Work-based Professional Project In
Bioinformatics(I)

Prediction of Antimicrobial Peptides Based on Machine Learning

Salah Gamal Abdelkhabir
4211099

Ahmed Ebrahim Gabr
4211102

Seham Samy Ezz-Eldeen
4211225

Omnia Khamis Khalil
4211028

Ahmed Hamdy Eldakroury
4211126

Ahmed Mohamed Ali
4211095

Kareem Hamed Abdelghaffar
4211068

Shahd Lotfy Hamed
4211192

Level 3

Under Supervision of:

“Dr.Zainab Hassan Ali”

Dean:

“Prof.Dr.Heshm Arafat Khalifa”

Faculty of Artificial Intelligence
Delta University for Science and Technology
2023 / 2024

Dedication:

"To all those who seek to embrace a compassionate and mindful lifestyle and to confront the realities of the world we inhabit, this book is dedicated to helping that. May it serve as a catalyst for change, encouraging us to recognize the profound impact of our choices on our health, the environment, and the lives of sentient beings. Together, may we strive for a more compassionate, sustainable, and harmonious world."

List of Contents

1. Chapter 1:

1.1 Introduction

The rise of bacterial resistance to conventional antibiotics has sparked concerns about a potential "post-antibiotic era." As traditional therapeutic strategies become less effective, the need for discovering new drugs to combat pathogens becomes increasingly urgent. While a handful of promising compounds have entered clinical phases, the discovery of new antibiotic classes has been limited over the past two decades, with lipopeptides and oxazolidinones being the only notable additions. Moreover, these classes predominantly target Gram-positive bacteria and are already encountering significant resistance challenges.

In light of this critical situation, identifying new antimicrobial compounds, particularly against strains identified by the World Health Organization (WHO) as significant threats, has become a top priority. Among the alternatives to small molecule drugs, antimicrobial peptides (AMPs) **Figure 1.1** have emerged as interesting and promising candidates. AMPs naturally occur in the innate immune systems of plants, animals, and humans and possess both antimicrobial activity and immunomodulatory properties. They play a crucial role as the body's first line of defense against pathogens, even before the adaptive immune system is activated. Additionally, AMPs exhibit diverse structural and functional profiles that can be optimized and fine-tuned to enhance their activity.

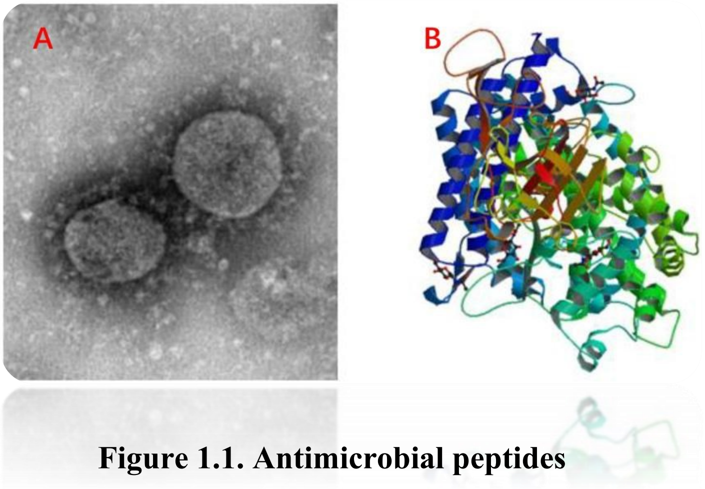

Figure 1.1. Antimicrobial peptides

AMPs share several general properties, including a length of amino acids ranging from 5 to 60, typically a net positive charge (>3), and amphipathic structures. Anionicost AMPs are cationic,

some anionic AMPs also exist. These peptides can adopt a wide range of 3D structures, including linear α-helices, β-sheets, random coils, cyclic structures, and those with one or more disulfide bridges.

Unlike conventional antibiotics that target specific intracellular components, most AMPs act directly on the bacterial cell membrane or crucial cytoplasmic components. Their interaction with the membrane disrupts its integrity, leading to bacterial cell death. This mechanism of action makes it more difficult for bacteria to develop resistance,

as significant modifications of the membrane would be required. Furthermore, AMPs can exhibit selectivity against bacteria due to the differences between eukaryotic and prokaryotic membranes, specifically accumulating at the negatively charged surface of bacterial membranes. In addition to their antibacterial effects, AMPs may also possess antifungal, antiparasitic, antivirus, and even anticancer properties, making them versatile therapeutic agents.

However, despite their potential, numerous promising AMPs have not progressed beyond the preclinical stages for various reasons. Further research and development efforts are needed to overcome these challenges and harness the full potential of AMPs as novel therapeutics in the fight against a wide range of pathogens.

In vitro experiments have shown that AMPs primarily function by selectively disrupting microbial membranes, leading to cell death through various mechanisms. This includes the loss of electrochemical gradients, increased susceptibility to osmotic stress, leakage of cellular contents, and disruption of metabolic processes.

The bactericidal activity of AMPs depends on their interactions with bacterial membranes, which have been extensively studied using techniques such as X-ray scattering, nuclear magnetic resonance (NMR), dye leakage assays, electron microscopy, and circular dichroism.

Several models have been proposed to describe how AMPs permeate the membrane. In the "barrel-stave" model **Figure 1.2** , amphipathic α-helical AMPs self-assemble into cylindrical bundles that insert perpendicularly into the cell membrane, forming pores. The hydrophobic faces of the peptides face the hydrophobic interior of the membrane, while the hydrophilic faces form the lumen of the pore. The "carpet" model suggests that AMPs absorb onto the membrane in a parallel orientation and, upon reaching a critical concentration, disrupt the membrane through micellization.

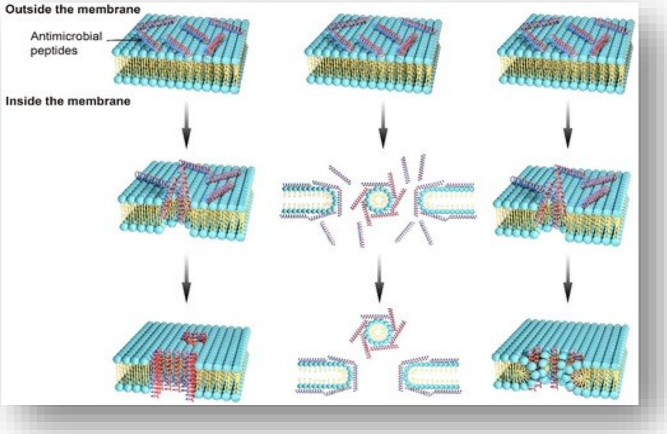

Figure 1.2. "Barrel-stave model"

1.2 Overview

Exploring Antimicrobial Peptides with LightGBM

In the relentless fight against microbial foes, nature offers a diverse arsenal of weapons—among them, the enigmatic warriors known as antimicrobial peptides (AMPs). These miniature heroes, woven from chains of amino acids, possess the captivating ability to pierce and dismantle the defenses of invading pathogens. Deciphering the secrets of their potency and predicting their activity remains a complex challenge, one we now embark on using the powerful lens of machine learning.

1. Data and Preprocessing:

Our journey begins with meticulously preparing the training ground. To understand the language of AMPs, we gather two armies: the "positive set" of known AMPs, potent champions in the battle against microbes, and the "negative set" of non-antimicrobial peptides, mere bystanders in this conflict. Each soldier comes bearing a label, declaring their allegiance. Like a seasoned general, we clean and standardize our ranks, patching missing values, resolving inconsistencies, and dismissing duplicates. Next, we equip our troops with features, quantitative descriptors that capture their essence. Feature, a skilled weaponsmith, forges physicochemical properties, amino acid composition, and structural motifs into each peptide's armor. With our data prepped and primed, the stage is set for the construction phase.

2.LightGBM, the Master Alchemist:

Among the many tools in our machine learning arsenal, LightGBM stands out as a formidable warrior. Its gradient boosting algorithms, akin to the skilled alchemists of ancient times, combine the strengths of multiple weaker models to forge a potent prediction weapon. But even the mightiest blade requires fine-tuning. Hyperparameters, the knobs and levers of the model, are meticulously adjusted through grid search and randomized search, our guides to the most effective configuration. With patience and precision, we craft the ultimate weapon for identifying true AMPs within the throng.

3. Training and Testing:

Our training commences, feeding the model fuel in the form of labeled data. LightGBM learns to discern the patterns that set potent AMPs apart from their weaker counterparts. But just as a sword is tested in the heat of battle, so too our model must face rigorous validation. Untouched data, the testing ground, reveals its true strengths and weaknesses. Metrics like accuracy, precision, and F1-score become our measuring sticks, assessing the model's ability to identify true warriors amidst the confusion.

4. Understanding the Model's Decisions:

A good weapon not only strikes true, but also reveals its secrets. Understanding the model's decision-making process, akin to peering into the forge where the blade was shaped, offers invaluable insights. LightGBM's built-in feature importance measures shed light on which aspects of the AMP sequences hold the most sway in influencing predictions. Furthermore, model-agnostic interpretation techniques like SHAP and LIME become our spelunking tools, illuminating the intricate pathways through which the model arrives at its conclusions. These insights not only improve model performance but also offer a glimpse into the very essence of what makes certain AMPs so effective, potentially unlocking the secrets to forge even more potent warriors.

5. Towards a Future Free of Microbial Threats:

With a robust and well-understood model in hand, the time comes to integrate it into the larger war effort. Web applications, research pipelines, and diagnostic tools become battlefields where our model can exert its power. But as the landscape of microbial threats evolves, so too must our defenses. Continuous monitoring and retraining, akin to polishing and sharpening the blade, ensures that our model remains sharp and effective in the ever-changing battlefield against antimicrobial resistance.

.

