

# **Prediction of small antimicrobial peptides using Machine learning**

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE

OF

Master of Science

In

Biotechnology

Submitted by:

**Ashish Kukreti**

**2K20/MSCBIO/02**

Under the supervision of:

**Prof. Yasha Hasija**



DEPARTMENT OF BIOTECHNOLOGY  
DELHI TECHNOLOGICAL UNIVERSITY  
(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

## **Abstract**

Antimicrobial resistance (AMR) is a concern to public health, prompting the development of novel strategies for combating AMR. While the use of machine learning (ML) to AMR is in its infancy, it has made significant progress as a diagnosis tool, owing to the growing availability of phenotypic/genotypic datasets and much faster computational power. While applying ML in AMR research is viable, its use is limited. It has been used to predict antimicrobial susceptibility genotypes/phenotypes, discover novel antibiotics, and improve diagnosis when combined with spectroscopic and microscopy methods. ML implementation in healthcare settings has challenges to adoption due to concerns about model interpretability and data integrity. The focus of this thesis is to outline the significant benefits and drawbacks of ML in AMR, with emphasis on models built for the prediction of antimicrobial peptides, along with the salient trends reported in recent studies.

## List of Figures

1. Comparison of AI vs ML vs DL.
2. Various mechanisms through which microbes acquire antimicrobial resistance.
3. Overview of the workflow for antibiotic resistant genes prediction and AMR species identification using machine learning and deep neural networks.
4. Overview of the workflow of antibiotic/antimicrobial peptides discovery by using machine learning and deep learning algorithm. The researchers first constructed a training dataset of compounds capable of inhibiting the growth of common bacterial species such as *E. coli*. They improved their model by specifying phenotypic and genotypic features, adjusting hyperparameters, and assembling the ML model, which resulted in the final machine learning or deep learning algorithm via iterative model re-training. In comparison to the traditional approach, this in-silico approach enabled the researchers to systematically scan over a 100 million compounds, many of which had molecular structures that differed from known antibiotics. The number of screened compounds is approximately 100 times more than the typical approach, which is more inexpensive at this scale.
5. Combinations of certain features are selected for the novel and/or effective antibiotics/AMPs determination.
6. Workflow of Methodology conducted in the study.
7. Dataset for Short AMPs filtered out from AmPEP database.
8. The above Python codes was used to filter redundant sequences in the dataset using CD-HIT.
9. A total of 19 composition-based feature classes provided by Pfeature. However, in this study only feature i.e., amino acid composition has been used to build ML model.
10. Python codes for applying a certain feature i.e., amino acid composition (aac\_wp) from Pfeature in the dataset. aac\_wp stands for amino acid composition of whole protein/peptide.
11. Data matrix showing results obtained after applying feature selection.
12. Python code for building 29 ML models. The dataset was split into training and testing dataset in the ratio 80:20.
13. Python code for plotting ML model performance for training and testing dataset.
14. Comparison of various ML algorithms vs their respective accuracy scores.

15. Comparison of various ML algorithms vs their respective MCC scores.

## **List of Tables**

1. Comparison of commonly used AI algorithms for AMR.
2. AI-based ARGs Databases.
3. AI-based AMPs Databases.
4. Performance of various ML algorithms in which DT classifier gave strong performance.



# Contents

Candidate's Declaration	1
Certificate	2
Acknowledgements	3
Abstract	4
List of Figure	5
List of Tables	6
<b>Chapter 1: Introduction</b>	<b>9</b>
<b>Chapter 2: Antimicrobial Resistance</b>	<b>12</b>
<b>Chapter 3: Overview of Machine Learning and its types of algorithms</b>	<b>14</b>
3.1: Supervised Learning	15
3.1.1: Linear Regression	15
3.1.2: Logistic Regression	15
3.1.3: Naïve Bayes Classifier	15
3.1.4: Support Vector Machine	16
3.1.5: Decision Trees	16
3.1.6: Random Forest	17
3.2: Unsupervised Learning	18
3.3: Semi-supervised Learning	18
3.4: Reinforcement Learning	18
<b>Chapter 4: Applications of ML in AMR</b>	<b>19</b>
4.1: AMR Genes Prediction	19
4.2: Assist in Diagnostic Methods	20
4.3: Antibiotics/Antimicrobial Peptides Discovery	24
<b>Chapter 5: Prediction of small antimicrobial peptides using Machine learning</b>	<b>26</b>
5.1: Literature review	29
5.2: Methodology	31
5.2.1: Prerequisites	32

5.2.2: AMPs Dataset Preparation	32
5.2.3: Removal of Redundant AMPs sequences using CD-HIT	33
5.2.4: Feature Selection using Pfeature	33
5.2.5: Building ML Models	35
5.2.6: Plotting ML performance	36
<b>Chapter 6: Results and Discussion</b>	<b>37</b>
6.1: Discussion	40
<b>Conclusion</b>	<b>41</b>
<b>Future Scope of Research</b>	<b>41</b>
<b>References</b>	<b>43</b>

# Chapter 1

## Introduction

AMR has emerged as a major public health hazard in the 21st century. It is increasing globally and is related to increased morbidity and mortality in hospital and community settings [1]. The expansion of antibiotic resistance to new environmental niches, as well as the emergence of superbugs, has challenged effective control methods. For a successful outcome, early and accurate diagnosis of the illness and its drug resistance profile is crucial. Over the years, some alternatives to standard antibiotic treatment have been presented to avoid the problem of bacterial resistance. Artificial intelligence (AI), specifically ML, has been widely used in biomedical field because of its larger capacity for interpreting information from several different sources and the development of novel methodologies and algorithms for the prediction of experimentally acquired information.

ML is a branch of AI that allows a computer to predict outcomes using a learnt model and a large number of empirical observations, referred to as training data [Fig. 1]. ML can aid in the identification and design of novel antibacterial drugs by generating models based on empirical knowledge accessible in compound databases. Furthermore, ML approaches can be utilized to examine the pharmacokinetics and toxicity features of potential antibiotics to improve their efficacy [2].

Any technique that employs ML models necessitates sufficient input data to create a 'training set' for guiding the ML model and a 'testing set' for evaluating the model's performance. An important consideration in the study of AMR is the availability of data sets including correct genotypic information connected to carefully selected samples of the AMR gene. For this reason, pre-processing of genotypic data is required before they can be used as input for ML

models. This procedure is called "feature selection" in the ML terminology. The studies took one of two approaches to this problem: using gene annotations as "features" or using k-mers (short DNA strings are constructed by combining individual nucleotides).

Most of the algorithms discussed here used supervised ML, where the algorithm is trained on data with labels to construct a learned model. Hence, these algorithms could analyse test data and execute self-annotation using their label. They are employed for classification and regression-based analysis. Regression algorithms and Decision Trees Classifier are some of the most used supervised algorithms. While some studies have employed unsupervised learning and deep learning algorithms.

Unsupervised learning makes use of unlabelled training data. To put it differently, observations are classified without information of the data sample. Clustering can be performed using unsupervised techniques like hierarchical cluster analysis and principal component analysis (PCA).

Other AI approach such as deep learning, is based on how the biological neural systems interpret information and is becoming increasingly popular. Artificial Neural Networks (ANNs) are constituted of an input layer of 'neurons,' which is interconnected to one or more hidden layers of neurons, which are then linked to an output layer [3].

This thesis will provide an overview of widely used ML and DL techniques in antimicrobial resistance research, particularly in AMR gene prediction, novel antibiotics/AMPs discovery and implementation in several diagnostic tools and techniques. Further, it will delve into the prediction of AMPs through amino acid composition by using and comparing several ML algorithms. The primary goal of AMP prediction is to produce novel peptide sequences that have antimicrobial and therapeutic properties. Despite the fact that peptide design is beyond the focus of this thesis, we believe that knowing the importance of residues, their

characteristics, and their placements in the sequence will be critical for the de novo design of an AMP.

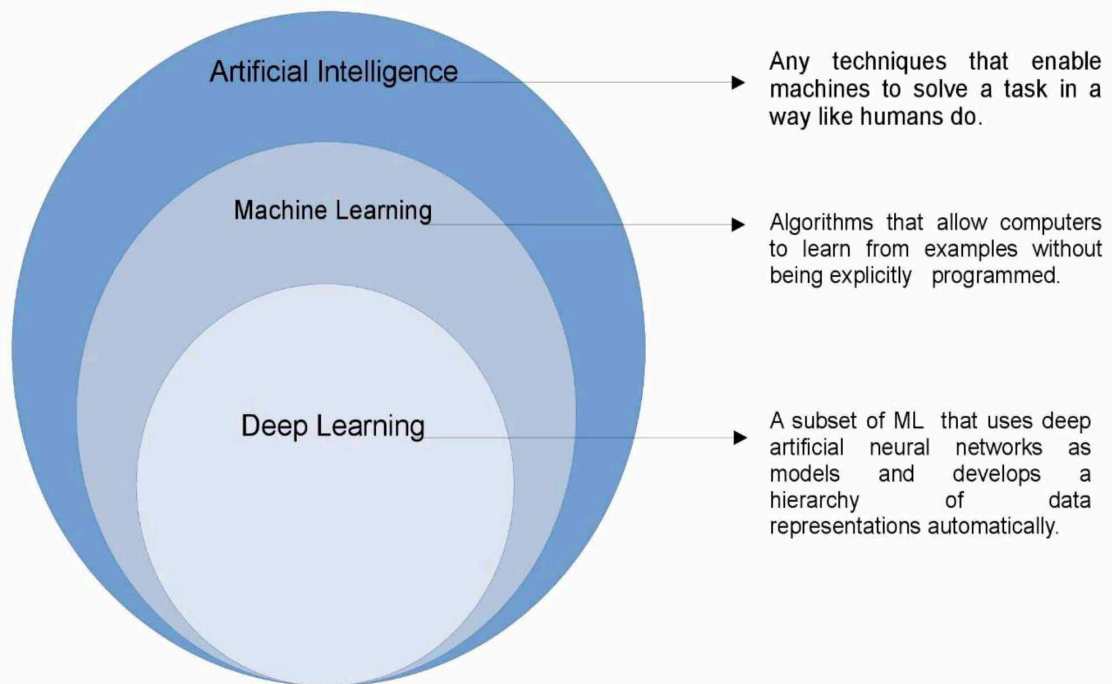


Fig.1: Comparison of AI vs ML vs DL.

## Chapter 2

### Antimicrobial Resistance

Fleming's unexpected discovery of antibiotics in 1928 ushered in the current era of antibiotics. Antibiotics have helped a lot of people who have been infected since then. Antibiotics' widespread clinical use, however, has resulted in the evolution of drug resistance as well as the prospect of super-resistant microorganisms or commonly known as "Superbugs". Antimicrobial resistance (AMR) is anticipated to kill 10 million people each year by 2050, with an economic cost of \$100 billion[1].

Antibiotic resistance mechanisms in bacteria can be divided into three categories:

- (i) blocking drug access to its target.
- (ii) enzymatic alteration and/or inactivation of the drug.
- (iii) alterations in the target molecule that degrade drug binding [Fig. 2].

The first method occurs when bacteria modify porin channels or their expression on membranes, causing therapeutic access to the cell to be restricted. Alternatively, efflux pumps can expel the medication from the cell; for example, Gram-negative bacteria extrude beta-lactams, and carbapenems reduce permeability [4]. Beta-lactamases, which are produced by Gram-negative bacteria such as *Enterobacter* species, *Pseudomonas aeruginosa*, and *Klebsiella pneumonia*, are one type of enzymatic drug inactivation. Finally, in Gram-positive bacteria, modifications in therapeutic targets can be significant. Methicillin-resistant *Staphylococcus aureus* (MRSA) was found two decades after penicillinase-producing *Staphylococcus aureus* was discovered in the 1940s. Some bacterial species require special attention because of their ability to express some or all of the above-mentioned mechanisms of

resistance, especially when considering their abundance in hospitals and other healthcare settings.

To avoid the problem of bacterial resistance, some alternatives to standard antibiotic treatment have been developed over the years. The use of bacteriophages, antibodies, and AMPs obtained from natural sources, most of which are in clinical studies, is one of the most promising [5]. Given the alarming state of resistant microorganisms around the world, as well as the unavailability of short-term options, the discovery of novel antibiotics/AMPs has garnered considerable attention. Under these conditions, existing antibiotics must be modified or novel peptides with antimicrobial activities must be discovered. However, drug design is a complex, time-consuming, and expensive process that requires huge multidisciplinary expert teams as well as a plethora of various experimental and computational methods.

MLT, in particular, can aid in the identification and design of novel AMPs by generating models based on empirical knowledge accessible in compound databases [6-9]. Furthermore, MLT are employed to evaluate the pharmacokinetics and toxicity features of potential antibiotics in order to improve their efficacy.

