBoost, was evaluated. Since AMP-GSM selected 10 features, for the remaining feature selection methods, we obtained results using their top 10 features. The top 10 features chosen by the four above-mentioned approaches were used to evaluate the effectiveness of numerous classifiers using different metrics. Table 4.6 shows, compared with other feature selection methods, how

the XGBoost and IG techniques enhanced the accuracy, sensitivity, specificity, F1 measure, and AUC values of the tested classifiers, applied on the Gram-negative dataset of Dataset 1. With the same data, it was possible to deduce that the mRMR and CMIM feature selection methods resulted in a low accuracy and a high sensitivity, as well as signs of poor fitting across evaluated models. On the other hand, on the Gram-negative dataset of Dataset 1, AMP-GSM performed better than all tested feature selection methods, all tested classifiers, in terms of the area under curve performance evaluation metric (as shown in Table 4.6, Figure 4.3).

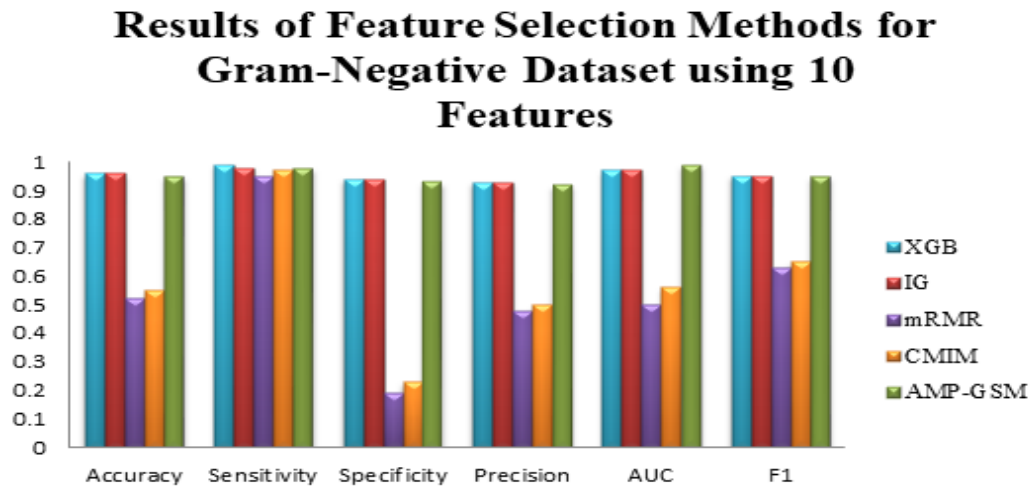


Figure 4.3 Performance evaluation of different feature selection techniques and the AMP-GSM approach on the Gram-negative dataset of Dataset 1 using 10 features and 100-fold MCCV.

Table 4.6 Performance metrics of different feature selection techniques with 10 features on the Gram-negative dataset of Dataset 1, using 100-fold MCCV.

Results for the Gram-Negative Dataset (10 Features, 100-fold MCCV)							
ML Method	**FS Method**	**Accuracy**	**Sensitivity (Recall)**	**Specificity**	**Precision**	**AUC**	**F1**
LogitBoost	XGB	**0.96 ± 0.03**	**0.99 ± 0.03**	**0.94 ± 0.06**	**0.93 ± 0.07**	0.97 ± 0.02	**0.95 ± 0.04**
LogitBoost	IG	0.96 ± 0.04	0.98 ± 0.03	0.94 ± 0.06	0.93 ± 0.07	0.97 ± 0.02	0.95 ± 0.04
Adaboost	MRMR	0.52 ± 0.09	0.95 ± 0.09	0.19 ± 0.22	0.48 ± 0.07	0.50 ± 0.13	0.63 ± 0.04
RF	CMIM	0.55 ± 0.11	0.97 ± 0.08	0.23 ± 0.23	0.50 ± 0.09	0.56 ± 0.14	0.65 ± 0.05
RF	AMP-GSM	0.95 ± 0.04	0.98 ± 0.03	0.93 ± 0.07	0.92 ± 0.08	**0.99 ± 0.006**	**0.95 ± 0.05**

* ML: Machine Learning, FS: Feature Selection, AUC: Area Under Curve.

Table 4.7 shows that compared with other feature selection methods, the IG technique enhanced the accuracy, sensitivity, specificity, F1 measure, and AUC values of the tested classifiers, applied on the Gram-positive dataset of Dataset 1. Although not

as good as IG, XGB provided a good estimation result on the overall. The mRMR and CMIM feature selection approaches resulted in low accuracy and high recall values, as well as indications of poor fitting across examined models on the same data. On the other hand, on the Gram-positive dataset of Dataset 1, AMP-GSM performed better than all tested feature selection methods, all tested classifiers in terms of different performance evaluation metrics except sensitivity score (as shown in Table 4.7, Figure 4.4).

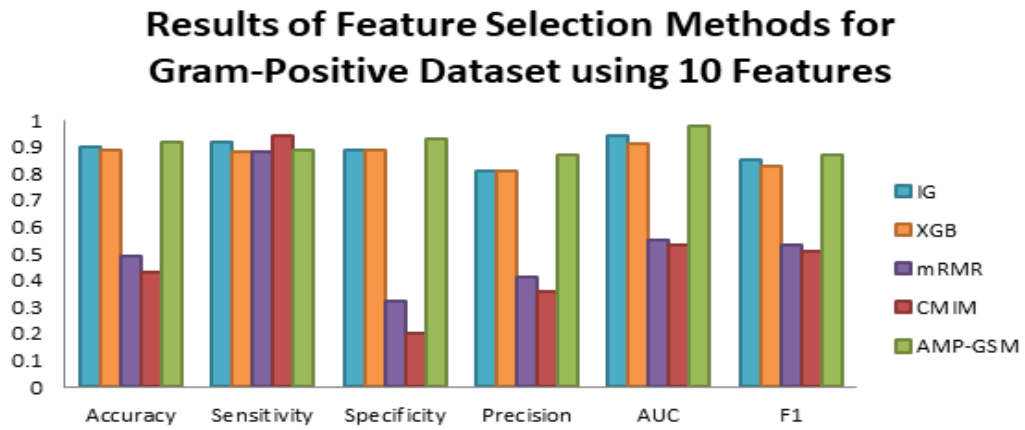


Figure 4.4 Performance evaluation of different feature selection techniques and the AMP-GSM approach on the Gram-positive dataset of Dataset 1 using 10 features and 100-fold MCCV.

Table 4.7 Performance metrics of different feature selection techniques with 10 features on the Gram-positive dataset of Dataset 1, using 100-fold MCCV.

Results for the Gram-Positive Dataset (10 Features, 100-fold MCCV)							
ML Method	**FS Method**	**Accuracy**	**Sensitivity (Recall)**	**Specificity**	**Precision**	**AUC**	**F1**
RF	IG	0.90 ± 0.04	0.92 ± 0.09	0.89 ± 0.07	0.81 ± 0.11	0.94 ± 0.03	0.85 ± 0.06
RF	XGB	0.89 ± 0.05	0.88 ± 0.10	0.89 ± 0.08	0.81 ± 0.12	0.91 ± 0.05	0.83 ± 0.07
RF	MRMR	0.49 ± 0.17	0.88 ± 0.15	0.32 ± 0.31	0.41 ± 0.14	0.55 ± 0.14	0.53 ± 0.07
RF	CMIM	0.43 ± 0.14	**0.94 ± 0.11**	0.20 ± 0.24	0.36 ± 0.08	0.53 ± 0.11	0.51 ± 0.05
RF	AMP-GSM	**0.92 ± 0.04**	0.89 ± 0.10	**0.93 ± 0.05**	**0.87 ± 0.09**	**0.98 ± 0.02**	**0.87 ± 0.06**

* ML: Machine Learning, FS: Feature Selection, AUC: Area Under Curve.

In Table 4.8, 10 features obtained from the feature selection method (IG-RF pairwise for Gram-positive and XGB-Logitboost pairwise for Gram-negative) that gave the best results and the 10 most important features selected by AMP-GSM are compared. While all of the 10 features identified by the AMP-GSM method belong to the physico-chemical group, it can be observed from Table 4.8 that the features detected

by the feature selection methods belong to different groups for both Gram-negative and Gram-positive datasets.

Table 4.8 Comparison of the most important 10 features found in the first two groups in the AMP-GSM method with the 10 most informative features identified by the feature selection methods for Gram-negative and Gram-positive datasets.

Gram-Negative Dataset			
FS/CLSF* Method	**Features Identified by FS Methods**	**Features Identified by AMP-GSM**	**Common Features Between the Top Features of the FS Method and AMP-GSM**
XGB/Logitboost	Net Charge MoranAuto_AvFlexibility8 Tilt Angle Normalized Hydrophobic Moment MoranAuto_Hydrophobicity15 MoranAuto_ResidueVol5 _ChargeC1 QSO_ex29 Isoelectric Point tausw2	SequenceLength Normalized Hydrophobic Moment Normalized Hydrophobicity Net Charge Isoelectric Point Penetration Depth Tilt Angle Disordered Conformation Propensity Linear Moment Propensity to in vitro Aggregation	Net Charge Tilt Angle Normalized Hydrophobic Moment Isoelectric Point
Gram-Positive Dataset			
IG/RF	Isoelectric Point Net Charge Disordered Conformation Propensity Normalized Hydrophobicity _ChargeC1 _PolarityC3 tausw9 taugrant6 _PolarityT13 tausw6	SequenceLength Normalized Hydrophobic Moment Normalized Hydrophobicity Net Charge Isoelectric Point Penetration Depth Tilt Angle Disordered Conformation Propensity Linear Moment Propensity to in vitro Aggregation	Normalized Hydrophobicity Net Charge Isoelectric Point Disordered Conformation Propensity

* FS: Feature Selection, CLSF: Classification.

4.3.5 Testing AMP-GSM on Different Datasets, Comparative Evaluation with Existing Approaches

Veltri et al. proposed a model consisting of a deep neural network (DNN) structure including convolution and LSTM layers [56]. They tested their proposed DNN on Dataset 2, the further details can be found in [56].

