İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ
2010
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

Estimation of the Synergistic Effect of Antimicrobial Peptides and Antibiotics by Machine Learning Models

Submitted to the Graduate School of Natural and Applied Sciences in partial fulfillment of the requirements for the degree of

Master of Science

in Biomedical Engineering

by

Başak Olcay

ORCID 0000-0002-0020-6055

June, 2022

This is to certify that we have read the thesis **Estimation of the Synergistic Effect of Antimicrobial Peptides and Antibiotics by Machine Learning Models** submitted by **Başak Olcay**, and it has been judged to be successful, in scope and in quality, at the defense exam and accepted by our jury as a MASTER'S THESIS.

APPROVED BY:

Advisor: **Assoc. Prof. Dr. Ozan Karaman**
İzmir Kâtip Çelebi University

Committee Members:

Assoc. Prof. Dr. Ozan Karaman
İzmir Kâtip Çelebi University

Assoc. Prof. Dr. Utku Kürşat Ercan
İzmir Kâtip Çelebi University

Prof. Dr. Muhammed Bahattin Tanyolaç
Ege University

Date of Defense: June 15, 2022

Declaration of Authorship

I, **Başak Olcay**, declare that this thesis titled **Estimation of the Synergistic Effect of Antimicrobial Peptides and Antibiotics by Machine Learning Models** and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for the Master's degree at this university.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this university or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. This thesis is entirely my own work, with the exception of such quotations.
- I have acknowledged all major sources of assistance.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 15.06.2022

Estimation of the Synergistic Effect of Antimicrobial Peptides and Antibiotics by Machine Learning Models

Abstract

Urinary catheters are widely used in patients who cannot empty their bladder. However, catheter surfaces are suitable for the adhesion of microorganisms, and this causes various complications in the following periods. Various catheter materials and different surface modifications have been tried to prevent complications, but most of these approaches have failed, and some have shown various side effects. Therefore, new methods are needed for the prevention and treatment of complications. Antimicrobial peptides have recently become popular for their use because they have advantages such as a wide range of activity, and not causing drug resistance. However, they suffer from disadvantages such as stability and reduced activity when linked to the surface. Therefore, the combined use of two antimicrobial agents has become one of the research topics of interest. In this study, the synergistic effects of antimicrobial peptides and antibiotics on each other were investigated. Within the scope of the research, machine learning modeling was carried out, and it was aimed to predict the agents whose synergistic effects have not been proven.

Keywords: Urinary catheters, antimicrobial peptides, antibiotics, machine learning, synergistic effect

Antimikrobiyal Peptid ve Antibiyotiklerin Sinerjistik Etkisinin Makine Öğrenmesi Modelleriyle Tahminlenmesi

Öz

Üriner kateterler mesanesini boşaltamayan hastalarda yaygın olarak kullanılmaktadır. Fakat, kateter yüzeyleri mikroorganizma adezyonuna uygunlardır ve bu durum, ilerleyen dönemlerde çeşitli komplikasyonlara neden olmaktadır. Komplikasyonların önlenmesine yönelik çeşitli kateter malzemeleri ve farklı yüzey modifikasyonları denenmiş, fakat bu yaklaşımların çoğu başarısız olmuş ve bazıları çeşitli yan etkiler göstermiştir. Bu nedenle, komplikasyonların önlenmesi ve tedavisi için yeni yöntemlere ihtiyaç duyulmaktadır. Antimikrobiyal peptidler geniş çaplı aktivite göstermeleri, ilaç direncine neden olmamaları gibi avantajlara sahip olmalarından dolayı kullanımları son zamanlarda popüler hale gelmiştir. Fakat, stabilite ve yüzeye bağlandıklarında aktivitelerinin azalması gibi dezavantajlardan muzdariplerdir. Bu nedenle, iki antimikrobiyal ajanın birlikte kullanımı ilgi çeken araştırma konularından biri haline gelmiştir. Bu çalışmada, antimikrobiyal peptidlerin ve antibiyotiklerin birbirleri üzerindeki sinerjistik etkileri araştırılmış, araştırmalar dahilinde makine öğrenmesi modellemesi yapılmış, ve sinerjistik etkileri kanıtlanmamış ajanların tahminlenmesi amaçlanmıştır.

Anahtar Kelimeler: Üriner kateterler, antimikrobiyal peptidler, antibiyotikler, makine öğrenmesi, sinerjistik etki

Acknowledgment

First of all, I would like to thank my advisor Assoc. Prof. Dr. Ozan Karaman for the knowledge he has added to me, and for his support. The extensive knowledge he had has been very valuable to me in my work.

I would like to thank Assoc. Prof. Dr. Utku Kürşat Ercan for the knowledge he has added to me both in my undergraduate education and in my graduate project.

I would like to thank Res. Asst. Gizem Dilara Özdemir, Akif Özdemir, Günnur Pulat and Ziyşan Buse Yaralı Çevik, for their supportive attitude in the laboratory environment and for teaching me step by step every experiment and procedure I have done.

Finally, I want to thank my family. I would not have been able to achieve all this if I did not have such a strong support group. I have felt their faith in me every moment of my life. I owe everything to my father Ali Rıza Olcay and my mother Zeynep Olcay, who never lacked their support, never left my side and made many sacrifices for me.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CAUTI	Catheter-associated Urinary Tract Infections
SRP	Slow Release Polymers
HDP	Host Defense Peptide
AMP	Antimicrobial Peptide
FIC	Fractional Inhibitory Concentration
AI	Artificial Intelligence
QM	Quantum Machines
ANN	Artificial Neural Networks
SVM	Support Vector Machines
DA	Discriminant Analysis
HMM	Hidden Markov Models
RF	Rain Forest
MIC	Minimum Inhibitory Concentration
CFU	Colony Forming Unit
LSTM	Long short-term memory
MDR	Multidrug-resistant
PAPs	Probable Antimicrobial Proteins
IG	Information gain
AAC	Amino Acid Composition
PseAAC	Pseudo Amino Acid Composition
ROC	Receiver Operating Characteristic
LR	Logistic Regression
LDA	Linear Discriminant Analysis

GPC	Gaussian Process Classifier
XGBC	Extreme Gradient Boosting Classifier
LGBMC	Light Gradient Boosted Machine Classifier
KNN	K-Nearest Neighbor
DTC	Decision Tree Classifier
ETC	Extra Tree Classifier
GNB	Gaussion Naive Bayes
BNB	Bernoulli Naive Bayes
BC	Bagging Classifier
ABC	AdaBoost Classifier
HGBC	Histogram Gradient Boosting Classifier
RFC	Random Forest Classifier
GBC	Gradient Boosting Classifier
MLPC	Multilayer Perceptron Classifier
AUC	The Area Under the Curve

Chapter 1

Introduction

1.1 Urinary Catheters and Associated Infections

Urinary catheters are hollowed and moderately flexible tubes. They are designed with the intention of draining urine from the bladder Despite the precautions aimed at avoiding contamination, catheters are prone to infections because they allow uropathogens to enter the urinary system, compromising the bladder's local host defense mechanisms [1,2]. Infections and complications including encrustation, bacteriuria, bladder stones, septicemia, and endotoxic shock are caused by opportunistic pathogens, which are primarily fecal or skin microbiota from subjects that can get into the bladder via the catheter lumen or through the catheter — urethra interface[3–5]. Catheter-associated urinary tract infections (CAUTIs) account for 27 percent of hospital infections in industrialized nations, with over 1 million cases occurring in the United States and Europe [6,7].

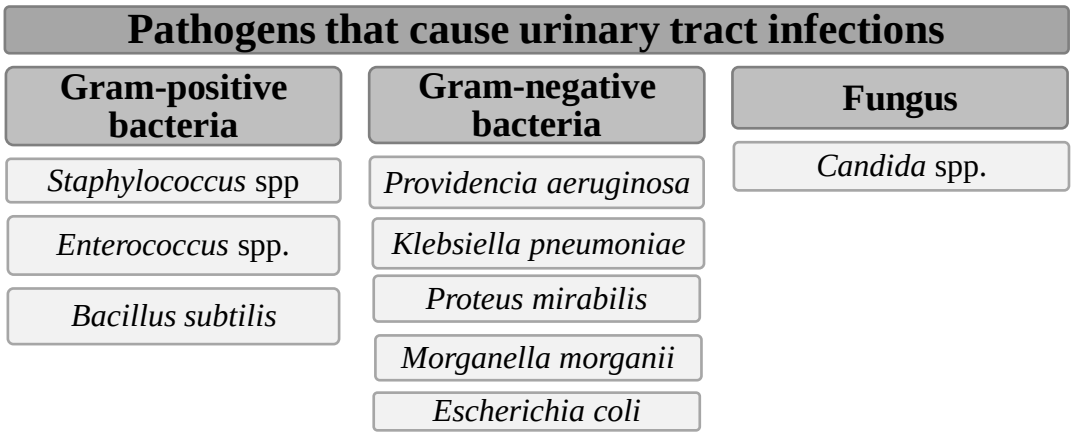

Figure 1.1 Pathogens that cause Urinary Tract Infections

The formation of biofilms is essential in infection development and progression. After the catheter is inserted, a film of organic molecules, electrolytes, and host proteins accumulates on the surface of the catheter, serving as a target for bacterial fimbrial attachment. Bacteria that are free-flowing, also called planktonic, can attach to the catheter surface [8,9]. The first contact between surface and bacteria is reversible because it is motivated by poor hydrophobic forces. Nevertheless, over time, the adhesion becomes irreversible because bacterial adhesins attach to their targets that are on the surface in addition to bacterial exopolysaccharide secretion, laying the foundation for the biofilm [10]. Quorum sensing through bacterial cell-to-cell signaling leads to mature biofilm formation with intricate structures and intertwined channels that allow homeostasis [11]. Following that, the bacteria secured in the biofilm or portions of the biofilm can detach, turn into planktonic state and migrate elsewhere, colonize new environments and reinitiate the biofilm formation [12]. Biofilms are difficult to destroy because of their matrix chemistry and can promote transfer of genes between resistant and non-resistant bacteria, leading to a higher risk of antibiotic resistance development in biofilms than that in planktonic cells. As a result, biofilms serve as reservoirs for the proliferation of pathogens, infections, and the development of resistance [13,14]. Furthermore, biofilms provide survival benefits to bacteria by allowing them to avoid shear stresses and evade phagocytosis.

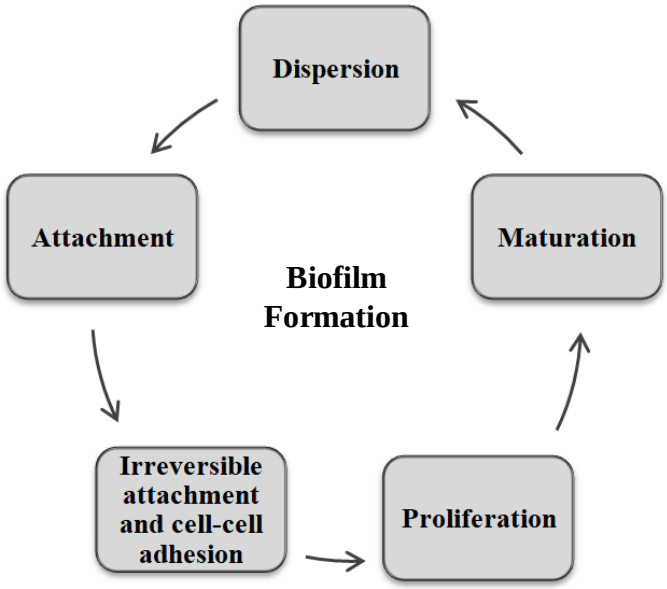

Figure 1.2 Cycle of biofilm formation

1.2 Methods Tried to Prevent Biofilm Formation and Infections

Microbial colonization on catheters is quite prevalent. To avoid such problems, scientists have developed many strategies for designing antimicrobial biomaterials. These methods are broadly classified as surfaces that repel protein and bacterial adsorption [15,16] and surfaces that are conjugated with antimicrobial compounds to induce cell death [14]. But, these methods have several limitations, such as inadequate antimicrobial agent concentration due to the biomaterials' low affinity for antimicrobial compounds, unspecific chemical reaction techniques for conjugation, narrow activity range, and cytotoxicity [17].

Silver has been shown to slow or prevent the formation of a biofilm. Particles of silver that enter the bacteria induce denaturation of cell proteins, leading to the dissociation of iron-sulfur clusters [18]. As a result, the iron component causes oxidative stress on pathogens, resulting in cell death. Multiple clinical studies have been conducted to date on the antibacterial efficiency of silver-coated catheters. However, studies revealed that silver-coated catheters were inefficient at preventing infections. Also, argyria as a result of extended use is one of the possible problems with silver coatings [19].

For biofilm prevention, urinary catheters coated with several antibiotics such as gentamicin, nitrofurazone, vancomycin, and rifampin have been developed. While this strategy was demonstrated to be successful for short-term application, it was complicated by the uncontrolled release profile of the antibiotic, which resulted in the release of high concentrations which may first harm the cells, and then it is followed by non-inhibitory concentrations. The release of antibiotics at suboptimal concentrations may raise the probability of drug-resistant microorganisms [20]. Given that this would not effectively kill all the bacteria, it will lead to a future infection that would be harder to eradicate because of resistance development. For these reasons, the application of antibiotics alone has limited effectiveness in preventing catheter-associated infections [6,21,22].

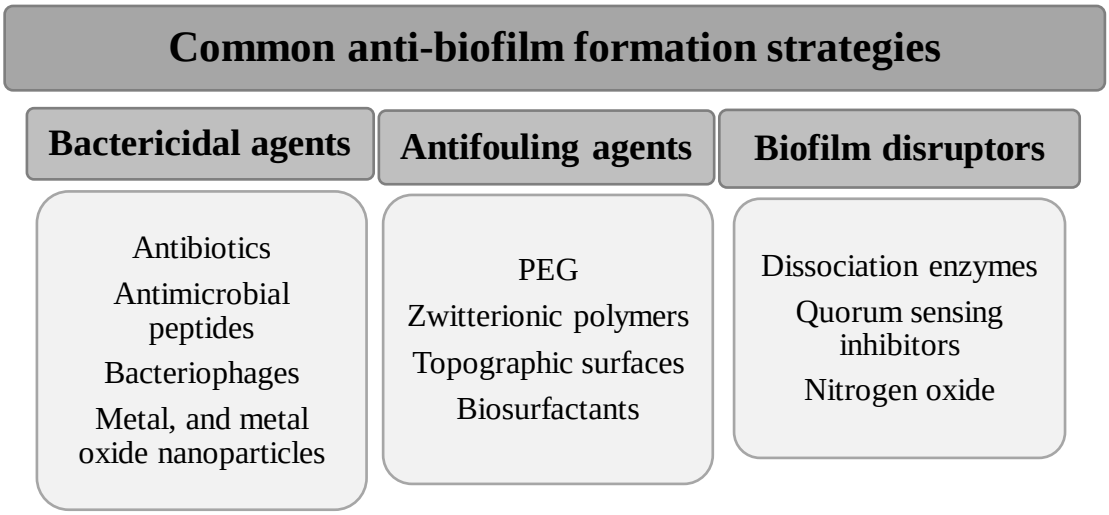

Figure 1.3 Common anti-biofilm formation strategies

Enzymes as key elements of antimicrobial coatings are being tested. Although the results demonstrated various advantages, antimicrobial enzymes have drawbacks such as protein denaturation under harsh conditions, and higher manufacturing and preparation costs as compared to antibiotic and silver coatings [18].