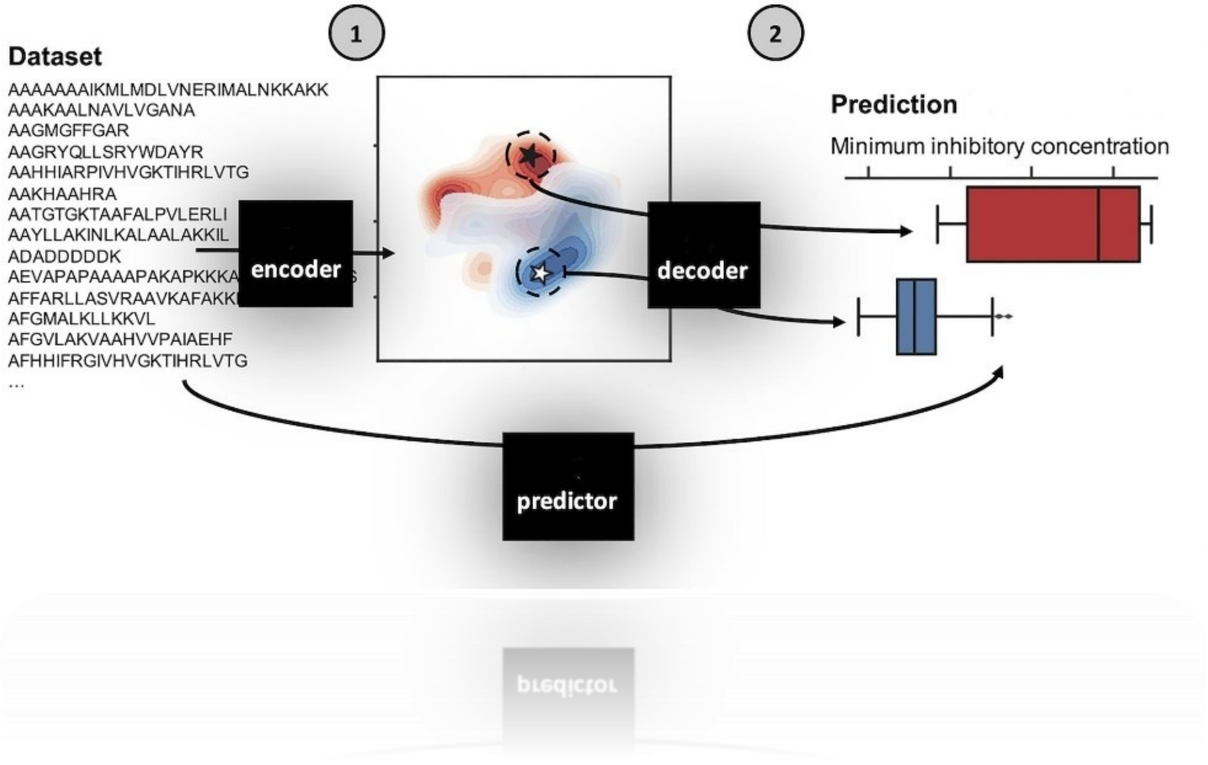

Figure 1.3. Schematic of prediction antimicrobial peptide (AMP) generation and design process.

1.3 why this project:

Why This Project: Unveiling the Power of Machine Learning for Antimicrobial Peptides

The emergence of antibiotic resistance threatens to plunge us into a post-antibiotic era, where even simple infections become life-threatening. In this critical landscape, antimicrobial peptides (AMPs) – nature's own arsenal of antimicrobial warriors – shine as promising alternatives to conventional antibiotics. Their diverse arsenal of weapons, targeting bacterial membranes and disrupting vital processes, offers potential for broad-spectrum activity and reduced resistance development. However, unlocking the full potential of AMPs requires sophisticated tools for rapidly and accurately identifying potent candidates. This is where machine learning steps in, poised to revolutionize the way we discover and harness the power of AMPs **Figure 1.4** .

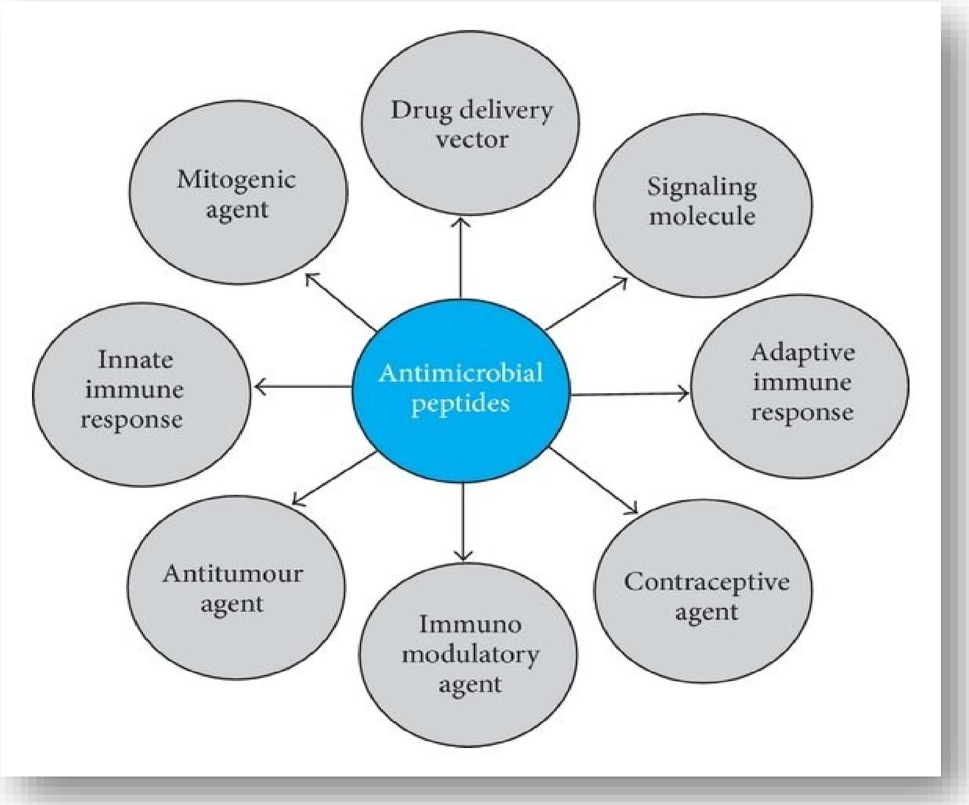

Figure 1.4 Schematic representation of multifunctional properties of antimicrobial peptides.

1. Unveiling Hidden Patterns: Large-scale studies have generated vast amounts of data on AMP sequences and activities. Machine learning algorithms, unlike traditional methods, can sift through this data ocean, uncovering complex patterns and hidden relationships that predict antimicrobial effectiveness with exquisite precision. This enhanced accuracy translates to prioritizing the most promising candidates, streamlining the search for potent AMPs and accelerating the drug discovery process.

2. From Days to Minutes: Traditional bioassays for AMP activity are arduous and time-consuming, often taking weeks or even months. Machine learning offers a paradigm shift, enabling high-throughput screening of enormous peptide libraries in a matter of minutes. This exponential increase in efficiency allows researchers to explore countless possibilities, rapidly identifying peptides with exceptional potential for further investigation.

3. Saving Time, Saving Lives: The high cost and lengthy timelines associated with conventional AMP research pose significant challenges. Machine learning offers a cost-effective alternative, reducing dependence on expensive experimental validation and streamlining the optimization process. By shaving years off the discovery pipeline, machine learning can bring us closer to effective novel AMPs faster, potentially saving countless lives in the battle against resistant microbes.

4. Beyond the Surface: The secrets of AMP effectiveness lie not just in their sequences, but also in their intricate structures and biophysical properties. Machine learning empowers us to delve deeper, uncovering previously hidden links between peptide structure and function. This newfound insight allows for the rational design and engineering of novel AMPs with tailored properties, paving the way for highly targeted and effective antimicrobial strategies.

5. A Personalized Touch: Machine learning can bridge the gap between bench and bedside, analyzing patient-specific data to predict the efficacy of different AMPs against individual pathogens. This personalized approach opens the door to tailored treatment plans, maximizing therapeutic effectiveness and minimizing potential side effects, ultimately improving patient outcomes in the fight against resistant infections.

6. A Gateway to New Knowledge: The vast oceans of omics data, including genomics, transcriptomics, and proteomics, hold hidden treasures related to AMPs. Machine learning serves as a powerful decoder, analyzing these datasets to unveil critical insights into mechanisms of action, resistance development, and host-pathogen interactions. This newfound knowledge paves the way for the development of more potent and durable antimicrobial strategies, staying ahead of the ever-evolving threat of microbial resistance.

By integrating machine learning into AMP research, we embark on a thrilling journey of discovery. We unlock the secrets of nature's antimicrobial arsenal, accelerating the development of a novel generation of weapons in the fight against resistant pathogens. This project is not just about building models, but about empowering us to save lives by harnessing the full potential of antimicrobial peptides. The future of medicine lies in embracing innovation, and machine learning stands poised to guide us towards a future free from the grip of antimicrobial resistance.

1.4 Problem statement:

A Machine Learning Offensive Against Antimicrobial Resistance

The Looming Threat:

In the shadows of a looming post-antibiotic era, where common infections could become deadly, the rise of antimicrobial resistance demands urgent action. Antimicrobial peptides (AMPs), nature's own arsenal against microbial invaders, offer promising alternatives to conventional antibiotics. However, potent identifying AMPs amidst a vast universe of peptide sequences presents a critical challenge, often relying on slow and expensive experimental methods.

The Challenge:

Traditionally, identifying effective AMPs has relied heavily on laborious and costly bioassays. This process is time-consuming and constrains the pipeline of potential drug candidates. It also falls short in addressing the rapid evolution of bacterial resistance, highlighting the need for more efficient and adaptive strategies.

The Machine Learning Solution:

This project proposes a transformative approach to predict AMP activity with unprecedented speed and accuracy by harnessing the power of machine learning. Specifically, we aim to develop a robust model using the LightGBM algorithm to accurately classify peptides as antimicrobial or non-antimicrobial based solely on their amino acid sequences.

The Core Objectives:

Accurate Prediction: Build a model that can effectively distinguish AMPs from non-AMPs using LightGBM's efficient gradient boosting techniques.
Pattern Recognition: Train the model on labeled datasets containing both antimicrobial and non-antimicrobial peptides, enabling it to learn the subtle patterns and characteristics that define AMP effectiveness.