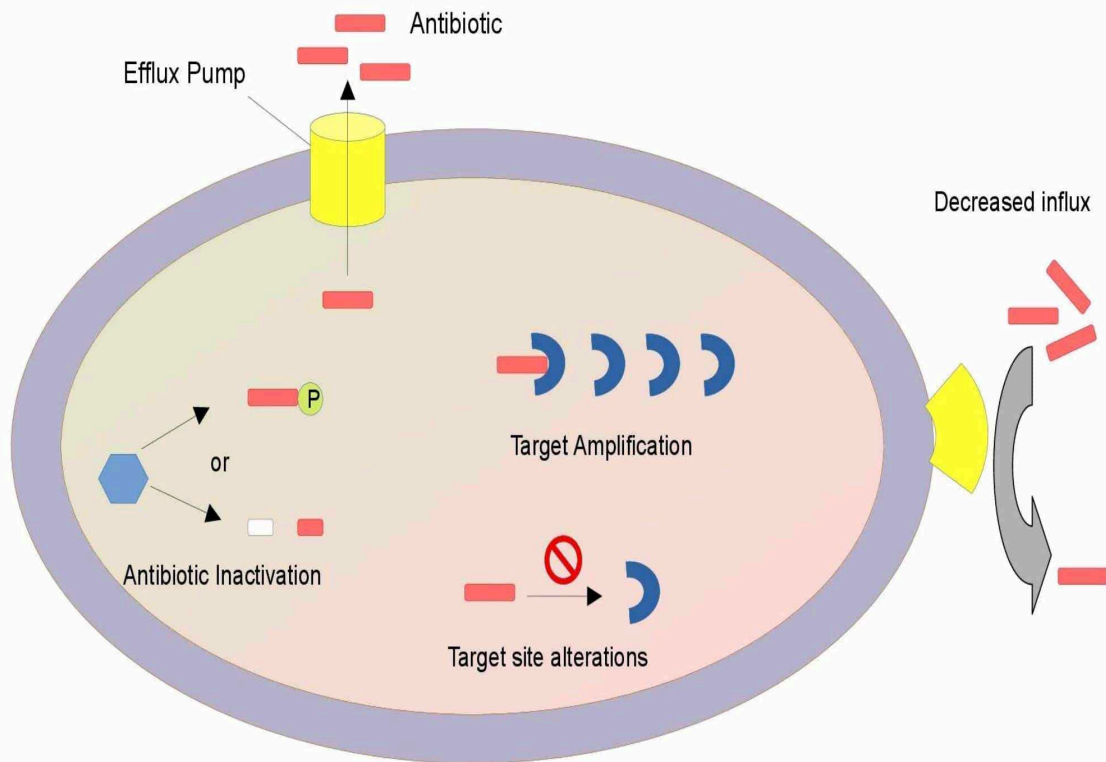


Fig. 2: Various mechanisms through which microbes acquire antimicrobial resistance.



## Chapter 3

### Overview of Machine Learning and its types of algorithms

Artificial intelligence (AI), and particularly its subdiscipline ML, have emerged as major trends that are gaining the interest of biologists. AI is a comprehensive term that can be understood as the notion and adoption of digital systems to execute complicated activities that ordinarily require human intellect, like decision-making, visual perception, natural language processing, and language processing. ML is a computational discipline that enables computers to learn from data without really being explicitly programmed [10]. This terminology was first used in 1959, but its possibilities remained limited until about the 1980s given the lack of computational capabilities, adequate and accurate datasets, storage capacity, and suitable applications. As of now, ML is permeating every domain including drug discovery, and healthcare sector, driven by the faster and cost-effective computers, an abundance of data and information produced by a more data-thriving world, and wider availability and adaptability of open-source software [10].

ML can be categorised into 4 types based on the amount of data and supervision they receive during model training:

- (i) Supervised learning
- (ii) Unsupervised learning
- (iii) Semi-supervised learning
- (iv) Reinforcement learning

## **3.1 Supervised Learning**

Data sets are collected, and training data is labelled in supervised learning. Having knowledge about the specific outcomes, predictions can be made. It is frequently employed in classification and regression problems [Table 1]. The most common algorithms used in supervised learning for AMR are mentioned below:

### **3.1.1 Linear regression**

The most well-known and well-understood technique in the domain of machine learning is linear regression, which is used to characterise the connection between a dependent variable and one or more predictors [10].

By building a linear model with features and a random error, it optimizes the residual sum of squares between the values based on the linear model and the observed values in the dataset [10].

### **3.1.2 Logistic regression**

A logistic function, which is a sigmoid function with a typical S-shaped curve that outputs a number between 0 and 1, can be used to convert a Linear Regression to a Logistic Regression[10].

However, when non-linear correlations are taken into consideration, both regression algorithms underperform and are not equipped to handle more complicated patterns.

### **3.1.3 Naïve Bayes Classifier (NB)**

Naïve Bayes classifier is a probabilistic approach based on Bayes' theorem and each feature's independent assumptions. The input/output combined probability distribution is produced for a given training dataset. The model is easy to build and therefore does not involve iterative parameter estimate, making it extremely useful in the healthcare sector.

In recent investigations, the naïve Bayes model has been used to track AMR. Rezaei-Hachesu et al.,[11] for example, employed naïve Bayes and a probabilistic algorithm to extract resistant patterns and discover the essential element of resistance. Choisy et al.[12] also employed naïve Bayes to calculate the chances of ineffective therapy due to AMR.

### **3.1.4 Support Vector Machine (SVM)**

The Support Vector Machine (SVM) is a binary classifier model that finds a partitioned hyperplane in the sample space to classify samples [13]. SVM models have frequently been used in recent studies to predict AMR phenotypes, such as 'resistant' or 'susceptible.' Her et al.,[14] for example, used SVM models to predict antibiotic resistant *E.coli*.

The results showed that this model was able to provide average accurate predictions of up to 0.95 (AUC curve analysis). Furthermore, Liu et al.[15] used SVM models to evaluate the several antibiotic drugs, with the results indicating that the model's accuracy is at least 90%. As a result, SVM models could be useful for AMR monitoring and clinical diagnosis.

### **3.1.5 Decision Trees (DT)**

For classification, the decision tree is commonly employed. A decision tree's learning process normally involves three steps: feature selection, decision tree generation, and decision tree pruning.

By calculating the burden of AMR, the decision tree model is frequently used to distribute medical resources effectively. Reynolds et al.,[16] for example, used a decision tree model for predicting healthcare utilisation and costs for AMR, suggesting that lowering AMR or improving antibiotic selection could result in significant cost savings. To guide antibiotic use, Voermans et al.[17] developed a procalcitonin (PCT)-based decision tree model, which resulted in a reduced treatment period and significantly smaller dosages.

### 3.1.6 Random Forest (RF)

Random forest is indeed an ensemble approach that improves accuracy and dependability by combining multiple decision trees [18]. It employs two fundamental concepts rather than merely averaging the prediction outcomes of all trees. The first is random selection from the training set, which means that only certain samples will appear in the tree several times. The other is a random subset of features, which means that while the efficiency of a single tree in a random forest could be degraded, the random forest normally converges to a lower generalisation error as the number of trees increases.

Antibiotic combinations are frequently predicted using random forest models. Incorporating chemo-genomics data and orthology, Chandrasekaran et al.[19] used a random forest model to predict efficient antibiotic combination therapy. However, because chemo-genomics data are insufficient, Mason et al.[20] used the molecular fingerprint as a feature to increase the predictive power of the previously mentioned models.

Methods	Advantages and Disadvantages
NB	Fast and simple to use, this method is ideal for datasets having missing data.
DT	The decision tree's results are simple to understand, and depending on the tree's complexity, it can also be used to analyse datasets having missing information.
RF	This approach works well with huge datasets containing a variety of features. It is, however, not accurate where there is an outlier data.
SVM	SVM can deal with complex issues by using kernel functions, however it is slower in processing and require specifying each and every features.
ANN	A large number of neural networks can be used to learn a range of complicated issues. The accuracy of the model will improve as the depth of the model grows, although the learning rate may be slow.

Table 1: Comparison of commonly used AI algorithms for AMR.

### **3.2 Unsupervised Learning**

The training data in unsupervised learning are not labelled. To put it differently, observations are categorised without any prior knowledge of the data sample. It recognises the data's class via prior knowledge about features when newer data is introduced.

Unsupervised methods can be used for clustering (e.g., k-means clustering, hierarchical cluster analysis), visualisation, and dimensionality reduction (e.g., principal component analysis).

### **3.3 Semi-supervised Learning**

Semi-supervised learning algorithms combine the benefits of both supervised and unsupervised learning techniques. It might be useful in ML and DL if there is already unlabelled data and acquiring the annotated data is a time-consuming operation. It is frequently used in medical image classification.

### **3.4 Reinforcement Learning**

In reinforcement learning, an entity (i.e., the learning system) learns which actions to do in order to maximise the cumulative reward or optimise the outcome of an approach (i.e., a policy). The two main concepts that govern reinforcement learning are trial and error searches and delayed results.

It assists in the detection of public health risks by spotting trends and modelling disease progression.

## Chapter 4

### Applications of ML in AMR

#### 4.1 AMR Genes Prediction

The most common way of developing antibiotic resistance is the transfer of antibiotic resistance genes (ARG) among the bacterial species. Current methods are incapable which rely on inefficient phenotypic data. Therefore, it becomes an important step to predict these ARGs accurately for a better understanding of their transmission from the environment. Most bioinformatics tools have been based on either assembly-based methods or read-based methods to identify or annotate resistance genes [Table 2]. With ever-increasing genomic data, ML models have been built to learn the statistical patterns of ARGs and may potentially identify novel ones as they detect certain features rather than using sequence similarity [Fig. 3]. Some ML methods have reported models to identify novel ARGs from pan-genome and metagenomic data [21,22]. Although the results were satisfactory, these models used limited features and did not include any feature selection method to remove redundant data. Subsequently, Li et al. [23] developed HMD-ARG (Hierarchical Multi-task Deep learning framework for prediction of the ARG) coupled with deep neural networks. The framework provides detailed information on annotated ARGs based on their biochemical properties and covers three significant aspects: resistance class, gene mobility, and mechanism. In another study, Chowdhury et al. [24] proposed a model called PARGT (Prediction of Antimicrobial Resistance via Game Theory), which can identify ARGs from bacterial species. The model utilized the supervised ML algorithm. These two methods [23,24] validated their results for feature selection. However, in the future, these methodologies should be made compatible with current sequencing technologies, which work on short reads like nanopore sequencing rather than on assembled sequences.

<b>Tool</b>	<b>Type of AI method</b>	<b>Description</b>
<b>ABCRpred</b>	Supervised ML (Random Forest)	Predict ceftazidime resistance/susceptibility of beta-lactamase protein sequence
<b>GenTB</b>	Supervised ML (Random Forest)	Predict antibiotic resistance against <i>Mycobacterium tuberculosis</i>
<b>VAMPr</b>	Supervised ML (Decision Tree)	Utilize next-generation sequencing data to determine antibiotic resistance
<b>DeepARG</b>	Deep learning (Artificial Neural Networks)	Utilize metagenomic data to characterize ARG

Table 2: AI-based ARGs Databases.

## 4.2 Assist in Diagnostic Methods

When designing a new AST (Antibiotic Susceptibility Testing) method, several nonautomated procedures are considered gold standards for comparative reasons. These methods include agar dilution, broth microdilution, disc diffusion, and the E-test. While these tests are cost-effective, they must be carried out manually and take 18-24 hours to complete. Although these are not considered quick AST techniques, they provide valid MIC values. Scientific research over the last few years has not only resulted in the development of novel AST platforms but also improve existing platforms. The integration of ML algorithms has resulted in a significant progression of AST approaches [Fig. 3].

MALDI- TOF MS has proven to be a fast, inexpensive, and accurate tool that is being used in the identification of antibiotic-resistant microorganisms by generating characteristic mass spectral fingerprints which are unique to microbes at genus and species level, which are then compared with a reference library database comprising of well- identified organisms for each isolate [25]. Unlike, conventional methods MALDI-TOF offers species-level identification

with reduced turn-around time and far more accurate results. Recently, ML algorithms have been employed in the optimization of MALDI- TOF to enhance species identification which can reveal unknown or novel information hidden in the mass spectra and, in antibiotic resistance profiling of closely related bacterial species [26]. However, there is an increased interest in utilizing MALDI-TOF for AST. Data retrieved from traditional methods may serve as user input for ML algorithms.