Bacteriophages are bacteria's natural predators. Bacteriophages selectively infect bacteria and disrupt numerous metabolic routes. Lytic phages penetrate, rapidly divide, and lyse bacteria. However, bacteriophages have a limited activity spectrum, and bacteria may develop resistance to them [23,24].

Slow release polymers (SRP) have been examined as a potential antibacterial agent source at sustained levels [25]. Because these substances are entirely soluble in water in their glassy form, they can release any components integrated into them over time [25]. However, they cause nonuniform release of the antimicrobial agent.

Heavy metals, quaternary ammonium salts, and antimicrobial peptides are all potential candidates for bactericidal compounds loaded on drug carriers [26]. Heavy metals and quaternary ammonium salts, on the other hand, may have limited efficiency, a narrow antibacterial spectrum, significant cytotoxicity, and the potential for drug resistance. Antimicrobial peptides can overcome the restrictions outlined above [26–28].

1.3 Antimicrobial Peptides

Antimicrobial peptides (AMPs) are a subcategory of host defense peptides (HDPs). HDPs can demonstrate a wide range of actions, sometimes within the same short peptide [29]. The majority of these actions provide direct such as anti-biofilm and antimicrobial or indirect such as immunomodulatory/anti-inflammatory defense against pathogens. Despite their diverse origins and functions, most natural HDPs exhibit common characteristics [29,30]. The sequence is usually made up of 12–50 amino acids. Their structure is related to a large number of hydrophobic residues and a net positive charge because of the presence of numerous cationic residues such as Arginine and Lysine, enabling them to fold into amphipathic forms [30]. AMPs can engage with bacterial membranes non-specifically due to their amphiphilic nature and positive charge, and AMPs have a low potential to induce drug resistance in bacteria [31].

The Barrel-stave
• The peptides pierce the membrane and create a pore, with the hydrophobic portions of the peptides facing the lipid core as well as the hydrophilic portions outlining the pore's interior.
The Toroidal pore
• The peptides cause the molecules of phospholipid layer to bend inwards, resulting in a pore formation with the hydrophilic portions of the peptides and the lipid parts of the layer molecules facing the pore's interior.
The Carpet
• Because of the parallel positioning of peptides regarding the lipid bilayer surface as well as the peptide carpet formation, the membrane is disrupted.

Figure 1.4 The main models of the action mechanisms of AMPs

Maintaining peptide activity and stability under physiological conditions is a critical need for optimum efficacy. The AMP stability under physiological conditions is determined by their susceptibility to enzyme degradation and inhibition by proteins, salts, and ions found in the environment [32–34]. Bacteria may protect themselves from AMPs by producing peptide degrading enzymes. Furthermore, mammalian

digestive enzymes like chymotrypsin and trypsin can lead to amino acid cleavage that are critical for the function of peptides [33].

AMPs can be coated or incorporated on the surfaces via ionic, covalent, physical trapping, and hydrogen binding interactions. Under hydrophobic conditions, free-form AMP with an amphipathic structure exhibits its highest antimicrobial activity. However, when the AMP is chemically bonded to the surface, it becomes less effective [35]. With varying degrees of effectiveness, several AMPs have been covalently linked onto the surface of diverse biomaterials. Nevertheless, the antimicrobial activities of many peptides are significantly reduced after being covalently linked onto the surface [14].

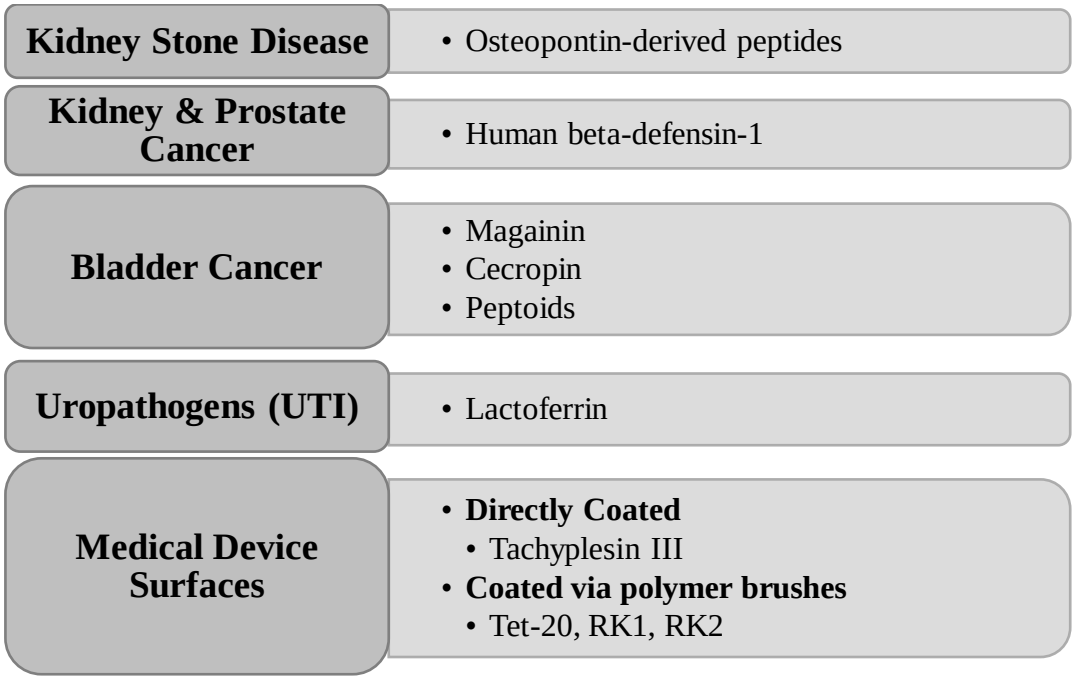

Figure 1.5 Widely used AMPs in medicine

Combining AMPs with antibiotics, other AMPs, or entirely different antimicrobial agents may allow the targeting of various aspects of biofilm development, improving treatment efficacy, reducing the effective concentration of antimicrobial agents that the treatment requires, and overall reducing the treatment cost [36].

The fractional inhibitory concentration (FIC) test is used to evaluate the interaction between drugs that are planned to be used together. The goal is to test new antimicrobial compounds in combination with existing ones to see if they have a synergistic, additive, or antagonistic effect. FIC method is easy to perform, is

adaptable to automated or semi-automated platforms, is cost-effective, and repeatable [37].

1.4 Artificial Intelligence

The introduction of modern technology advances in artificial engineering has altered the prospects of biomedicine. The scale of human error has stimulated further investment in technological development in the areas of diagnostics and intelligent designs. In biomedical engineering, Artificial Intelligence (AI) subfields are being applied to solve complicated problems [38].

AI is the intelligence displayed by machines as opposed to natural intelligence displayed by human beings and other living things. In simpler terms, AI term is used when a machine demonstrates cognitive functions associated with human qualities such as learning, problem-solving, perception, reasoning and more [39].

1.5 Machine Learning

Machine learning is a subfield of AI and is described as the algorithm studies that enable machines to make decisions and operate without being specifically programmed to do so. The goal of machine learning is to create algorithms that guide machines on how to access data and utilize it to learn a task [38].

In accordance with the 5-step rule of Chou [40], the following five principles should be followed to develop a predictor :

1. Creating or selecting a dataset in order to train & test the predictor
2. Formulating the samples with a mathematical expression that properly reflects their intrinsic association with the target that will be predicted
3. Introducing or developing an algorithm in order to operate the prediction
4. Performing cross-validation tests to assess the predictor's expected accuracy
5. Creating a public-accessible web service for the predictor

The learning process begins with observing or data, which is then used to build a knowledge base and then using it to detect patterns and make decisions for problems

brought to it [41]. Learning is the most crucial part of this process. Based on the training set utilized and how it is interpreted for the learning process, learning can be categorized into two categories such as supervised, and unsupervised learning [38].

Table 1.1 Machine Learning Techniques and Principles

Technique	Working Principle

Supervised Learning	• Uses labeled example data from previous experiences to predict future events with new data.
	• A known dataset is investigated throughout the training process.
	• The learning algorithm generates a function to anticipate output values for the given problems.
	• Supervised learning can be divided into two kinds of problems: Classification and Regression.
	  ○ **Classification Algorithms :** Linear Classifiers, Decision Trees, Support Vector Machines, and Random Forest.
	  ○ **Regression Algorithms :** Logistic Regression, Linear Regression, and Polynomial Regression.

Unsupervised Learning	• Used for providing a form to random data and finding meaning in such data. • For training data, the algorithm learns from unlabeled examples and related target responses that include numerical values or text labels. • When faced with new problems, the scheme attempts to restructure the new data in the form of previously processed data, generating the same patterns as scanned in the training set to achieve the problem output.

Machine learning has been commonly used in the case of structural and functional characterizations of proteins and peptides [42,43]. The properties obtained from the primary structures of proteins and/or peptides are used to predict secondary/tertiary structures, and diverse functions of biomolecules such as anticancer, antibacterial, and biofilm inhibition. As protein/peptide descriptors for classification, grouping, and regression, a number of sequence-based features have been presented. The physicochemical characteristics such as polarity, hydrophobicity, molecular weight, and polarizability have been used in a variety of applications [44].

1.6 Computational Prediction of AMPs

Progress in AMP studies has fueled ongoing efforts to develop computational approaches for accurate AMP prediction, with the goal of significantly reducing the effort and time required for experimental identification [45,46]. Computational prediction of AMPs provides a supportive technique to the time-consuming and labor-intensive experimental characterization of AMPs by shortlisting potential AMP candidates for later experimental validation [47]. To date, various computational methods for the assessment, prediction, and design of new AMPs have been developed. AMPer [48], AntiBP [49], CAMP [50], AVPpred [51], iAMP-2L [52], AntiBP2 [53], BIPEP [54], EFC-FCBF [55], AmPEP [56], ClassAMP [57], and DBAASP [58] are a few examples. The dataset size, quality of data, primary algorithms, extraction of features, feature selection methods, and evaluation techniques used by these systems vary significantly [47].

Some databases are devoted to certain AMP classes. For instance, PhytAMP [59], AMSdb, PenBase [60], and BACTIBASE [61] are AMP databases from plants, eukaryotes, shrimps, and bacteria, respectively. While RAPD [62] is focused on recombinant AMPs, SAPD [63] is focused on synthesized AMPs. The Peptaibol Database [64] and the Defensin Knowledgebase [65] are focused on peptaibols, and defensins, respectively. These databases are quite helpful when searching for AMPs that belong to certain classes.

The AntiBP server makes predictions about active AMPs utilizing Quantum Machines (QM), Artificial Neural Networks (ANN), and Support Vector Machines

(SVM) based on peptide sequence binary patterns [49]. The AntiBP2 server uses SVM to predict AMPs based on amino acid composition (AAC)-based properties [53].

CAMP is a pioneering tool that incorporates prediction algorithms such as SVM, RF, ANN, and Discriminant Analysis (DA) into the database [50]. For the training of the aforementioned predictors, the CAMP employs a variety of physicochemical features. CAMP has approximately three times the amount of sequences than APD [66] and provides extra information about taxonomy and activity. CAMP's data is split into datasets such as experimentally validated and predicted. On the test datasets, the prediction models achieved accuracies of 93.2 % in the case of RF, 91.5 % in the case of SVM, and 87.5 % in the case of DA [67]. CAMP's search features allow you to search across all or each of the datasets.

The AMPer has built Hidden Markov Models (HMMs) for each AMP class, including cathelicidins, defensins, and cecropins. The model is trained by using existing data from known class members [48].

The AVPpred web server is the first to predict antiviral peptides. This algorithm utilizes peptide sequence descriptors such as alignment and motif related characteristics, AAC, and physicochemical properties to train an SVM classifier [51].

The ClassAMP algorithm employs RF and SVM to assess a peptide sequence's propensity for antifungal, antibacterial, and antiviral activities [57]. Aminoacid, dipeptide, tripeptide compositions, and other physicochemical properties were utilized as features to predict the activities, and three one-against-all classifiers were constructed [47].

The iAMP-2L is a two-level classifier based upon the pseudo amino acid composition (PseAAC) structure and fuzzy K-nearest neighbor algorithm. It first determines active AMP sequences and then assigns peptide functionality [52].

C-PAmP is a high-scoring database that involves predicted AMPs for a great number of plant species. C-PAmP includes 15,174,905 peptides that are 5 to 100 amino acids long and are derived from more than 33 thousand proteins [68]. This database

identified AMPs by using the PseAAC and five quantitative descriptors converted from 237 physicochemical amino acid descriptors [69–71].

ADAM is a public AMP database that was created to establish extensive associations between peptide sequences and the structures and to make it easy to view their relations. It contains 7,007 distinct peptide sequences and 759 structures. Also, it offers two computational tools for predicting AMPs based on HMM and SVM [72].

iAMPpred is a tool to predict antiviral, antibacterial, and antifungal peptides based on three categories of features such as compositional, structural, and physicochemical. It is built based on three SVM models. The information gain for each feature was computed to determine the significance of each feature in predicting antiviral, antibacterial or antifungal peptides [73].

MLAMP is a two-level AMP predictor based on unbalanced data sets. The predictor employs ML-SMOTE and gray PseAAC to predict AMPs and their functions. The RF algorithm is used at the first prediction level to predict if a peptide is an AMP or not. A classifier based on the RF algorithm is employed for the second level. ML-SMOTE, an oversampling model, is applied to account for imbalanced functional types of AMPs [74].

MAMPs-Pred is another two-level model that uses the RF algorithm to predict AMPs and their functional characteristics. SVM-Prot was used to calculate 188 features for AMP prediction, which were based on eight types of amino acid characteristics and physicochemical properties. For data balance, random undersampling and weighted random sampling methods were utilized [75].

AMPfun is a web server for identifying AMPs and their activities based on their functions. It is a two-stage structure with three steps in each stage such as the calculation of the features, the selection of the features, and classification algorithm applications. For feature selection, the sequential forward selection technique was utilized, while RF was utilized as the prediction strategy [76].