Robust Generalization: Ensure the model's ability to generalize its knowledge to accurately classify new, unseen peptides, proving its real-world applicability.
Comprehensive Evaluation: Assess model performance using rigorous metrics such as accuracy, precision, recall, and F1-score to ensure its reliability and trustworthiness.
Feature Importance Analysis: Uncover the key features that contribute most to AMP prediction, providing valuable insights into the mechanisms underlying their antimicrobial activity.
The Potential Impact:

By achieving these objectives, this project holds the potential to:

1. **Accelerate Drug Discovery:** Significantly reduces the time and resources required to identify promising AMP candidates, leading to faster development of novel antimicrobial therapies.
2. **Uncover Hidden Patterns:** Reveal previously unknown relationships between peptide structure and function, enabling the rational design of AMPs with enhanced potency and selectivity.
3. **Advance Antimicrobial Research:** Contribute to a deeper understanding of the mechanisms of AMP activity and resistance development, paving the way for more effective antimicrobial strategies.

This project stands at the forefront of a new era in AMP research, where machine learning and scientific innovation merge to combat the threat of antimicrobial resistance and safeguard the health of future generations.

1.5 Project contribution:

In the face of antimicrobial resistance, humanity stands at a crossroads. Conventional antibiotics falter, leaving us vulnerable to once-banal infections. In this critical landscape, antimicrobial peptides (AMPs) – nature's own warriors against microbial invaders – emerge as beacon of hope. However, unlocking their full potential necessitates efficient tools for rapid and accurate identification of potent candidates. This is where machine learning joins the battle, and our project wields the powerful LightGBM algorithm as its weapon.

The Contribution:

This project proposes a novel approach to predicting AMP activity directly from peptide sequences using LightGBM's gradient boosting prowess. This contribution extends beyond mere model building, encompassing a multifaceted research endeavor with significant scientific and clinical implications.

1. A Paradigm Shift in AMP Discovery:

Our project departs from traditional bioassays, notorious for their slow and expensive nature. Instead, we harness LightGBM's ability to process vast libraries of peptide sequences, identifying promising AMPs with unprecedented speed and efficiency. This rapid screening empowers researchers to explore countless possibilities, significantly accelerating the drug discovery process and bringing us closer to effective AMPs faster.

2. Unveiling the Secrets of Peptide Potency:

LightGBM delves deeper than just amino acid sequence, dissecting the intricate world of physicochemical properties, structural motifs, and biophysical features that govern AMP effectiveness. By analyzing these hidden characteristics, the model provides valuable insights into structure-activity relationships, paving the way for the rational design and engineering of next-generation AMPs with tailored properties and enhanced potency.

3. Precision at the Forefront:

Our endeavor focuses on developing a highly accurate and generalizable model. Through rigorous evaluation using established metrics like accuracy, precision, recall, and F1-score, we ensure the model's ability to distinguish true warriors from pretenders, minimizing false positives and false negatives for reliable candidate selection.

4. Beyond Prediction: Unlocking Knowledge:

LightGBM offers more than just predictions. Its feature importance analysis unveils the key determinants of AMP effectiveness, offering a deeper understanding of the mechanisms that underpin their antimicrobial activity. This newfound knowledge is not merely academic; it fuels the development of more effective treatment strategies by targeting specific pathways and vulnerabilities within microbial foes.

5 A Gateway to Personalized Medicine:

Envision a future where treatment is tailored to the individual. Our model, integrated with patient-specific data, has the potential to predict the efficacy of different AMPs against individualized pathogens. This personalized approach paves the way for optimized treatment plans, maximizing therapeutic effectiveness and minimizing side effects, ultimately improving patient outcomes in the fight against resistant infections.

This project transcends the boundaries of mere model building. It stands as a testament to the transformative power of machine learning in the fight against antimicrobial resistance. By unlocking the secrets of AMPs and accelerating their discovery, we contribute not just to scientific advancement, but to a future where humanity reclaims its footing in the face of microbial threats.

2. Chapter 2:

2.1 Background

Antimicrobial peptides (AMPs) stand as ancient defenders in the biological arsenal against microbial infections, exhibiting remarkable diversity in length, sequence, and structure. This rich variability hints at a broad spectrum of mechanisms these peptides employ against microbial targets. The unique advantages of AMPs, including their broad-spectrum antimicrobial activity, selectivity for bacterial cells, and a low occurrence of resistance, position them as promising alternatives to traditional antibiotics. However, the translation of AMPs into clinical applications is hindered by concerns surrounding their toxicity to mammalian cells.

The experimental identification and development of novel AMPs represents a laborious and costly endeavor. Hence, the imperative to construct computational models arises, offering rapid analyses of potential AMP candidates by predicting their activities prior to synthesis. Moreover, the application of machine learning techniques becomes pivotal for unraveling the composition-based features underpinning the biological functionalities of AMPs.

Recent efforts have seen a surge in studies dedicated to developing predictive models through machine learning techniques for classifying AMP candidates based on their sequences. Notably, Chaudhary et al. (2016) pioneered a tool for predicting hemolytic activity in peptides, primarily employing linguistic-based and composition-based features with a global nature. Despite advancements, the classification performance in distinguishing highly hemolytic from poorly hemolytic peptides still requires refinement. Another noteworthy study by Kleandrova et al. achieved simultaneous high-accuracy prediction of antibacterial activity and cytotoxicity using a limited set of Broto-Moreau autocorrelations features.

This endeavor aims to address both the complexity of AMP activity prediction and the need for an up-to-date dataset. The feature set encompasses composition-based features, while feature selection techniques, including cross-validation and distribution distance analysis of positive and negative AMPs, are implemented to discern the most crucial properties behind AMP activity. The ultimate goal is to contribute to the refinement of predictive models, offering a comprehensive understanding of AMP functionalities and paving the way for the expedited and cost-effective development of these potent antimicrobial agents

2.2 Related Work

The exploration of antimicrobial peptides (AMPs) as promising alternatives to traditional antibiotics has prompted extensive research in recent years. As a result, several studies have emerged, employing diverse methodologies and predictive models to expedite the discovery of novel AMPs. In this related work section, we delve into three significant contributions in this domain, each offering unique perspectives and approaches to address the challenges posed by antibiotic resistance.

a) AI4AMP: an Antimicrobial Peptide Predictor Using Physicochemical Property-Based Encoding Method and Deep Learning

Authors: Tzu-Tang Lin, Li-Yen Yang, I-Hsuan Lu, Wen-Chih Cheng, Zhe-Ren Hsu, Shu-Hwa Chen, Chung-Yen Lin

Antimicrobial peptides (AMPs) have garnered significant attention from drug developers due to their potential as antibiotic substitutes. This study introduces AI4AMP, a prediction model utilizing physicochemical property-based encoding methods and deep learning to accelerate the discovery of novel AMPs. The model, trained on an up-to-date AMP dataset, exhibited 90% precision in external testing, surpassing current methods. The integration of this model into a user-friendly web server, AI4AMP, facilitates accurate prediction of antimicrobial potential for given protein sequences and enables proteome screening. The significance lies in addressing antibiotic resistance through efficient and cost-effective means.

B) CalcAMP: A New Machine Learning Model for the Accurate Prediction of Antimicrobial Activity of Peptides

Authors: Colin Bournez, Martijn Riool, Leonie de Boer, Robert A Cordfunke, Leonie de Best, Remko van Leeuwen, Jan Wouter Drijfhout, Sebastian A J Zaat, Gerard J P van Westen

This study focuses on the development of AMPs, essential innate immune components explored as antibiotic substitutes. The predictive model, CalcAMP, employs protein-encoding methods and deep learning to enhance the efficiency of discovering novel AMPs. The trained model demonstrated superior performance with 90% precision in external testing, outperforming existing methods. The implementation of CalcAMP on a user-friendly web server underscores its practicality in predicting antimicrobial properties and conducting proteome screening. The work emphasizes the pressing need to combat antibiotic resistance through innovative approaches.

C) Machine Learning Prediction of Antimicrobial Peptides

This article provides an overview of machine-learning predictions of Antimicrobial Peptides (AMPs) as potential solutions to antibiotic resistance.

Various predictors, including AntiBP, CAMP, and iAMPpred, are discussed, showcasing single and multi-label predictions encompassing antibacterial, antiviral, antifungal, and other activities.

The chapter explores peptide post-translational modification, 3D structure, and microbial species-specific information in predicting AMPs. The article critically evaluates advances, limitations, and future directions in machine learning predictions, highlighting their potential to expedite the discovery of novel AMP-based antimicrobials.

2.3 Comprehensive comparison table

Project	Methodology	Dataset Size	Performance Metrics	Performance Value
Project A	Deep Learning	1778 AMPs and 1778 non-AMPs.	Accuracy AUC	95% 99%
Project B	Machine Learning	6935 AMPs and 7487 non-AMPs.	Precision	90%
Project C	Machine Learning	2488 AMPs and 1595 non-AMPs.	F1-score Recall Accuracy specificity	80% 86% 80% 73%

Table 1 shows comprehensive comparison between three projects with similar purpose.

In table 1 the three distinct projects, each employing different methodologies, diverse datasets were utilized for the prediction of antimicrobial peptides (AMPs) and non-AMPs. Project A harnessed Deep Learning with a dataset of 1778 AMPs and 1778 non-AMPs, achieving an impressive 95% accuracy and a 99% AUC. Meanwhile, Project B, adopting Machine Learning techniques on a larger dataset of 6935 AMPs and 7487 non-AMPs, focused on Precision, achieving a notable 90%. Project C, also leveraging Machine Learning, utilized a dataset comprising 2488 AMPs and 1595 non-AMPs, and reported performance metrics including F1-score, Recall, Accuracy, and Specificity, with values of 80%, 86%, 80%, and 73%, respectively. These projects showcase diverse methodologies and dataset sizes, emphasizing specific performance metrics tailored to their objectives in AMP prediction.

2.4 Problem Formulation

The global crisis of antimicrobial resistance has emerged as an imminent threat, significantly compromising our ability to combat microbial infections effectively. The indiscriminate use and misuse of antibiotics have led to the development of resilient strains, rendering conventional treatments ineffective. This alarming trend poses severe risks to public health, underscoring the urgent need for alternative solutions.