Table 4.9 compares the performance metrics of one of the DNN models proposed in [56] with the AMP-GSM model and with other feature selection methods for Dataset 2. The methods being compared are listed in column 1 of Table 4.9, along

with the five performance metrics listed in columns 3 through column 7. The results in bold in Table 4.9 represent the best results for a particular metric. According to Table 4.9, the AMP-GSM model performs the best in terms of accuracy, sensitivity, F1 measure and AUC metrics.

Table 4.9 Performance evaluation of AMP-GSM with a DNN model for Dataset 2 [56].

Method	Evaluation	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC (%)	F1 Measure
DNN Model	10-fold CV	88.81 (±3.53)	**94.21 (±2.68)**	91.51 (±0.89)	96.58 (±0.66)	-
CMIM-DT	10-fold MCCV	51.34 ± 0.17	51.40 ± 0.17	51.37 ± 0.03	51.37 ± 0.03	50.02 ± 0.09
IG-RF	10-fold MCCV	88.70 ± 0.02	91.40 ± 0.02	90.05 ± 0.01	96.36 ± 0.007	89.91 ± 0.01
mRMR-RF	10-fold MCCV	33.70 ± 0.18	67.41 ± 0.21	50.56 ± 0.03	50.80 ± 0.05	37.80 ± 0.14
AMP-GSM Model	10-fold MCCV	**91.01 (±0.23)**	92.97(±0.03)	**91.71 (±0.13)**	**97.07 (±0.06)**	**91.59 (±0.15)**

Manavalan et al. proposed a model consisting of a random forest classifier and feature selection methods [61]. They used amino acid composition, amino acid index, dipeptide composition, physico-chemical properties, and distribution of amino acid patterns as peptide features. They compare their results with commonly used machine learning models, such as SVM, k-NN, and extremely randomized trees (ERT).

Table 4.10 compares the performance metrics of the method proposed in [61] with AMP-GSM and other feature selection methods for Dataset 3. The methods being compared are listed in column 1 of Table 4.10 along with the four performance metrics listed in columns 3 through column 6. The results in bold in Table 4.10 indicate the best results for a particular metric. AMP-GSM considerably outperforms CMIM, IG, and mRMR feature selection methods for predicting AIPs. There is a significant difference for all performance metrics.

Table 4.10 Performance evaluation of AMP-GSM with other traditional feature selection and classification models for Dataset 3 [61].

Method	Evaluation Set	Accuracy	Sensitivity	Specificity	AUC
AIPpred	5-Fold CV	0.73	0.75	0.71	0.80
ERT	5-Fold CV	0.73	0.73	0.72	0.79

SVM	5-Fold CV	0.65	0.64	0.67	0.70
k-NN	5-Fold CV	0.64	0.51	0.77	0.69
CMIM-LogitBoost	5-Fold MCCV	0.54	0.67	0.40	0.55
IG-AdaBoost	5-Fold MCCV	0.69	0.66	0.72	0.73
mRMR-LogitBoost	5-Fold MCCV	0.50	0.79	0.20	0.50
AMP-GSM Model	5-Fold MCCV	**0.99**	**1**	**0.99**	**1**

4.4 Discussions

In this study, we propose AMP-GSM, a novel approach that is built on the grouping and ranking of peptide features. The method relies on grouping the features according to their biological characteristics, and then scoring those groups according to their importance in terms of distinguishing antimicrobial peptides from non-antimicrobial peptides. Traditional methods often use the properties of antimicrobial peptides together rather than grouping them. On the other hand, feature selection methods select the features that they identify as important, and then develop the classification models using the selected features. However, such a selection is not a group-based selection. Based on all the attributes, traditional feature selection methods select the most important ones. Studies in this area are mostly aimed at classification by taking feature groups individually or collectively, or by using a set of features selected by traditional feature selection methods [30,139,140].

In this study, structure-based, sequence-based, and physico-chemical features were grouped and their effects on classification performance were evaluated. We analyzed a comprehensive set of features, including amino acid composition, dipeptide composition, pseudo amino acid composition, CTD of physico-chemical properties, various autocorrelations, quasi-sequence-order descriptors, and sequence order coupling number. These features are generated for each peptide within the Gram-positive and Gram-negative datasets separately. Separately for the Gram-positive and Gram-negative datasets, we compared the performances of the models that apply the proposed AMP-GSM technique and alternative feature selection strategies. As shown in Figures 4.3 and 4.4, AMP-GSM resulted in higher AUC values on both Gram-negative and Gram-positive dataset.

An AMP prediction system has a very high number of potential input features, and the decisions made regarding which features to use for antimicrobial prediction greatly affect prediction performance in terms of accuracy and AUC. Finding novel antimicrobial descriptors that may be connected to physico-chemical properties could reduce the wide accuracy range of the prediction algorithms, and aid in identifying the real significance of characteristics related to antimicrobial activities.

Ten of the 1508 factors displayed statistically significant variations in positive and negative datasets separately. Compared with the known feature selection algorithms, these ten features are effective antimicrobial peptide descriptors that produce higher accuracy when used with the AMP-GSM approach. As seen in Table 4.8, all 10 features belong to the physico-chemical group. Additionally, when we removed the physico-chemical group from the dataset and run our approach, it was observed that accuracy and AUC values significantly decreased for both Gram-negative and Gram-positive datasets.

We also used two other benchmark datasets in order to make a comparison between different approaches. In Dataset 2, there are 1778 AMPs and 1778 non-AMPs. Using the whole dataset with 10-fold MCCV, for some performance metrics, AMP-GSM outperformed the DNN model with LSTM and convolutional layers, as proposed in [56]. As seen in Table 4.9, we obtained higher performance metrics for sensitivity, accuracy, and AUC compared to the DNN model with LSTM and convolutional layers [56]. Another dataset that we analyzed (Dataset 3) was provided by Manavalan et al. in [61]. This dataset consists of anti-inflammatory peptides (AIP). It includes 1258 AIPs and 1887 non-AIPs. Their model consists of a feature selection part with RF. Using Dataset3, we obtained higher performance metrics (99% accuracy, 100% AUC) compared with their method and traditional machine learning approaches, such as SVM and k-NN (as seen in Table 4.10). Hence, we can conclude that the novel approach developed in this study can be used to predict not only antimicrobial peptides, but also anti-inflammatory peptides by considering group characteristics.

Our technique performs well and provides better categorization of AMPs based on different types of information (physico-chemical, sequence-based, etc.). However, AMPs can be hazardous and inefficient as a medicine, which is undesirable. Studies on the synthesis and modification of AMPs have shown that even small modifications can impact how well they work. This approach does not take into account the functional traits of AMPs, but instead, it can only identify AMPs. It is possible to undertake

additional studies in accordance with the roles played by AMPs, which will improve our comprehension of their method of action and our ability to forecast their behaviors.

Another issue regarding the design of AMPs is that toxicity, stability, and bacterial resistance must all be addressed concurrently in the rational design of AMP-based therapeutics [141]. To achieve this, it is essential to determine the key attributes that a peptide contains in order to be effective against various bacterial species. To rationally develop antimicrobial peptides that target certain bacteria, this study offers a feature selection method based on grouping that is specific to bacteria. It will be crucial to test our study using larger datasets active against bacteria.

Chapter 5

5.Antimicrobial Peptide Design Using Match Score Motif Representation

5.1 Motivation

In this study, a machine learning model with motif matching score has built to create novel AMP sequences that may have antibacterial activity against both Gram-positive and Gram-negative bacteria individually. Different classification models were trained and used to generate datasets of novel sequences that were classified as AMP or non-AMP. Most new sequences are validated with the "DBAASP:strain-specific antibacterial prediction based on machine learning approaches and data on AMP sequences" tool . The study presented in this paper advances the field of computational research by making it easier to create or modify AMPs in wet lab settings.

5.2 Model Construction

We divided our model into three steps. For the first step, we extracted motifs for Gram-negative and Gram-positive dataset separately from MEME motif finding web server [142]. The negative(non-AMP) and positive(AMP) sequences in each of our datasets were given separately to the MEME motif finding program. We selected parameters as 50 motifs of each, varying in length from 5 to 12 (100 motifs in total). For each motif, match score is calculated between sequence and motif and given to the machine learning algorithms as features. We divide our dataset 90% as training and 10% as testing. We used 100 Monte Carlo Cross Validation (MCCV) [106]. Feature importances libraries were built using Scikit-learn [143]. According to feature importances' results, the first 3 most important features are selected and extracted from regular expression mode and we get a new feature set for step 2. We expand the regular expression(with new matching scores) and give it as features to the dataset (totally 368 features for Gram-positive and 68 features for Gram-negative). Then, we run classification algorithms again and get new feature importances' scores for our third step. We have selected the five most important features according to the results. For

creating new peptides and adding old dataset, triple combinations of the first 5 motifs (motifs from AMPs) were generated, resulting in 60 new peptides for each data set. New peptides created were added to the existing dataset and 10 physicochemical features (mentioned above in Section 2.2) were generated for each sequence. The generated dataset was given as input to the classification algorithms and the prediction score (being AMP or non-AMP) of each sequence was calculated. Those most likely to become AMPs are usable for further experimental processes.

Our approach is put into practice using the Konstanz Information Miner (KNIME) platform [85].