Table 1.2 Computational approaches for AMP prediction

YEAR	APPROACH	REFERANCE
2007	AMPer	[48]
2010	CAMP	[67]
	Porto *et* al.	[77]
2011	Song *et* al.	[78]
	Torrent *et* al.	[79]
2012	ClassAMP	[57]
	CS-AMPPred	[80]
	Veltri *et* al.	[81]
	Fernandes *et* al.	[82]
2013	C-PAmP	[68]
	iAMP-2L	[52]
	Randou *et* al.	[81]
2014	DBAASP	[83]
2015	ADAM	[72]
	Camacho *et* al.	[84]
	Ng *et* al.	[85]
2016	MLAMP	[74]
2017	MOEA-FW	[86]
	iAMPpred	[73]
2018	AmPEP	[56]
	AMP Scanner V2	[87]
2019	APIN	[88]
	AMAP	[89]
	MAMPs-Pred	[75]
	dbAMP	[90]
	AMPfun	[76]
2020	Deep-AMPEP30	[91]
	AMPlify	[92]
	Fu *et* al.	[93]

The use of catheters leads to bacterial colonization. The colonizing bacteria form a biofilm layer, which causes infections. Many methods have been tried to control and prevent CAUTIs, but these methods were unable to achieve the expected success as well as have shown various negative effects, such as the emergence of resistant organisms and toxicity. In addition, AMPs can exert a strong antimicrobial effect on pathogens that have multidrug resistance and cause nosocomial infections. But it has been proven in various studies that it is difficult to maintain the stability of AMPs in physical conditions and also that their activity decreases when they are bound to a

surface. For these reasons, it is expected that treatment will be more successful with the combined use of two antimicrobial agents and that the side effects caused by the use of high concentrations will decrease.

Although there are many machine learning algorithms that predict various functions of AMPs, after a review of the literature conducted by us, it was found that there is no machine learning algorithm that predicts the synergistic effects of antibiotics and AMPs. Considering the lack of this issue in the literature, this study aims to predict the synergistic effect of various antibiotics with various AMPs by predicting the FIC index. The existence of such a model is thought to save researchers from wasting time and resources in the laboratory on an antibiotic - AMP combination that would not work.

In this study, data on the interaction between antibiotics and AMPs were collected. Various preprocessing applications were performed on the data. Based on the final data, different classifiers and machine learning models were tested, and the final model was developed by choosing the classifier and the normalization method with the highest accuracy. The accuracy of the model was evaluated by interpreting it in terms of various parameters. As a result of the analyzes made, it was concluded that the predictive success of the model was high and that it gave promising results for future studies.

Chapter 2

Materials & Methods

2.1 Data Collection

All data were collected from the **DBAASP** and **DrugBank** database. When collecting data from DBAASP database;

1. Sequences containing intrachain and coordination bonds were excluded.
2. Sequences containing unusual amino acids were excluded.
3. The C terminal was determined to be H (without modification), and the N terminal was determined to be amide (NH2). Sequences with different terminal modifications were excluded.
4. Sequences that are 90 percent or more similar to each other were excluded in order not to decrease the accuracy of the algorithm.

Table 2.1 Selections made in the DBAASP peptide database

Selection Criteria	**Selection**
Synthesis Type	Synthetic & Ribosomal
N Terminus	Without N Terminus Modification
C Terminus	Amide (NH_2)
Unusual Amino Acid	Without Modification
Intrachain Bond	Without Intrachain Bond
Coordination Bond	Without Coordination Bond
Synergy	All with data on Synergy

Table 2.2 Input names and Types

INPUTS (Predictors)	TYPE
Sequence Length	Numerical
Molecular Weight of the Sequence	Numerical
Normalized Hydrophobicity	Numerical
Net Charge	Numerical
Isoelectric Point	Numerical
Penetration Depth	Numerical
Tilt Angle	Numerical
Disordered Conformation Propensity	Numerical
Linear Moment	Numerical
Amphiphilicity Index	Numerical
Average Hydrophilicity	Numerical
Ratio of hydrophilic residues to total number of residues	Numerical
Target Species	Nominal
Gram stain of the Target Species	Nominal
Antibiotic Name	Nominal
Gram stain of the species in which the antibiotic is active	Nominal
Class of the Antibiotic	Nominal
Molecular weight of the Antibiotic	Numerical
Charge of the Antibiotic	Numerical
LogP	Numerical
Water Solubility	Numerical
pKa	Numerical
Mechanism of Action	Nominal
Activity of the Peptide Alone (MIC)	Numerical
Activity of the Antibiotic Alone (MIC)	Numerical

Table 2.3 Output names and types

OUTPUTS (Outcomes)	TYPE
Fractional Inhibitory Concentration (FIC) Index	Numerical

The predictor variables adopted for this study are divided into three categories. The first category includes antimicrobial peptide characteristics such as sequence length, molecular weight, normalized hydrophobicity, net charge, isoelectric point, penetration depth, tilt angle, disordered conformation propensity, linear moment, amphiphilicity index, average hydrophilicity, and ratio of hydrophilic residues / total number of residues (%). The second category includes antibiotic characteristics such as molecular weight, class, physiological charge, logP, water solubility, pKa and mechanism of action. Lastly, the third category includes bacteria characteristics such as Gram type, and MIC.

Table 2.4 Predictor Categories and Characteristics

Predictor Categories	**Predictor Characteristics**
Antimicrobial Peptide	Length Molecular Weight Normalized Hydrophobicity Net Charge Isoelectric Point Penetration Depth Tilt Angle Disordered Conformation Propensity Linear Moment Amphiphilicity Index Average Hydrophilicity Ratio Of Hydrophilic Residues / Total Number Of Residues (%).
Antibiotic	Molecular Weight Class Physiological Charge LogP Water Solubility pKa Mechanism of Action
Bacteria	Gram Type MIC

Table 2.5 Numerical Values and Units

Numerical Values	Units
Molecular Weight of the Sequence	g / mol
Molecular Weight of the Antibiotic	g / mol
Activity (MIC)	µg / ml
Activity of the Peptide Alone (MIC)	µg / ml
Activity of the Antibiotic Alone (MIC)	µg / ml
Water Solubility of the Antibiotic	mg / ml

For two antimicrobial agents (A and B) ;

$$FIC_A = \frac{MIC_{(A\ in\ the\ presence\ of\ B)}}{MIC_{(A\ alone)}} \quad (2.1)$$

$$FIC_B = \frac{MIC_{(B\ in\ the\ presence\ of\ A)}}{MIC_{(B\ alone)}} \quad (2.2)$$

$$FIC_{Index} = FIC_A + FIC_B \quad (2.3)$$

Table 2.6 Fractional Inhibitory Concentration Index Values and Their Indications

Fractional Inhibitory Concentration (FIC) Index	
≤ 0.5	Synergism
> 0.5	No Interaction

2.2 Data Preprocessing

2.2.1 Normalization

Normalization is a scaling method in machine learning used during data preparation to adjust the values of numeric inputs in the dataset in order to use a similar scale [94].

Despite the fact that there are many normalization methods in machine learning, Min-Max scaling and Standardization scaling are the most commonly used. The Min-Max scaling technique helps the dataset in shifting and rescaling the values of their attributes such that they range between 0 and 1. Standardization scaling, also known as Z-score normalization, is a method in which the values are centralized around the mean with a unit standard deviation, resulting in the attribute becoming zero and the resultant distribution having a unit standard deviation [95].

Maximum Absolute (Max Abs) scaling operates by dividing each value by the largest value in that feature, regardless of its sign. This transformation provides a distribution with values ranging from -1 to 1 [96].

Robust data scaling or Robust standardization is a method for normalizing input variables in the presence of outliers. This method ignores the outliers from the calculations of the mean and the standard deviation, then scales the variable using the calculated values [96].

Data normalization was performed to improve model performance since magnitude range vary and can impact model optimization during training. With the four normalization methods mentioned above, models were developed separately and compared. Their effects on the accuracy of the model were evaluated.

2.2.2 One Hot Encoding

Most of the machine learning algorithms cannot operate with nominal data, also called categorical variables. These data must be converted into numerical values. One Hot Encoding is a binary representation of categorical variables. To begin, this step requires translating the values to integer values. Later, integer values are represented in the form of binary vectors, with all values being zero except the integer index, which is labeled as 1 [94]. One Hot Encoding makes categorical data representation easier and more expressive.

2.2.3 Resampling

If there is an imbalance in the instance numbers that constitutes a class in a dataset, the expected outcomes will be affected when used as training data for machine learning. To fix the imbalance in the training data, resampling is commonly employed, which balances the number of instances [97].

The Synthetic Minority Oversampling Technique (SMOTE) is a technique to increase the number of instances in a balanced manner in a dataset. The component generates new instances from existent minority instances that are provided as input [98].

2.3 Data Splitting

Data splitting is a method that is widely used in machine learning. In order to train the model and test the performance, the data is split as training and test sets [94].

The goal of the rational splitting algorithms is to choose the most representative group for the training set. In order to select samples, the similarity between the data points or data distribution is used [99].

Random splitting algorithms pick a number of samples randomly as the training set, while the remaining samples are used as the test set [99]. In this study, the dataset was randomly split into two sets: a training set (containing 75% of the data) that was used to train the model and a test set (containing 25% of the data) that was used to test the accuracy of the model.

2.4 Model Development

Decision Tree is a classifier in which internal nodes represent the features of datasets, branches represent decision rules, and leaf nodes represent the outputs. The purpose of employing a decision tree is building a training model which can predict the class and/or value of a target variable by learning basic decision rules from the training data [100]. Decision trees categorize samples by descending the tree from

the root to leaf/terminal nodes, with the leaf/terminal nodes providing the example's categorization. Each node represents a test case for variable, and each edge going down from the node represents one of the test case's possible answers. This recursive approach is repeated for each subtree [101].

Light GBM (LGBM) is a gradient boosting framework based on the decision tree algorithm that can be used for classification, ranking, and a variety of other machine learning applications [102].

The Adaptive Boosting or shortly AdaBoost algorithm is a boosting approach used in Machine Learning as an ensemble method. The algorithm employs one-level decision trees called weak learners, which are successively added to the ensemble. Each model aims to correct the predictions generated by the model preceding it. This is accomplished by weighting the training dataset in order to focus more on training examples where previous models made prediction mistakes [103]. Adaptive Boosting is so named because the weights are re-assigned to each example, with larger weights applied to mistakenly categorized instances. Boosting method is regularly used to reduce bias and variation [104].

Extreme Gradient Boosting (XGBoost) is a Machine Learning technique that employs a gradient boosting (GB) framework. It may be used to solve problems including classification, ranking, regression and user-defined prediction [105]. Gradient-boosted decision trees train an ensemble of superficial decision trees repeatedly, with each iteration utilizing the prior model's error residuals to fit the next model. The final prediction is the weighted total of all predictions. Boosting reduces bias and underfitting [106].

Rain Forest is a well-known machine learning algorithm, and it is a member of the supervised learning approach. It can be applied to regression and classification issues. RF uses decision trees on different subsets of the given dataset and averages the results to improve the accuracy of the prediction of the dataset [107].

SVM technique is applicable to both classification and regression problems. The SVM algorithm's goal is to identify a hyperplane in an N-dimensional space and clearly classify the data points [108].

Despite its name, logistic regression (LR) is a classification model rather than a regression model. It is also known as maximum-entropy classification (MaxEnt), the log-linear classifier, or logit regression. A logistic function is used in this model to describe the probability defining the probable outcomes of a single experiment [109].

Linear Discriminant Analysis (LDA) is a classifier that has a linear decision surface. The LDA classifier is appealing because it provides a closed-form solution that is simply computed, is intrinsically multiclass, has been demonstrated to operate well in practice, and does not have hyperparameters to modify [110].

The concept behind nearest neighbor approaches is to identify a preset number of training samples that are closest in proximity to a new point and anticipate the label based on them. The number of samples might be fixed, which is called k-nearest neighbor learning (KNN) [111], or variable depending on the density of points, which is called radius-based neighbor learning.

Gaussian processes are a supervised learning approach that can be used to tackle regression and probabilistic classification issues. The GaussianProcessClassifier (GPC) uses Gaussian processes for classification, especially probabilistic classification, where predictions are in the form of class possibilities. GPC can perform one-versus-one or one-versus-rest based training and prediction for multi-class classification [112].

Naive Bayes techniques are a type of supervised learning algorithm that employ Bayes' theorem with the "naive" assumptions of conditional independence between feature pairs given the class variable value. They require little training data for estimating the required parameters, and these classifiers can be very fast when compared to more sophisticated algorithms. [113]. There are many types of naive bayes, including Gaussian Naive Bayes (GNB) and Bernoulli Naive Bayes (BNB).

Bagging classifiers (BC) are ensemble meta-estimators that fit base classifiers on the original dataset subsets and aggregate their predictions through voting or averaging to generate a final prediction. BC is often used to minimize the deviation of a black-box estimator by incorporating randomization into its building mechanism and then constructing an ensemble from it [114].

Various models were developed using the algorithms described above. Accuracy scores were calculated and compared. Lastly, the algorithm with the highest score was chosen to develop the final model.

2.5 Hyperparameter Tuning

The process of determining the correct combination of hyperparameters that maximize the performance of a model is known as hyperparameter tuning. It works by running several trials in one training process. Once completed, the method will provide the set of hyperparameter values that are most suited for the model to provide optimal results [115].

2.6 Model Evaluation

The Confusion Matrix was created to visualize the correct and incorrect predictions. The receiver operating characteristic (ROC) curve was drawn according to the True Positive Rate and False Positive Rate. The success of the model was evaluated with the F1 score, Accuracy, Recall, and Precision measures. Additionally, the importance of the features was examined and the most important features were determined.

Chapter 3

Results

3.1 Data Collection

The data were collected from the DBAASP site by applying the necessary extraction criteria. Information on antibiotics was obtained from the DrugBank site. The FIC Index value was determined as the output.

Rows with missing values were removed. In total, 407 rows of data were collected. Some values were given in µM, while others were given in µg/ml. Values given as µM were converted to µg/ml by calculations for unit integrity. Inputs were arranged so that all data in a column have the same unit.

3.2 Data Preprocessing

Due to the fact that some of the entries were nominal, the nominal data were converted to numerical data using the One Hot Encoding method. Since the big gap between the smallest and largest values would affect the accuracy of the model, the values were rescaled by the normalization method. To train the model and test its performance, data were divided into training and test sets by the data splitting method.

Data distributions before SMOTE;

While the number of those who showed synergism (≤ 0.5) was 199, the number of those who did not interact (>0.5) was 208. 9 new instances were generated with SMOTE.