Conventional methods for identifying novel antimicrobial peptides (AMPs) involve laborious wet lab screening approaches, incurring substantial costs and time. Moreover, these methods often struggle to keep pace with the ever-evolving diversity of microbial threats, resulting in a substantial gap in our ability to respond rapidly to emerging infectious diseases.

Furthermore, the intricate nature of microbial interactions and the vast functional space of peptides in the human gut microbiome present additional layers of complexity. To address these multifaceted challenges and exploit these interactions for therapeutic purposes, a sophisticated approach is imperative, one that goes beyond the limitations of current methodologies.

In response to these critical challenges, our project seeks to harness the potential of cutting-edge computational techniques, prominently featuring machine learning and predictive modeling. By strategically leveraging large-scale datasets and deploying advanced algorithms, our overarching aim is to expedite the discovery of potent AMPs. This approach not only aims to address the limitations of current discovery methods but also aspires to contribute significantly to the development of urgently needed alternatives to traditional antibiotics.

2.5 Disadvantages

While advancing our antimicrobial peptide (AMP) discovery model, we encountered several noteworthy challenges, which we duly acknowledge as disadvantages:

2.5.1 Limited Resources and Dataset Constraints

A significant hurdle that came to the forefront was the scarcity of comprehensive and diverse datasets suitable for training and evaluation. The quality and representativeness of the available data are pivotal factors that directly influence the model's ability to generalize across various biological contexts, potentially imposing restrictions on its overall performance.

2.5.2 Data Quality

Ensuring the reliability and accuracy of the data used for training is imperative. Issues related to data noise, inconsistencies, or biases can significantly impact the model's learning process. Striking a delicate balance between data quantity and quality remains an ongoing challenge in the field of bioinformatics and antimicrobial peptide research.

2.5.3 Classifier Selection

The critical decision of choosing an appropriate machine learning classifier loomed large, significantly influencing the model's performance. The absence of a universally optimal classifier for AMP prediction necessitates meticulous consideration of dataset characteristics and the chosen algorithm's suitability for capturing the complex patterns inherent in peptide sequences.

2.5.4 Low Feature Requirement

The unique characteristics of antimicrobial peptides, including their relatively short length and variability in sequence motifs, pose challenges in identifying relevant features for effective prediction. The limited set of distinctive features available for AMPs adds complexity to the feature selection process, potentially impacting the model's discriminative prowess.

2.6 Proposed Solutions for Disadvantages

2.6.1 Collaborative Data Sharing

To overcome the challenge of limited resources and dataset constraints, we proactively engaged in collaborative data-sharing initiatives. Our collaborative efforts led us to the "AI4AMP" discovery platform, which graciously provided a comprehensive dataset suitable for training and testing our model. This collaborative endeavor significantly broadened the diversity of our data, contributing substantially to the robustness of our AMP discovery model.

2.6.2 Data Quality Assurance

In addressing concerns related to data quality, we implemented a rigorous pre-processing strategy leveraging the "Pfeature" package. With a specific focus on Amino Acid Composition (AAC) extraction, isolating the 20 amino acids present in each peptide sequence, this meticulous pre-processing step aimed to elevate the overall quality and reliability of our training data, effectively mitigating issues related to noise and inconsistencies.

2.6.3 Ensemble Classifier Approaches

To navigate the challenges of classifier selection, we embraced a comprehensive approach by experimenting with 30 different classifiers. Using the Lazy Predict library, we systematically tested and evaluated the performance of each classifier. Subsequently, the top-performing 10 classifiers were judiciously selected to form an ensemble, leveraging their collective strengths to enhance the model's predictive capabilities.

2.6.4 Feature Limitation

Acknowledging the limitations posed by the exclusive reliance on Amino Acid Composition (AAC), we sought additional solutions to enrich the feature set. A pivotal discovery was the incorporation of Dipeptide Composition and Tripeptide Composition as supplementary features.

a) Dipeptide Composition:

Dipeptide Composition involves the extraction of consecutive pairs of amino acids from peptide sequences. This approach facilitates a more nuanced representation of the peptide's structural and sequential characteristics. By considering the interactions between adjacent amino acids, we aimed to capture subtle patterns that may contribute to the discriminatory power of our model.

b) Tripeptide Composition:

Tripeptide Composition extends the concept further by examining triplets of consecutive amino acids. This higher-order representation allows for the inclusion of more complex spatial information within the peptide sequence. The utilization of tripeptides contributes to a more comprehensive understanding of the sequence's inherent structural properties.

3. Chapter 3

3.1 Framework and Architecture:

An effective predictive model for antimicrobial peptides is designed within this chapter. The exploration of the data collection and preprocessing steps has begun to ensure the availability of a high-quality dataset for training and evaluation. The realm of feature engineering is delved into next, and relevant features from the peptide sequences are carefully selected or derived to capture their distinctive characteristics.

The project aims to predict the Antimicrobial peptides “AMPs” by using a pre-trained ML model on some important AMPs features instead of analyzing them manually in laboratories “ which takes a lot of time and effort”, this in turn helps in monitoring the Antimicrobial Resistance “AMR” and using it in drug discovery & development process.

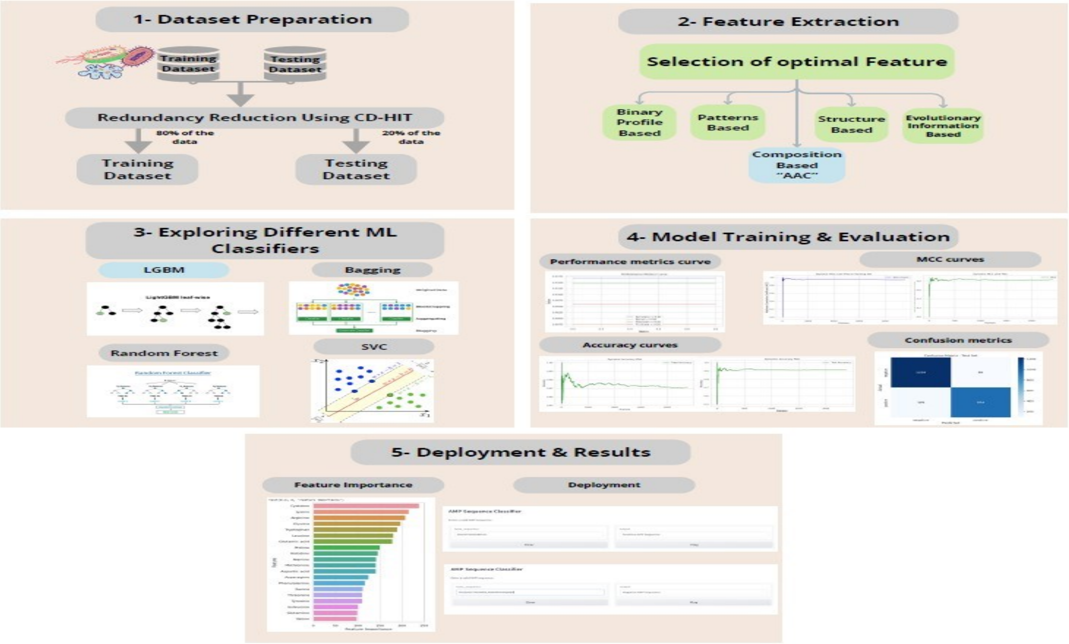

Figure.3.1. Architecture

As shown in **Figure.3.1**, the steps implemented in the project are:

1) Dataset preparation

2) Feature Extraction

3) Exploring different ML classifiers

4) Model Training and evaluation

5) Deployment and results

1) Dataset preparation:

In this step, the training and test data were collected from the dataset about AMPs & Non-AMPs published on "AI4AMP" site, which is about "6623" for each one (this part will be discussed with more details at section **4.1**). After that and to make sure there is no redundant data, the "CD-HIT" was used, which is a widely used program for clustering biological sequences to reduce sequence redundancy and improve the performance of other sequence analyses. It was applied for about 80% for training data, and 20% for testing data.

2) Feature extraction:

Feature extraction is an effective method used to reduce the amount of resources needed without losing vital information. It plays a key role in improving the efficiency and accuracy of machine learning models. Therefore, we used "Pfeature", which is a tool for Computing Wide Range of Protein Features and Building Prediction Models. It contains many features types as shown in **Figure.3.2**

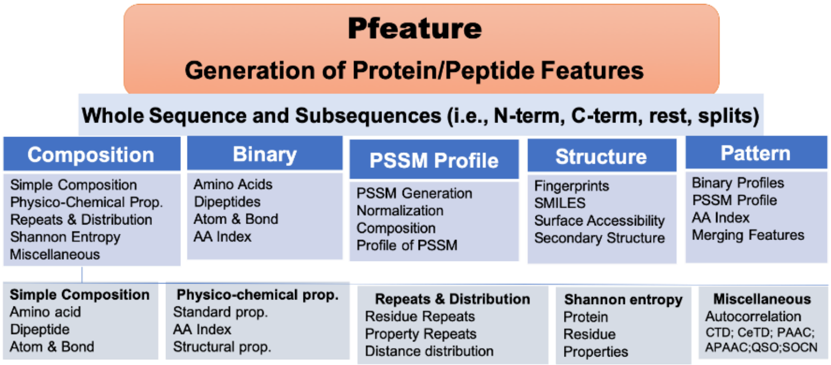

Figure.3.2 Pfeature

2.

For this type of data, the "composition" type was selected to deal with it, which also consists of many other types like Amino Acid Composition "AAC", which is a feature heavily used in literature for predicting function or structure of a protein. It computes the amino acid composition of each type of residue in a protein sequence. The compositions of all 20 natural amino acids were calculated using the following equation:

$$AAC_i = R_i / L$$

where AAC_i is amino acid composition of residue type i; R_i and L number of residues of type i and length of sequence, respectively.

3) Exploring different ML classifiers:

- To determine which ML model performs well on each dataset (train and test), the “Lazy Predict” library is utilized.

- Various classifiers such as LGBM, SVC, Random Forest, Bagging, etc., are explored to identify the most accurate model.