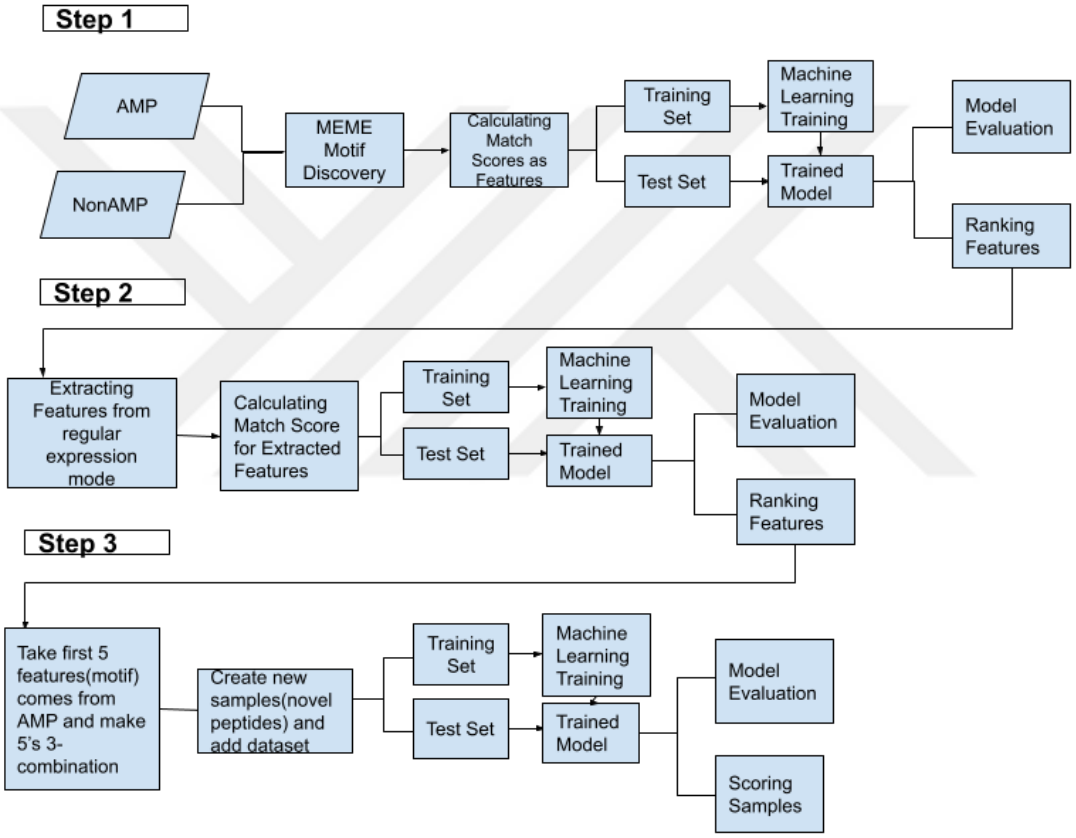


Figure 5.1 Model Construction.

5.2.1 Motif Parameters

In this study, a small sequence of amino acids is called a sequence motif. Finding such small sequences within a larger area of sequences is the process of motif discovery. In our work, we use the MEME suite web server for motif finding (https://meme-suite.org/meme/tools/meme). The algorithm is based on and operates by repeatedly looking for input sequences that contain ungapped sequence motifs. The outcomes are presented by MEME web server as regular expressions.Alternative

aminoacids are shown in bracketed sequences; without brackets, only the supplied aminoacid is abundantly present in all compiled sequences that make up the motif. Sequence logos provide more visual depictions of these motifs.

The negative and positive sequences in each of our datasets were given separately to the MEME motif finding program [142]. We selected parameters as 50 motifs of each, varying in length from 5 to 12 (100 motifs in total). An example motif representation is shown in Table 5.1. First column represents motifs belonging to AMP and second column corresponds motifs belonging to NonAMP.

Table 5.1 Motif representation that is found by MEME motif program for both AMP and NonAMP sequences.

AMP	NonAMP
L[KR][KR][FL]G[KR]K[VI]KKAX	GKEFK
HL[LR]R[IP]	IW[DS][AS]I
I[GV]Q[KR]IKDF[LF][RQ][NK]	[IL]N[QY][AN]W
TRGRW	LNV[CN]R
R[IK]H[KR]H	F[CY][ST]YI
…	…

The match score of these motifs in the sequences are calculated. Matching score values obtained from each motif were calculated as follows. The matching score, which gives the highest score by sliding one row on the series as long as the length of each motif here, was observed.

•Match Score=Number of matching aminoacid/The length of motif

•For example for below table: Match Score=5/12=0.41

Table 5.2 Example of match score between a motif and a part of a sequence.

Motif	**Sequence**													
	Alignment													
Reg.Exp.		[FV]	[LK]	[HD]	[ST]	[AL]	[KG]	K	F	[GA]	K	[GA]	F	
Seq.Window	…	F	K	G	A	S	K	V	F	P	A	V	F	…
Match Score		1	1	0	0	0	1	0	1	0	0	0	1	

Table 5.3 An example representation of how to use motif match scores as features.

Sequence	**Motif1**	**Motif2**	**Motif3**	**…**	**class**

Sequence1	0.12	0.58	1.0		pos
Sequence2	0.24	0.42	0.35	…	pos
Sequence3	0.15	1.0	0.65		neg
…					

5.3 Results

5.3.1 Classification Results for Step 1

The classification algorithms' results for the first step are shown in Table 5.4 for Gram–negative dataset and in Table 5.5 for Gram-positive dataset. When we use 100 motif match score as features we get 93% accuracy and 95% Area Under Curve(AUC) using Random Forest (RF) algorithm for Gram-negative dataset and 82% accuracy and 84% AUC using LogitBoost algorithm for Gram-positive dataset.

Table 5.4 Classification Results for Gram-negative dataset (Step 1).

Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1
Adaboost	0.91±0.05	0.89±0.10	0.93±0.08	0.92±0.09	0.93±0.05	0.90±0.06
DT	0.85±0.07	0.81±0.11	0.88±0.10	0.85±0.11	0.84±0.07	0.82±0.08
LogitBoost	0.93±0.06	0.89±0.10	0.96±0.08	0.95±0.08	0.94±0.05	0.91±0.07
RF	0.93±0.05	0.92±0.09	0.95±0.07	0.94±0.07	0.95±0.05	0.92±0.06
SVM_opt	0.92±0.05	0.89±0.09	0.94±0.06	0.93±0.08	0.94±0.05	0.91±0.06
Stack_SVM_Kmeans	0.92±0.05	0.87±0.10	0.96±0.06	0.95±0.07	0.95±0.04	0.91±0.06
Stack_SVM_Logitboost	0.92±0.05	0.89±0.09	0.94±0.06	0.93±0.08	0.94±0.05	0.91±0.06

Table 5.5 Classification Results for Gram-positive dataset (Step 1).

Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1
Adaboost	0.80±0.09	0.79±0.13	0.81±0.14	0.69±0.14	0.81±0.09	0.72±0.09
DT	0.55±0.21	0.76±0.21	0.46±0.40	0.47±0.16	0.63±0.09	0.53±0.08
LogitBoost	0.82±0.08	0.83±0.11	0.82±0.14	0.72±0.15	0.84±0.07	0.75±0.08
RF	0.80±0.09	0.82±0.12	0.79±0.15	0.69±0.14	0.84±0.07	0.73±0.08

SVM_opt	0.74±0.10	0.83±0.13	0.71±0.18	0.60±0.14	0.77±0.09	0.68±0.08
Stack_SVM_Kmeans	0.75±0.10	0.72±0.17	0.76±0.20	0.66±0.19	0.78±0.07	0.65±0.08
Stack_SVM_Logitboost	0.74±0.10	0.83±0.13	0.71±0.18	0.60±0.14	0.77±0.09	0.68±0.08

In our approach each algorithm has its own feature rank score as a result. Since random forest for Gram-negative and logitboost for Gram-positive give the best performance results, we have based the results of these two algorithms on their feature importance. Table 5.6 shows the motifs which are the most important for the Gram-positive dataset, while Table 5.7 shows the motifs which are the most important for the Gram-negative dataset.

Table 5.6 Results for ranked first 3 features of Gram-positive dataset.

Motif Number	Model	Score	Motif
51	LogitBoost	0.87	R[AV][GV]LQ[FW]P[VI]G[RK][VIL][HLV]
71	LogitBoost	0.62	W[AR][AG][HN][GK][SV]V[HS]RY
59	LogitBoost	0.62	C[KR][GR]W[LQ][CW]

Table 5.7 Results for ranked first 3 features of Gram-negative dataset.

Motif Number	Model	Score	Motif
53	RF	1	L[KR][KR][FL]G[KR]K[VI]KKAX
100	RF	0.09	HL[LR]R[IP]
51	RF	0.09	I[GV]Q[KR]IKDF[LF][RQ][NK]

5.3.2 Results for Step 2

In Step 2, we expand the three most important motifs with regular expression(with new matching scores) which we get feature ranking part of Step 1 for both datasets and give it as features to the datasets (368 features for gram-positive and 68 features for gram-negative). We rerun workflow with new features for both datasets. Our aim here was to find out which motifs are more important in prediction rather than getting a good prediction result. We get 68% accuracy for Gram-positive dataset (shown in Table 5.8) and 78% accuracy for Gram-negative dataset (shown in Table 5.9).

Table 5.8 Classification Results for Gram-positive dataset (Step 2).

Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1
Adaboost	0.67±0.12	0.77±0.15	0.62±0.22	0.52±0.14	0.67±0.11	0.60±0.08
DT	0.51±0.20	0.79±0.21	0.38±0.39	0.43±0.14	0.61±0.09	0.52±0.07
LogitBoost	0.68±0.12	0.77±0.15	0.64±0.23	0.54±0.15	0.69±0.09	0.61±0.07
RF	0.66±0.12	0.83±0.14	0.58±0.22	0.52±0.15	0.70±0.08	0.61±0.07
SVM_opt	0.55±0.16	0.83±0.17	0.42±0.30	0.44±0.14	0.57±0.12	0.54±0.07
Stack_SVM_Kmeans	0.48±0.18	0.85±0.19	0.32±0.34	0.41±0.16	0.57±0.09	0.51±0.05
Stack_SVM_Logitboost	0.55±0.16	0.83±0.17	0.42±0.30	0.44±0.14	0.57±0.12	0.54±0.07

Table 5.9 Classification Results for Gram-negative dataset (Step 2).

Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1
Adaboost	0.73±0.11	0.81±0.13	0.67±0.24	0.69±0.14	0.73±0.10	0.72±0.08
DT	0.69±0.16	0.79±0.15	0.61±0.35	0.67±0.17	0.72±0.10	0.70±0.09
LogitBoost	0.74±0.10	0.82±0.13	0.68±0.22	0.70±0.14	0.76±0.09	0.74±0.07
RF	0.78±0.12	0.87±0.10	0.71±0.17	0.72±0.12	0.81±0.07	0.78±0.06
SVM_opt	0.70±0.16	0.84±0.12	0.59±0.24	0.64±0.14	0.72±0.09	0.71±0.07
Stack_SVM_Kmeans	0.69±0.18	0.87±0.13	0.56±0.28	0.65±0.16	0.76±0.10	0.72±0.07
Stack_SVM_Logitboost	0.70±0.16	0.84±0.12	0.59±0.24	0.64±0.14	0.72±0.09	0.71±0.07

In this step, it was important for us to find the top 5 motifs with the best predictive significance score and to create new peptide sequences by taking their triple combinations. Therefore, we took the first 5 motifs that are most important for Gram-negative and Gram-positives (shown in Table 5.10) and generated 60 new peptide sequences with their triple combinations.

Table 5.10 Results of Ranked Features(Step 2).

Gram-Positive			**Gram-Negative**		
Motif(Feature)	**Model**	**Score**	**Motif(Feature)**	**Model**	**Score**
CKRWLW	LogitBoost	0.66	HLRRP	RF	0.90
RAGLQFPIGKLV	LogitBoost	0.53	HLRRI	RF	0.79
WRAHGVVHRY	LogitBoost	0.44	LRKLGRKIKKA	RF	0.61
RAGLQWPIGRLL	LogitBoost	0.43	LKRLGRKIKKA	RF	0.53

CKGWQW	LogitBoost	0.43	LKRFGRKIKKA	RF	0.50

5.3.3 Motif Combination(Step3)

60 new peptides for both datasets are created by taking a triple combination of the first five motifs for both datasets.

Example sequences for gram-negative: HLRRPHLRRILRKLGRKIKKA, HLRRPHLRRILKRLGRKIKKA, HLRRPHLRRILKRFGRKIKKA, HLRRPLRKLGRKIKKAHLRRI, HLRRPLRKLGRKIKKALKRLGRKIKKA, HLRRPLRKLGRKIKKALKRFGRKIKKA

Along with the newly produced 60 peptides, 30 non-AMP and 5 AMP peptides from the initially existing peptides were removed from the Gram-negative and 40 non-AMP and 5 AMP peptides from the initially existing peptides were removed from the Gram-positive dataset separately for use in the test dataset. The remaining peptides from the extracted peptides were also used as a train dataset. At the end, we have totally 84 negatively labeled nonAMP and 85 positively labeled AMP peptides for a train set and 65 positively labeled (60 of newly created) and 30 negatively labeled peptides for a test set of Gram-negative dataset. For the Gram-positive dataset, we have 154 negatively labeled and 84 positively labeled peptides, totaling 238 peptides for a train set and 65 positively labeled (60 of newly created) and 40 negatively labeled peptides, totaling 105 peptides for a test set**.**

We extracted 10 physico-chemical features from DBAASP web server that are mentioned in Section 2.2. After extracting 10 physicochemical properties of each sequence, the model was rerun to have prediction performance results. The prediction results of our datasets created with newly added peptides are 98% accuracy and 99% AUC using LogitBoost algorithm for Gram-negative dataset and 96% accuracy and 98% AUC using Random Forest algorithm for Gram-positive dataset.

Table 5.11 Results for Gram-negative dataset(Step3).

Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1
Adaboost	0.98	0.98	1	1	0.99	0.99
DT	0.95	0.93	1	1	0.96	0.96
LogitBoost	0.98	1	1	0.98	0.99	0.99
RF	0.96	0.98	1	0.96	0.99	0.97

Table 5.12 Results for Gram-positive dataset(Step3).

Model	Accuracy	Recall	Specificity	Precision	Area Under Curve	F1
Adaboost	0.93	0.95	0.92	0.92	0.98	0.93
DT	0.81	0.82	0.8	0.80	0.81	0.81
LogitBoost	0.95	1	0.82	0.91	0.98	0.95
RF	0.96	1	0.87	0.93	0.98	0.96

Vishnepolsky et al. have created an AMP predictor based on microbial strain-specific [144]. Using this predictor, it is determined whether a given peptide will be active against a list of bacteria. Since we selected "*Escherichia coli ATCC 25922*" as one of the Gram-negative bacteria while creating our dataset, we have seen whether the newly created peptides against this strain are active or not, with this tool. Also, for observing whether the newly created peptides against Gram-positive bacteria are active or not, we selected "*Staphylococcus aureus ATCC 25923*" as one of the Gram-positive bacteria. the length of sequences should not exceed 30 amino acid. Therefore, we eliminated sequences with a sequence length of more than 30 amino acids. As a result of elimination, we have 35 new sequences for the Gram-negative dataset, while this number has increased to 54 for Gram-positive dataset. According to the results, while all of the new peptides formed against gram negatives were active, 49 of the 54 peptides formed against gram positives were predicted as active. The prediction results are added to the Appendix Table A2 and Table A3.

Chapter 6

6.Conclusions and Future Prospects

6.1 Conclusions

The main contribution of the first study in this thesis is the development of two accurate classification models for the prediction of antimicrobial peptides active against (i) Gram-negative and (ii) Gram-positive bacteria, separately. 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 RF 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 dataset and second most important feature for Gram-positive dataset. Moreover, for the Gram-positive dataset, the pI was found as the most important feature, while it was determined as the second most important feature for the 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 [119]. Hence, our findings are not contradictory with previous results which suggest that net charge and pI are the main factors for strong antimicrobial activity, and this situation further proves the validity of the computational models created in this study. The PCA visualization is applied on the Gram-negative and the Gram-positive dataset, and some outlier samples have been observed. Based on the distribution of the positive and negative labeled 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 AUC results increased by 7% for both the Gram-negative and Gram-positive dataset.

We repeated our experiments using an extended feature set including amino acid composition, pseudo amino acid composition, sequence order, autocorrelation, composition, distribution, and transition of physico-chemical properties. When we run

our workflow on these extended feature sets, the performance metrics did not improve, and even lowered slightly. For the Gram-negative dataset, while the extended set of features yielded 98% AUC with LogitBoost, physico-chemical features yielded 99% AUC with RF. For the Gram-positive dataset, while the model using an extended set of features achieved 95% AUC with RF, the generated model using only ten physico-chemical features achieved 97% AUC. When we compared the performance metrics obtained using physico-chemical properties (10 features) with an extended set of features (1507 features), we observed that rather than using a large selection of features, a small number of features yielded better results on both Gram-negative and Gram-positive datasets.

Different feature selection methods are applied on the extended dataset for removing redundant features. It is worthwhile to note that for the Gram-positive dataset, among 1507 different descriptors belonging to the structure-based, linguistic-based, sequence-based, and physico-chemical-based classes in the extended dataset, all 3 selected features (isoelectric point, net charge, disordered conformation propensity) are physico-chemical descriptors. After the feature selection is applied on the extended dataset including 1507 features, the AUC values of the models using the top 3 scoring features decreased only by 1% and 2% for the Gram-positive and Gram-negative datasets, respectively. When we compare the performance metrics before and after feature selection is applied, we can deduce that using only 3 features yields satisfactory performance results (96% and 94% AUC) for Gram-negative and Gram-positive datasets, respectively. However, for both of the Gram-negative and Gram-positive datasets, the performance of the models using 10 physico-chemical features (99% and 97% AUC values respectively) was still higher than the performance of the extended feature set, and higher than the performance of the extended feature set after feature selection.

To conclude, AMPs 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 the first study of thesis paves the way to the precise prediction and the design of antimicrobial peptides that are highly effective against specific 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.

As the second work of the thesis, we create a novel approach based on grouping, scoring, and modeling to accurately predict the antimicrobial peptides. To determine key properties involved in the prediction of antimicrobial peptides, we used different types of feature groups. Each group has its own feature set. The group including physico-chemical features is identified as the best group in terms of predicting AMP activity. We observed that estimating antimicrobial peptides using only physico-chemical properties generated the best score. It has been demonstrated that physico-chemical properties play a significant impact in peptide prediction, and should be taken into account while developing novel models.

It is crucial to compare our novel approach with benchmark datasets in this area. Our findings demonstrate how effective and discriminating the AMP-GSM model is. In-depth evaluations of AMP-GSM against other traditional feature selection techniques for AMP prediction place it among one of the best predictors.

To sum up, AMPs are thought to be the most promising antibiotic substitutes. Consequently, precise antimicrobial peptide prediction aids in the development of cheaper, more efficient peptides. Additionally, they gained popularity in this industry since computational prediction approaches minimize losses during production phases. This study's grouping methodology will be beneficial to precise prediction and the design of antimicrobial peptides that are extremely efficient against particular bacterial infections. Although the categorization method we have created here is only applicable to antimicrobial and anti-inflammatory peptides, it could be used in future research to predict antifungal, antiviral, antiprotozoal, and anticancer drugs. Additionally, it is possible to expand our current work by adding more groups for future studies.

As the last study of this thesis, in order to apply machine learning techniques to peptides, a new sequence representation based on conserved motifs was proposed. This feature extraction method which is based on the match score aims to capture important motifs occurring in AMPs. According to feature importance scores, we combine the first five best motifs to create novel peptides. The new peptides and some other peptides extracted from the initial dataset were added to the test dataset. Our models obtained 98% accuracy and 99% AUC for Gram-negative; and 96% accuracy and 98% AUC for Gram-positive datasets. We used the predictor created by Visnepolsky [144] to understand whether the peptides that we have designed were active against Gram-

negative and Gram-positive bacteria. This tool agreed that most of the new peptides that we have created were active against Gram-negative and Gram-positive bacteria. Hence, we concluded that these peptides that are predicted to be active against Gram-negative and Gram-positive bacteria can be used for laboratory experiments.