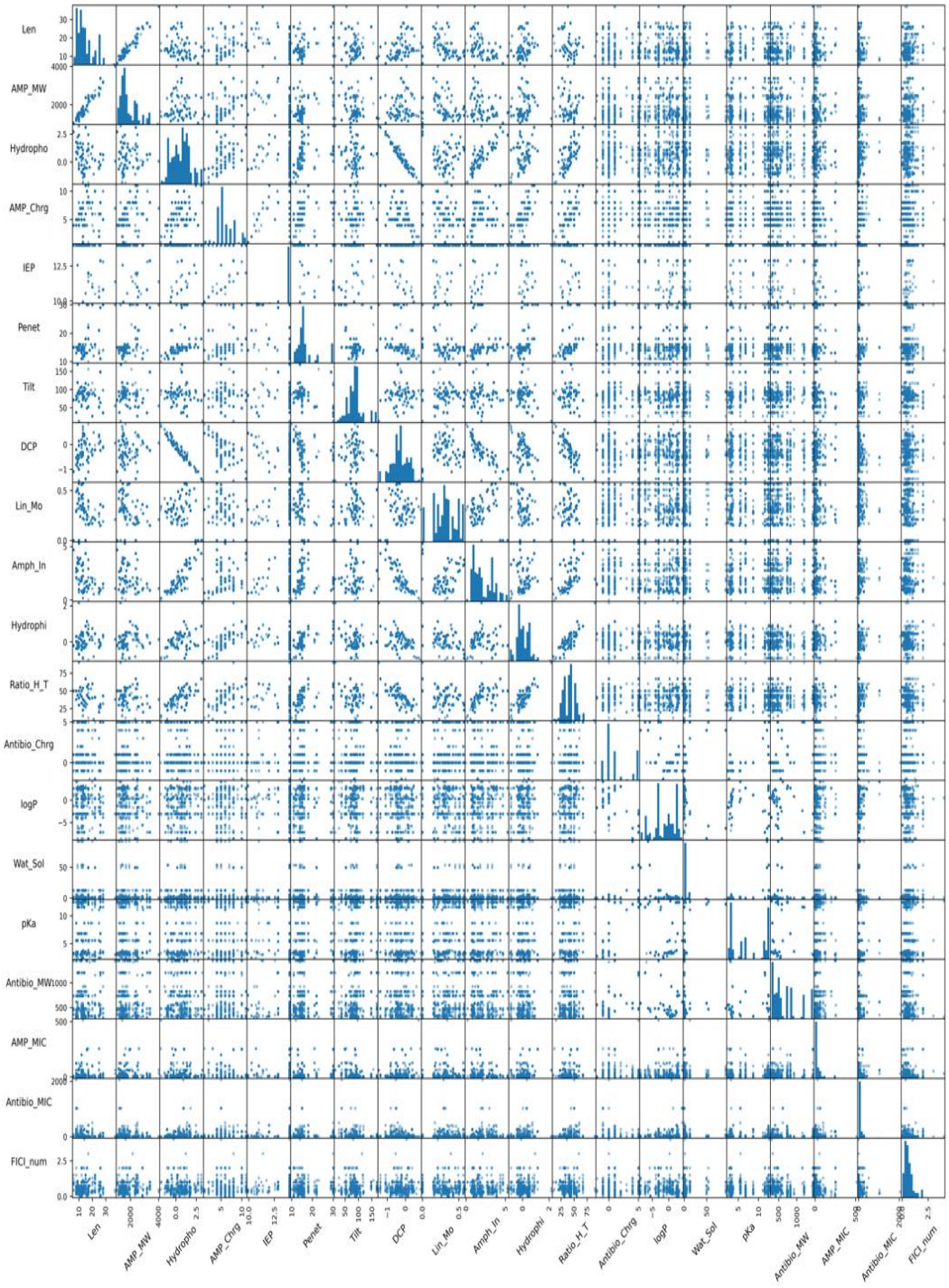

Figure 3.1 Scatter Plot Matrix

A scatter plot is a graph in which each value is illustrated by a dot. Scatter plots use dots to show how one variable affects another or the relation between them. Scatter plots plot data points on the x and y axes [116].

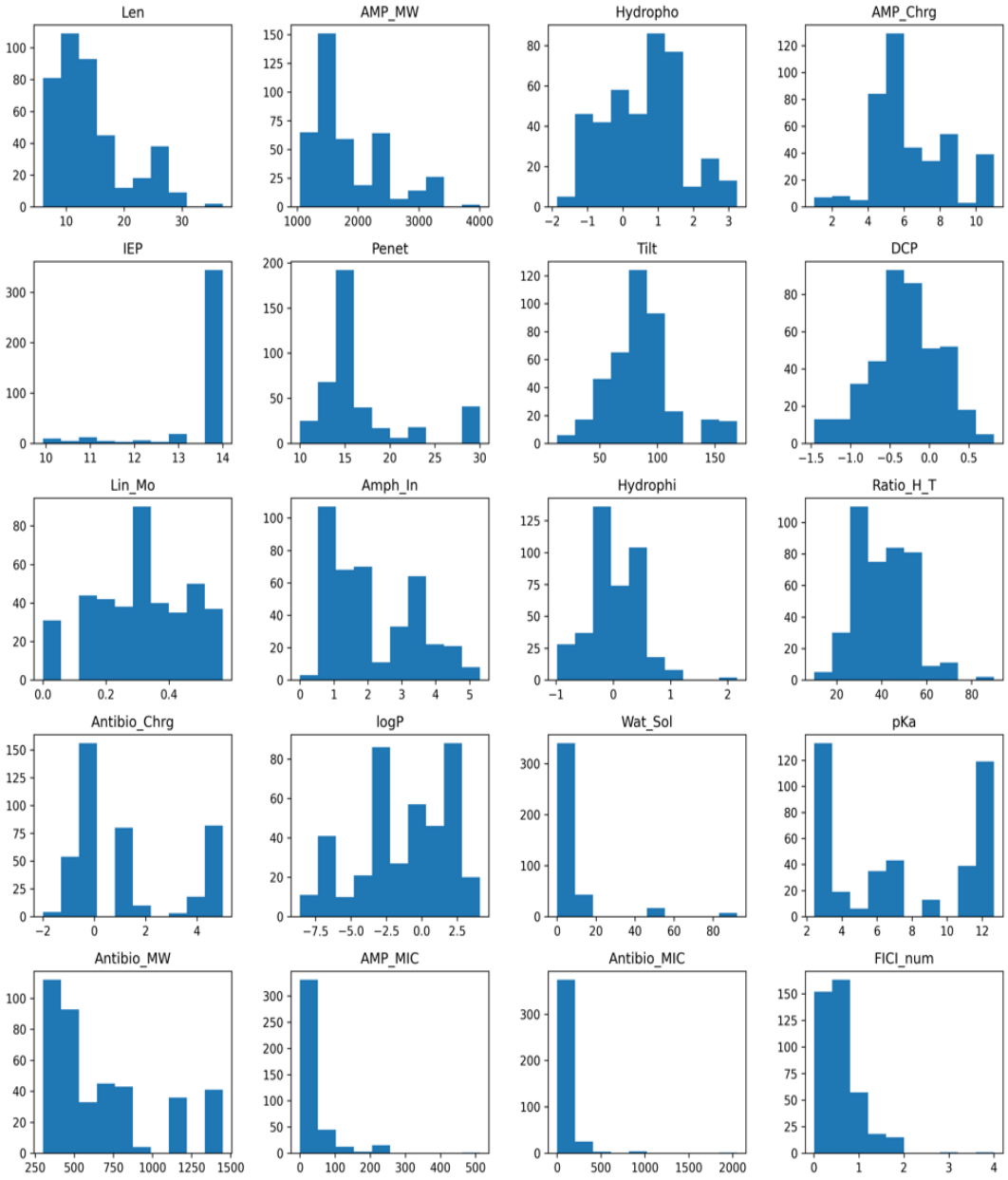

Figure 3.2 Training data histogram

A dimension is utilized to construct a histogram on each diagonal plot of the scatter matrix. A histogram is a type of 0D visualization that illustrates the data distribution that is on a single dimension. It displays the data distribution far more clearly and accurately than a plot of the dimension of data against itself. Histograms are useful for observing data distribution [117]. As can be seen in Figures 3.1 and 3.2, the original data distributions are wide-ranging and imbalanced.

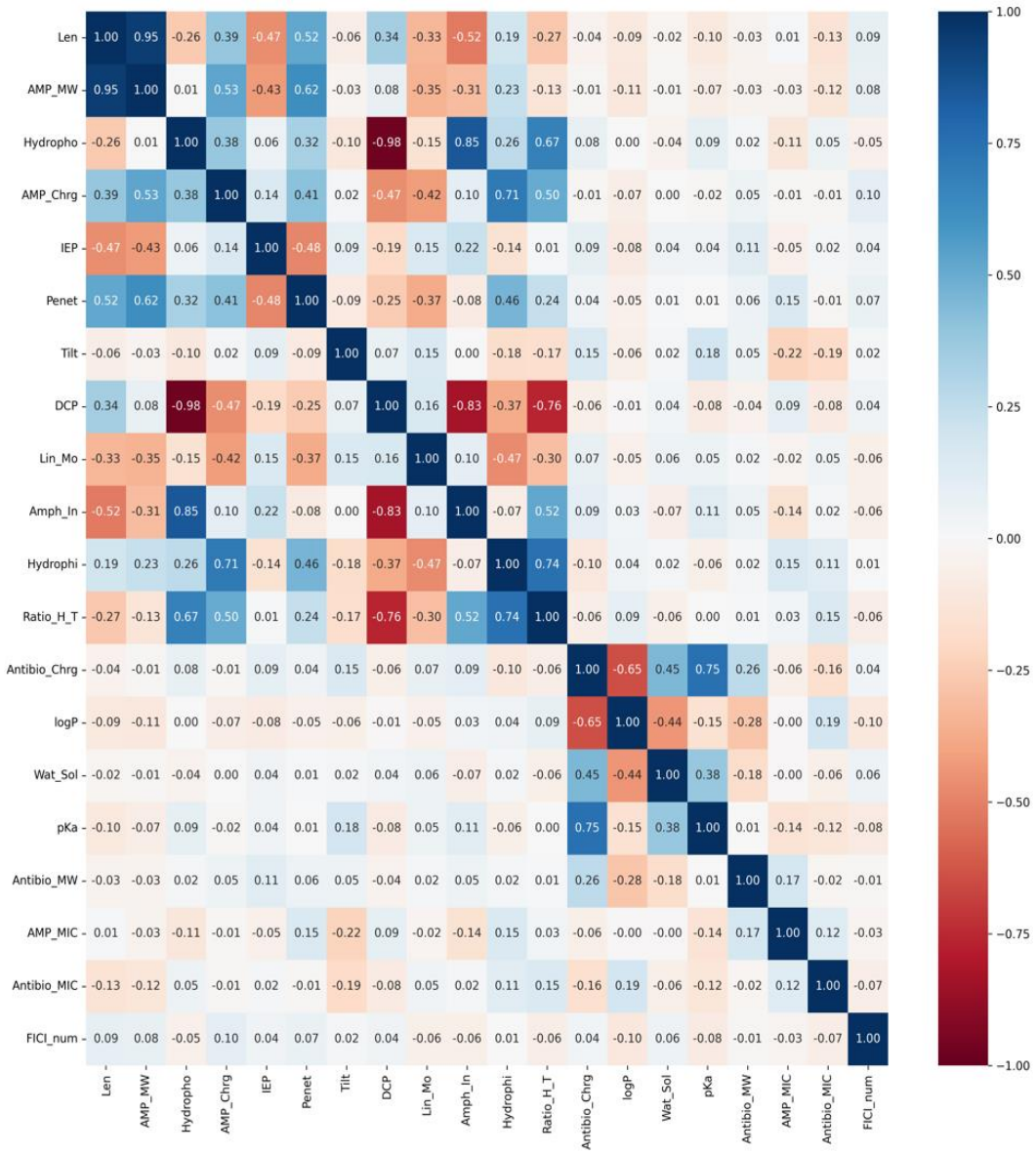

Figure 3.3 Correlation matrix

A correlation matrix is a table that illustrates the correlation coefficients for variables. The matrix shows the correlation between all possible pairings in a table. It is a strong tool for summarizing large datasets as well as identifying and visualizing patterns in the data. As can bee seen from Figure 3.3, it comprises of rows and columns displaying the variables. The correlation coefficient is contained in each cell of the table. Coefficients range between -1 and 1 where -1 represents a totally negative linear correlation, 0 represents that there is no linear correlation, and 1 represents a totally positive linear correlation. This means that the farther the coefficient from zero, the stronger the relationship between two variables. Looking at the figure, a strong positive correlation (0.95) was seen between the molecular

weight and length of AMP. However, a strong negative correlation (-0.98) was seen between the disordered conformation propensity and the normalized hydrophobicity of AMP, and also between the disordered conformation propensity and the amphiphilicity index of AMP (-0.83). There was a strong negative correlation (-0.65) between the charge and the logP value of the antibiotic, while a strong positive correlation (0.75) was observed between the charge and the pKa value of the antibiotic. In addition, a strong positive correlation was observed between the normalized hydrophobicity and amphiphilicity index of AMP (0.85), and between the charge and average hydrophilicity of AMP (0.71).