Huang et al. [27] evaluated five ML algorithms (RF, logistic regression, NB, NN, and SVM) to check their antibiotic susceptibility of *Klebsiella pneumoniae* against carbapenem. A total of 100 spectra peaks from *K. pneumoniae* isolates were used as the training dataset for their classification. RF algorithm surpasses the other algorithms achieving an overall classification accuracy of 97%. Sogawa and colleagues [28] tested a prediction model using a supervised ML algorithm to classify 50 isolates of both *Methicillin-susceptible S. aureus* (MSSA) and MRSA. The accuracy rates were 90% and 87.5% respectively. This study has proposed rapid detection from one colony in 5 minutes, however, the accuracy rate was not 100% which is crucial for clinical diagnosis. Wang et al. [29] included 787 *Group-B streptococci* (GBS) isolates in their analysis where they have used supervised ML to build models for the prediction of 5 different serotypes achieving up to 87.5 % accuracy while other studies [27,29] included less than or equal to 50 isolates only which is quite a small sample size for ML models that require larger sample size to feed. Although most of the studies showed good results, lack of external validation and poor reproducibility dampen progress towards this approach.

Raman spectroscopy analyses biochemical composition by using multichromatic emitters in the ultraviolet/infrared/visible spectrum. Raman scattering can be used to measure small quantities of material, such as single bacterial cells, and has several microbiological applications. Laser light can be used to investigate the physicochemical characteristics of the probed sample. Molecular bonds then inelastically scatter the photons which are analyzed by



spectrophotometer and a Raman spectrum is plotted to depict the intensity of the inelastic scattering [30]. These Raman spectra and their associated information can be utilized for the identification of bacteria and for other microbial diagnoses by analyzing through multivariate statistics and ML algorithms. Finally, these class-specific Raman spectral models, after iterations by ML algorithms, can be utilized to distinguish bacterial classes of interest [30].

Ullah et al. [31] employed unsupervised ML techniques to classify 60 tuberculosis (TB) serum samples (30- TB positive and 30- TB negative) based on the variation in biochemical concentration. The findings indicate a significant difference in Raman spectra in both TB positive and TB negative groups as well as in the control group too. Similarly, Moawad et al. [32] developed an SVM model in combination with PCA to identify *Burkholderia mallei* and related species. The optimized model identification accuracy reached above 90%. However, the model's reliability was limited due to the occurrence of misclassification of *B. thailandensis* with *B. mallei*. For which, the authors suggested that this misclassification occurred due to the less representation of *B. thailandensis* samples in the training dataset. Rebrošová et al. [33] successfully identified 277 staphylococci strains from 16 species by utilizing Raman spectroscopy with supervised ML algorithms showing better results achieving an accuracy of 99%. Meanwhile, Ho et al. [34] implemented DL, along with logistic regression for the rapid classification of (n=30) bacterial isolates, for antibiotic resistance and empirical treatment. For (n=30) bacterial species classification, the DL classifier comprised of 25 1-D convolutional layers along with some residual connections which achieved an average isolate-level accuracy of 82%. However, most of the misclassifications occurred at the genus level. Moreover, a binary DL classifier was built to classify MRSA and MSSA which achieved an accuracy of 89%, and another DL model categorized known bacterial isolates into several groups based on the common empirical antibiotic treatment that achieved an accuracy of 97%.

Additionally, Yi et al. [35] devised a rapid AST based on Raman scattering, which detects an activity and stability of certain metabolites in the presence of antibiotics via single-cell Raman spectroscopy (SCRS). SCRS works by detecting the biomolecules' vibrations within a cell/bacterium, hence determining their biochemical attributes or phenotype. In this study, FRAST was applied to (n=8) bacterial species, consisting of 4 Gram-positive and 4 Gram-negative bacterial species along with ML algorithms to train a model to classify training dataset which achieved sensitivities of 98.8% for Gram-positive bacteria and 94.3% for Gram-negative bacteria at the unicellular level. The model only took less than 30 minutes for the results. The model was then validated using (n=6) bacterial isolates consisting of 3 Gram-positive and 3 Gram-negative bacterial species, which achieved an overall sensitivity of over 90%, which was then confirmed by 16S rRNA sequencing.

However, the current limitations of implementing Raman spectroscopy in clinical labs are that first, they generate huge amount of data shortly, which could be surpassed by employing ML techniques. Second, training datasets consisting of a small sample size result in lower accuracy by the learning model that might be resolved by feeding the training model with large datasets encompassing varieties of antibiotic-susceptible and resistant bacteria. Finally, the improper optimization of training data and lack of external validation of the analysis results in misclassifications amongst similar bacterial species.

Several other diagnostic tools with the help of ML algorithms have been employed to improve accuracy and turnaround time in AST. For instance, Inglis et al. [36] implemented AST with the flow cytometry method (FAST). The method utilized a decision tree algorithm to deliver the results within 3 hr. Based on a multivariate analysis of microcolony images, Maeda et al. [37] employed a novel technology known as "colony fingerprinting" to distinguish five *Staphylococcus* species. The method used supervised ML and deep learning algorithms which showed high performance and generated the results within 11 hr. Moreover, Smith et al. [38]

developed an AST platform, based on microscopy (MAST) which used deep neural networks for classification and could determine the AST after incubation of 2 hr. Further, Lechowicz et al. [39] used the combination of deep neural networks and infrared spectroscopy to classify 109 uropathogenic E. coli strains against cephalothin. The method generated quick results within 30 min. which are much faster than conventional AST methods (24 hr.).

### **4.3 Antibiotics/Antimicrobial Peptides Discovery**

This will be discussed in detail in next chapter.

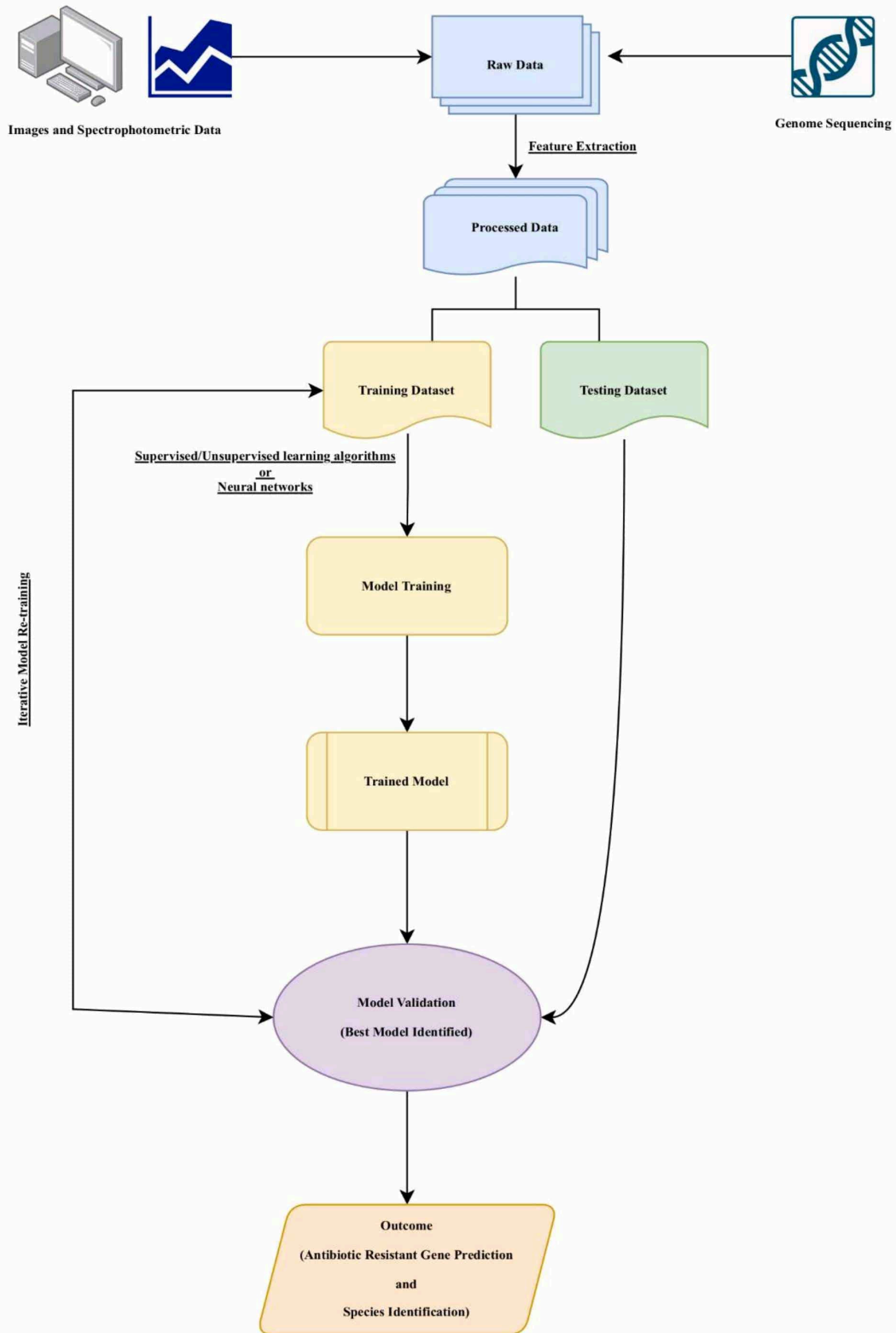


Fig.3: Overview of the workflow for antibiotic resistant genes prediction and AMR species identification using machine learning and deep neural networks.

## Chapter 5

### Prediction of small antimicrobial peptides using Machine learning

Antimicrobial peptides (AMPs) are considered a promising antibiotic replacement candidate because they are key part of the innate immunity with a broad spectrum of activities that can protect the host from a spectrum of pathogenic microorganisms such as viruses, bacteria, parasites, and fungi.

Nonetheless, identifying and extracting AMPs is costly and time-consuming. To mitigate these dangers, in silico ML can be used for early screening to shortlist proteins with possible antimicrobial activity before proceeding to the lab testing phase [40].

The preliminary stage in AMPs design is to acquire a training dataset (input) that comprises non-AMPs and AMPs databases with specific protein composition features to achieve a suitable balance of desirable and undesirable inputs. Although some databases, such as the Antimicrobial Peptide Database (APD), the Collection of Antimicrobial Peptides (CAMPR3), and Yet Another Database of Antimicrobial Peptides (YADAMPs), are publicly available, the non-AMPs were excluded in most studies since there is no experimentally tested data available in such non-AMPs databases [41,42]. [Table 3].

Database	Type of ML algorithm	Description
<b>INDIGO</b> (INferring Drug Interactions using chemo-Genomics and Orthology)	Supervised learning (Random Forest)	Exploits chemogenomic data of model organisms for the prediction of antibiotic combinations
<b>CAMP</b> (Collection of Anti-Microbial Peptides)	Semi-supervised learning (Hidden Markov Model)	Detailed information on 8164 AMR sequences with 752 AMR 3-D structure
<b>YADAMP</b> (Yet Another Database of Anti-Microbial Peptides)	Supervised learning	Contain comprehensive information on 2525 AMPs against common bacterial species

Table 3: AI-based AMPs Databases.

Following that, the appropriate peptide descriptors must be chosen based on the expected final result. These descriptors are used to assess the value of specific features in the design of antimicrobial protein characteristics. 2D Quantitative Structure–Activity Relationship (QSAR) descriptors, 3D QSAR descriptors, inductive descriptors, and other descriptors like the number of residues between amino acids are the four most prevalent types of descriptors. Then, an algorithm will be devised or determined based on the accuracy of the desired outcomes and the stability of the input dataset [43]. The algorithm will provide a list of AMPs, which will then be analyzed for antibacterial activity and toxicity.

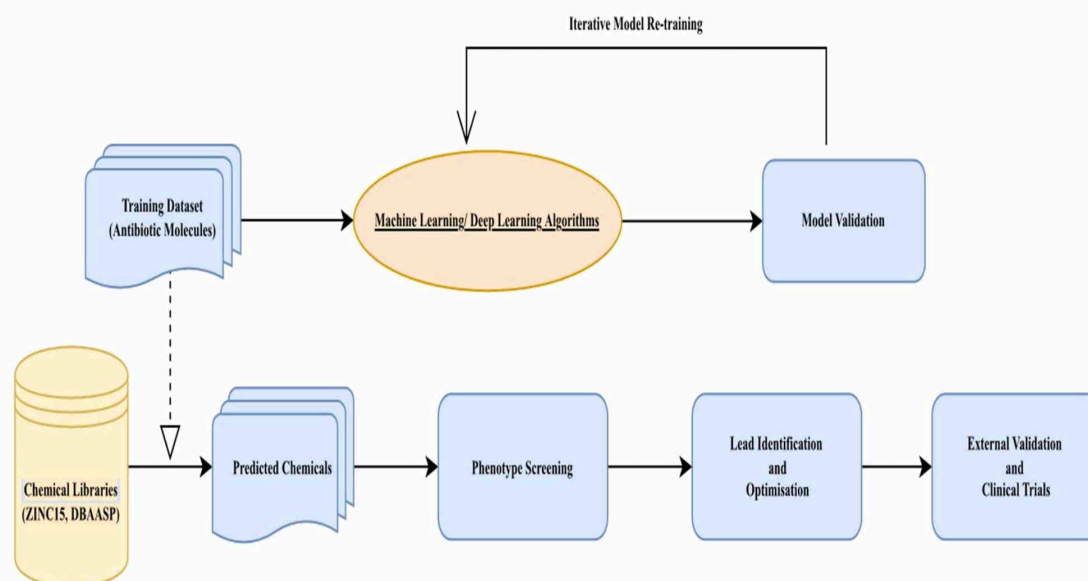


Fig.4: Overview of the workflow of antibiotic/antimicrobial peptides discovery by using machine learning and deep learning algorithm. The researchers first constructed a training dataset of compounds capable of inhibiting the growth of common bacterial species such as *E. coli*. They improved their model by specifying phenotypic and genotypic features, adjusting hyperparameters, and assembling the ML model, which resulted in the final machine learning or deep learning algorithm via iterative model re-training. In comparison to the traditional approach, this in-silico approach enabled the researchers to systematically scan over a 100 million compounds, many of which had molecular structures that differed from known antibiotics. The number of screened compounds is approximately 100 times more than the typical approach, which is more inexpensive at this scale.