6.2 Societal Impact and Contribution to Global Sustainability

According to WHO statistics in 2020, one of the major risks to modern development, food security, and global health is antibiotic resistance. Medicines known as antibiotics are used to both prevent and treat bacterial infections. When bacteria adapt to the use of antibiotics, antibiotic resistance develops. Antibiotic resistance causes increased mortality, longer hospital stays, and higher medical expenses.

Innate immunity produces antimicrobial peptides (AMPs), which are naturally occurring antibiotics and are encoded by particular genes. They are created by a variety of human, plant, and animal tissues and cell types. Typically, these antimicrobial peptides have 12 to 50 amino acids. Currently, antibiotic resistance is quickly rising in parallel with the increased usage of antibiotics. Antimicrobial resistance is reportedly increasing globally and new resistance mechanisms are developing, according to the World Health Organization (WHO). As a result, we may soon live in a time when infections cannot be cured with medications. It is necessary to develop new antimicrobial agents that can be used in treatment because there are more and more germs that are resistant to antibiotics. Detailing research on the characteristics of antimicrobial peptides is a crucial area for medication development. Although Gram-positive and Gram-negative bacteria are the principal targets of AMPs, they can also be employed to fight mycobacteria, viruses, and cancerous cells. Since they have a lesser potential of developing resistance, AMPs are viewed as a potent alternative to antibiotics in this regard. Hence, developing new antimicrobial peptides became a key area of research. Prior to the laborious, expensive, and challenging production processes, it is crucial to accurately predict the activity of candidate peptides. Several computational techniques have been suggested for predicting the antimicrobial activity of AMPs and for identifying promising AMP candidates without involving costly wet-lab experiments. The application of machine learning approaches became common among many computational techniques for the estimation of antimicrobial peptides. Therefore, our main goal in this thesis is to develop antimicrobial peptide prediction

using machine learning methods and to produce new peptides at the least cost in order to send them to laboratory studies, which is the final stage.

6.3 Future Prospects

For each of the three studies carried out as part of this thesis work, we can briefly describe our prospects for the future as follows:

- Future prospects of study 1:

  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.

- Future prospects of study 2:

  The categorization method we have created here is only applicable to antimicrobial and anti-inflammatory peptides, it could be used in future research to predict antifungal, antiviral, antiprotozoal, and anticancer drugs. Additionally, it is possible to expand our current work by adding more groups for future studies.

- Future prospects of study 3:

  The novel peptides that are predicted active against Gram-negative and Gram-positive bacteria can be used for laboratory experiments as a future work.

BIBLIOGRAPHY

[1] Y. Jung, B. Kong, S. Moon, S. H. Yu, J. Chung, J., C. Ba, ... & D. H. Kweon, "Envelope-deforming antiviral peptide derived from influenza virus M2 protein." Biochemical and Biophysical Research Communications, 517, 507-512, (2019).

[2] V. Carnicelli, A. Lizzi, A. Ponzi, G. Amicosante, A. Bozzi, and A. Di Giulio, "Articolo su libro (2013)" (2015).

[3] N. K. Brogden and K. A. Brogden, "Will new generations of modified antimicrobial peptides improve their potential as pharmaceuticals?", Int. J. Antimicrob. Agents, S0924857911002342, (2011).

[4] F. Xie et al., "The SapA Protein Is Involved in Resistance to Antimicrobial Peptide PR-39 and Virulence of Actinobacillus pleuropneumoniae", Front. Microbiol., 8, 811 (2017).

[5] D. Neubauer et al., "Retro analog concept: comparative study on physico-chemical and biological properties of selected antimicrobial peptides", Amino Acids, 49, 1755–1771 (2017).

[6] M. Erdem Büyükkiraz and Z. Kesmen, "Antimicrobial peptides (AMPs): A promising class of antimicrobial compounds", J. Appl. Microbiol., (2021).

[7] B. Mishra and G. Wang, "Ab Initio Design of Potent Anti-MRSA Peptides Based on Database Filtering Technology", ACS Publications, (2012).

[8] D. Faccone et al., "Antimicrobial activity of de novo designed cationic peptides against multi-resistant clinical isolates", Eur. J. Med. Chem., 71, 31–35 (2014).

[9] C. H. Chen et al., "Simulation-Guided Rational de Novo Design of a Small Pore-Forming Antimicrobial Peptide", J. Am. Chem. Soc., (2019).

[10] B. Vishnepolsky et al., "De Novo Design and In Vitro Testing of Antimicrobial Peptides against Gram-Negative Bacteria", Pharmaceuticals,12, 82, (2019).

[11] C. Loose, K. Jensen, I. Rigoutsos, and G. Stephanopoulos, "A linguistic model for the rational design of antimicrobial peptides", Nature, 443, 867–869 (2006).

[12] D. Nagarajan et al., "Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria", J. Biol. Chem., 293, 3492–3509 (2018).

[13] 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., (2018).

[14] E. S. Cândido et al., "Short Cationic Peptide Derived from Archaea with Dual Antibacterial Properties and Anti-Infective Potential", ACS Infect. Dis., 5, 1081–1086 (2019).

[15] I. C. M. Fensterseifer et al., "Selective antibacterial activity of the cationic peptide PaDBS1R6 against Gram-negative bacteria", Biochim. Biophys. Acta BBA - Biomembr., 1861, 1375–1387 (2019).

[16] 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., (2019).

[17] 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., 77, 48–56 (2011).

[18] G. Maccari et al., "Antimicrobial Peptides Design by Evolutionary Multiobjective Optimization", PLoS Comput. Biol., 9, e1003212 (2013).

[19] W. F. Porto et al., “In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design”, Nat. Commun., 9, 1490 (2018).
[20] M. Yoshida et al., “Using Evolutionary Algorithms and Machine Learning to Explore Sequence Space for the Discovery of Antimicrobial Peptides”, Chem, 4, 533–543 (2018).
[21] 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., 23, 4–12 (2017).
[22] 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., 436, 168–177 (2013).
[23] A. C. Schierz, “Virtual screening of bioassay data”, J. Cheminformatics, 1, 21 (2009).
[24] 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., 8, 1697 (2018).
[25] S. Lata, N. K. Mishra, and G. P. Raghava, “AntiBP2: improved version of antibacterial peptide prediction”, BMC Bioinformatics, 11, S19 (2010).
[26] D. Dhall, R. Kaur, and M. Juneja, “Machine Learning: A Review of the Algorithms and Its Applications”, in Proceedings of ICRIC 2019, Cham, 47–63 (2020).
[27] 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, 7, 20160153 (2017).
[28] M. Burdukiewicz et al., “Proteomic Screening for Prediction and Design of Antimicrobial Peptides with AmpGram”, Int. J. Mol. Sci., 21, 4310 (2020).
[29] C.-R. Chung et al., “Characterization and Identification of Natural Antimicrobial Peptides on Different Organisms”, Int. J. Mol. Sci., 21, 986 (2020).
[30] P. Wang et al., “Prediction of Antimicrobial Peptides Based on Sequence Alignment and Feature Selection Methods”, PLoS ONE, 6, e18476 (2011).
[31] P. Agrawal and G. P. S. Raghava, “Prediction of Antimicrobial Potential of a Chemically Modified Peptide From Its Tertiary Structure”, Front. Microbiol., 9, 2551 (2018).
[32] S. Gull, N. Shamim, and F. Minhas, “AMAP: Hierarchical multi-label prediction of biologically active and antimicrobial peptides”, Comput. Biol. Med., 107, 172–181 (2019).
[33] M. Torrent, V. M. Nogués, and E. Boix, “A theoretical approach to spot active regions in antimicrobial proteins”, BMC Bioinformatics, 10, 373 (2009).
[34] 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., 44, D1094–D1097 (2016).
[35] W. Lin and D. Xu, “Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types”, Bioinformatics, 32, 3745–3752 (2016).
[36] J. Yan et al., “Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning”, Mol. Ther. - Nucleic Acids, 20, 882–894 (2020).
[37] X. Su, J. Xu, Y. Yin, X. Quan, and H. Zhang, “Antimicrobial peptide identification using multi-scale convolutional network”, BMC Bioinformatics, 20, 730 (2019).
[38] P. Schneider et al., “Hybrid Network Model for “Deep Learning” of Chemical Data: Application to Antimicrobial Peptides”, Mol. Inform., 36, 1600011 (2017).
[39] J. Witten and Z. Witten, “Deep learning regression model for antimicrobial peptide design”, Bioinformatics, preprint, (2019).

[40] J. A. Beltran, L. Aguilera-Mendoza, and C. A. Brizuela, “Optimal selection of molecular descriptors for antimicrobial peptides classification: an evolutionary feature weighting approach”, BMC Genomics, 19, 672 (2018).
[41] H. Fu, Z. Cao, M. Li, and S. Wang, “ACEP: improving antimicrobial peptides recognition through automatic feature fusion and amino acid embedding”, BMC Genomics, 21, 597 (2020).
[42] A. T. Müller et al., “Sparse Neural Network Models of Antimicrobial Peptide-Activity Relationships”, Mol. Inform., 35, 606–614 (2016).
[43] M.-N. Hamid and I. Friedberg, “Identifying antimicrobial peptides using word embedding with deep recurrent neural networks”, Bioinformatics, 35, 2009–2016 (2019).
[44] C. Li et al., “AMPlify: attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens”, BMC genomics, 23,1-15 (2022).
[45] S. Liu, J. Bao, X. Lao, and H. Zheng, “Novel 3D Structure Based Model for Activity Prediction and Design of Antimicrobial Peptides”, Sci. Rep., 8, 11189 (2018).
[46] A. Capecchi, X. Cai, H. Personne, T. Köhler, C. van Delden, and J.-L. Reymond, “Machine learning designs non-hemolytic antimicrobial peptides”, Chem. Sci., 12, 9221–9232 (2021).
[47] B. Vishnepolsky et al., “Predictive Model of Linear Antimicrobial Peptides Active against Gram-Negative Bacteria”, J. Chem. Inf. Model., 58, 1141–1151 (2018).
[48] 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, 5608 (2018).
[49] F. Plisson, O. Ramírez-Sánchez, and C. Martínez-Hernández, “Machine learning-guided discovery and design of non-hemolytic peptides”, Sci. Rep., 10, 16581 (2020).
[50] 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., 10, 12619 (2020).
[51] 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., 3, 163–175 (2008).
[52] W. Li and A. Godzik, “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences”, Bioinformatics, 22, 1658–1659 (2006).
[53] B. Vishnepolsky and M. Pirtskhalava, “Comment on: “Empirical comparison of web-based antimicrobial peptide prediction tools”“, Bioinformatics, 35, 2692–2694 (2019).
[54] J. H. Lee et al., “Transcriptome Analysis of Psacothea hilaris: De Novo Assembly and Antimicrobial Peptide Prediction”, Insects, 11, (2020).
[55] 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., 98, 280–287 (2012).
[56] D. Veltri, U. Kamath, and A. Shehu, “Deep learning improves antimicrobial peptide recognition”, Bioinformatics, 34, 2740–2747 (2018).
[57] A. Gautam et al., “Development of Antimicrobial Peptide Prediction Tool for Aquaculture Industries”, Probiotics Antimicrob. Proteins, 8, 141–149 (2016).
[58] M. N. Gabere and W. S. Noble, “Empirical comparison of web-based antimicrobial peptide prediction tools”, Bioinformatics, 33, 1921–1929 (2017).