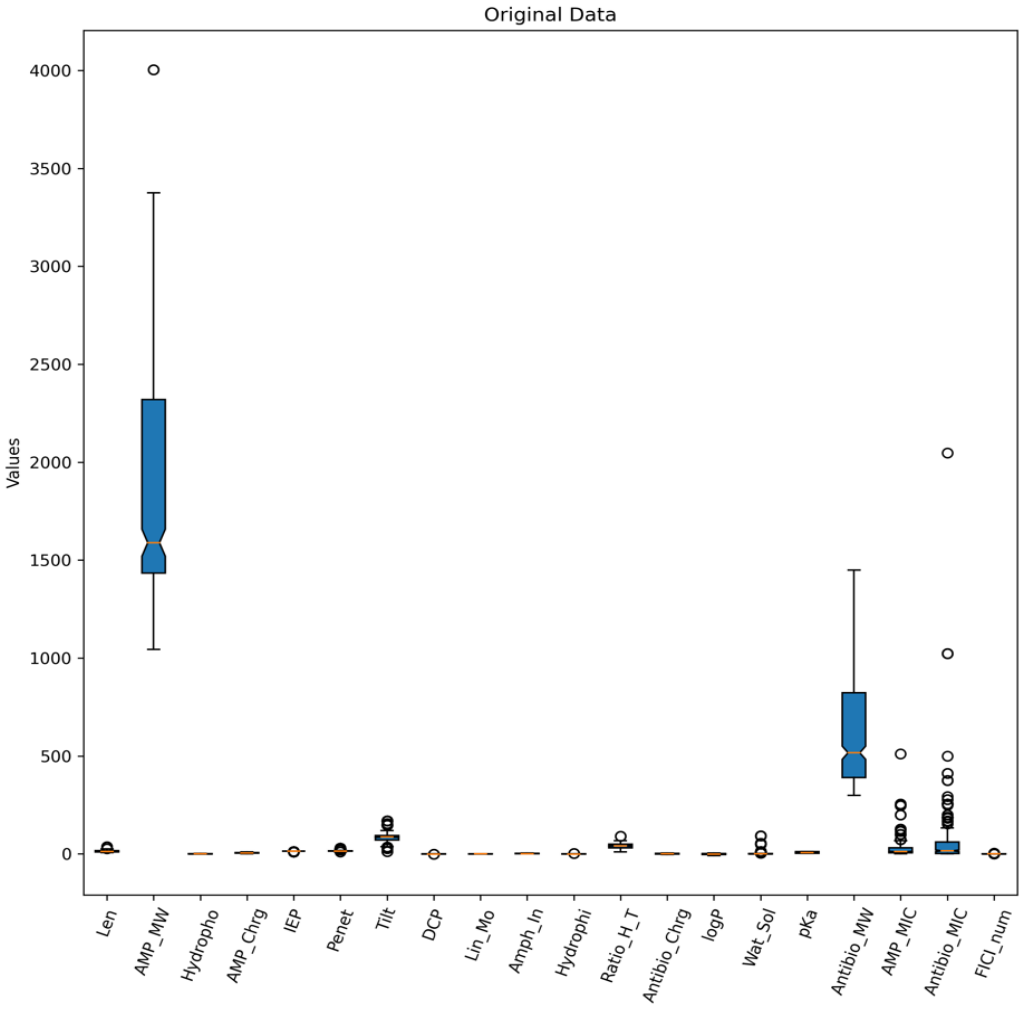

Figure 3.4 Original data value distribution

As can be seen from Figure 3.4, the original values were distributed over a wide range. Four different normalization methods, namely Z-score, Min-Max, Max-Abs, and Robust, were tried, and the data were fit into certain intervals specific to the method.

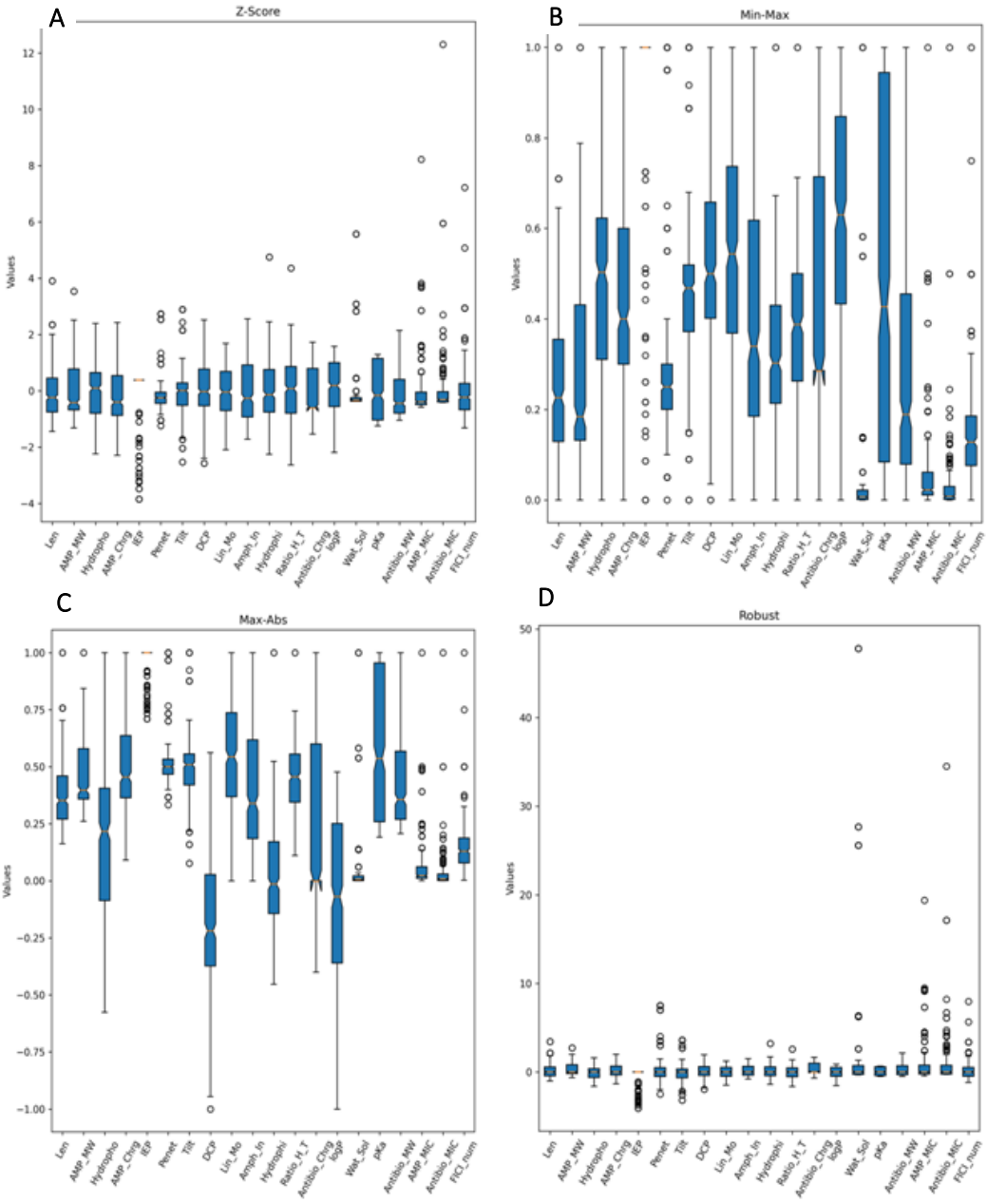

Figure 3.5 Normalized data by **A)** Z-score **B)** Min-Max **C)** Max-Abs **D)** Robust methods

Looking at Figure 3.5D, it can be seen that the method that can normalize the data distribution to the narrowest area is the Robust method. The extremes of the values decrease the accuracy of the model. Therefore, the closer the values are to each other, in other words, if they are distributed over a narrow area, the more successful the model will be. For these reasons, the Robust method was chosen in model development.

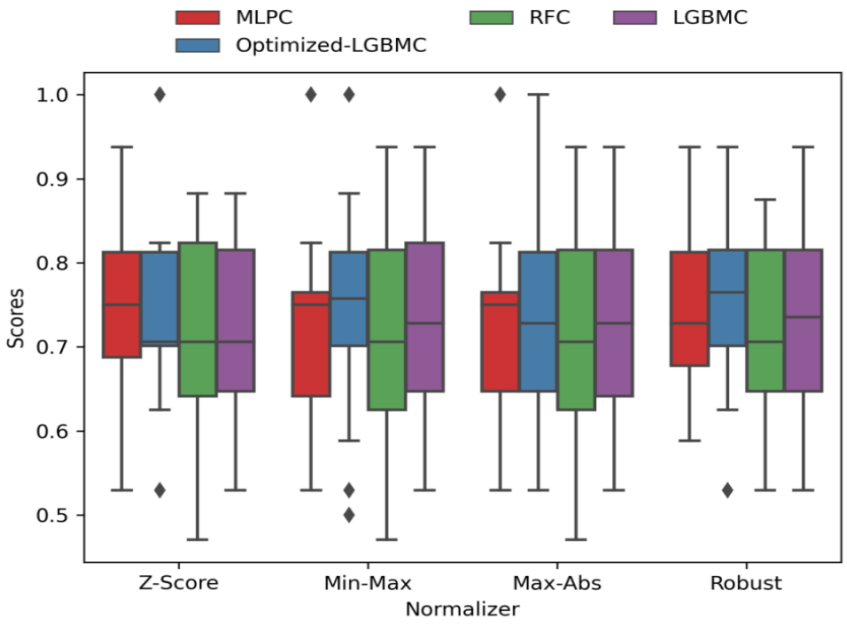

Figure 3.6 Normalizer results

Table 3.1 Normalizer accuracy scores

Normalization Strategy	**Classifier**	**Accuracy Scores**
Z-Score	MLPC	0.732 ± 0.095
	Optimized-LGBMC	0.732 ± 0.096
	RFC	0.717 ± 0.115
	LGBMC	0.723 ± 0.098
Min-Max	MLPC	0.718 ± 0.109
	Optimized-LGBMC	0.744 ± 0.119
	RFC	0.717 ± 0.120
	LGBMC	0.730 ± 0.107
Max-Abs	MLPC	0.714 ± 0.108
	Optimized-LGBMC	0.729 ± 0.104
	RFC	0.714 ± 0.119
	LGBMC	0.732 ± 0.107
Robust	MLPC	0.739 ± 0.090
	Optimized-LGBMC	**0.753 ± 0.097**
	RFC	0.720 ± 0.095
	LGBMC	0.739 ± 0.106

Looking at Figure 3.6 and Table 3.1, it was seen that the model using the LGMBC classifier and the Robust normalization method gave the highest accuracy score. Therefore, while developing the model, the model with the highest accuracy score was selected.

3.3 Model Development

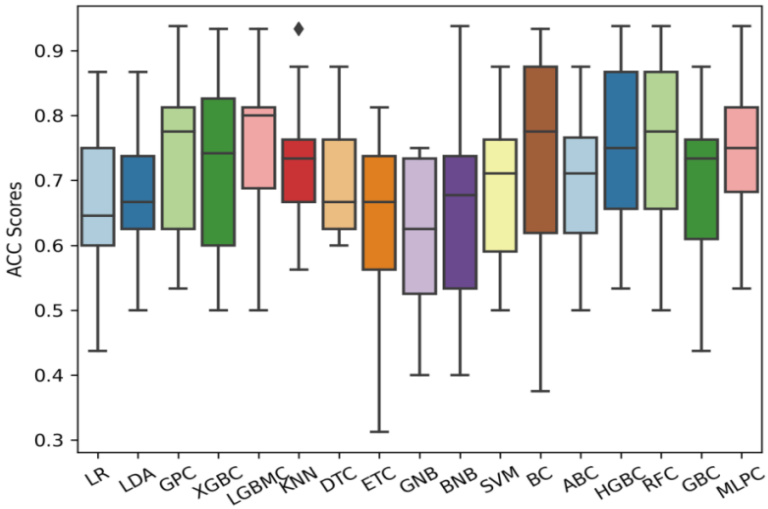

Figure 3.7 Accuracy scores of different classifiers

17 different classifiers were tried, and the accuracy scores of the models developed with these classifiers were measured. It can be seen from Figure 3.7 and Table 3.1 that the classifier with the highest accuracy score was LGBMC. For this reason, the LGBMC classifier was chosen for the model development.

Table 3.2 Accuracy scores of the classifiers

Model	**ACC Scores**
LR	0.667 ± 0.111
LDA	0.682 ± 0.089
GPC	0.739 ± 0.111
XGBC	0.727 ± 0.129
LGBMC	**0.757 ± 0.103**
KNN	0.728 ± 0.094
DTC	0.705 ± 0.094
ETC	0.638 ± 0.116
GNB	0.619 ± 0.107
BNB	0.659 ± 0.150
SVM	0.689 ± 0.116
BC	0.741 ± 0.152
ABC	0.701 ± 0.112
HGBC	0.746 ± 0.124
RFC	0.749 ± 0.121
GBC	0.696 ± 0.116
MLPC	0.753 ± 0.099

3.4 Model Evaluation

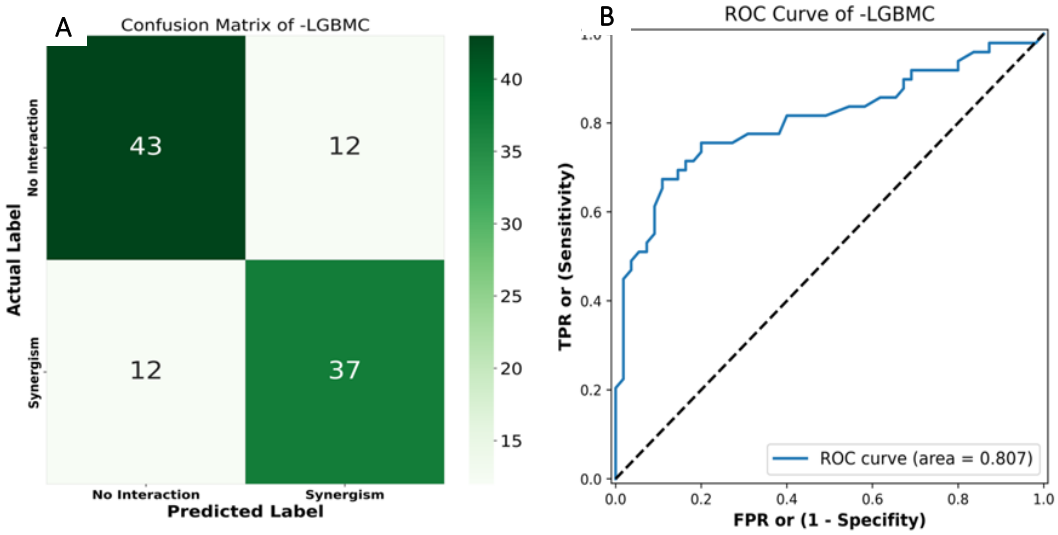

Figure 3.8 **A)** Confusion matrix and **B)** ROC curve of the LGBMC model

As can be seen from the Figure 3.8A, there were 55 data with an FIC index greater than 0.5 (No Interaction). The model predicted 43 of them correctly (true positive) and 12 of them incorrectly (false positive). Also, there were 49 data with an FIC index less than or equal to 0.5 (Synergism). The model predicted 37 of them correctly (true positive) and 12 of them incorrectly (false positive). If expressed as a percentage, the model correctly predicted 78.2% of the data with no interaction, and correctly predicted 75.5% of the data with synergistic effects.

ROC curve is a graph that depicts model performance at all classification thresholds. This curve depicts two parameters: True Positive Rate (TPR), and False Positive Rate (FPR) [118]. AUC is an abbreviation for "Area Under the ROC Curve." AUC, in other words, measures the whole two-dimensional area under the ROC curve. AUC values vary from 0 to 1 [118]. As can be seen from the Figure 3.8B, the AUC value in this study was 0.807.

The effect of the importance of predictors on the performance of the model was measured in terms of information gain. The initial accuracy of the model was 72.4%. After hyperparameter tuning was performed, the accuracy of the model reached 75.4%.