## 5.1 Literature review

Despite the clear need for more antibiotics, very few antibiotics have been available in the market. The last entire class of antibiotics was discovered in the late 1980s. Many pharmaceutical companies have gradually shied away from antibiotic research and development due to the several challenges posed in the discovery of novel antibiotics as their development is time-consuming and expensive. The cost-benefit ratio is much less favourable than for other drugs. Subsequently, physicians avoid prescribing new antibiotics to delay antibiotic resistance and are usually used as "last resort drugs" when traditional medicines failed to do their work. Hence, these industries shift their focus on developing profitable long-term treatments for chronic infections. These limitations have been addressed progressively with the employment of various ML techniques that discover newer antibiotics and potential lead compounds which makes their identification less laborious and cost-effective. For instance, Stokes et al. [44] implemented deep neural networks to identify eight novel broad-spectrum antibiotic molecules. These identified molecules were structurally distinct from all known antibiotics. In this study, more than 107 million molecules from the ZINC15 database were assembled to build a training dataset of 2335 molecules for the prediction of potential molecules showing the inhibitory action on *Escherichia coli*. The researchers were able to find out the potential candidates who met a predetermined threshold score and various other exclusion criteria [Fig. 4]. The study successfully identified "halicin" as a potent growth inhibitor of *E. coli* and efficient against other bacterial infections in animal models. In recent years, most of the studies employed ML algorithms to identify novel antimicrobial peptide (AMP) drugs [45]. These AMPs are the class of small peptides that can directly kill pathogenic microbes as well as can indirectly modulate the host defence system [46]. In a study conducted by Capecchi et al. [47], deep neural networks and supervised learning algorithms were utilized to generate a classifier model to predict non-haemolytic AMPs for *Methicillin Resistant*



*Staphylococcus Aureus* (MRSA), *P. aeruginosa*, and *A. baumannii*. The training and testing dataset were assembled from 4774 peptides found in “DBAASP” (Database of Antimicrobial Activity and Structure of Peptides) [Some other databases are discussed in Table 3]. The study showed promising results by identifying eight non-haemolytic AMPs that met a predetermined threshold value of minimum inhibitory concentration (MIC). Moreover, Boone et al. [48] successfully employed supervised ML and a codon-based genetic algorithm to identify an active AMP against *S. epidermidis*, a normal commensal found on skin. Several studies have been implementing ML to find a treatment for COVID-19 [49]. Kowalewski and Ray [50] developed models for the prediction of potential drugs against 65 target human proteins, including the ACE-2 receptor by employing supervised ML algorithms. The researchers gathered 14 million compounds from the ZINC database and applied ML models for the prediction of certain features like binding affinity and toxicity to classify molecules and identify compounds with similar chemical space [Fig. 5]. Altogether, a thorough investigation of literature reviews briefs us about the application of ML for AMPs design. With sufficient prior information about known AMPs, ML can be applied to discover novel AMPs by which the development of novel antibiotics would become cost-effective and time-efficient while achieving more efficacy than conventional methods.

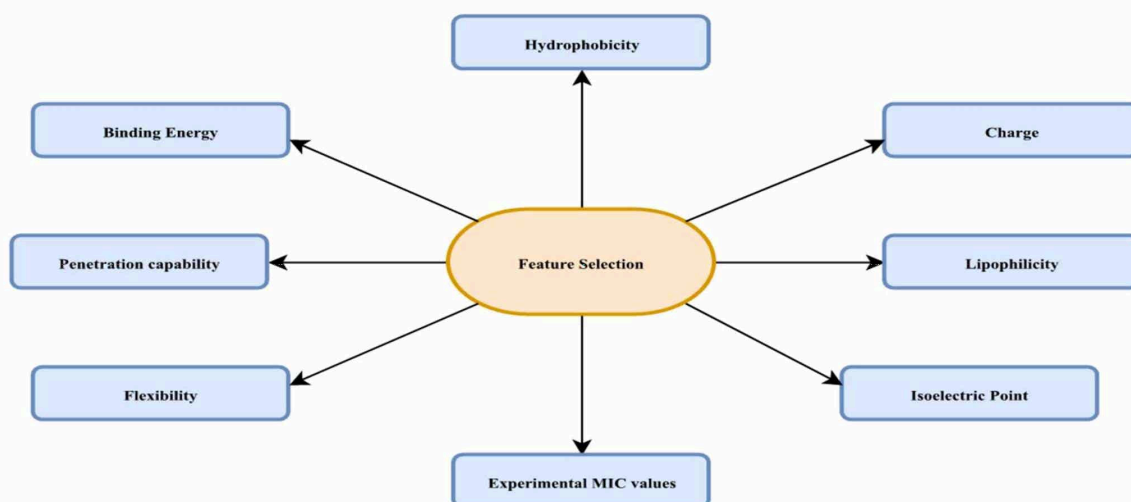


Fig. 5: Combinations of certain features are selected for the novel and/or effective antibiotics/AMPs determination.

## 5.2 Methodology

Machine Learning algorithms depend upon the quality and quantity of dataset. For AMPs design, this in-silico approach can be utilized to screen candidates that show antimicrobial activities based on certain pre-defined selected features. The overview of the workflow is provided in fig. which involves the following steps:

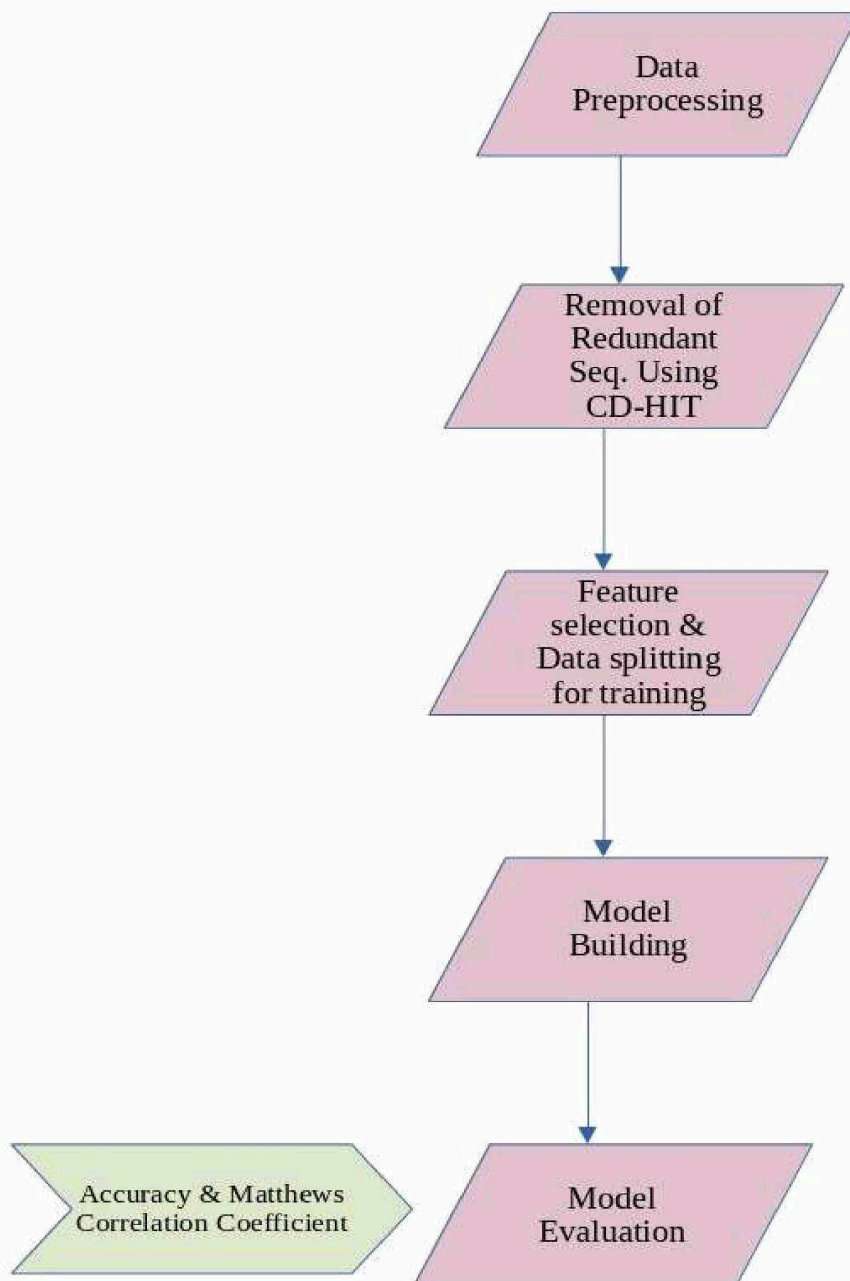


Fig. 6: Workflow of Methodology conducted in the study.

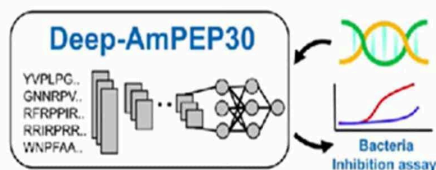
### 5.2.1 Prerequisites

For building and comparing various ML models for the prediction of amino acid composition in short AMPs, a web-based workspace called “Google colab” [51] was used. It allows and executes Python [52] in a browser with comprehensive configurations and free of charge access to GPUs. It contains Jupyter [53] notebook which is a prerequisite for writing and executing relevant codes in Python.

### 5.2.2 AMPs Dataset Preparation

The dataset for short AMPs was filtered from AmPEP database, which includes peptide sequences having only 5-30 amino acids in length [Fig. 7]. The dataset consists of 1529 positive and 1529 negative AMPs in Fasta format.

Short Anti-Microbial Peptides



Data is filtered from our AmPEP dataset, include sequences only with 5-30 AA in length. This dataset is used for constructing the Deep-AmPEP30 and RF-AmPEP30 prediction models. An independent dataset was constructed as benchmark to compare model performances with other existing methods.

Fig. 7: Dataset for Short AMPs filtered out from AmPEP database.

### 5.2.3 Removal of Redundant AMPs sequences using CD-HIT

CD-HIT is a common web-based tool that allow to filter out any redundancy in the peptide sequences meaning that similar peptides are removed and a non-redundant or a unique subset of peptides is obtained for better results in the ML model [54] [Fig. 8]. After executing CD-HIT in the dataset, 108 redundant or similar peptide sequences were removed. The non-redundant dataset contained 1421 unique short AMPs.

```
Remove redundant sequences using CD-HIT

[ ] ! cd-hit -i train_po.fasta -o train_po_cdhit.txt -c 0.99

[ ] ! cd-hit -i train_ne.fasta -o train_ne_cdhit.txt -c 0.99

[ ] ! ls -l

[ ] ! grep ">" train_po_cdhit.txt | wc -l

[ ] ! grep ">" train_po.fasta | wc -l

[ ] ! grep ">" train_ne.fasta | wc -l

[ ] ! grep ">" train_ne_cdhit.txt | wc -l
```

Fig.8: The above Python codes was used to filter redundant sequences in the dataset using CD-HIT.

### 5.2.4 Feature Selection using Pfeature

Pfeature is a web-based server that uses an amino acid sequence to compute a variety of protein and peptide properties [55] [Fig. 9]. It is useful in annotating certain features to a peptide/protein sequence [Fig. 10,11].