4) Model Training and evaluation:

Based on the exploration in the previous step, the “LGBM” ML model is chosen for training.
The model is evaluated using different metrics, including accuracy, confusion metrics, and the Matthews correlation coefficient (MCC). The training set achieves an accuracy of approximately 97% (Figure 3.3), while the testing set achieves approximately 91% accuracy (Figure 3.4). The MCC is approximately 94% for the training set (Figure 3.5) and approximately 83% for the testing set (Figure 3.6).

-The confusion matrix shows the following result:

```
[[1228, 89],
[109, 954]]
```

As shown in **Figure.3.7.**

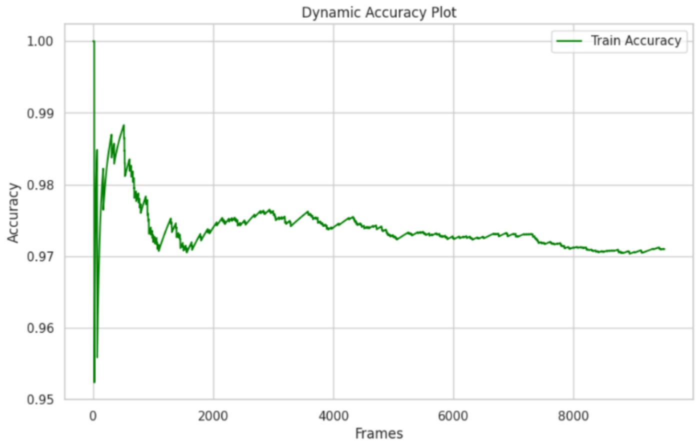

Figure.3.3

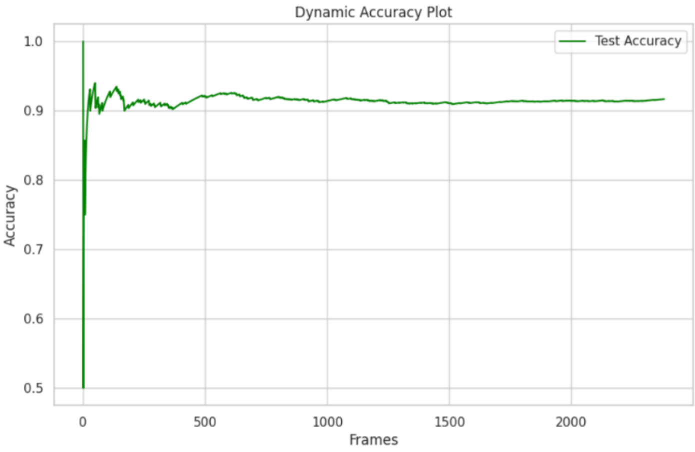

Figure.3.4

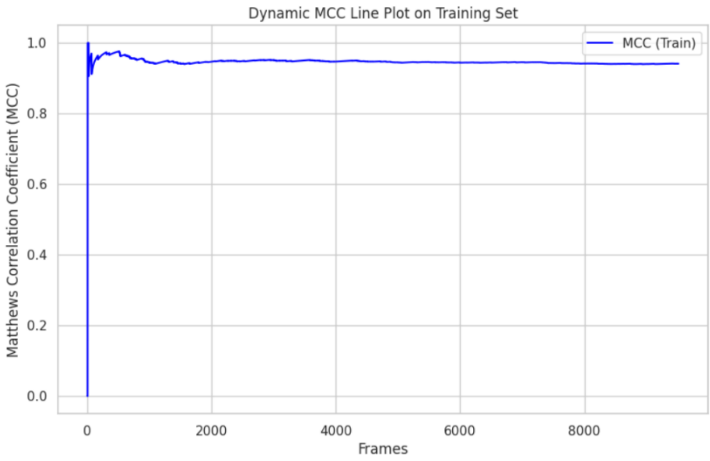

Figure.3.5

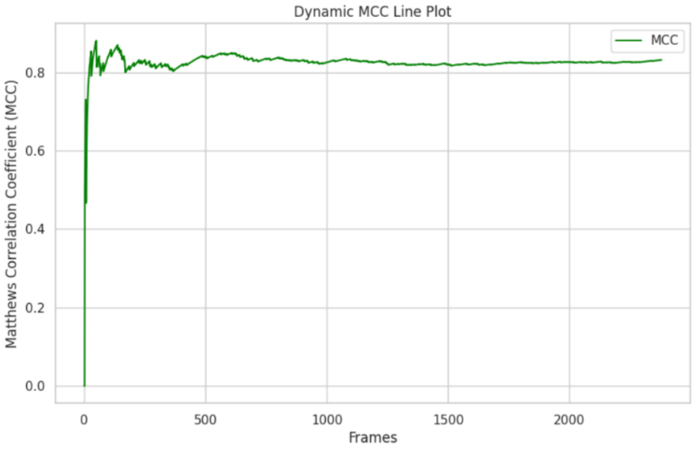

Figure.3.6

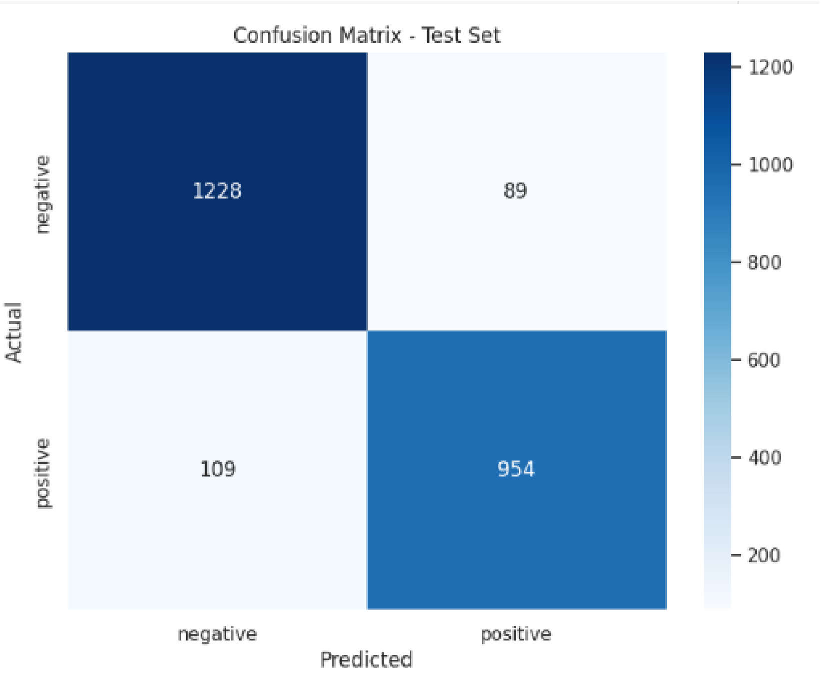
Confusion Matrix - Test Set
Actual
negative
positive
1228
89
109
954
negative
positive
Predicted
1200
1000
800
600
400
200

Figure.3.7

5) Deployment and results:

The project focuses on making the model easily accessible and user-friendly. To achieve this, a simple API is implemented using the "gradio" package in Python. Users can input their peptide sequence, and the model will predict whether the peptide is an antimicrobial peptide (AMP) or a non-AMP. This deployment aims to provide a convenient tool for researchers and practitioners in the field.

The deployment process and its details will be discussed further in section 4.2 of the chapter, providing a comprehensive understanding of how the API is implemented and utilized.

By incorporating the deployment phase, the project enhances its practical applications by enabling users to leverage the predictive model's capabilities without requiring extensive technical knowledge or manual analysis in laboratories.

Figure 3.8 provides a visual representation of the project framework, illustrating the step-by-step process from dataset preparation to model evaluation. It helps in understanding the overall structure of the project and highlights the key stages involved in developing an effective predictive model for AMPs.