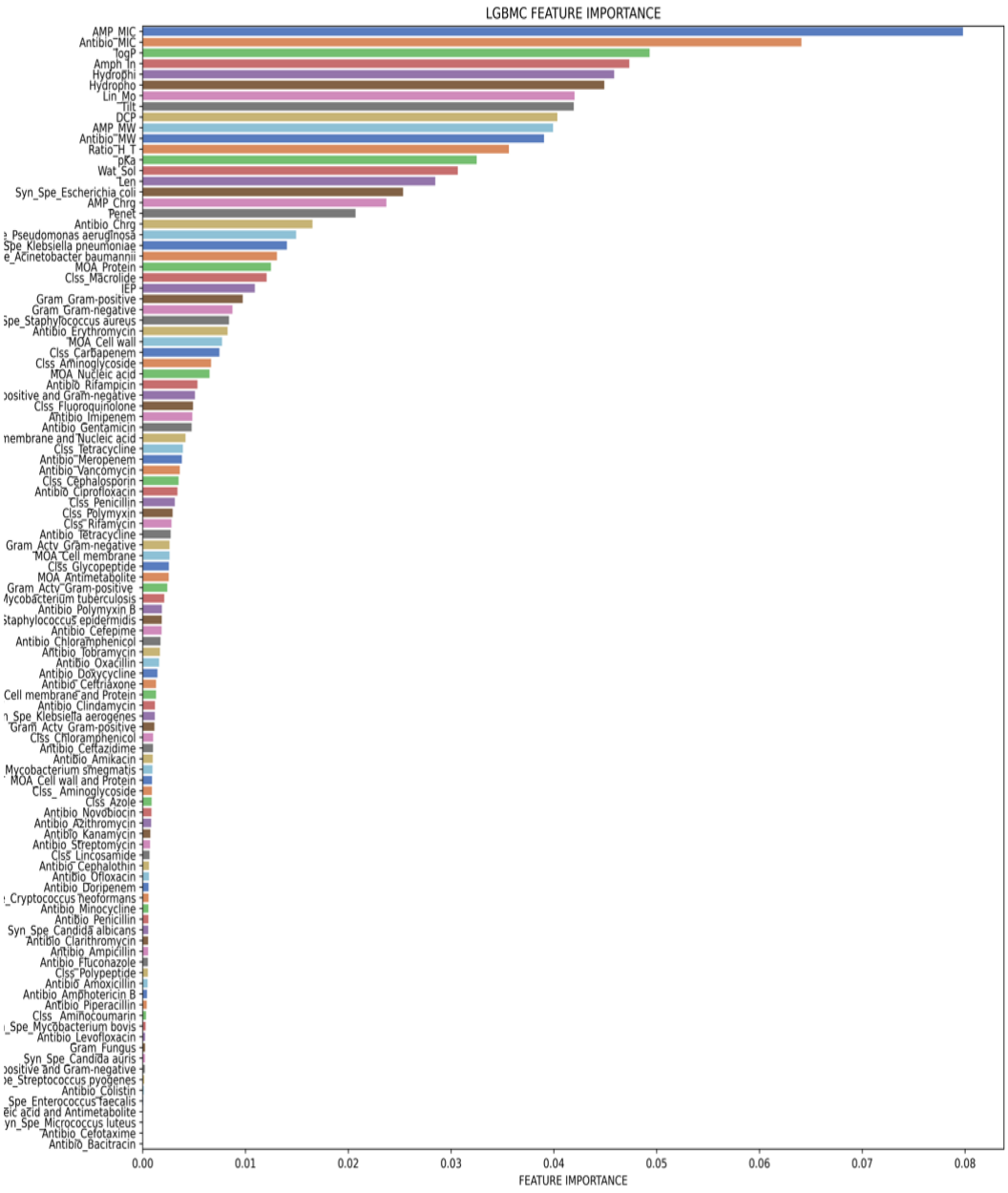

Figure 3.9 Feature importance after one hot encoding

The term "feature importance" refers to a class of strategies for giving scores to features in a predictive model, indicating the relative significance of features when making predictions. With the One Hot Encoding, all nominal inputs are converted to numerical inputs with the 1-0 encoding method. When One Hot Encoding is done, each line is converted to a column for encoding. In other words, each row becomes a feature. That is the reason there are so many features in Figure 3.9. The sum of the importance values of all the features in the Figure 3.9 is equal to 1.

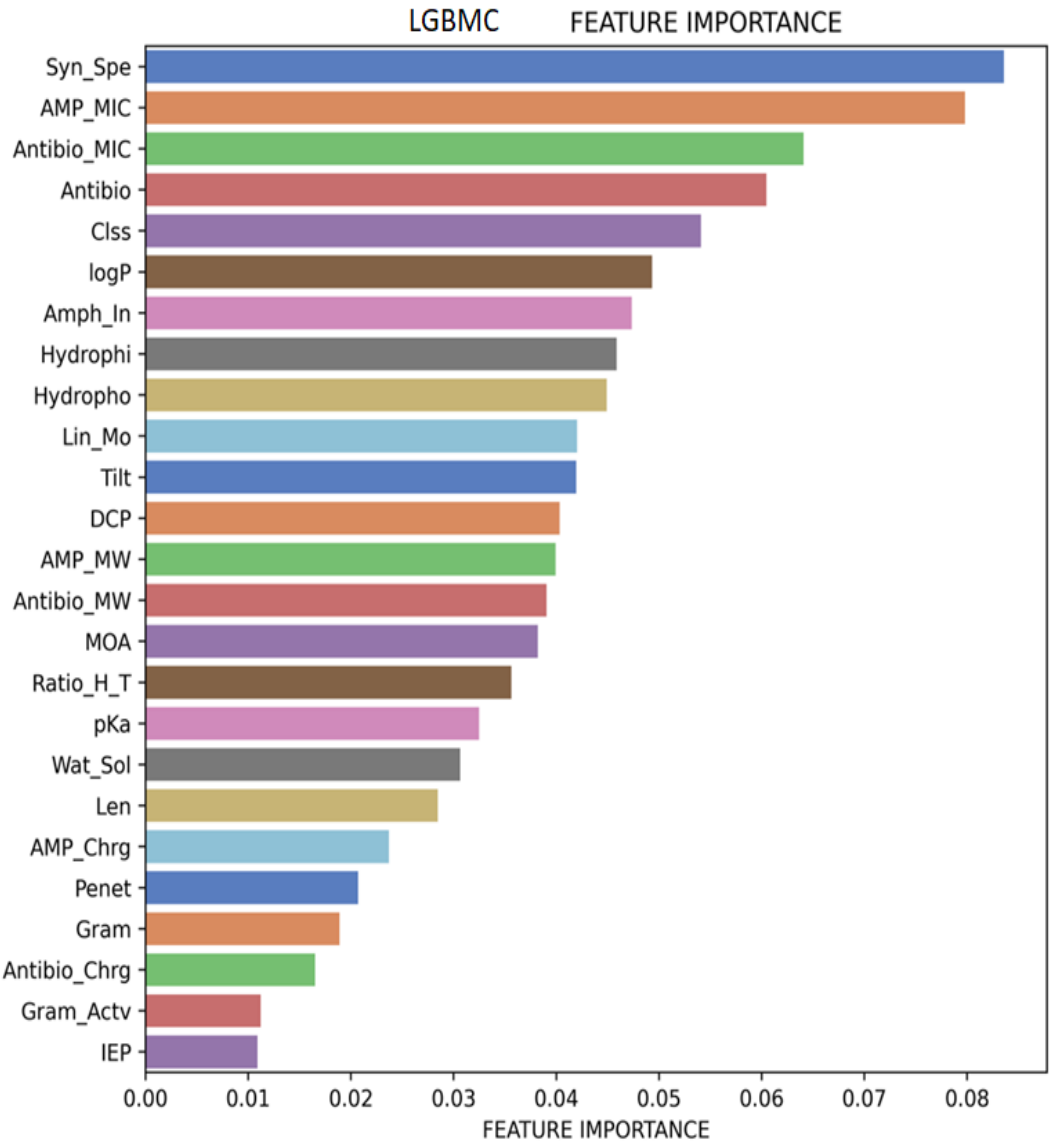

Figure 3.10 Feature importance

The sum of the feature importance values is equal to 1. As can be seen from Figure 3.10, it was seen that the most important feature was the species in which the synergistic effect of antibiotic and AMP was investigated. The second and third most important features were the MIC values of the peptide and antibiotic when used alone. It was also seen that the least important features were the isoelectric point of the AMP, the charge of the antibiotic, and the gram type of the pathogen in which the antibiotic was active on. In addition, it was seen that the features of the antibiotic were more important.

Developed model was evaluated by calculating the Accuracy, Presicion, Recall (sentivity), and F1 Score values.

Accuracy is a measure that describes the proportion of correct predictions among all predictions.

$$Accuracy = \frac{N.of\ Correct\ Predictions}{N.of\ All\ Predictions} = \frac{N.of\ Correct\ Predictions}{Size\ of\ Dataset} \quad (3.1)$$

Where the number of correct predictions is equal to 80 and the size of the dataset is equal to 104 ;

$$Accuracy\ = \frac{80}{104} = 0{,}7692\ \approx \mathbf{0,77}$$

Precision is a measure of how many of the positive predictions made are correct, in other words, true positives.

$$Precision = \frac{N.of\ Correctly\ Predicted\ Positive\ Instances}{N.of\ Total\ Positive\ Predictions} \quad (3.2)$$

In the case of Synergism, there are 37 correct predictions and 12 incorrect predictions;

$$Precision = \frac{37}{49} = 0{,}755\ \approx 0{,}76$$

In the case of No Interaction, there are 43 correct predictions and 12 incorrect predictions;

$$Precision = \frac{43}{55} = 0{,}78$$

$$Precision_{Average} = \frac{0{,}76 + 0{,}78}{2} = \mathbf{0,77}$$

Recall, also known as sensitiviy, is a measure of how many positive instances the classifier predicted correctly out of all the positive instances in the data.

$$Recall = \frac{N.of\ Correctly\ Predicted\ Positive\ Instances}{N.of\ Total\ Positive\ Instances\ in\ the\ Dataset} \quad (3.3)$$

In the case of Synergism, there are 37 correct predictions and 12 incorrect positive predictions;

$$Recall = \frac{37}{49} = 0{,}755 \approx 0{,}76$$

In the case of No Interaction, there are 43 correct predictions and 12 incorrect positive predictions;

$$Recall = \frac{43}{55} = 0{,}78$$

$$Recall_{Average} = \frac{0{,}76 + 0{,}78}{2} = \mathbf{0,77}$$

F1-Score is a metric that combines precision and recall results. It is commonly referred to as the harmonic mean of the precision and recall.

$$F1\ Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \quad (3.4)$$

$$F1\ Score = 2\ x\ \frac{0{,}77\ x\ 0{,}77}{0{,}77 + 0{,}77} = \frac{1{,}186}{1{,}54} = \mathbf{0,77}$$

Chapter 4

Discussions

Fields et al. [119] have created a pipeline for the creation and testing of bacteriocin-derived compounds that combines sequence-free bacteriocin prediction with machine learning and a biophysical feature filter to identify peptides that contain 20 amino acids and can be synthesized and tested for activity. They generated a total of 28,895 20-mer potential peptides and rated them based on charge, as well as hydrophobic moment. They chose 16 sequences for synthesiss, then tested their antibacterial, cytotoxic, and hemolytic actions. Peptides with the highest biophysical criteria scores demonstrated strong antimicrobial efficacy against Pseudomonas aeruginosa and Escherichia coli. Their combination strategy incorporates biophysical-based minimum region determination and machine learning to develop a novel methodology for discovering bacteriocin candidates suitable for rapid synthesis and assessment for therapeutic use.

Because of their hemolytic toxicity, most AMPs used in clinical studies are administered topically. Plisson et al. [120] constructed machine learning algorithms and outlier detection techniques to guarantee robust predictions for AMP discovery and the design of new peptides with lower hemolytic activity. Their best model predicted the hemolytic tendency of any peptide sequence with an accuracy of 95-97%. Using multivariate outlier detection models, researchers discovered that 273 AMPs could not be reliably predicted. Their combined strategy led to the design of 507 peptides, identification of 34 AMPs that are not hemolytic, and the development of non-hemolytic peptide design guidelines.

Li et al. [121] sought to identify factors regulating selectivity by correlating peptide sequence information with bioactivity data using the random forest algorithm. Out-of-bag prediction generated satisfactory predictive models with accuracies in excess

of 0.80. Model interpretation using variable significance metrics and partial dependency plots revealed that the distribution patterns and composition of molecular charge and solubility-related factors strongly influenced selectivity. Furthermore, because it appears to be similar selective mechanism based on charge-solubility properties, the investigated target species had a great influence on how selectivity was achieved.

Nagarajan et al. [122] employed a language model with long short-term memory (LSTM) to comprehend the arrangement as well as the frequencies of amino acids in known peptide sequences. They generated 10 peptides based on LSTM network output and tested these peptides against pathogens. All peptides demonstrated broad-spectrum antimicrobial activity, confirming the validity of their LSTM-based design approach. Their two most effective AMPs were shown to be effective against multidrug-resistant (MDR) clinical isolates of Acinetobacter baumannii, Escherichia coli, Pseudomonas aeruginosa, Klebsiella pneumoniae, and Staphylococcus aureus. Peptides interacted with cell membranes and disrupted them, resulting in secondary gene regulatory effects.

Xu et al. [47] gave a detailed survey of current methodologies for AMP identification and highlighted the variations between these methods. Furthermore, they assessed the prediction performance of the tools using an independent test datasets of 1536 AMPs and non-AMPs. They also constructed six validation datasets based on six popular AMP databases and made comparisons between different computational approaches using these data sets. The results showed that amPEPpy outperformed the other evaluated methods in terms of prediction performance. Because the prediction performances of different approaches are impacted by the datasets used, they also performed cross-validation tests in order to compare several traditional machine learning algorithms on the same dataset. Cross-validation findings showed that SVM, RF, and XG Boosting outperformed other machine learning techniques and were frequently the algorithms of choice for several AMP prediction tools.

There are several computational approaches for predicting AMPs. However, Vishnepolsky et al. [58] discovered that most of these approaches could only forecast if a peptide would have any antimicrobial potency, there are no tools that could predict antimicrobial efficacy against specific strains. They introduced a semi-

supervised learning strategy using a clustering algorithm to predict the activity of linear AMPs against specific Gram-negative strains. The algorithm was capable of distinguishing peptides that are active against specific strains from those that are active but not against the specific strain under consideration. The present AMP prediction technologies were incapable of performing this task.

Gull et al. [89] developed AMAP, a machine learning-based model to predict the biological activity of peptides, with an emphasis on antimicrobial activity predictions. AMAP improves on previous state-of-art methodologies by using multi-label classification for predicting 14 distinct types of activities of a given peptide sequence. They conducted performance studies of the suggested method. AMAP was benchmarked using newly published experimentally validated peptides that were not part of their training set, in addition to performance comparisons with current AMP predictors. They also examined the features employed in this study, and their findings demonstrated that the suggested predictor is capable of accurately predicting the biological activity of new peptide sequences.