Feature class	Description	Function
AAC	Amino acid composition	aac_wp
DPC	Dipeptide composition	dpc_wp
TPC	Tripeptide composition	tpc_wp
ABC	Atom and bond composition	atc_wp, btc_wp
PCP	Physico-chemical properties	pcp_wp
AAI	Amino acid index composition	aai_wp
RRI	Repetitive Residue Information	rri_wp
DDR	Distance distribution of residues	ddr_wp
PRI	Physico-chemical properties repeat composition	pri_wp
SEP	Shannon entropy	sep_wp
SER	Shannon entropy of residue level	ser_wp
SPC	Shannon entropy of physicochemical property	spc_wp
ACR	Autocorrelation	acr_wp
CTC	Conjoint Triad Calculation	ctc_wp
CTD	Composition enhanced transition distribution	ctd_wp
PAAC	Pseudo amino acid composition	paac_wp
APAAC	Amphiphilic pseudo amino acid composition	apaac_wp
QSO	Quasi sequence order	qos_wp
SOC	Sequence order coupling	soc_wp

Fig.9: A total of 19 composition-based feature classes provided by Pfeature. However, in this study only feature i.e., amino acid composition has been used to build ML model.

```
[19] import pandas as pd

# Amino acid composition (AAC)

from Pfeature.pfeature import aac_wp

def aac(input):
    a = input.rstrip('txt')
    output = a + 'aac.csv'
    df_out = aac_wp(input, output)
    df_in = pd.read_csv(output)
    return df_in

aac('train_po_cdhit.txt')
```

Fig.10: Python codes for applying a certain feature i.e., amino acid composition (aac\_wp) from Pfeature in the dataset. aac\_wp stands for amino acid composition of whole protein/peptide.



	AAC_A	AAC_C	AAC_D	AAC_E	AAC_F	AAC_G	AAC_H	AAC_I	AAC_K	AAC_L	AAC_M	AAC_N	AAC_P	AAC_Q	AAC_R	AAC_S	AAC_T	AAC_V	AAC_W	AAC_Y	class
0	27.27	0.00	9.09	0.00	9.09	9.09	0.00	0.00	0.00	0.00	0.00	9.09	9.09	0.00	0.00	0.00	0.00	18.18	0.0	9.09	positive
1	0.00	54.55	0.00	9.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	18.18	18.18	0.00	0.0	0.00	positive
2	0.00	0.00	9.09	18.18	9.09	9.09	9.09	0.00	0.00	9.09	9.09	0.00	18.18	0.00	0.00	0.00	0.00	9.09	0.0	0.00	positive
3	0.00	0.00	0.00	18.18	9.09	9.09	9.09	0.00	0.00	9.09	9.09	9.09	18.18	0.00	0.00	0.00	0.00	9.09	0.0	0.00	positive
4	0.00	0.00	9.09	18.18	18.18	9.09	0.00	0.00	0.00	9.09	9.09	9.09	18.18	0.00	0.00	0.00	0.00	0.00	0.0	0.00	positive
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1417	13.33	0.00	3.33	3.33	0.00	13.33	0.00	3.33	26.67	10.00	3.33	3.33	3.33	3.33	0.00	0.00	3.33	0.00	10.0	0.00	negative
1418	6.67	0.00	0.00	3.33	6.67	20.00	3.33	3.33	13.33	13.33	3.33	0.00	3.33	10.00	0.00	6.67	3.33	3.33	0.0	0.00	negative
1419	6.67	20.00	0.00	3.33	0.00	10.00	0.00	6.67	10.00	3.33	0.00	3.33	6.67	0.00	0.00	6.67	6.67	10.00	0.0	6.67	negative
1420	6.67	0.00	3.33	3.33	6.67	0.00	3.33	3.33	6.67	3.33	16.67	3.33	6.67	10.00	6.67	0.00	13.33	0.0	3.33	negative	
1421	0.00	20.00	6.67	6.67	0.00	6.67	6.67	6.67	6.67	10.00	10.00	0.00	6.67	0.00	3.33	3.33	0.00	3.33	0.0	3.33	negative

Fig.11: Data matrix showing results obtained after applying feature selection.

### 5.2.5 Building ML Model

The dataset was then split into training and testing dataset in 80:20 ratio. Three Python libraries (lazypredict, sklearn, matplotlib) were installed[Fig. 12]. Lazy Predict allows you to easily construct ML models at scale, choose the best suitable mode without writing a code from scratch, and determine which models perform better without having to tweak any parameters [56]. Sklearn is a free ML building software and library for Python [57]. Matplotlib is a graph-plotting library for Python for visualization [58].

```

# Import libraries
import lazypredict
from lazypredict.Supervised import LazyClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import matthews_corrcoef

# Load dataset
X = feature.drop('class', axis=1)
y = feature['class'].copy()

# Data split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state =42, stratify=y)

# Defines and builds the lazyclassifier
clf = LazyClassifier(verbose=0,ignore_warnings=True, custom_metric=matthews_corrcoef)
models_train,predictions_train = clf.fit(X_train, X_train, y_train, y_train)
models_test,predictions_test = clf.fit(X_train, X_test, y_train, y_test)

```

100% 29/29 [00:07<00:00, 3.94it/s]  
100% 29/29 [00:00<00:00, 7963.52it/s]

Fig.12: Python code for building 29 ML models. The dataset was split into training and testing dataset in the ratio 80:20.

## 5.2.6 Plotting ML Performance

ML models' performance was evaluated on the basis of three parameters: Accuracy, Matthew's correlation coefficient (MCC) and time taken for the prediction [Fig. 13]. MCC seems to be a more valid and reliable rate that yields a high score only if the prediction performed well in all 4 confusion matrix classes, proportionately to both the size of positive and negative values in the data set [59]. It has a range of +1 to -1, with +1 representing the highest correlation among expected and actual values.

```
# Prints the model performance (Training set)
models_train

# Prints the model performance (Test set)
models_test

# Plot of Accuracy
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(5, 10))
sns.set_theme(style="whitegrid")
ax = sns.barplot(y=models_test.index, x="Accuracy", data=models_test)
ax.set(xlim=(0, 1))

# Plot of MCC
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(5, 10))
sns.set_theme(style="whitegrid")
ax = sns.barplot(y=models_test.index, x="matthews_corrcoef", data=models_test)
ax.set(xlim=(0, 1))
```

Fig.13: Python code for plotting ML model performance for training and testing dataset.

## Chapter 6

### Results and Discussion

Initially, the dataset contained 1529 short AMPs sequences which was then reduced to 1421 unique sequences by using CD-HIT software. It is observed through table that DT algorithms perform better than rest of the ML model applied [Table 4]. The time taken by DT classifier is 0.07 sec. and it has a MCC of 0.86 that is close to +1 which indicated a strong model performance [Fig. 15,16].

Model	Accuracy	Matthews corrcoeff.	Time Taken
RandomForestClassifier	0.93	0.86	0.59
LabelSpreading	0.93	0.86	0.68
LabelProgression	0.93	0.86	0.58
DecisionTreeClassifier	0.93	0.86	0.07
ExtraTreeClassifier	0.93	0.86	0.05
ExtraTreesClassifier	0.93	0.84	0.41
LGBMClassifier	0.92	0.84	0.21
BaggingClassifier	0.92	0.83	0.20
MuSVC	0.86	0.72	0.51
SVC	0.82	0.65	0.44
KNeighboursClassifier	0.82	0.64	0.29
XGBClassifier	0.80	0.61	0.22
AdaBoostClassifier	0.76	0.51	0.34
QDC	0.75	0.51	0.08
RidgeClassifier	0.74	0.48	0.08
RidgeClassifierCV	0.74	0.48	0.30
LinearSVC	0.74	0.47	0.31
LDA	0.73	0.47	0.09
LogisticRegression	0.73	0.47	0.09
CalibratedClassifierCV	0.73	0.41	0.09
NearestCentroid	0.72	0.46	0.06
GaussianNB	0.71	0.45	0.06
BernoulliNB	0.7	0.42	0.06
SGDClassifier	0.67	0.33	0.09
Perception	0.65	0.29	0.08
PassiveAggressiveClassifier	0.64	0.30	0.07
DummyClassifier	0.49	-0.01	0.05

Table 4: Performance of various ML algorithms in which DT classifier gave strong performance.



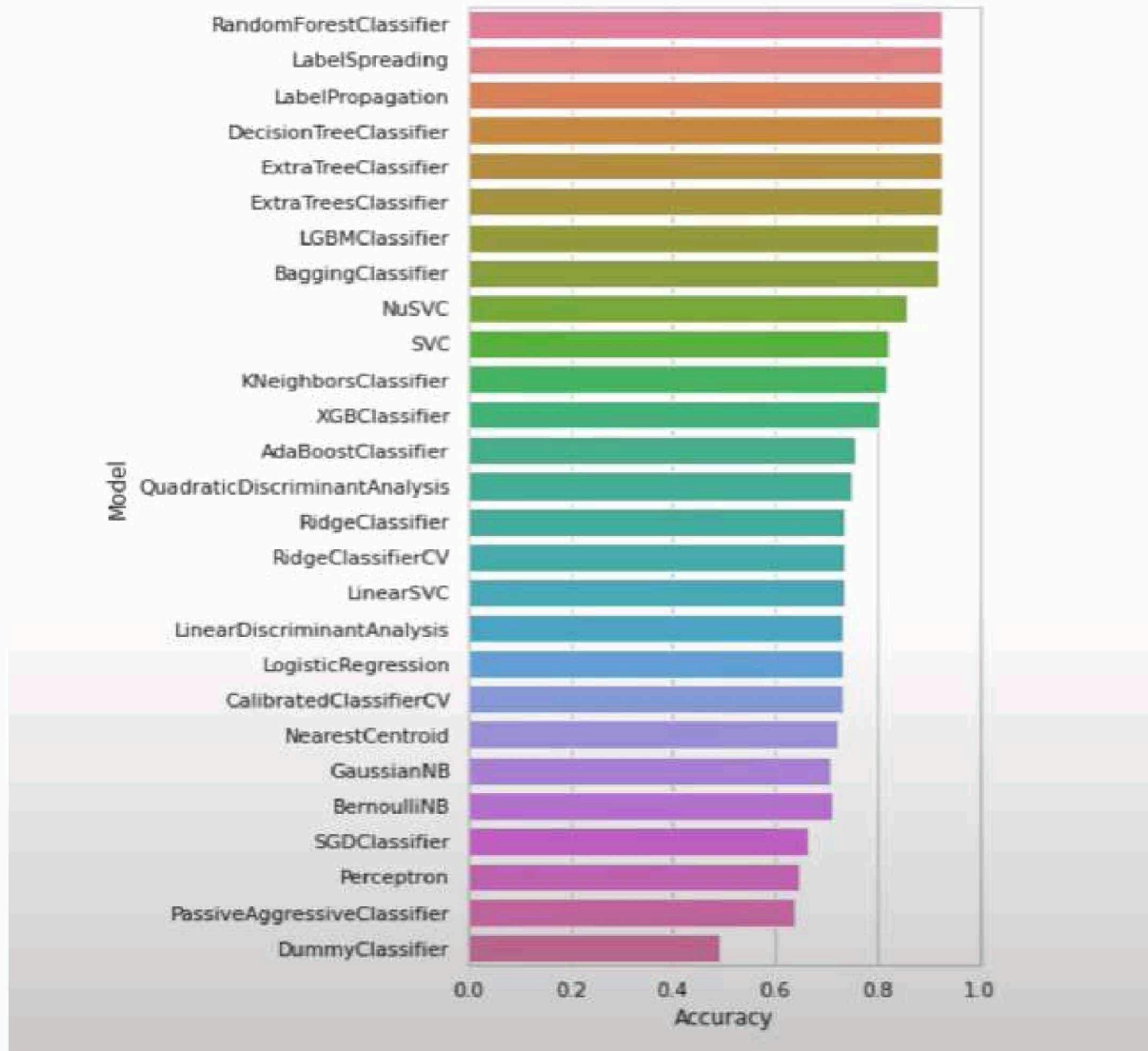


Fig.14: Comparison of various ML algorithms vs their respective accuracy scores.