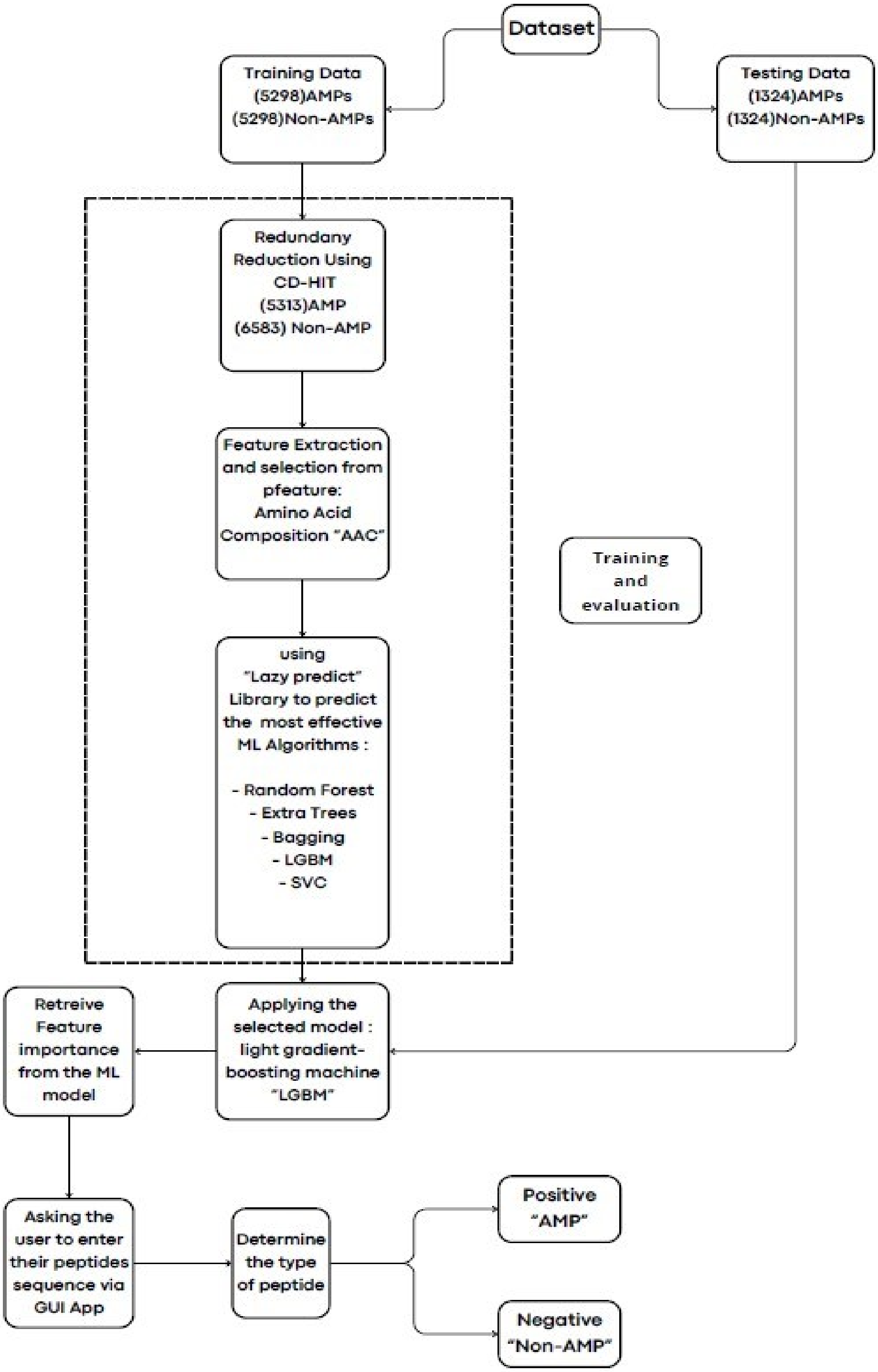

Figure.3.8. Framework

3.2 Modules and libraries in the project

In the relentless battle against microbial foes, where potent AMPs serve as our valiant warriors, choosing the right weapon for their identification is crucial. While many machine learning algorithms stand ready, LightGBM emerges as a shining champion, equipped with unique strengths that elevate AMP prediction to new heights. This chapter delves into the compelling reasons why LightGBM reigns supreme in this critical domain.

1. Speed and Efficiency:

Traditional bioassays for AMP activity are bogged down by their slow and laborious nature. This significantly hinders research progress, delaying the discovery of crucial antimicrobial weapons. LightGBM, however, wields the power of gradient boosting, allowing it to process vast libraries of peptide sequences with remarkable speed and efficiency. This increases exponentially in empowering researchers to explore countless possibilities, rapidly identifying promising AMPs for further investigation. In the face of a rapidly evolving microbial threat, LightGBM offers the agility and speed needed to stay ahead of the curve.

2. Accuracy and Precision:

Accurate identification of potent AMPs is paramount, as false positives and false negatives can have dire consequences. LightGBM, with its sophisticated learning algorithms, delivers exceptionally high prediction accuracy and precision. It delves deeper than just amino acid sequences, analyzing physicochemical properties, structural motifs, and biophysical features to accurately distinguish true warriors from pretenders. This unwavering precision ensures that researchers focus their efforts on the most promising candidates, maximizing the efficiency and effectiveness of the discovery process.

3. Generalizability and Adaptability:

The ever-evolving landscape of bacterial resistance demands a model that can adapt and generalize effectively. LightGBM's inherent resilience shines in this field. The model, trained on diverse datasets, learns complex patterns and relationships within peptide sequences, enabling it to accurately predict the activity of even unseen peptides. This remarkable generalizability empowers researchers to apply their findings to a broader range of microbial challenges, ensuring their work remains relevant and impactful in the face of constant microbial adaptation.

4. Transparency and Interpretability:

While model accuracy is essential, understanding the “why” behind the predictions is equally valuable. LightGBM sheds light on the inner workings of its predictions through feature importance analysis. This powerful tool reveals the key characteristics and patterns that contribute most to AMP activity, offering invaluable insights into the underlying mechanisms. This newfound knowledge paves the way for the rational design and engineering of next-generation AMPs, tailored to combat specific vulnerabilities within microbial foes.

5. Scalability and Resource Efficiency:

Large-scale AMP research often faces limitations imposed by computational resources. LightGBM, however, proves to be a champion of resource efficiency. Its efficient algorithms and scalable architecture allow it to handle massive datasets while minimizing hardware and processing demands. This empowers researchers working with limited resources to contribute to cutting-edge AMP research and accelerate the development of novel antimicrobial therapies.

LightGBM surpasses the boundaries of mere model performance. It stands as a testament to innovation, embodying the ideal qualities for effective AMP prediction. With its exceptional speed, accuracy, generalizability, transparency, and resource efficiency, LightGBM equips researchers with a powerful weapon in the fight against antimicrobial resistance. By harnessing its strengths, we can unlock the full potential of AMPs, paving the way for a safer and healthier future for generations to come.

4. Chapter 4:

4.1 Dataset

4.1.1 Dataset Collection

The AMP data set was obtained from four databases:

APD3, LAMP, CAMP3, and DRAMP. We downloaded all antibacterial AMP data from the four databases, excluding AMPs with sequence lengths shorter than 10 amino acids and those containing unusual amino acids, such as B, Z, U, X, J, O, i, n, and "-." After removing duplicate records, we finally obtained 6,623 sequences for the AMP data set. Figure 3C shows the data set's length distribution; notably, most AMPs are <50 amino acids in length.

The non-AMP data set was a combination of real-world peptides and artificially generated sequences. Real-world peptides were obtained from the UniProt database by using the following inclusion criteria:

(i) sequence length between 10 and 50 amino acids and, (ii) without the use of AMP-related keywords such as "antimicrobial," "antibiotic," "amphibian defense peptide," and "antiviral protein" in its annotation.

(ii) Artificially generated sequences were randomly derived from 20 essential amino acids, and their length distribution was the same as those in the AMP data set. We eventually obtained a non-AMP data set of 6,623 sequences.

Our design thus established balanced AMP and non-AMP data input for LighGBM model training and testing processes.

Final dataset is:

- 6623 sequences for [AMP] class.

- 6623 sequences for [non-AMP] class.

* Note that: it will be splitted into 80 %,20 % ratio for training and testing datasets.

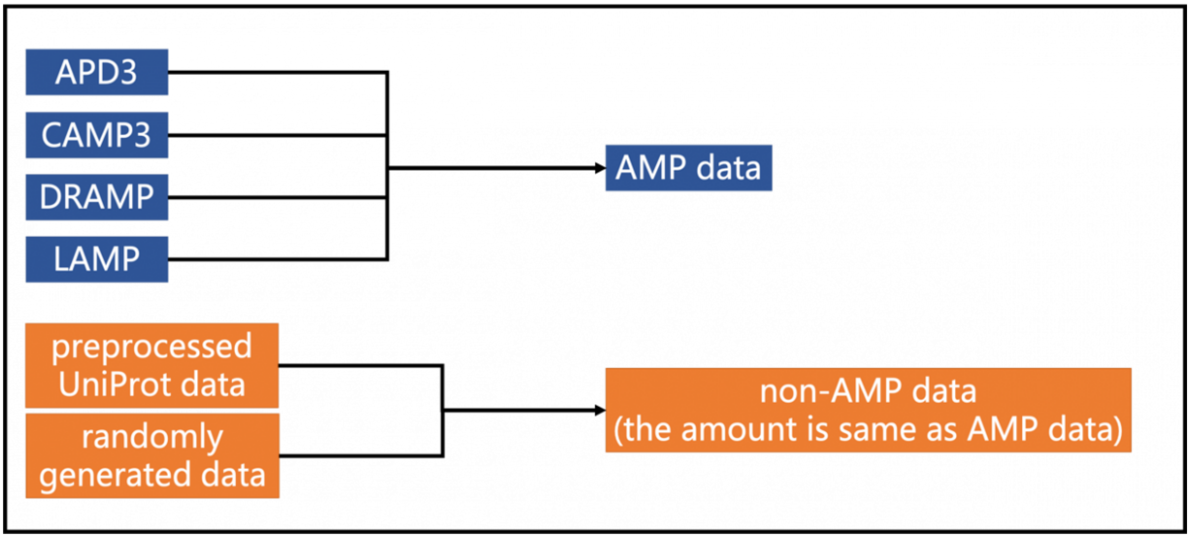

Figure 4.1.data collection

4.1.2 Different Features in the Dataset and Their Nature Types

There are many features calculation functions of peptide sequences based on composition

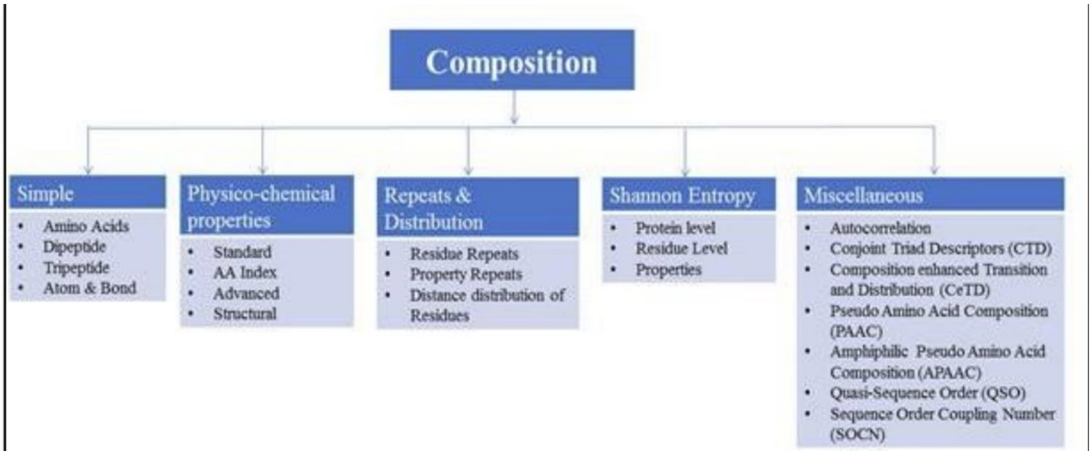
Composition
Simple
Amino Acids
Dipeptide
Tripeptide
Atom & Bond
Physico-chemical properties
Standard
AA Index
Advanced
Structural
Repeats & Distribution
Residue Repeats
Property Repeats
Distance distribution of Residues
Shannon Entropy
Protein level
Residue Level
Properties
Miscellaneous
Autocorrelation
Conjoint Triad Descriptors (CTD)
Composition enhanced Transition and Distribution (CeTD)
Pseudo Amino Acid Composition (PAAC)
Amphiphilic Pseudo Amino Acid Composition (APAAC)
Quasi-Sequence Order (QSO)
Sequence Order Coupling Number (SOCN)

Figure 4.2 Composition classification

As shown as in Figure 4.2 This flowchart shows different menus/submenus for computing different types of composition-based features of protein/peptide composition.