Sharma et al. [123] developed the AniAMPpred model by taking into account all of the available AMPs from the animal world with lengths ranging from 10-200. The model identified probable antimicrobial proteins (PAPs) in animal genomes using a SVM algorithm. The findings demonstrated that the suggested model outperformed previous state-of-art classifiers, had high confidence in predictions, and could accurately classify AMPs and non-AMPs for a wide range of lengths. They identified 436 PAPs in the Helobdella robusta genome. They also discovered similarities between PAPs and antimicrobial proteins from various animal species through detailed analysis.

Antiviral peptides are a kind of AMP which has the potential to combat virus infection. Pang et al. [124] suggested a two-stage classification approach to predict antiviral peptides and their antiviral functional activities. The initial step was to distinguish the antiviral peptides from a large array of peptides that included not only non-AMPs but also AMPs that did not have antiviral functions. The second step was in charge of identifying one or more virus families as well as species that the antiviral peptide targets. Imbalanced learning was used to improve prediction performance. The model employed machine learning strategies that are based on

Shapley value to analyze how the descriptors affected the antiviral activities and used numerous descriptors to precisely display the peptide features. Lastly, the proposed model's evaluation performance indicated its capacity to forecast antiviral actions and their prospective functions against different virus families.

In this study, the interaction between antibiotics and antimicrobial peptides was evaluated in terms of FIC index value. To summarize the steps followed, data were collected from various databases. Missing values have been removed. With the resampling, more specifically SMOTE technique, the sets with unbalanced data numbers were equalized in number. The wide data distribution was narrowed by the data normalization method. Nominal variables were converted to numeric values with one hot encoding method. Data were separated by data splitting method as training and test sets. Finally, various modeling algorithms were tried and the algorithm with the highest accuracy was selected. The model was developed with the selected algorithm and the success of the model was evaluated.

A correlation matrix was created. If the values are very close to 1 and -1 in the correlation matrix, it is called the dependent variable. In other words, the two variables are highly dependent on each other. Dependent variables are undesirable as they will negatively affect the accuracy of the model and are usually removed from the model. All were included in this study. If the dependent variables are also removed, the high accuracy score already obtained in the model will increase even more.

The original values were distributed over a wide range. There were both very high and very small values, which is something that would negatively affect the accuracy of the model. Therefore, normalization was required to narrow these ranges. The normalization method with the highest accuracy value was the robust method. For that reason, normalization was done by choosing the robust method.

A model with 100% incorrect predictions has an AUC value of 0.0; one with 100% correct predictions has an AUC value of 1.0. That is to say, the closer the AUC value to 1, the better the performance of the model [118]. In our model, this value was

measured as 0.807. Since it is close to 1, it can be said that the success of the developed model is high.

Looking at the feature importance data, it was seen that the most important feature was the pathogen type, as expected. Afterwards, the second and third most important features were the MIC values of AMP and antibiotic, respectively. This was also expected because the FIC index was determined as output, and according to the FIC index formula, the MIC values of the antimicrobial agents used are included in the formula. However, the surprising thing was that although the MIC values of the two agents were included in the formula, the correlation between them and the FIC index was low, even very close to zero. As the FIC value was calculated using the MIC values, they were expected to be dependent variables.

It was also expected that the values were negative because the MIC values are in the denominator in the formula. In other words, an increase in MIC values causes a decrease in FIC value.

To summarize, the predictive success of the model is high based on accuracy score and other calculated values. As mentioned, the success of the model will increase if the dependent variables are eliminated.

Chapter 5

Conclusions

Although the use of a urinary catheter becomes a necessity for bladder-related problems, infections cause many problems, such as blockage of catheters and kidney-related problems if infections progress. Many methods have been tried to prevent these problems, but none of the methods have achieved satisfactory success.

The use of antimicrobial agents in high concentrations can cause a lot of side effects. It has been proven in various studies that antimicrobial agents can have a synergistic effect on each other. For this reason, it is expected that the combined use of two antimicrobial agents, which have a synergistic effect on each other may prevent the toxic effects that may be caused by the use of antimicrobial agents in high concentrations.

In this study, it is predicted that the use of antibiotics and antimicrobial peptides together will prevent the stability and decreased antimicrobial activity problems of antimicrobial peptides and the emergence of drug-resistant microorganisms caused by antibiotics.

With the machine learning method, it is aimed to create a model with high predictive success and to prevent the loss of time and resources spent in laboratory experiments.

Models were developed using the 17 different classifiers, and their accuracy was tested and compared. As a result of the comparisons made, LGBMC was determined as the classifier with the highest accuracy.

The model was created with the LGBMC algorithm using the robust scaling method. The accuracy result of the model was calculated as AUC. According to the results, it was seen that the accuracy of the model was high. In line with the satisfactory results obtained as a result of various tests, it is anticipated that this study will shed light on future studies.

References

1. Zhang S, Wang L, Liang X, Vorstius J, Keatch R, Corner G *et al*. Enhanced antibacterial and antiadhesive activities of silver-PTFE nanocomposite coating for urinary catheters. ACS Biomaterials Science & Engineering 2019; 5(6): 2804-14.

2. Jacobsen SM, Stickler DJ, Mobley HLT, Shirtliff ME. Complicated catheter-associated urinary tract infections due to Escherichia coli and Proteus mirabilis. Clinical Microbiology Reviews 2008; 21(1): 26-59.

3. Ramstedt M, Ribeiro IA, Bujdakova H, Mergulhão FJ, Jordao L, Thomsen P *et al*. Evaluating efficacy of antimicrobial and antifouling materials for urinary tract medical devices: Challenges and recommendations. Macromolecular Bioscience 2019; 19(5): 1800384.

4. Cortese YJ, Wagner VE, Tierney M, Devine D, Fogarty A. Review of catheter-associated urinary tract infections and in vitro urinary tract models. Journal of healthcare engineering 2018; 2018.

5. Stickler DJ. Clinical complications of urinary catheters caused by crystalline biofilms: something needs to be done. Journal of Internal Medicine 2014; 276(2): 120-9.

6. Singha P, Locklin J, Handa H. A review of the recent advances in antimicrobial coatings for urinary catheters. Acta biomaterialia 2017; 50: 20-40.

7. Saint S, Gaies E, Fowler KE, Harrod M, Krein SL. Introducing a catheter-associated urinary tract infection (CAUTI) prevention guide to patient safety (GPS). American Journal of Infection Control 2014; 42(5): 548-50.

8. Ong C-LY, Ulett GC, Mabbett AN, Beatson SA, Webb RI, Monaghan W *et al*. Identification of type 3 fimbriae in uropathogenic Escherichia coli reveals a role in biofilm formation. Journal of bacteriology 2008; 190(3): 1054-63.

9. Donlan RM, Costerton JW. Biofilms: survival mechanisms of clinically relevant microorganisms. Clinical Microbiology Reviews 2002; 15(2): 167-93.

10. Tenke P, Köves B, Nagy K, Hultgren SJ, Mendling W, Wullt B *et al*. Update on biofilm infections in the urinary tract. World Journal of Urology 2012; 30(1): 51-7.

11. Davies DG, Parsek MR, Pearson JP, Iglewski BH, Costerton JW, Greenberg EP. The involvement of cell-to-cell signals in the development of a bacterial biofilm. Science 1998; 280(5361): 295-8.

12. Costerton JW, Lewandowski Z, Caldwell DE, Korber DR, Lappin-Scott HM. Microbial biofilms. Annual Review of Microbiology 1995; 49(1): 711-45.

13. Kho K, Cheow WS, Lie RH, Hadinoto K. Aqueous re-dispersibility of spray-dried antibiotic-loaded polycaprolactone nanoparticle aggregates for inhaled anti-biofilm therapy. Powder Technology 2010; 203(3): 432-9.

14. Costa F, Carvalho IF, Montelaro RC, Gomes P, Martins MCL. Covalent immobilization of antimicrobial peptides (AMPs) onto biomaterial surfaces. Acta Biomaterials 2011; 7(4): 1431-40.

15. Buck ME, Breitbach AS, Belgrade SK, Blackwell HE, Lynn DM. Chemical modification of reactive multilayered films fabricated from poly(2-alkenyl azlactone)s: design of surfaces that prevent or promote mammalian cell adhesion and bacterial biofilm growth. Biomacromolecules 2009; 10(6): 1564-74.

16. Harbers GM, Emoto K, Greef C, Metzger SW, Woodward HN, Mascali JJ *et al.*. A functionalized poly(ethylene glycol)-based bioassay surface chemistry that facilitates bio-immobilization and inhibits non-specific protein, bacterial, and mammalian cell adhesion. Chemistry of Materials: A Publication of the American Chemical Society 2007; 19(18): 4405-14.

17. Mishra B, Basu A, Chua RRY, Saravanan R, Tambyah PA, Ho B *et al.* Site specific immobilization of a potent antimicrobial peptide onto silicone catheters: evaluation against urinary tract infection pathogens. Journal of Materials Chemistry. B 2014; 2(12): 1706-16.

18. Majeed A, Sagar F, Latif A, Hassan H, Iftikhar A, Darouiche RO, *et al*. Does antimicrobial coating and impregnation of urinary catheters prevent catheter-associated urinary tract infection? A review of clinical and preclinical studies. Expert Review of Medical Devices 2019; 16(9): 809-20.

19. Wang R, Neoh KG, Kang E-T, Tambyah PA, Chiong E. Antifouling coating with controllable and sustained silver release for long-term inhibition of infection and encrustation in urinary catheters. Journal of Biomedical Materials Research Part B: Applied Biomaterials 2015; 103(3): 519-28.

20. Walder B, Pittet D, Tramèr MR. Prevention of bloodstream infections with central venous catheters treated with anti-infective agents depends on catheter type and insertion time: evidence from a meta-analysis. Infection Control & Hospital Epidemiology 2002; 23(12): 748-56.

21. Zhu Z, Wang Z, Li S, Yuan X. Antimicrobial strategies for urinary catheters. Journal of Biomedical Materials Research Part A. 2019; 107(2): 445-67.

22. Hanna H, Bahna P, Reitzel R, Dvorak T, Chaiban G, Hachem R, *et al.* Comparative in vitro efficacies and antimicrobial durabilities of novel antimicrobial central venous catheters. Antimicrobial Agents and Chemotherapy 2006; 50(10) : 3283-3288.

23. Morris Jr JG, Sulakvelidze A, Alavidze Z. Bacteriophage Therapy. Antimicrobial agents and Chemotherapy 2001; 45(3): 649-59.

24. Carson L, Gorman SP, Gilmore BF. The use of lytic bacteriophages in the prevention and eradication of biofilms of Proteus mirabilis and Escherichia coli. FEMS Immunology & Medical Microbiology 2010; 59(3): 447-55.

25. Gilchrist T, Healy DM, Drake C. Controlled silver-releasing polymers and their potential for urinary tract infection control. Biomaterials 1991; 12(1): 76-8.

26. Bagheri M, Beyermann M, Dathe M. Immobilization reduces the activity of surface-bound cationic antimicrobial peptides with no influence upon the activity spectrum. Antimicrobial agents and chemotherapy 2009; 53(3): 1132-41.

27. Ferreira L, Zumbuehl A. Non-leaching surfaces capable of killing microorganisms on contact. Journal of Materials Chemistry 2009; 19(42): 7796-806.

28. Willcox MDP, Hume EBH, Aliwarga Y, Kumar N, Cole N. A novel cationic-peptide coating for the prevention of microbial colonization on contact lenses. Journal of applied microbiology 2008; 105(6): 1817-25.

29. Sun E, Belanger CR, Haney EF, Hancock RE. Host defense (antimicrobial) peptides. Peptide applications in biomedicine, biotechnology and bioengineering. Elsevier; 2018 ; 253-85.

30. Haney EF, Mansour SC, Hancock REW. Antimicrobial Peptides: An Introduction. Methods in Molecular Biology 2017; 1548: 3-22.

31. Dostert M, Belanger CR, Hancock REW. Design and Assessment of Anti-Biofilm Peptides: Steps Toward Clinical Application. Journal of Innate Immunity 2019; 11(3): 193-204.

32. Maisetta G, Di Luca M, Esin S, Florio W, Brancatisano FL, Bottai D, vd. Evaluation of the inhibitory effects of human serum components on bactericidal activity of human beta defensin 3. Peptides 2008; 29(1): 1-6.

33. Mansour SC, Pena OM, Hancock RE. Host defense peptides: front-line immunomodulators. Trends in immunology 2014; 35(9): 443-50.

34. Bowdish DM, Davidson DJ, Lau YE, Lee K, Scott MG, Hancock RE. Impact of LL-37 on anti-infective immunity. Journal of leukocyte biology. 2005; 77(4): 451-9.

35. Wang S-H, Tang TW-H, Wu E, Wang D-W, Liao Y-D. Anionic surfactant-facilitated coating of antimicrobial peptide and antibiotic reduces biomaterial-associated infection. ACS Biomaterials Science & Engineering 2020; 6(8): 4561-72.

36. Carmona-Ribeiro AM, de Melo Carrasco LD. Novel formulations for antimicrobial peptides. International journal of molecular sciences 2014; 15(10): 18040-83.

37. Hall MJ, Middleton RF, Westmacott D. The fractional inhibitory concentration (FIC) index as a measure of synergy. Journal of Antimicrobial Chemotherapy 1983; 11(5): 427-33.

38. Bhardwaj KK, Banyal S, Sharma DK. Artificial intelligence based diagnostics, therapeutics and applications in biomedical engineering and bioinformatics. Internet of Things in Biomedical Engineering. Elsevier 2019 ; 161-87.

39. Poole D, Mackworth A, Goebel R. Computational Intelligence: Oxford University Press, New York; 1998.

40. Chou K-C. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of theoretical biology 2011; 273(1): 236-47.

41. Hudson DL, Cohen ME. Neural networks and artificial intelligence for biomedical engineering. Wiley Online Library; 2000.

42. Bonetta R, Valentino G. Machine learning techniques for protein function prediction. Proteins: Structure, Function, and Bioinformatics 2020; 88(3): 397-413.

43. Lee EY, Lee MW, Fulan BM, Ferguson AL, Wong GC. What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning? Interface focus 2017; 7(6): 20160153.

44. Kavousi K, Bagheri M, Behrouzi S, Vafadar S, Atanaki FF, Lotfabadi BT, *et al*. IAMPE: NMR-assisted computational prediction of antimicrobial peptides. Journal of Chemical Information and Modeling 2020; 60(10): 4691-701.

45. Lande R, Gregorio J, Facchinetti V, Chatterjee B, Wang Y-H, Homey B, *et al*. Plasmacytoid dendritic cells sense self-DNA coupled with antimicrobial peptide. Nature 2007; 449(7162): 564-9.

46. Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. Nucleic acids research 2016; 44(D1): D1087-93.