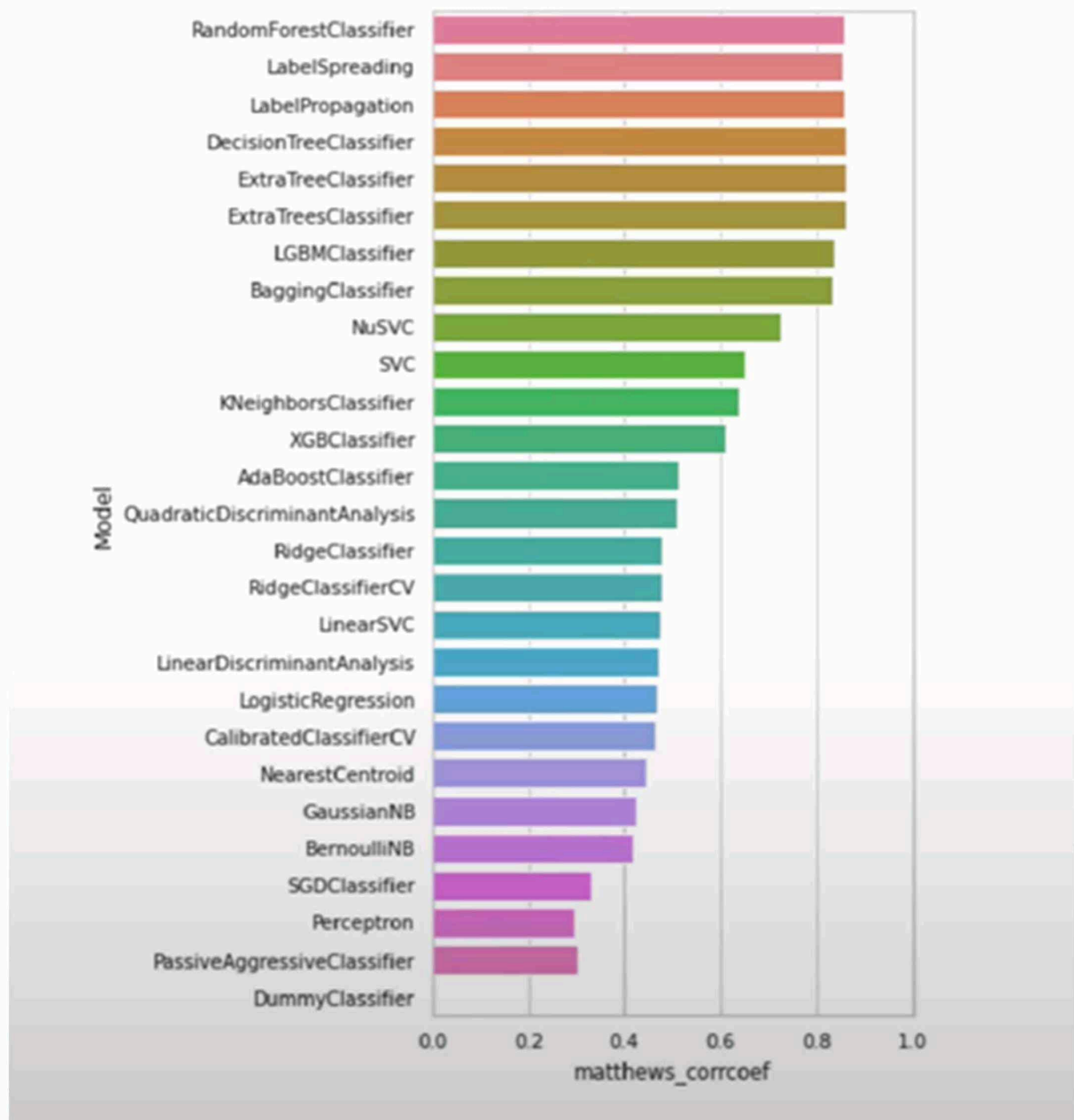


Fig. 15: Comparison of various ML algorithms vs their respective MCC scores.

## 6.1 Discussion

Alternative therapies for infectious diseases, such as AMPs and antibiotic combinations, are being developed to address the AMR problem. Despite some success with AI technologies for AMPs, there are still many challenges to resolve [60]. First, because collecting AMPs and antibiotic combination data is quick and inexpensive with the progression of experimental techniques, substantial and accurate databases with combined higher data quality and recurring updates are possible, which can improve the predictive accuracy of AMPs models [60]. Secondly, current ML-based AMP prediction models employ binary classification approaches, which could only determine whether or not AMPs are active, but not how active they are [60]. Lastly, despite the fact that doses, antibiotic characteristics, and infections all influence antibiotic combinations, existing ML-based algorithms only consider one of these issues, resulting in reduced prediction performance [60]. However, advances in ML and AL technologies can alleviate these issues if a large amount of high-quality data is available to validate the efficacy of the AMP design.

## **Conclusion**

The application of ML to AMR is in its early phases, however due to the increasing availability of genetic information, the most immediate applications of ML to AMR are expected to be laboratory-based, such as AST phenotypic prediction. Despite the obvious benefits of ML in improving overall laboratory productivity and optimising diagnostic procedures, integration into everyday practise remains difficult due to concerns about model interpretability and data quality. There is a need to improve laboratory personnel' understanding of the broader ML ecosystem. It is critical to understand that ML will not work immediately, but as a critical supporting tool. Finally, ML has significantly improved AMR identification, antibiotic development, and discovery procedures by minimizing resources, time, and effort as compared to traditional methods.

## **Future Scope of Research**

Despite obtaining great success in combating antibiotic resistance using ML approaches, it has certain inevitable flaws. Inefficient data is one of the core issues that must be solved. The success of these strategies is dependent on the complete quality of the databases containing vast clinical data that has been put into the ML models for training [60]. Lack of uniform standardisation and intermittent data updates in AMR databases prevent ML models from training efficiently in the above-mentioned AMR prediction and classification scenarios. As a result, several ML models performed poorly. These databases should be carefully curated in order to obtain reliable information on the genotype-phenotype relationships of microbial species, as inaccuracies in the training dataset will result in faulty ML models. The majority of AMR-prediction methods are binary classifiers, which can only predict whether AMPs are active against a certain bacterium. There appears to be no information in those studies about the level of their antibacterial action. Furthermore, a lack of interpretability results from the

"black box" approach of ML algorithms, which extracts crucial information in a way that makes data interpretation difficult for analysts and researchers. To make things easier for doctors and lab professionals, the interface of ML algorithms and models should be constructed in such a way that they can grasp the mechanism and its output. The open sourcing of high-quality machine learning algorithms and models can lead to faster adoption. Furthermore, laboratory professionals can conduct scientific research freely and without supervision. Furthermore, some research based on AMR diagnosis did not provide the validation of the ML models, implying that the dataset (which was not included in the training) was not used to test the efficiency of ML models. This "external" validation verifies that the model is not impacted by other biological processes. Furthermore, validation in studies verifies that the projected results are meaningful and that they can be implemented in the clinical situation. To do this, diagnostic methods such as MALDI-TOF and Raman spectroscopy should be standardised across such clinical labs and associated locally constructed ML models should be adapted to datasets from other geographical locations, particularly those in low and middle-income nations. Data privacy and adaptation are two more basic concerns when deploying ML in clinical contexts. The first concern is the exchange of sensitive information between researchers and medical personnel. The developers have the ability to falsify the data in order to mislead ML models for the benefit of studies, which raises ethical questions about patient data confidentiality, autonomy, and informed permission. The latter concern is that laboratory experts who will be dealing with ML models in the future must embrace this growing and revolutionary technology and be actively involved in its development and application. Otherwise, without their involvement, the developers or owners may distort the outcomes in order to benefit their research. In conclusion, AI has been a huge boost in the detection of AMR and new antibiotics because to the tremendous reduction in time and effort required in compared to traditional methodologies.

## References

1. O'Neill, "Antimicrobial Resistance; Tackling a Crisis for the Health and Wealth of Nations," J. Review on Antimicrobial Resistance, 2014.
2. J. C. Gertrudes, V. G. Maltarollo, R. A. Silva, P. R. Oliveira, K. M. Honório, and A. B. F. da Silva, "Machine learning techniques and drug design," *Curr. Med. Chem.*, vol. 19, no. 25, pp. 4289–4297, 2012.
3. T. Ching et al., "Opportunities and obstacles for deep learning in biology and medicine," *J. R. Soc. Interface*, vol. 15, no. 141, p. 20170387, 2018.
4. J. M. A. Blair, M. A. Webber, A. J. Baylay, D. O. Ogbolu, and L. J. V. Piddock, "Molecular mechanisms of antibiotic resistance," *Nat. Rev. Microbiol.*, vol. 13, no. 1, pp. 42–51, 2015.
5. C. Ghosh, P. Sarkar, R. Issa, and J. Haldar, "Alternatives to conventional antibiotics in the era of antimicrobial resistance," *Trends Microbiol.*, vol. 27, no. 4, pp. 323–338, 2019.
6. N. Stephenson et al., "Survey of machine learning techniques in drug discovery," *Curr. Drug Metab.*, vol. 20, no. 3, pp. 185–193, 2019.
7. A. N. Lima, E. A. Philot, G. H. G. Trossini, L. P. B. Scott, V. G. Maltarollo, and K. M. Honorio, "Use of machine learning approaches for novel drug discovery," *Expert Opin. Drug Discov.*, vol. 11, no. 3, pp. 225–239, 2016.
8. J. Panteleev, H. Gao, and L. Jia, "Recent applications of machine learning in medicinal chemistry," *Bioorg. Med. Chem. Lett.*, vol. 28, no. 17, pp. 2807–2815, 2018.
9. S. Basith, B. Manavalan, T. Hwan Shin, and G. Lee, "Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening," *Med. Res. Rev.*, vol. 40, no. 4, pp. 1276–1314, 2020.

10. S. De Bruyne, M. M. Speeckaert, W. Van Biesen, and J. R. Delanghe, “Recent evolutions of machine learning applications in clinical laboratory medicine,” *Crit. Rev. Clin. Lab. Sci.*, vol. 58, no. 2, pp. 131–152, 2021.
11. P. Rezaei-Hachesu et al., “The design and evaluation of an antimicrobial resistance surveillance system for neonatal intensive care units in Iran,” *Int. J. Med. Inform.*, vol. 115, pp. 24–34, 2018.
12. M. Choisy, N. Van Cuong, T. D. Bao, B. T. Kiet, B. V. Hien, and H. V. Thu, “Assessing antimicrobial misuse in small-scale chicken farms in Vietnam from an observational study, *BMC Vet.*,” *BMC Vet. Res.*, vol. 15, 2019.
13. J. A. Suykens and J. Vandewalle, “Least squares support vector machine classifiers, *Neural.*,” *Neural. Process. Lett.*, vol. 9, 1999.
14. H.-L. Her and Y.-W. Wu, “A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the *Escherichia coli* strains,” *Bioinformatics*, vol. 34, no. 13, pp. i89–i95, 2018.
15. Z. Liu, D. Deng, H. Lu, J. Sun, L. Lv, and S. Li, “Evaluation of machine learning models for predicting antimicrobial resistance of *Actinobacillus pleuropneumoniae* from whole genome sequences, *Front.*,” *Front. Microbiol.*, vol. 11, 2020.
16. C. A. Reynolds, J. A. Finkelstein, G. T. Ray, M. R. Moore, and S. S. Huang, “Attributable healthcare utilization and cost of pneumoniae due to drug-resistant *Streptococcus pneumoniae*: a cost analysis, *Antimicrob.*,” *Antimicrob. Resist. Infect. Control*, vol. 3, 2014.
17. A. M. Voermans, J. C. Mewes, M. R. Broyles, and L. M. G. Steuten, “Cost-effectiveness analysis of a procalcitonin-guided decision algorithm for antibiotic stewardship using real- World U.S. Hospital Data,” *OMICS, J. Integr. Biol.*, vol. 23, 2019.

18. V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, “Random Forest: a classification and regression tool for compound classification and QSAR modeling,” *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 1947–1958, 2003.
19. S. Chandrasekaran, M. Cokol-Cakmak, N. Sahin, K. Yilancioglu, H. Kazan, and J. J. Collins, “Chemogenomics and orthology-based design of antibiotic combination therapies, *Mol.*,” *Mol. Syst. Biol.*, vol. 12, 2016.
20. D. J. Mason et al., “Prediction of antibiotic interactions using descriptors derived from molecular structure,” *J. Med. Chem.*, vol. 60, no. 9, pp. 3902–3912, 2017.
21. D. Moradigaravand, M. Palm, A. Farewell, V. Mustonen, J. Warringer, and L. Parts, “Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data,” *PLoS Comput. Biol.*, vol. 14, no. 12, p. e1006258, 2018.
22. G. Arango-Argoty, E. Garner, A. Pruden, L. S. Heath, P. Vikesland, and L. Zhang, “DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data,” *Microbiome*, vol. 6, no. 1, 2018.
23. Y. Li et al., “HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes,” *Microbiome*, vol. 9, no. 1, p. 40, 2021.
24. A. S. Chowdhury, D. R. Call, and S. L. Broschat, “PARGT: a software tool for predicting antimicrobial resistance in bacteria,” *Sci. Rep.*, vol. 10, no. 1, p. 11033, 2020.
25. N. Singhal, M. Kumar, P. K. Kanaujia, and J. S. Viridi, “MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis,” *Front. Microbiol.*, vol. 6, p. 791, 2015.
26. J. C. V. Weis, C. R. Jutzeler, and K. Borgwardt, “Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review,” *Clin. Microbiol. Infect.*, vol. 26, no. 10, pp. 1310–1317, 2020.