[59] 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., 42, D1154–D1158 (2014).
[60] 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”, 5 (2021).
[61] B. Manavalan, T. H. Shin, M. O. Kim, and G. Lee, “AIPpred: Sequence-Based Prediction of Anti-inflammatory Peptides Using Random Forest”, Front. Pharmacol., 9, 276, (2018).
[62] Q. Zhang et al., “Immune epitope database analysis resource (IEDB-AR)”, Nucleic Acids Res., 36, W513-W518, (2008).
[63] W. Fleri et al., “The Immune Epitope Database and Analysis Resource in Epitope Discovery and Synthetic Vaccine Design”, Front. Immunol., 8, (2017).
[64] S. Spänig and D. Heider, “Encodings and models for antimicrobial peptide classification for multi-resistant pathogens”, BioData Min., 12, 7 (2019).
[65] 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, 22, 549 (2021).
[66] A. Moretta et al., “A bioinformatic study of antimicrobial peptides identified in the Black Soldier Fly (BSF) Hermetia illucens (Diptera: Stratiomyidae)”, Sci. Rep., 10, 16875 (2020).
[67] B. Vishnepolsky and M. Pirtskhalava, “Prediction of Linear Cationic Antimicrobial Peptides Based on Characteristics Responsible for Their Interaction with the Membranes”, J. Chem. Inf. Model., 54, 1512–1523 (2014).
[68] N. Thakur, A. Qureshi, and M. Kumar, “AVPpred: collection and prediction of highly effective antiviral peptides”, Nucleic Acids Res., 40, W199–W204 (2012).
[69] F. Lira, P. S. Perez, J. A. Baranauskas, and S. R. Nozawa, “Prediction of Antimicrobial Activity of Synthetic Peptides by a Decision Tree Model”, Appl. Environ. Microbiol., 79, 3156–3159 (2013).
[70] K. Pane et al., “Antimicrobial potency of cationic antimicrobial peptides can be predicted from their amino acid composition: Application to the detection of “cryptic” antimicrobial peptides”, J. Theor. Biol., 419, 254–265 (2017).
[71] Z. Chen et al., “iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data”, Brief. Bioinform., 21, 1047–1057 (2020).
[72] R. Muhammod, S. Ahmed, D. Md Farid, S. Shatabda, A. Sharma, and A. Dehzangi, “PyFeat: a Python-based effective feature generation tool for DNA, RNA and protein sequences”, Bioinformatics, 35, 3831–3833 (2019).
[73] R. Nikam and M. M. Gromiha, “Seq2Feature: a comprehensive web-based feature extraction tool”, Bioinformatics, 35, 4797–4799 (2019).
[74] D.-S. Cao, Q.-S. Xu, and Y.-Z. Liang, “propy: a tool to generate various modes of Chou’s PseAAC”, Bioinformatics, 29, 960–962 (2013).
[75] Z. Chen et al., “iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences”, Bioinformatics, 34, 2499–2502 (2018).
[76] J. Dong et al., “PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions”, J. Cheminformatics, 10, 16 (2018).
[77] S. M. H. Mahmud, W. Chen, H. Meng, H. Jahan, Y. Liu, and S. M. M. Hasan, “Prediction of drug-target interaction based on protein features using undersampling and feature selection techniques with boosting”, Anal. Biochem., 589, 113507 (2020).

[78] S.-J. Yeh, J.-F. Lin, and B.-S. Chen, “Multiple-Molecule Drug Design Based on Systems Biology Approaches and Deep Neural Network to Mitigate Human Skin Aging”, Molecules, 26, 3178 (2021).
[79] S.-J. Yeh, Y.-C. Chung, and B.-S. Chen, “Investigating the Role of Obesity in Prostate Cancer and Identifying Biomarkers for Drug Discovery: Systems Biology and Deep Learning Approaches”, Molecules, 27, 900 (2022).
[80] M. A. Wani, P. Garg, and K. K. Roy, “Machine learning-enabled predictive modeling to precisely identify the antimicrobial peptides”, Med. Biol. Eng. Comput., 59, 2397–2408 (2021).
[81] K. Pearson, “LIII. On lines and planes of closest fit to systems of points in space”, Lond. Edinb. Dublin Philos. Mag. J. Sci., 2, 559–572, (1901).
[82] 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, 7, e51444 (2012).
[83] M. Shu et al., “Predicting the Activity of Antimicrobial Peptides with Amino Acid Topological Information”, Med. Chem., 9, 32–44 (2013).
[84] L. Moll, E. Badosa, M. Planas, L. Feliu, E. Montesinos, and A. Bonaterra, “Antimicrobial Peptides With Antibiofilm Activity Against Xylella fastidiosa”, Front. Microbiol., 12, 753874 (2021).
[85] 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., 30, 369–376 (2016).
[86] 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, 1200–1205 (2015).
[87] 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, 8, 144 (2007).
[88] 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, 9, (2020).
[89] 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, 1479, 215–224 (2021).
[90] M. Yousef, L. Abdallah, and J. Allmer, “maTE: discovering expressed interactions between microRNAs and their targets”, Bioinformatics, 35, 4020–4028 (2019).
[91] 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., 7, e336 (2021).
[92] 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, 9, e11458 (2021).
[93] M. Yousef, G. Goy, and B. Bakir-Gungor, “miRModuleNet: Detecting miRNA-mRNA Regulatory Modules”, 835 (2022).
[94] 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, 1479, 205-214 (2021).

[95] M. Yousef, A. Kumar, and B. Bakir-Gungor, “Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data”, Entropy, 23, 2 (2020).
[96] P. Hanchuan, L. Fuhui, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy”, IEEE Trans. Pattern Anal. Mach. Intell., 27, 1226–1238, (2005).
[97] G. Brown, A. Pocock, M.-J. Zhao, and M. Lujan, “Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection”, The journal of machine learning research, 13, 27-66, (2012).
[98] T. Chen and T. He, “xgboost: eXtreme Gradient Boosting”, 1, 1-4, (2015).
[99] J. T. Kent, “Information gain and a general measure of correlation”, Biometrika, 70, 163–173, (1983)
[100] T. K. Ho, “Random decision forests”, in Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, Que., Canada, 1, 278–282 (1995).
[101] C. Cortes and V. Vapnik, “Support-vector networks”, Mach. Learn., 20, 273–297 (1995).
[102] Y. Freund and R. E. Schapire, “A Short Introduction to Boosting”, 14 (1999).
[103] 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., 28 (2000).
[104] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, “Classification And Regression Trees”, 1st ed. Routledge, (2017).
[105] E. Fix and J. L. Hodges, “Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties”, Int. Stat. Rev. Rev. Int. Stat., 57, 238–247, (1989).
[106] Q.-S. Xu and Y.-Z. Liang, “Monte Carlo cross validation”, Chemom. Intell. Lab. Syst., 56, 1–11 (2001).
[107] S. Thudumu, P. Branch, J. Jin, and J. Singh, “A comprehensive survey of anomaly detection techniques for high dimensional big data”, J. Big Data, 7, 42 (2020).
[108] L. M. Manevitz and M. Yousef, “One-Class SVMs for Document Classification”, J. Mach. Learn. Res., 139-154 (2001).
[109] L. Manevitz and M. Yousef, “One-class document classification via Neural Networks”, Neurocomputing, 70, 1466–1481 (2007).
[110] L. Abdallah, M. Badarna, W. Khalifa, and M. Yousef, “MultiKOC: Multi-One-Class Classifier Based K-Means Clustering”, Algorithms, 14, 134 (2021).
[111] L. Abedalla, M. Badarna, W. Khalifa, and M. Yousef, “K – Means Based One-Class SVM Classifier”, in Database and Expert Systems Applications, 1062, 45-53 (2019).
[112] M. Yousef, W. Khalifa, and L. AbedAllah, “Ensemble Clustering Classification compete SVM and One-Class classifiers applied on plant microRNAs Data.”, J. Integr. Bioinforma., 13, 304 (2016).
[113] M. Pirtskhalava and M. Grigolava, “Transmembrane and Antimicrobial Peptides. Hydrophobicity, Amphiphilicity and Propensity to Aggregation”, 24 (2013).
[114] P. Kumar, J. Kizhakkedathu, and S. Straus, “Antimicrobial Peptides: Diversity, Mechanism of Action and Strategies to Improve the Activity and Biocompatibility In Vivo”, Biomolecules, 8, 4 (2018).
[115] Y. Shai, “Mode of action of membrane active antimicrobial peptides”, Pept. Sci., 66, 236–248 (2002).
[116] D. Osorio, P. Rondón-Villarreal, and R. Torres Sáez, “Peptides: A Package for Data Mining of Antimicrobial Peptides”, R J., 7, 4–14 (2015).
[117] 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., 270, 2805–2813 (2003).