47. Xu J, Li F, Leier A, Xiang D, Shen H-H, Marquez Lago TT, *et al*. Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides. Briefings in Bioinformatics 2021; 22(5): bbab083.

48. Fjell CD, Hancock RE, Cherkasov A. AMPer: a database and an automated discovery tool for antimicrobial peptides. Bioinformatics 2007; 23(9): 1148-55.

49. Lata S, Sharma BK, Raghava GP. Analysis and prediction of antibacterial peptides. BMC bioinformatics 2007; 8(1): 1-10.

50. Yeaman MR, Yount NY. Mechanisms of antimicrobial peptide action and resistance. Pharmacological reviews 2003; 55(1): 27-55.

51. Thakur N, Qureshi A, Kumar M. AVPpred: collection and prediction of highly effective antiviral peptides. Nucleic acids research 2012; 40(W1): W199-204.

52. Xiao X, Wang P, Lin W-Z, Jia J-H, Chou K-C. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Analytical biochemistry 2013; 436(2): 168-77.

53. Lata S, Mishra NK, Raghava GP. AntiBP2: improved version of antibacterial peptide prediction. BMC bioinformatics 2010; 11(1): 1-7.

54. Fallah Atanaki F, Behrouzi S, Ariaeenejad S, Boroomand A, Kavousi K. BIPEP: Sequence-based prediction of biofilm inhibitory peptides using a combination of nmr and physicochemical descriptors. ACS omega 2020; 5(13): 7290-7.

55. Veltri D, Kamath U, Shehu A. Improving recognition of antimicrobial peptides and target selectivity through machine learning and genetic programming. IEEE/ACM transactions on computational biology and bioinformatics. 2015; 14(2): 300-13.

56. Bhadra P, Yan J, Li J, Fong S, Siu SW. AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. Scientific reports 2018; 8(1): 1-10.

57. Joseph S, Karnik S, Nilawe P, Jayaraman VK, Idicula-Thomas S. ClassAMP: a prediction tool for classification of antimicrobial peptides. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2012; 9(5): 1535-8.

58. Vishnepolsky B, Gabrielian A, Rosenthal A, Hurt DE, Tartakovsky M, Managadze G, *et al*. Predictive model of linear antimicrobial peptides active against gram-negative bacteria. Journal of chemical information and modeling 2018; 58(5): 1141-51.

59. Hammami R, Ben Hamida J, Vergoten G, Fliss I. PhytAMP: a database dedicated to antimicrobial plant peptides. Nucleic acids research 2009; 37: D963-8.

60. Gueguen Y, Garnier J, Robert L, Lefranc M-P, Mougenot I, De Lorgeril J, *et al*. PenBase, the shrimp antimicrobial peptide penaeidin database: sequence-based classification and recommended nomenclature. Developmental & Comparative Immunology 2006; 30(3): 283-8.

61. Hammami R, Zouhir A, Ben Hamida J, Fliss I. BACTIBASE: a new web-accessible database for bacteriocin characterization. BMC microbiology 2007; 7(1): 1-6.

62. Li Y, Chen Z. RAPD: a database of recombinantly-produced antimicrobial peptides. FEMS microbiology letters 2008; 289(2): 126-9.

63. Wade D, Englund J. Synthetic antibiotic peptides database. Protein and peptide letters 2002; 9(1): 53-7.

64. Whitmore L, Wallace B. The Peptaibol Database: a database for sequences and structures of naturally occurring peptaibols. Nucleic Acids Research 2004; 32: D593-4.

65. Seebah S, Suresh A, Zhuo S, Choong YH, Chua H, Chuon D, *et al*. Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides. Nucleic acids research. 2007; 35: D265-8.

66. Wang Z, Wang G. APD: the antimicrobial peptide database. Nucleic acids research. 2004; 32: D590-2.

67. Thomas S, Karnik S, Barai RS, Jayaraman VK, Idicula-Thomas S. CAMP: a useful resource for research on antimicrobial peptides. Nucleic Acids Research 2010; 38: D774-780.

68. Niarchou A, Alexandridou A, Athanasiadis E, Spyrou G. C-PAmP: Large Scale Analysis and Database Construction Containing High Scoring Computationally Predicted Antimicrobial Peptides for All the Available Plant Species. PLOS ONE 2013; 8(11): e79728.

69. Chou K-C. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics 2001; 43(3): 246-55.

70. Chou K-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 2005; 21(1): 10-9.

71. Liu B, Liu F, Wang X, Chen J, Fang L, Chou K-C. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic acids research 2015; 43(W1): W65-71.

72. Lee H-T, Lee C-C, Yang J-R, Lai JZ, Chang KY. A large-scale structural classification of antimicrobial peptides. BioMed research international 2015.

73. Meher PK, Sahu TK, Saini V, Rao AR. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. Scientific reports 2017; 7(1): 1-12.

74. Lin W, Xu D. Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types. Bioinformatics 2016; 32(24): 3745-52.

75. Lin Y, Cai Y, Liu J, Lin C, Liu X. An advanced approach to identify antimicrobial peptides and their function types for penaeus through machine learning strategies. BMC Bioinformaticsn 2019; 20(8): 291.

76. Chung C-R, Kuo T-R, Wu L-C, Lee T-Y, Horng J-T. Characterization and identification of antimicrobial peptides with different functional activities. Brief Bioinformatics 2019; bbz043.

77. Porto WF, Fernandes FC, Franco OL. An SVM model based on physicochemical properties to predict antimicrobial activity from protein sequences with cysteine knot motifs. Brazilian Symposium on Bioinformatics. Springer 2010; 59-62.

78. Wang P, Hu L, Liu G, Jiang N, Chen X, Xu J, *et al*. Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. PloS one. 2011; 6(4): e18476.

79. Torrent M, Andreu D, Nogués VM, Boix E. Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. PloS one. 2011; 6(2): e16968.

80. Porto WF, Pires ÁS, Franco OL. CS-AMPPred: an updated SVM model for antimicrobial activity prediction in cysteine-stabilized peptides. PLoS One. 2012; 7(12): e51444.

81. Randou EG, Veltri D, Shehu A. Systematic analysis of global features and model building for recognition of antimicrobial peptides. 2013 IEEE 3rd International Conference on Computational Advances in Bio and medical Sciences (ICCABS). IEEE; 2013; 1-6.

82. Fernandes FC, Rigden DJ, Franco OL. Prediction of antimicrobial peptides based on the adaptive neuro-fuzzy inference system application. Peptide Science 2012; 98(4): 280-7.

83. Vishnepolsky B, Pirtskhalava M. Prediction of linear cationic antimicrobial peptides based on characteristics responsible for their interaction with the membranes. Journal of chemical information and modeling 2014; 54(5): 1512-23.

84. Camacho FL, Torres R, Pollán RR. Classification of antimicrobial peptides with imbalanced datasets. 11th International Symposium on Medical Information Processing and Analysis. International Society for Optics and Photonics; 2015; 96810T.

85. Ng XY, Rosdi BA, Shahrudin S. Prediction of antimicrobial peptides based on sequence alignment and support vector machine-pairwise algorithm utilizing LZ-complexity. BioMed research international 2015.

86. Beltran JA, Aguilera-Mendoza L, Brizuela CA. Feature weighting for antimicrobial peptides classification: a multi-objective evolutionary approach. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2017; 276-83.

87. Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. Bioinformatics 2018; 34(16): 2740-7.

88. Su X, Xu J, Yin Y, Quan X, Zhang H. Antimicrobial peptide identification using multi-scale convolutional network. BMC bioinformatics 2019; 20(1): 1-10.

89. Gull S, Shamim N, Minhas F. AMAP: Hierarchical multi-label prediction of biologically active and antimicrobial peptides. Computers in Biology and Medicine 2019; 107: 172-81.

90. Jhong J-H, Chi Y-H, Li W-C, Lin T-H, Huang K-Y, Lee T-Y. dbAMP: an integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data. Nucleic acids research 2019; 47(D1): D285-97.

91. Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning . Nucleic Acids 2020; 20: 882-894.

92. Li C, Sutherland D, Hammond SA, Yang C, Taho F, Bergman L, *et al*. AMPlify: attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens. BMC genomics 2022; 23(1): 1-15.

93. Fu H, Cao Z, Li M, Xia X, Wang S. Prediction of anuran antimicrobial peptides using AdaBoost and improved PSSM profiles. Proceedings of the Fourth International Conference on Biological Information and Biomedical Engineering. 2020; 1-6.

94. Mirzaei M, Furxhi I, Murphy F, Mullins M. A machine learning tool to predict the antibacterial capacity of nanoparticles. Nanomaterials 2021; 11(7): 1774.

95. Fukunaga K. Introduction to statistical pattern recognition. Elsevier 2013.

96. Ahsan MM, Mahmud MAP, Saha PK, Gupta KD, Siddique Z. Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies 2021; 9(3): 52.

97. Sasada T, Liu Z, Baba T, Hatano K, Kimura Y. A resampling method for imbalanced datasets considering noise and overlap. Procedia Computer Science 2020; 176: 420-9.

98. Fernandez A, Garcia S, Herrera F, Chawla NV. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research 2018; 61: 863-905.

99. Birba DE. A Comparative study of data splitting algorithms for machine learning model selection 2020.

100. Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society 2004; 18(6): 275-85.

101. Quinlan JR. Learning decision tree classifiers. ACM Computing Surveys (CSUR) 1996; 28(1): 71-2.

102. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, vd. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2017.

103. Ying C, Qi-Guang M, Jia-Chen L, Lin G. Advance and prospects of AdaBoost algorithm. Acta Automatica Sinica. 2013; 39(6): 745-58.

104. Kim T-H, Park D-C, Woo D-M, Jeong T, Min S-Y. Multi-class classifier-based adaboost algorithm. International conference on intelligent science and intelligent data engineering. Springer 2011; 122-7.

105. Sheridan RP, Wang WM, Liaw A, Ma J, Gifford EM. Extreme gradient boosting as a method for quantitative structure–activity relationships. Journal of chemical information and modeling 2016; 56(12): 2353-60.

106. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, vd. Xgboost: extreme gradient boosting. R package version 04-2. 2015; 1(4): 1-4.

107. Al-Mistarehi BW, Alomari AH, Imam R, Mashaqba M. Using Machine Learning Models to Forecast Severity Level of Traffic Crashes by R Studio and ArcGIS. Frontiers in Built Environment 2022.

108. Suthaharan S. Support vector machine. Machine learning models and algorithms for big data classification. Springer 2016; 207-35.

109. Wright RE. Logistic regression. Reading and understanding multivariate statistics. Washington, DC, US: American Psychological Association; 1995; 217-44.

110. Balakrishnama S, Ganapathiraju A. Linear Discriminant Analysis - A Brief Tutorial. Institute for Signal and information Processing 1998; 1-8.

111. Jiang L, Cai Z, Wang D, Jiang S. Survey of Improving K-Nearest-Neighbor for Classification. Fourth International Conference on Fuzzy Systems and Knowledge Discovery. 2007; 679-83.

112. Hensman J, Matthews A, Ghahramani Z. Scalable Variational Gaussian Process Classification. Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics 2015; 351-60.

113. Webb G. Naïve Bayes. 2016; 1-2.

114. Skurichina M, Duin RPW. Bagging for linear classifiers. Pattern Recognition. 1998; 31(7): 909-30.

115. Schratz P, Muenchow J, Iturritxa E, Richter J, Brenning A. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecological Modelling 2019; 406: 109-20.

116. Touchette PE, MacDonald RF, Langer SN. A Scatter Plot for Identifying Stimulus Control of Problem Behavior. Journal of Applied Behavior Analysis. 1985; 18(4): 343-51.

117. Cui Q, Ward M, Rundensteiner E. Enhancing Scatterplot Matrices for Data with Ordering or Spatial Attributes 2006.

118. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 1997; 30(7): 1145-59.

119. Fields FR, Freed SD, Carothers KE, Hamid MN, Hammers DE, Ross JN, *et al*. Novel antimicrobial peptide discovery using machine learning and biophysical selection of minimal bacteriocin domains. Drug Development Research 2020; 81(1): 43-51.

120. Plisson F, Ramírez-Sánchez O, Martínez-Hernández C. Machine learning-guided discovery and design of non-hemolytic peptides. Scientific Reports 2020; 10(1): 16581.

121. Li H, Tamang T, Nantasenamat C. Toward insights on antimicrobial selectivity of host defense peptides via machine learning model interpretation. Genomics. 2021; 113(6): 3851-63.

122. Nagarajan D, Nagarajan T, Roy N, Kulkarni O, Ravichandran S, Mishra M, *et al*. Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria. Journal Biology Chemistry 2018; 293(10): 3492-509.

123. Sharma R, Shrivastava S, Kumar Singh S, Kumar A, Saxena S, Kumar Singh R. AniAMPpred: artificial intelligence guided discovery of novel antimicrobial peptides in animal kingdom. Briefings in Bioinformatics 2021; 22(6): bbab242.

124. Pang Y, Yao L, Jhong J-H, Wang Z, Lee T-Y. AVPIden: a new scheme for identification and functional prediction of antiviral peptides based on machine learning approaches. Briefings in Bioinformatics 2021; 22(6): bbab263.

Appendix

Publications from the Thesis

Conference Papers

1. Antimicrobial Peptide Conjugation on the Catheter Surfaces for the Control and Prevention of Catheter Associated Urinary Tract Infections , 5th International Conference on Medical Devices.

Curriculum Vitae

Name Surname : Başak Olcay
E-mail : basak.olcayy@hotmail.com

Education:

2011 – 2014 Izmir Sasalı Anatolian Teacher High School

2014 – 2015 Atakent Anatolian High School

2018 (01 - 06) University Of Oulu, Faculty of Biochemistry and Molecular Medicine

2015 – 2020 İzmir Kâtip Çelebi University, Dept. of Biomedical Engineering

2020 – 2022 İzmir Kâtip Çelebi University, Dept. of Biomedical Engineering

Work Experience:

2018 (05 – 06) Internship - University of Oulu Faculty of Biochemistry and Molecular Medicine, Developmental Biology Laboratory, Oulu / FINLAND

2019 (06 – 07) Internship - Gulhane Health Sciences Unıversity Medical Design and Production Center (METUM), Ankara / TURKEY

Publications (if any):

1. ÇEVIK, Ziyşan Buse Yarali; OLCAY, Başak; KARAMAN, Ozan. Determination of Optimum Concentration of NGR Peptide With Anticancer Effect On Breast Cancer Microtissue. In: *2020 Medical Technologies Congress (TIPTEKNO)*. IEEE, 2020. p. 1-3.

2. Antimicrobial Peptide Conjugation on the Catheter Surfaces for the Control and Prevention of Catheter Associated Urinary Tract Infections, *Journal Of Intelligent Systems With Application,* (At the printing stage)