27. T.-S. Huang, S. S.-J. Lee, C.-C. Lee, and F.-C. Chang, “Detection of carbapenem-resistant *Klebsiella pneumoniae* on the basis of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using supervised machine learning approach,” *PLoS One*, vol. 15, no. 2, p. e0228459, 2020.
28. K. Sogawa et al., “Rapid Discrimination between Methicillin-Sensitive and Methicillin-Resistant *Staphylococcus aureus* Using MALDI-TOF Mass Spectrometry,” *Biocontrol Sci.*, vol. 22, no. 3, pp. 163–169, 2017.
29. H.-Y. Wang et al., “Rapid classification of group B *Streptococcus* serotypes based on matrix-assisted laser desorption ionization-time of flight mass spectrometry and machine learning techniques,” *BMC Bioinformatics*, vol. 20, no. Suppl 19, p. 703, 2019.
30. S. Stöckel, J. Kirchhoff, U. Neugebauer, P. Rösch, and J. Popp, “The application of Raman spectroscopy for the detection and identification of microorganisms: Raman spectroscopy for microorganism detection and identification,” *J. Raman Spectrosc.*, vol. 47, no. 1, pp. 89–109, 2016.
31. R. Ullah, S. Khan, I. I. Chaudhary, S. Shahzad, H. Ali, and M. Bilal, “Cost effective and efficient screening of tuberculosis disease with Raman spectroscopy and machine learning algorithms,” *Photodiagnosis Photodyn. Ther.*, vol. 32, no. 101963, p. 101963, 2020.
32. A. A. Moawad et al., “A machine learning-based Raman spectroscopic assay for the identification of *Burkholderia mallei* and related species,” *Molecules*, vol. 24, no. 24, p. 4516, 2019.
33. K. Rebrošová et al., “Rapid identification of staphylococci by Raman spectroscopy,” *Sci. Rep.*, vol. 7, no. 1, p. 14846, 2017.

34. C.-S. Ho et al., “Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning,” *Nat. Commun.*, vol. 10, no. 1, p. 4927, 2019.
35. X. Yi et al., “Development of a fast Raman-assisted antibiotic susceptibility test (FRAST) for the antibiotic resistance analysis of clinical urine and blood samples,” *Anal. Chem.*, vol. 93, no. 12, pp. 5098–5106, 2021.
36. T. J. J. Inglis, T. F. Paton, M. K. Kopczyk, K. T. Mulroney, and C. F. Carson, “Same-day antimicrobial susceptibility test using acoustic-enhanced flow cytometry visualized with supervised machine learning,” *J. Med. Microbiol.*, vol. 69, no. 5, pp. 657–669, 2020.
37. Y. Maeda et al., “Colony fingerprint-based discrimination of *Staphylococcus* species with machine learning approaches,” *Sensors (Basel)*, vol. 18, no. 9, p. 2789, 2018.
38. K. P. Smith, D. L. Richmond, T. Brennan-Krohn, H. L. Elliott, and J. E. Kirby, “Development of MAST: A microscopy-based antimicrobial susceptibility testing platform,” *SLAS Technol.*, vol. 22, no. 6, pp. 662–674, 2017.
39. L. Lechowicz, M. Urbaniak, W. Adamus-Białek, and W. Kaca, “The use of infrared spectroscopy and artificial neural networks for detection of uropathogenic *Escherichia coli* strains’ susceptibility to cephalothin,” *Acta Biochim. Pol.*, vol. 60, no. 4, pp. 713–718, 2013.
40. E. Y. Lee, M. W. Lee, B. M. Fulan, A. L. Ferguson, and G. C. Wong, what can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning? *Interface Focus*. 2017.
41. M. H. Cardoso et al., “Computer-aided design of antimicrobial peptides: Are we generating effective drug candidates?,” *Front. Microbiol.*, vol. 10, p. 3097, 2019.

42. P. Rondón-Villarreal, D. A. Sierra, and R. Torres, “Machine learning in the rational design of antimicrobial peptides,” *Curr. Comput. Aided Drug Des.*, vol. 10, no. 3, pp. 183–190, 2014.
43. H. J. Lau, C. H. Lim, S. C. Foo, and H. S. Tan, “The role of artificial intelligence in the battle against antimicrobial-resistant bacteria,” *Curr. Genet.*, vol. 67, no. 3, pp. 421–429, 2021.
44. J. M. Stokes et al., “A deep learning approach to antibiotic discovery,” *Cell*, vol. 180, no. 4, pp. 688-702.e13, 2020.
45. E. Y. Lee, M. W. Lee, B. M. Fulan, A. L. Ferguson, and G. C. L. Wong, “What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning?,” *Interface Focus*, vol. 7, no. 6, p. 20160153, 2017.
46. A. A. Bahar and D. Ren, “Antimicrobial peptides,” *Pharmaceuticals (Basel)*, vol. 6, no. 12, pp. 1543–1575, 2013.
47. A. Capecchi, X. Cai, H. Personne, T. Köhler, C. van Delden, and J.-L. Reymond, “Machine learning designs non-hemolytic antimicrobial peptides,” *Chem. Sci.*, vol. 12, no. 26, pp. 9221–9232, 2021.
48. K. Boone, C. Wisdom, K. Camarda, P. Spencer, and C. Tamerler, “Combining genetic algorithm with machine learning strategies for designing potent antimicrobial peptides,” *BMC Bioinformatics*, vol. 22, no. 1, p. 239, 2021.
49. S. Mohanty, M. Harun Ai Rashid, M. Mridul, C. Mohanty, and S. Swayamsiddha, “Application of Artificial Intelligence in COVID-19 drug repurposing,” *Diabetes Metab. Syndr.*, vol. 14, no. 5, pp. 1027–1031, 2020.
50. J. Kowalewski and A. Ray, “Predicting novel drugs for SARS-CoV-2 using machine learning from a >10 million chemical space,” *Heliyon*, vol. 6, no. 8, p. e04639, 2020.
51. [https://colab.research.google.com/?utm\\_source=scs-index](https://colab.research.google.com/?utm_source=scs-index)

52. <https://www.python.org/>
53. <https://jupyter.org/>
54. W. Li, L. Jaroszewski, and A. Godzik, “Clustering of highly homologous sequences to reduce the size of large protein databases,” *Bioinformatics*, vol. 17, no. 3, pp. 282–283, 2001.
55. A. Pande et al., “Computing wide range of protein/peptide features from their sequence and structure,” *bioRxiv*, 2019.
56. <https://lazypredict.readthedocs.io/en/latest/index.html>
57. <https://scikit-learn.org/stable/>
58. <https://matplotlib.org/>
59. D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 1, p. 6, 2020.
60. J. Lv, S. Deng, and L. Zhang, “A review of artificial intelligence applications for antimicrobial resistance,” *Biosafety and Health*, vol. 3, no. 1, pp. 22–31, 2021.

PAPER NAME

ashish kukreti 2K20MSCBIO02 Thesis

ASD  
6/5/22

WORD COUNT

6619 Words

CHARACTER COUNT

37723 Characters

PAGE COUNT

36 Pages

FILE SIZE

1.6MB

SUBMISSION DATE

May 4, 2022 8:16 AM GMT+5:30

REPORT DATE

May 4, 2022 8:17 AM GMT+5:30

**● 7% Overall Similarity**

The combined total of all matches, including overlapping sources, for each database.

- 1% Internet database
- Crossref database
- 3% Submitted Works database
- 5% Publications database
- Crossref Posted Content database

**● Excluded from Similarity Report**

- Bibliographic material

## ● 7% Overall Similarity

Top sources found in the following databases:

- 1% Internet database
- 5% Publications database
- Crossref database
- Crossref Posted Content database
- 3% Submitted Works database

### TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	<b>Sander De Bruyne, Marijn M. Speeckaert, Wim Van Biesen, Joris R. Del...</b>	2%
	Crossref	
2	<b>Hul Juan Lau, Chern Hong Lim, Su Chern Foo, Hock Siew Tan. "The role...</b>	2%
	Crossref	
3	<b>University of Mines and Technology on 2021-07-28</b>	<1%
	Submitted works	
4	<b>No.55 High School of Beijing on 2021-11-14</b>	<1%
	Submitted works	
5	<b>University of Bath on 2018-09-10</b>	<1%
	Submitted works	
6	<b>University of Ulster on 2015-05-01</b>	<1%
	Submitted works	
7	<b>Napier University on 2017-08-13</b>	<1%
	Submitted works	
8	<b>UT, Dallas on 2011-08-08</b>	<1%
	Submitted works	

9	ueaeprints.uea.ac.uk Internet	<1%
10	Zicheng Fei, Fangfang Yang, Kwok-Leung Tsui, Lishuai Li, Zijun Zhang. ... Crossref	<1%
11	ebin.pub Internet	<1%
12	RMIT University on 2018-09-03 Submitted works	<1%
13	University of Ulster on 2015-05-29 Submitted works	<1%
14	Queen Mary and Westfield College on 2021-05-04 Submitted works	<1%
15	National College of Ireland on 2018-08-16 Submitted works	<1%
16	Yu Li, Zeling Xu, Wenkai Han, Huiluo Cao et al. "HMD-ARG: hierarchical ... Crossref	<1%
17	"Medical Image Computing and Computer Assisted Intervention – MIC... Crossref	<1%
18	University of Northumbria at Newcastle on 2018-06-01 Submitted works	<1%
19	"13th European Congress of Clinical Microbiology and Infectious Disea... Crossref	<1%
20	Xiaohu Tang, Zhifeng Liu, Taizhao Li, Wenbin Wu, Zhenhua Wei. "The A... Crossref	<1%

# *Machine Learning: A promising in-silico approach to curb antimicrobial resistance*

Ashish Kukreti

Department of Biotechnology  
Delhi Technological University  
Delhi, India  
ashkukreti2000@gmail.com

Yasha Hasija\*

Department of Biotechnology  
Delhi Technological University  
Delhi, India  
yashahasija06@gmail.com  
\*Corresponding Author

**Abstract**— Antimicrobial resistance (AMR) is a concern to public health, prompting the development of novel strategies for combating AMR. While the use of machine learning (ML) to AMR is in its infancy, it has made significant progress as a diagnosis tool, owing to the growing availability of phenotypic/genotypic datasets and much faster computational power. While applying ML in AMR research is viable, its use is limited. It has been used to predict antimicrobial susceptibility genotypes/phenotypes, discover novel antibiotics, and improve diagnosis when combined with spectroscopic and microscopy methods. ML implementation in healthcare settings has challenges to adoption due to concerns about model interpretability and data integrity. The focus of this review is to outline the significant benefits and drawbacks along with the salient trends reported in recent studies.

**Keywords**— antibiotics, antimicrobial resistance, artificial intelligence, COVID-19, deep learning, halicin, machine learning.

## I. INTRODUCTION

AMR has emerged as a major public health hazard in the 21st century. It is increasing globally and is related to increased morbidity and mortality in hospital and community settings [1]. The expansion of antibiotic resistance to new environmental niches, as well as the emergence of superbugs, has challenged effective control methods. For a successful outcome, early and accurate diagnosis of the illness and its drug resistance profile is crucial. Over the years, some alternatives to standard antibiotic treatment have been presented to avoid the problem of bacterial resistance. Among these, the use of artificial intelligence (AI), particularly ML, has been widely used in the medicinal chemistry field due to its high capacity for processing data from multiple sources as well as the development of new methods and algorithms for predicting experimental data. ML is an AI subset that gives a computer the ability to predict events using a learned model and a vast amount of experimental data, often known as training data. ML can aid in the identification and design of novel antibacterial drugs by generating models based on empirical knowledge accessible in compound databases. Furthermore, ML approaches can be utilized to examine the pharmacokinetics and toxicity features of potential antibiotics

to improve their efficacy [2]. Any technique that employs ML models necessitates sufficient input data to create a 'training set' for guiding the ML model and a 'testing set' for evaluating the model's performance. An important consideration in the study of AMR is the availability of data sets including correct genotypic information connected to carefully selected samples of the AMR gene. For this reason, pre-processing of genotypic data is required before they can be used as input for ML models. This procedure is called "feature selection" in the ML terminology. The studies took one of two approaches to this problem: using gene annotations as "features" or using k-mers (short DNA strings are constructed by combining individual nucleotides). Most of the algorithms discussed here used supervised ML, where the algorithm is trained on data with labels to construct a learned model. Hence, these algorithms could analyze test data and execute self-annotation using their label. Generally, they are employed for classification and regression-based analysis. Linear Regression, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVMs), Decision Trees (DTs), and Random Forests (RFs) are some of the most used supervised algorithms. While some studies have employed unsupervised learning and deep learning algorithms. Unsupervised learning makes use of unlabeled training data. To put it differently, observations are classified without information of the data sample. Clustering can be performed using unsupervised techniques like hierarchical cluster analysis and principal component analysis (PCA). Other AI approach such as deep learning, is based on how the biological neural systems interpret information and is becoming increasingly popular. Artificial Neural Networks (ANNs) are constituted of an input layer of 'neurons,' which is interconnected to one or more hidden layers of neurons, which are then linked to an output layer [3]. This review will not delve into detailed discussions on these algorithms. However, this review article will present an up-to-date summary of recent advances in antimicrobial resistance research utilizing the current machine-learning techniques.