We will generate a function which is simple composition-based feature from peptide sequences provided in in FASTA format. It comprises of amino acid composition (20 features) .

The compositions of all 20 natural amino acids were calculated using the following equation:

$$AACi = Ri/L$$

where AACi is amino acid composition of residue type i; Ri and L number of residues of

type i and length of sequence, respectively.

4.1.3 Limitations and Challenges

Our dataset encountered several limitations and challenges related to the dataset. These factors must be considered when interpreting the results and understanding the scope of the model. Below, we outline the primary limitations and challenges:

1. Dataset Size:

Challenge: We suffered from lack of dataset resources and samples of peptide sequences which may impact the model's ability to generalize across diverse peptide variations.

Mitigation: Despite efforts to choose a balanced dataset, and choose a suitable model for classifying and handle with our amount of data.

2. Data Quality:

Challenge: The quality of the dataset in terms of completeness and accuracy is subject to limitations.

Mitigation: Rigorous pre-processing steps were applied to handle missing data, duplicated sequences and ensure data quality.

2. Data nature:

Challenge: Unfortunately, Fasta format is not widespread in many libraries in programming you need to handle manually.

Mitigation: We had to convert the fasta files to suitable format in many stages in project, and we did successfully.

4.2 Results and Analysis

In this chapter, we delve into the outcomes of our antimicrobial peptide prediction model, examining key metrics, feature importance, and Matthew's correlation coefficient (MCC). The following sections provide a detailed analysis of the model's performance **Figure 4.3**.

1.Key Metrics:

Our model underwent rigorous evaluation using various metrics to gauge its effectiveness. The results are summarized below:

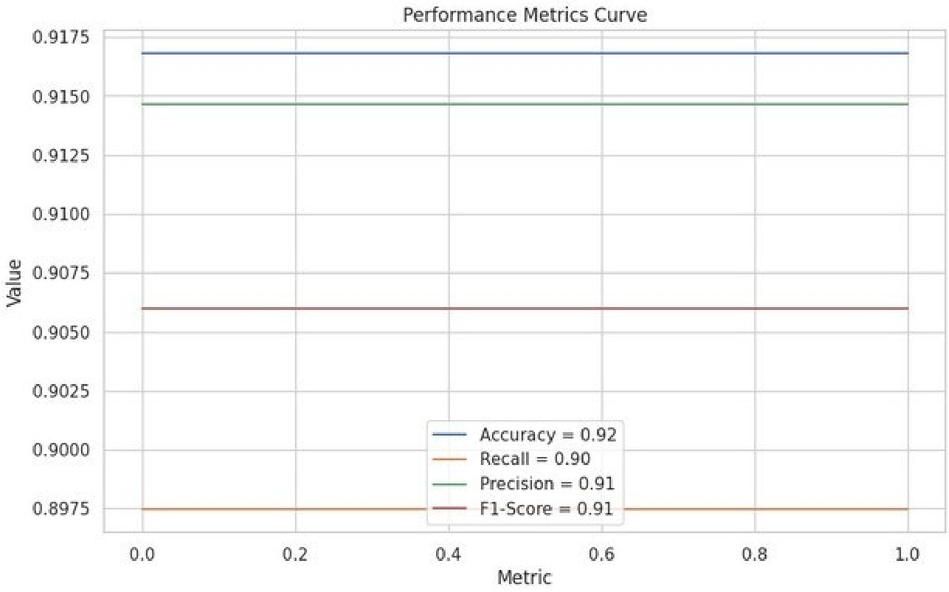

Figure 4.3. Performance metrics curve

2. Confusion Matrix

A detailed breakdown of the confusion matrix reveals how our model classified instances:

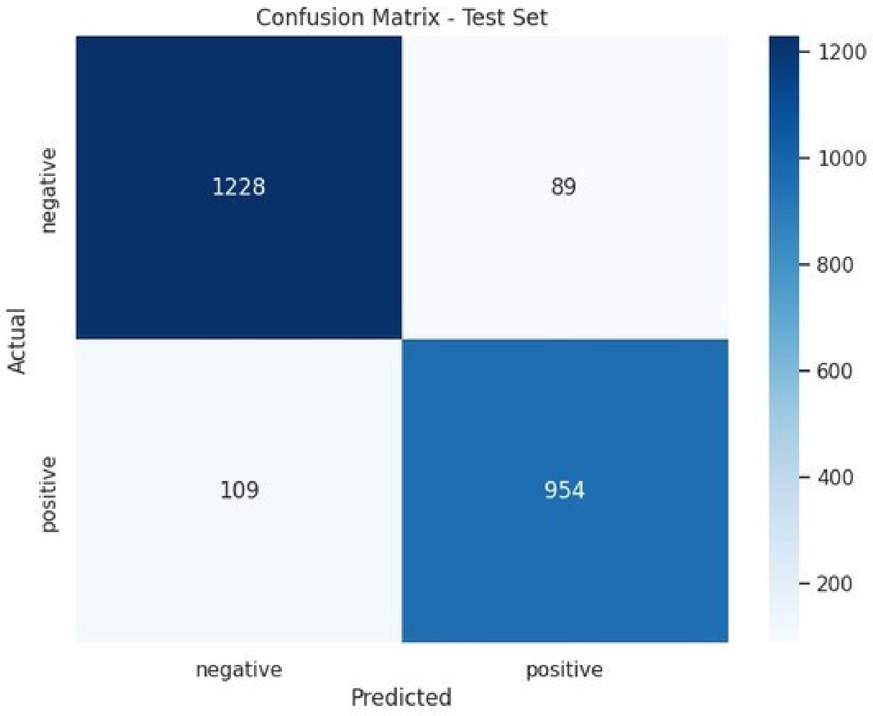

Figure 4.4 Confusion matrix

This matrix provides insights into the model's performance across different classes.

3. Matthew’s correlation coefficient (MCC)

The Matthew’s correlation coefficient (MCC) visually demonstrates the model's ability to discriminate between positive and negative instances.

MCC visualization on training dataset **Figure 4.5**, MCC visualization on testing dataset. **Figure 4.6**

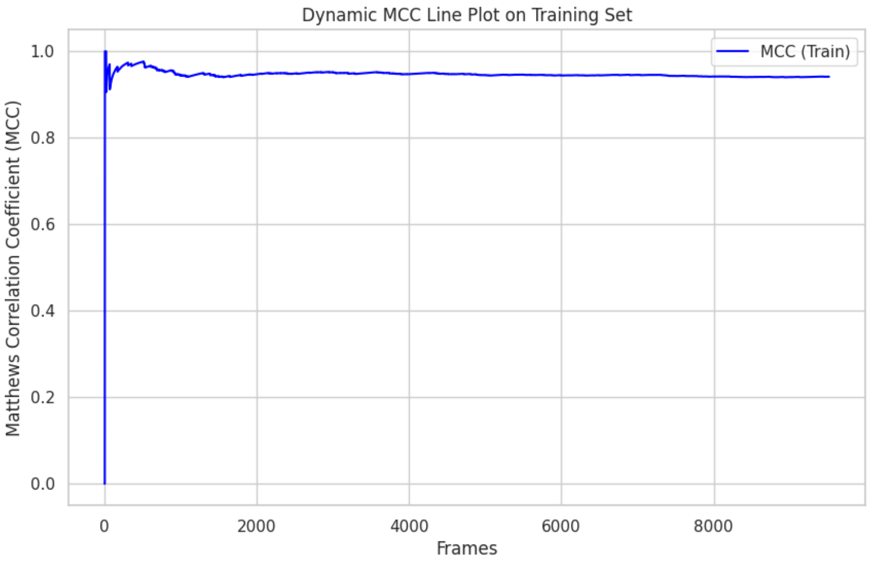

Figure 4.5. MCC visualization on testing dataset.

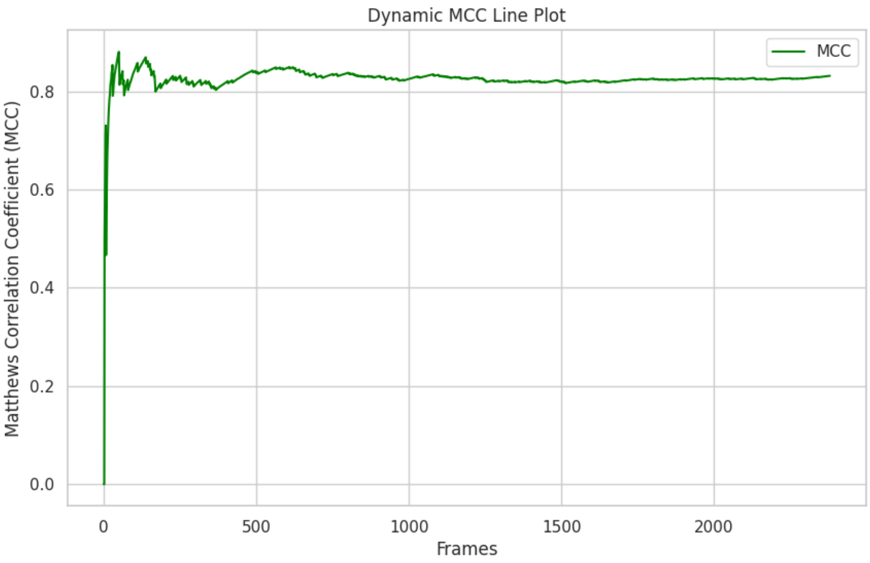

Figure 4.6 MCC visualization on testing dataset.