[118] I. M. Bezerra, L. C. Moreira, O. Chiavone-Filho, and S. Mattedi, “Effect of different variables in the solubility of ampicillin and corresponding solid phase”, Fluid Phase Equilibria, 459, 18–29 (2018).
[119] H. Le, L. Ting, C. Jun, and W. Weng, “Gelling properties of myofibrillar protein from abalone (Haliotis Discus Hannai Ino) muscle”, Int. J. Food Prop., 21, 277–288 (2018).
[120] N. Ni et al., “Gel properties and molecular forces of lamb myofibrillar protein during heat induction at different pH values”, Process Biochem., 49, 631–636 (2014).
[121] 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, 27, 640–648 (2006).
[122] M. Pirtskhalava, B. Vishnepolsky, and M. Grigolava, “Physicochemical Features and Peculiarities of Interaction of Antimicrobial Peptides with the Membrane”, 14, 471 (2021).
[123] N. Papo and Y. Shai, “Can we predict biological activity of antimicrobial peptides from their interactions with model phospholipid membranes?”, Peptides, 24, 1693–1703 (2003).
[124] V. Teixeira, M. J. Feio, and M. Bastos, “Role of lipids in the interaction of antimicrobial peptides with membranes”, Prog. Lipid Res., 51, 149–177 (2012).
[125] 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., 51, 1398–1406 (2007).
[126] D. Eisenberg, R. M. Weiss, and T. C. Terwilliger, “The helical hydrophobic moment: a measure of the amphiphilicity of a helix”, Nature, 299, 371–374 (1982).
[127] M. Yousef, D. Levy, and J. Allmer, “Species Categorization via MicroRNAs - Based on 3”UTR Target Sites using Sequence Features:”, in Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies, 112–118 (2018).
[128] M. Yousef, W. Khalifa, İ. E. Acar, and J. Allmer, “Distinguishing between MicroRNA Targets from Diverse Species using Sequence Motifs and K-mers”:, in Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies, 133–139 (2017).
[129] M. Yousef, W. Khalifa, İ. E. Acar, and J. Allmer, “MicroRNA categorization using sequence motifs and k-mers”, BMC Bioinformatics, 18, 170 (2017).
[130] M. Torrent, D. Andreu, V. M. Nogués, & E. Boix, “Connecting peptide physicochemical and antimicrobial properties by a rational prediction model”, PloS one, 6, e16968 (2011).
[131] K. Boone, K. Camarda, P. Spencer, C. Tamerler, “Antimicrobial peptide similarity and classification through rough set theory using physicochemical boundaries”, BMC Bioinformatics, 19, 469 (2018).
[132] S. Thomas, S. Karnik, R. S. Barai, V. K. Jayaraman, & S. Idicula-Thomas, “CAMP: a useful resource for research on antimicrobial peptides”, Nucleic acids research, 38(suppl_1), D774-D780 (2010).
[133] J. Xu, F. Li, A. Leier, D. Xiang, H. H. Shen, T. T. M. Lago, J. Li, D. J. Yu, J. Song, “Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides”, Brief. Bioinform., 22, 83 (2021).
[134] E. Khaledian, S.L. Broschat, “Sequence-Based Discovery of Antibacterial Peptides Using Ensemble Gradient Boosting”, In Proceedings of the 1st International Electronic Conference on Microbiology, Sciforum Online, 2, 6 (2020).

[135]P.B. Timmons, C.M. Hewage, “HAPPENN is a novel tool for hemolytic activity prediction for therapeutic peptides which employs neural networks”, Sci. Rep., 10, 10869 (2020).
[136]M. Yousef, F. Ozdemir, A. Jaber, J. Allmer, and B. Bakir-Gungor, “PriPath: identifying dysregulated pathways from differential gene expression via grouping, scoring, and modeling with an embedded feature selection approach”, BMC Bioinformatics, 24, 60, (2023).
[137]M. Unlu Yazici, J. S. Marron, B. Bakir-Gungor, F. Zou, and M. Yousef, “Invention of 3Mint for feature grouping and scoring in multi-omics”, Front. Genet., 14, (2023).
[138]R. Kolde, S. Laur, P. Adler, and J. Vilo, “Robust rank aggregation for gene list integration and meta-analysis”, Bioinformatics, 28, 573–580, (2012).
[139]H. Teimouri, A. Medvedeva, and A. B. Kolomeisky, “Bacteria-Specific Feature Selection for Enhanced Antimicrobial Peptide Activity Predictions Using Machine-Learning Methods”, J. Chem. Inf. Model., 63, 1723–1733, (2023).
[140] S. Joseph, S. Karnik, P. Nilawe, V. K. Jayaraman, and S. Idicula-Thomas, “ClassAMP: A Prediction Tool for Classification of 654 Antimicrobial Peptides”, IEEE/ACM Trans. Comput. Biol. Bioinform., 9, 1535–1538, (2012).
[141] A. L. Tornesello, A. Borrelli, L. Buonaguro, F. M. Buonaguro, and M. L. Tornesello, “Antimicrobial Peptides as Anticancer 657 Agents: Functional Properties and Biological Activities”, Molecules, 25, 2850, (2020).
[142] T. L. Bailey et al., “MEME SUITE: tools for motif discovery and searching”, Nucleic Acids Res., 37, W202–W208, (2009).
[143] F. Pedregosa et al. “Scikit-learn: Machine learning in Python.” the Journal of machine Learning research, 12, 2825-2830, (2011).
[144] B. Vishnepolsky et al., “Comparative analysis of machine learning algorithms on the microbial strain-specific AMP prediction”, Brief. Bioinform., 23, 4, (2022).

APPENDIX

Table A1. Ranking of the Groups by the RobustRankAggreg method on Gram-negative and Gram-positive datasets of Dataset 1.

Gram-negative Dataset				**Gram-positive Dataset**			
Group	**Count**	**Score**	**RobustRankAgg (p-value)**	**Group**	**Count**	**Score**	**RobustRankAgg (p-value)**
Physico-chemical	1000	10	2.36E-25	Physico-chemical	1000	10	1.2E-106
Physicochemical composition	839	8.39	3.17E-13	Physicochemical composition	844	8.44	4.8E-37
Quasi-sequence-order	757	7.57	1.1E-09	Quasi-sequence-order	745	7.45	1.7E-25
Amino acid composition	706	7.06	7.15E-09	Amino acid composition	700	7	2.83E-20
Physicochemical distribution	436	4.36	0.000159	Physicochemical transition	546	5.46	3.21E-11
Pseudo-amino acid composition	472	4.72	0.000159	Pseudo-amino acid composition	505	5.05	3.21E-11
Dipeptide composition	470	4.7	0.000159	Physicochemical distribution	396	3.96	1.1E-09
Physicochemical transition	496	4.96	0.000264	Sequence order coupling number	259	2.59	1.21E-06
Sequence order coupling number	205	2.05	1	Dipeptide Composition	369	3.69	2.54E-05
Moran autocorrelation	53	0.53	1	Normalized Moreau–Broto autocorrelation	99	0.99	1

Geary autocorrelation	43	0.43	1	Moran autocorrelation	25	0.25	1
Normalized Moreau–Broto autocorrelation	23	0.23	1	Geary autocorrelation	12	0.12	1

Table A2. The prediction results of DBAASP tool [48] for Gram-negative Dataset.

Strain Type	Class	Predictive value	Sequence
Escherichia coli ATCC 25922	Active	0.85	LKRLGRKIKKAEYRKLRDLKRFGRKIKKA
Escherichia coli ATCC 25922	Active	0.86	LKRLGRKIKKAEYRKLRDLKRFGRKVKKA
Escherichia coli ATCC 25922	Active	0.86	LKRLGRKIKKAEYRKLRDLRRLGRKIKKA
Escherichia coli ATCC 25922	Active	0.77	LKRLGRKIKKALKRFGRKIKKAEYRKLRD
Escherichia coli ATCC 25922	Active	0.77	LKRLGRKIKKALKRFGRKVKKAEYRKLRD
Escherichia coli ATCC 25922	Active	0.82	LKRLGRKIKKALRRLGRKIKKAEYRKLRD
Escherichia coli ATCC 25922	Active	0.86	EYRKLRDLKRLGRKIKKALKRFGRKIKKA
Escherichia coli ATCC 25922	Active	0.87	EYRKLRDLKRLGRKIKKALKRFGRKVKKA
Escherichia coli ATCC 25922	Active	0.87	EYRKLRDLKRLGRKIKKALRRLGRKIKKA
Escherichia coli ATCC 25922	Active	0.86	EYRKLRDLKRFGRKIKKALKRLGRKIKKA
Escherichia coli ATCC 25922	Active	0.84	EYRKLRDLKRFGRKIKKALKRFGRKVKKA
Escherichia coli ATCC 25922	Active	0.85	EYRKLRDLKRFGRKIKKALRRLGRKIKKA
Escherichia coli ATCC 25922	Active	0.86	EYRKLRDLKRFGRKVKKALKRLGRKIKKA
Escherichia coli ATCC 25922	Active	0.84	EYRKLRDLKRFGRKVKKALKRFGRKIKKA
Escherichia coli ATCC 25922	Active	0.9	EYRKLRDLKRFGRKVKKALRRLGRKIKKA
Escherichia coli ATCC 25922	Active	0.88	EYRKLRDLRRLGRKIKKALKRLGRKIKKA
Escherichia coli ATCC 25922	Active	0.9	EYRKLRDLRRLGRKIKKALKRFGRKIKKA
Escherichia coli ATCC 25922	Active	0.9	EYRKLRDLRRLGRKIKKALKRFGRKVKKA
Escherichia coli ATCC 25922	Active	0.82	LKRFGRKIKKALKRLGRKIKKAEYRKLRD
Escherichia coli ATCC 25922	Active	0.86	LKRFGRKIKKAEYRKLRDLKRLGRKIKKA
Escherichia coli ATCC 25922	Active	0.84	LKRFGRKIKKAEYRKLRDLKRFGRKVKKA
Escherichia coli ATCC 25922	Active	0.88	LKRFGRKIKKAEYRKLRDLRRLGRKIKKA
Escherichia coli ATCC 25922	Active	0.8	LKRFGRKIKKALKRFGRKVKKAEYRKLRD
Escherichia coli ATCC 25922	Active	0.84	LKRFGRKIKKALRRLGRKIKKAEYRKLRD
Escherichia coli ATCC 25922	Active	0.83	LKRFGRKVKKALKRLGRKIKKAEYRKLRD
Escherichia coli ATCC 25922	Active	0.87	LKRFGRKVKKAEYRKLRDLKRLGRKIKKA
Escherichia coli ATCC 25922	Active	0.84	LKRFGRKVKKAEYRKLRDLKRFGRKIKKA
Escherichia coli ATCC 25922	Active	0.88	LKRFGRKVKKAEYRKLRDLRRLGRKIKKA
Escherichia coli ATCC 25922	Active	0.81	LKRFGRKVKKALKRFGRKIKKAEYRKLRD
Escherichia coli ATCC 25922	Active	0.84	LKRFGRKVKKALRRLGRKIKKAEYRKLRD
Escherichia coli ATCC 25922	Active	0.85	LRRLGRKIKKALKRLGRKIKKAEYRKLRD
Escherichia coli ATCC 25922	Active	0.86	LRRLGRKIKKAEYRKLRDLKRLGRKIKKA
Escherichia coli ATCC 25922	Active	0.87	LRRLGRKIKKAEYRKLRDLKRFGRKIKKA
Escherichia coli ATCC 25922	Active	0.87	LRRLGRKIKKAEYRKLRDLKRFGRKVKKA
Escherichia coli ATCC 25922	Active	0.8	LRRLGRKIKKALKRFGRKIKKAEYRKLRD