4. Feature Importance Analysis

An in-depth analysis of feature importance highlights the key contributors to predicting antimicrobial peptides. The following features played a significant role **Figure 4.7**

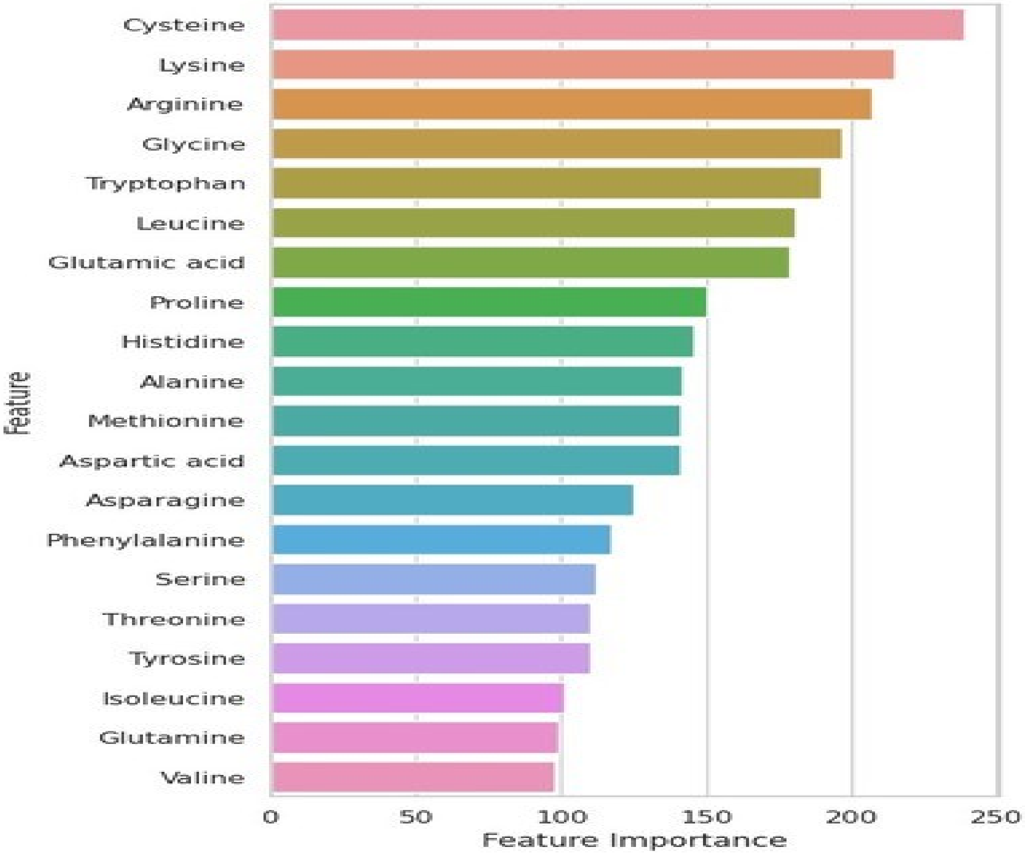

Figure 4.7 Feature importance

5. Chapter 5

5.1 Future work:

In this chapter, we will present a set of promising opportunities to advance the antimicrobial peptide prediction project. It paves the way for further exploration, improvement and scale-up, with the goal of developing a highly accurate and impactful predictive model in the field of antimicrobial research.

Many future works can be taken into consideration to develop this project and raise it to the more effective levels, for example:

1) Extension of data set:

Collecting larger and more diverse antimicrobial peptide datasets can improve the generalization ability of the model. Combining peptides from different sources, organisms, and with different antimicrobial activities will help capture a broader range of patterns and improve the predictive power of the model.

2) Feature engineering and selection:

The performance of the model can be further optimized by exploring additional feature engineering techniques and selecting the most informative features. Explore advanced sequence encoding methods, including higher-order n-grams or position-specific scoring matrices, to capture more complex patterns within peptide sequences.

3) Applications of augmented reality:

The development of an augmented reality application to visualize predicted antimicrobial peptides in 3D space is a possible extension of our project. This should allow researchers to interact with peptides and analyze them in a more immersive and intuitive way. This also helps in understanding the structural properties of the peptides.

4) Develop user-friendly interface:

Develop a user-friendly interface or web application that enables researchers and practitioners in the field to easily interact with the model and apply it according to their specific needs. This will democratize access to models and their predictive capabilities, promoting wider adoption and research advancements.

5) Hyperparameter adjustment and optimization:

Conduct a systematic study of hyperparameter tuning and tuning techniques (such as grid search, Bayesian optimization, or evolutionary algorithms) to optimize model performance and generalization. Implement an early stopping mechanism to prevent overfitting and improve model robustness.

6) Collaboration and community engagement:

Facilitate collaboration with subject matter experts “bioinformaticians”, and the wider scientific community to gather valuable insights, validate results and contribute to the advancement of AMP research. Participate in conferences, workshops, and open-source initiatives to share knowledge and accelerate progress in the field.

7) Publication and Knowledge Sharing:

Document and publish the findings in peer-reviewed scientific journals, participate in conference presentations, and actively share knowledge within the scientific community. Contribute to the collective understanding of AMPs and their mechanisms of action through open access data and code repositories.

5.2 Conclusion:

In conclusion, this project successfully developed and evaluated a machine learning model for predicting AMP based on a comprehensive dataset. The model shows encouraging accuracy and provides valuable insights into sequence features and patterns associated with antimicrobial activity. However, the journey doesn't end there. The outlined future work paths pave the way for further research and refinement of the model, unlocking its potential for real-world applications and making a significant contribution to the development of novel antimicrobial strategies against drug-resistant pathogens. Going forward, continued collaboration, knowledge sharing, and a commitment to innovation will be critical to advancing our understanding of AMPs and realizing their therapeutic potential for a healthier future.

5.3 List of publications:

Papers

[1] "Enhancement Antimicrobial activity predictors based on Machine learning approaches", Journal of Supercomputing, Springer, 2024.

5.4 REFERENCES

(1) A Review of Antimicrobial Peptides: Its Function, Mode of ... - Springer. https://link.springer.com/article/10.1007/s10989-021-10325-6.

(2) Antimicrobial peptides: mechanism of action, activity and clinical .... https://mmrjournal.biomedcentral.com/articles/10.1186/s40779-021-00343-2.

(3) The multifaceted nature of antimicrobial peptides: current synthetic.... https://pubs.rsc.org/en/content/articlelanding/2021/cs/d0cs00729c.

(4) Rediscovery of antimicrobial peptides as therapeutic agents. https://link.springer.com/article/10.1007/s12275-021-0649-z.

(5) Antimicrobial Peptides for Therapeutic Applications: A Review - MDPI. https://www.mdpi.com/1420-3049/17/10/12276.

(6) undefined. http://dramp.cpu-bioinfor.org/.

(7) en.wikipedia.org.https://en.wikipedia.org/wiki/Antimicrobial_peptides.

(8) CalcAMP: A New Machine Learning Model for the Accurate Prediction of Antimicrobial Activity of Peptides

(9) https://www.mdpi.com/2079-6382/12/4/725

(10) https://webs.iiitd.edu.in/raghava/pfeature/Pfeature_Manual.pdf

(11) https://www.frontiersin.org/articles/10.3389/fmicb.2019.03097/full

(12) https://axp.iis.sinica.edu.tw/AI4AMP/helppage.html

(13) https://onlinelibrary.wiley.com/doi/10.1002/med.21658

(14) Machine learning designs non-hemolytic antimicrobial peptides - Chemical Science (RSC Publishing)

(15) Treating infectious disease with the help of antimicrobial peptides | FORMAMP Project | Results in brief | FP7 | CORDIS | European Commission (europa.eu)

(16) Choi H, Rangarajan N, Weisshaar JC. Lights, camera, action! Antimicrobial peptide mechanisms imaged in space and time. Trends Microbiol. 2016.

(17) Simpson DH, Hapeshi A, Rogers NJ, Brabec V, Clarkson GJ, Fox DJ, Hrabina O, Kay GL, King AK, Malina J, et al. Metallohelices that kill Gram-negative pathogens using intracellular antimicrobial peptide pathways. Chem Sci. 2019.

(18) Yoshida M, Hinkley T, Tsuda S, Abul-Haija YM, McBurney RT, Kulikov V, Mathieson JS, Reyes SG, Castro MD, Cronin L. Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. Chem. 2018.

(19) Bechinger B, Gorr SU. Antimicrobial peptides: mechanisms of action and resistance. J Dent Res. 2017.

(20) Afshar A, Yuca E, Wisdom C, Alenezi H, Ahmed J, Tamerler C, Edirisinghe M. Next-generation antimicrobial peptides (AMPs) incorporated nanofbre wound dressings. Med Dev Sens. 2021.

(21) Xie SX, Boone K, VanOosten SK, Yuca E, Song LY, Ge XP, Ye Q, Spencer P, Tamerler C. Peptide mediated antimicrobial dental adhesive system. Appl Sci (Basel). 2019.

(22) Machine learning designs non-hemolytic antimicrobial peptides - Chemical Science (RSC Publishing)

(23) Adlerova L, Bartoskova A, Faldyna M (2008) Lactoferrin a review. https://doi.org/10.17221/1978-VETMED

(24) https://doi.org/10.1177/0022034516679973

(25)Berthelot K, Peruch F, Lecomte S (2016) Highlights on Hevea brasiliensis (pro) hevein proteins. https://doi.org/10.1016/j.biochi.2016.06.006

(26) Bruni N, Capucchio M, Biasibetti E, Pessione E, Cirrincione S, Giraudo L, Dosio F (2016) Antimicrobial activity of lactoferrin-related peptides and applications in human and veterinary medicine. https://doi.org/10.3390/molecules21060752

(27)Ciociola T, Giovati L, Conti S, Magliani W, Santinoli C, Polonelli L (2016) Natural and synthetic peptides with antifungal activity. Future Med Chem

(28)Fan K, An Y, Wang Z, Yin W, Sun N, Sun Y, Li H (2019) Antibacterial activity of recombinant porcine β-defensin 2.

(29)Hou X, Li S, Luo Q, Shen G, Wu H, Li M, Zhang Z (2019) Discovery and identification of antimicrobial peptides in Sichuan pepper (Zanthoxylum bungeanum Maxim) seeds by peptidomics and bioinformatics.

(30)Jain A, Yadav BK, Chugh A (2015) Marine antimicrobial peptide tachyplesin as an efficient nanocarrier for macromolecule delivery in plant and mammalian cells. The FEBS. https://doi.org/10.1111/febs.13178