Table A3. The prediction results of DBAASP tool [48] for Gram-positive Dataset.

Sequence	Strain Type	Class	Predictive value
CKRWLWWRAHGVVHRYCKGWQW	Staphylococcus aureus ATCC 25923	Active	0.58
CKRWLWCKGWQWWRAHGVVHRY	Staphylococcus aureus ATCC 25923	Active	0.59
WRAHGVVHRYCKRWLWCKGWQW	Staphylococcus aureus ATCC 25923	Active	0.58
WRAHGVVHRYCKGWQWCKRWLW	Staphylococcus aureus ATCC 25923	Active	0.52
CKGWQWCKRWLWWRAHGVVHRY	Staphylococcus aureus ATCC 25923	Active	0.56
CKGWQWWRAHGVVHRYCKRWLW	Staphylococcus aureus ATCC 25923	Active	0.52
CKRWLWRAGLQFPIGKLVCKGWQW	Staphylococcus aureus ATCC 25923	Active	0.55
CKRWLWRAGLQWPIGRLLCKGWQW	Staphylococcus aureus ATCC 25923	Active	0.5
CKRWLWCKGWQWRAGLQFPIGKLV	Staphylococcus aureus ATCC 25923	Not Active	0.55
CKRWLWCKGWQWRAGLQWPIGRLL	Staphylococcus aureus ATCC 25923	Not Active	0.6
RAGLQFPIGKLVCKRWLWCKGWQW	Staphylococcus aureus ATCC 25923	Active	0.61
RAGLQFPIGKLVCKGWQWCKRWLW	Staphylococcus aureus ATCC 25923	Active	0.54
RAGLQWPIGRLLCKRWLWCKGWQW	Staphylococcus aureus ATCC 25923	Active	0.5
RAGLQWPIGRLLCKGWQWCKRWLW	Staphylococcus aureus ATCC 25923	Active	0.5
CKGWQWCKRWLWRAGLQFPIGKLV	Staphylococcus aureus ATCC 25923	Not Active	0.51
CKGWQWCKRWLWRAGLQWPIGRLL	Staphylococcus aureus ATCC 25923	Not Active	0.54
CKGWQWRAGLQFPIGKLVCKRWLW	Staphylococcus aureus ATCC 25923	Active	0.53
CKGWQWRAGLQWPIGRLLCKRWLW	Staphylococcus aureus ATCC 25923	Not Active	0.56
CKRWLWRAGLQFPIGKLVWRAHGVVHRY	Staphylococcus aureus ATCC 25923	Active	0.68
CKRWLWWRAHGVVHRYRAGLQFPIGKLV	Staphylococcus aureus ATCC 25923	Active	0.65
CKRWLWWRAHGVVHRYRAGLQWPIGRLL	Staphylococcus aureus ATCC 25923	Active	0.64
CKRWLWRAGLQWPIGRLLWRAHGVVHRY	Staphylococcus aureus ATCC 25923	Active	0.64
RAGLQFPIGKLVCKRWLWWRAHGVVHRY	Staphylococcus aureus ATCC 25923	Active	0.71
RAGLQFPIGKLVWRAHGVVHRYCKRWLW	Staphylococcus aureus ATCC 25923	Active	0.68
RAGLQFPIGKLVWRAHGVVHRYCKGWQW	Staphylococcus aureus ATCC 25923	Active	0.62
RAGLQFPIGKLVCKGWQWWRAHGVVHRY	Staphylococcus aureus ATCC 25923	Active	0.64
WRAHGVVHRYCKRWLWRAGLQFPIGKLV	Staphylococcus aureus ATCC 25923	Active	0.62
WRAHGVVHRYCKRWLWRAGLQWPIGRLL	Staphylococcus aureus ATCC 25923	Active	0.62
WRAHGVVHRYRAGLQFPIGKLVCKRWLW	Staphylococcus aureus ATCC 25923	Active	0.67
WRAHGVVHRYRAGLQFPIGKLVCKGWQW	Staphylococcus aureus ATCC 25923	Active	0.59
WRAHGVVHRYRAGLQWPIGRLLCKRWLW	Staphylococcus aureus ATCC 25923	Active	0.65
WRAHGVVHRYRAGLQWPIGRLLCKGWQW	Staphylococcus aureus ATCC 25923	Active	0.51
WRAHGVVHRYCKGWQWRAGLQFPIGKLV	Staphylococcus aureus ATCC 25923	Active	0.55
WRAHGVVHRYCKGWQWRAGLQWPIGRLL	Staphylococcus aureus ATCC 25923	Active	0.53
RAGLQWPIGRLLCKRWLWWRAHGVVHRY	Staphylococcus aureus ATCC 25923	Active	0.69
RAGLQWPIGRLLWRAHGVVHRYCKRWLW	Staphylococcus aureus ATCC 25923	Active	0.7
RAGLQWPIGRLLWRAHGVVHRYCKGWQW	Staphylococcus aureus ATCC 25923	Active	0.64

RAGLQWPIGRLLCKGWQWWRAHGVVHRY	Staphylococcus aureus ATCC 25923	Active	0.56
CKGWQWRAGLQFPIGKLVWRAHGVVHRY	Staphylococcus aureus ATCC 25923	Active	0.61
CKGWQWWRAHGVVHRYRAGLQFPIGKLV	Staphylococcus aureus ATCC 25923	Active	0.5
CKGWQWWRAHGVVHRYRAGLQWPIGRLL	Staphylococcus aureus ATCC 25923	Active	0.5
CKGWQWRAGLQWPIGRLLWRAHGVVHRY	Staphylococcus aureus ATCC 25923	Active	0.54
CKRWLWRAGLQFPIGKLVRAGLQWPIGRLL	Staphylococcus aureus ATCC 25923	Active	0.62
CKRWLWRAGLQWPIGRLLRAGLQFPIGKLV	Staphylococcus aureus ATCC 25923	Active	0.66
RAGLQFPIGKLVCKRWLWRAGLQWPIGRLL	Staphylococcus aureus ATCC 25923	Active	0.66
RAGLQFPIGKLVRAGLQWPIGRLLCKRWLW	Staphylococcus aureus ATCC 25923	Active	0.67
RAGLQFPIGKLVRAGLQWPIGRLLCKGWQW	Staphylococcus aureus ATCC 25923	Active	0.6
RAGLQFPIGKLVCKGWQWRAGLQWPIGRLL	Staphylococcus aureus ATCC 25923	Active	0.55
RAGLQWPIGRLLCKRWLWRAGLQFPIGKLV	Staphylococcus aureus ATCC 25923	Active	0.63
RAGLQWPIGRLLRAGLQFPIGKLVCKRWLW	Staphylococcus aureus ATCC 25923	Active	0.67
RAGLQWPIGRLLRAGLQFPIGKLVCKGWQW	Staphylococcus aureus ATCC 25923	Active	0.62
RAGLQWPIGRLLCKGWQWRAGLQFPIGKLV	Staphylococcus aureus ATCC 25923	Active	0.54
CKGWQWRAGLQFPIGKLVRAGLQWPIGRLL	Staphylococcus aureus ATCC 25923	Active	0.53
CKGWQWRAGLQWPIGRLLRAGLQFPIGKLV	Staphylococcus aureus ATCC 25923	Active	0.53

CURRICULUM VITAE

2009 – 2014	B.Sc., Computer Engineering, Muğla Sıtkı Koçman University, Muğla, TURKEY
2015 – 2017	M.Sc., Computer Engineering, Abdullah Gul University, Kayseri, TURKEY
2017-2023	Ph.D., Electrical and Computer Engineering, Abdullah Gul University, Kayseri, TURKEY

SELECTED PUBLICATIONS AND PRESENTATIONS

J1) Ü. G. Söylemez, M. Yousef, Z. Kesmen, M E. Büyükkiraz, B. Bakir-Gungor Prediction of linear cationic antimicrobial peptides active against gram-negative and gram-positive bacteria based on machine learning models published in Applied Sciences (April 2022).

J2) Ü. G. Söylemez, M. Yousef, B. Bakir-Gungor, AMP-GSM: Prediction of Antimicrobial Peptides via a Grouping–Scoring–Modeling Approach published in Applied Sciences (April 2023).

C1) Ü. G. Söylemez, M. Yousef, B. Bakir-Gungor, Prediction of Antimicrobial Peptides Using Deep Neural Networks In Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023) .