## II. APPLICATIONS OF ML TECHNIQUES IN ANTIMICROBIAL RESISTANCE

### A. Antibiotics/Anti-microbial peptides Discovery

Despite the clear need for more antibiotics, very few antibiotics have been available in the market. The last entire class of antibiotics was discovered in the late 1980s. Many pharmaceutical companies have gradually shied away from antibiotic research and development due to the several challenges posed in the discovery of novel antibiotics as their development is time-consuming and expensive. The cost-benefit ratio is much less favourable than for other drugs. Subsequently, physicians avoid prescribing new antibiotics to delay antibiotic resistance and are usually used as "last resort drugs" when traditional medicines failed to do their work. Hence, these industries shift their focus on developing profitable long-term treatments for chronic infections. These limitations have been addressed progressively with the employment of various ML techniques that discover newer antibiotics and potential lead compounds which makes their identification less laborious and cost-effective. For instance, Stokes et al. [4] implemented deep neural networks to identify eight novel broad-spectrum antibiotic molecules. These identified molecules were structurally distinct from all known antibiotics. In this study, more than 107 million molecules from the ZINC15 database were assembled to build a training dataset of 2335 molecules for the prediction of potential molecules showing the inhibitory action on *Escherichia coli*. The researchers were able to find out the potential candidates who met a predetermined threshold score and various other exclusion criteria [Fig. 1]. The study successfully identified "halicin" as a potent growth inhibitor of *E. coli* and efficient against other bacterial infections in animal models. In recent years, most of the studies employed ML algorithms to identify novel antimicrobial peptide (AMP) drugs [5]. These AMPs are the class of small peptides that can directly kill pathogenic microbes as well as can indirectly modulate the host defence system [6]. In a study conducted by Capecchi et al. [7], deep neural networks and supervised learning algorithms were utilized to generate a classifier model to predict non-haemolytic AMPs for *Methicillin Resistant Staphylococcus Aureus* (MRSA), *P. aeruginosa*, and *A. baumannii*. The training and testing dataset were assembled from 4774 peptides found in "DBAASP" (Database of Antimicrobial Activity and Structure of Peptides) [Some other databases are discussed in Table 1]. The study showed promising results by identifying eight non-haemolytic AMPs that met a predetermined threshold value of minimum inhibitory concentration (MIC). Moreover, Boone et al. [8] successfully employed supervised ML and a codon-based genetic algorithm to identify an active AMP against *S. epidermidis*, a normal commensal found mainly on skin. Several studies have been implementing ML to find a possible treatment for COVID-19 [9]. Kowalewski and Ray [10] developed models for the prediction of potential drugs against 65 target human proteins, including the ACE-2 receptor by employing supervised ML algorithms. The researchers gathered 14 million compounds

from the ZINC database and applied ML models for the prediction of certain features like binding affinity and toxicity to classify molecules and identify compounds with similar chemical space [Fig. 2]. Altogether, with sufficient prior information about known AMPs, ML can be applied to discover novel AMPs by which the development of novel antibiotics would become cost-effective and time-efficient while achieving more efficacy than conventional methods.

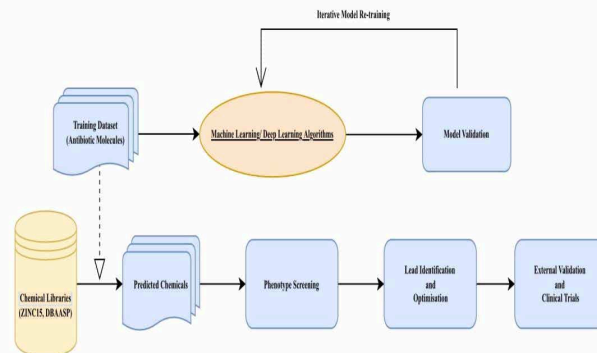


Fig. 1. Overview of the workflow of antibiotic/antimicrobial peptides discovery by using machine learning and deep learning algorithm. The researchers first constructed a training dataset of compounds capable of inhibiting the growth of common bacterial species such as *E. coli*. They improved their model by specifying phenotypic and genotypic features, adjusting hyperparameters, and assembling the ML model, which resulted in the final machine learning or deep learning algorithm via iterative model re-training. In comparison to the traditional approach, this in-silico approach enabled the researchers to systematically scan over a 100 million compounds, many of which had molecular structures that differed from known antibiotics. The number of screened compounds is approximately 100 times more than the typical approach, which is more inexpensive at this scale.

TABLE I. AI-BASED AMP DATABASES

Database	Type of ML algorithm	Description
INDIGO (INferring Drug Interactions using chemo-Genomics and Orthology)	Supervised learning (Random Forest)	Exploits chemogenomic data of model organisms for the prediction of antibiotic combinations
CAMP (Collection of Anti-Microbial Peptides)	Semi-supervised learning (Hidden Markov Model)	Detailed information on 8164 AMR sequences with 752 AMR 3-D structure
YADAMP (Yet Another Database of Anti-Microbial Peptides)	Supervised learning	Contain comprehensive information on 2525 AMPs against common bacterial species



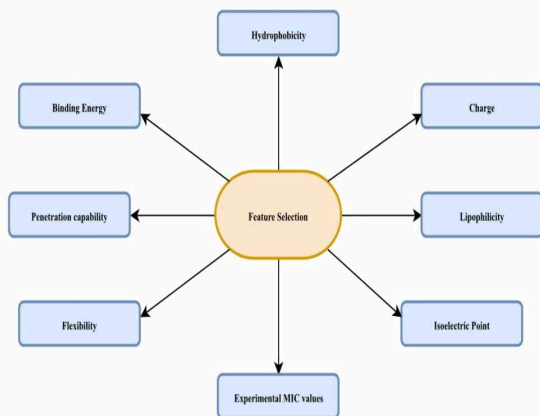


Fig. 2. Combinations of certain features are selected for the novel and/or effective antibiotics/AMPs determination.

### B. Assist in Diagnosis

When designing a new AST (Antibiotic Susceptibility Testing) method, several nonautomated procedures are considered gold standards for comparative reasons. These methods include agar dilution, broth microdilution, disc diffusion, and the E-test. While these tests are cost-effective, they must be carried out manually and take roughly 18-24 hours to complete. Although these are not considered quick AST techniques, they provide valid MIC values. Scientific research over the last few years has not only resulted in the development of novel AST platforms but also improve existing platforms. The integration of ML algorithms has resulted in a significant progression of AST approaches [Fig. 3].

MALDI-TOF MS has proven to be a fast, inexpensive, and accurate tool that is being used in the identification of antibiotic-resistant microorganisms by generating characteristic mass spectral fingerprints which are unique to microbes at genus and species level, which are then compared with a reference library database comprising of well-identified organisms for each isolate [11]. Unlike conventional methods MALDI-TOF offers species-level identification with reduced turn-around time and far more accurate results. Recently, ML algorithms have been employed in the optimization of MALDI-TOF to enhance species identification which can reveal unknown or novel information hidden in the mass spectra and, in antibiotic resistance profiling of closely related bacterial species [12]. However, there is an increased interest in utilizing MALDI-TOF for AST. Data retrieved from traditional methods may serve as user input for ML algorithms. Huang et al. [13] evaluated five ML algorithms (RF, logistic regression, naïve Bayes, NN, and SVM) to check their antibiotic susceptibility of *Klebsiella pneumoniae* against carbapenem. A total of 100 spectra peaks from 46 *carbapenem-resistant K. pneumoniae* and 50 *carbapenem-susceptible K. pneumoniae* isolates were used as the training dataset for their classification. RF algorithm surpasses the other algorithms achieving an overall classification accuracy of 97%. Mather et al. [14] successfully

employed a supervised ML algorithm to differentiate *vancomycin-intermediate Staphylococcus aureus (VISA)* and heterogeneous VISA (hVISA) from *vancomycin-susceptible S. aureus (VSSA)* achieving overall classification accuracy of 89%. Sogawa and colleagues [15] tested a prediction model using a supervised ML algorithm to classify 50 isolates of both *Methicillin-susceptible S. aureus (MSSA)* and MRSA. The accuracy rates were 90% and 87.5% respectively. This study has proposed rapid detection from one colony in 5 minutes, however, the accuracy rate was not 100% which is crucial for clinical diagnosis. Wang et al. [16] included 787 *Group-B streptococci (GBS)* isolates in their analysis where they have used supervised ML to build models for the prediction of 5 different serotypes achieving up to 87.5 % accuracy while other studies [13,14,15] included less than or equal to 50 isolates only which is quite a small sample size for ML models that require larger sample size to feed. Although most of the studies showed good results, lack of external validation and poor reproducibility dampen progress towards this approach.

Raman spectroscopy analyses biochemical composition by using multichromatic emitters in the ultraviolet/infrared/visible spectrum. Raman scattering can be used to measure small quantities of material, such as single bacterial cells, and has several microbiological applications. Laser light can be used to investigate the physicochemical characteristics of the probed sample. Molecular bonds then inelastically scatter the photons which are analyzed by spectrophotometer and a Raman spectrum is plotted to depict the intensity of the inelastic scattering [17]. These Raman spectra and their associated information can be utilized for the identification of bacteria and for other microbial diagnoses by analyzing through multivariate statistics and ML algorithms. Finally, these class-specific Raman spectral models, after iterations by ML algorithms, can be utilized to distinguish bacterial classes of interest [17]. Ullah et al. [18] employed unsupervised ML techniques to classify 60 tuberculosis (TB) serum samples (30- TB positive and 30- TB negative) based on the variation in biochemical concentration. The findings indicate a significant difference in Raman spectra in both TB positive and TB negative groups as well as in the control group too. Similarly, Moawad et al. [19] developed an SVM model in combination with PCA to identify *Burkholderia mallei* and related species. The optimized model identification accuracy reached above 90%. However, the model's reliability was limited due to the occurrence of misclassification of *B. thailandensis* with *B. mallei*. For which, the authors suggested that this misclassification occurred due to the less representation of *B. thailandensis* samples in the training dataset. Rebrošová et al. [20] successfully identified 277 staphylococci strains from 16 species by utilizing Raman spectroscopy with supervised ML algorithms showing better results achieving an accuracy of 99%. Meanwhile, Ho et al. [21] implemented a deep learning technique known as convolutional neural network (CNN), along with logistic regression for the rapid classification of (n=30) bacterial isolates, for antibiotic resistance and empirical treatment. For (n=30) bacterial species classification, the CNN classifier comprised of 25 1-D convolutional layers along with some residual connections which achieved an average isolate-level



accuracy of 82%. However, most of the misclassifications occurred at the genus level. Moreover, a binary CNN classifier was built to classify MRSA and MSSA which achieved an accuracy of 89%, and another CNN model categorized known bacterial isolates into several groups based on the common empirical antibiotic treatment that achieved an accuracy of 97%. Additionally, Yi et al. [22] devised a rapid AST based on Raman scattering, which detects an activity and stability of certain metabolites in the presence of antibiotics via single-cell Raman spectroscopy (SCRS). SCRS works by detecting the biomolecules' vibrations within a cell/bacterium, hence determining their biochemical attributes or phenotype. In this study, FRAST was applied to (n=8) bacterial species, consisting of 4 Gram-positive and 4 Gram-negative bacterial species along with ML algorithms to train a model to classify training dataset which achieved sensitivities of 98.8% for Gram-positive bacteria and 94.3% for Gram-negative bacteria at the unicellular level. The model only took less than 30 minutes for the results. The model was then validated using (n=6) bacterial isolates consisting of 3 Gram-positive and 3 Gram-negative bacterial species, which achieved an overall sensitivity of over 90%, which was then confirmed by 16S rRNA sequencing. However, the current limitations of implementing Raman spectroscopy in clinical labs are that first, they generate huge amount of data shortly, which could be surpassed by employing ML techniques. Second, training datasets consisting of a small sample size result in lower accuracy by the learning model that might be resolved by feeding the training model with large datasets encompassing varieties of antibiotic-susceptible and resistant bacteria. Finally, the improper optimization of training data and lack of external validation of the analysis results in misclassifications amongst similar bacterial species.

Several other diagnostic tools with the help of ML algorithms have been employed to improve accuracy and turnaround time in AST. For instance, Inglis et al. [23] implemented AST with the flow cytometry method (FAST). The method utilized a decision tree algorithm to deliver the results within 3 hr. Based on a multivariate analysis of microcolony images, Maeda et al. [24] employed a novel technology known as "colony fingerprinting" to distinguish five *Staphylococcus* species. The method used supervised ML and deep learning algorithms which showed high performance and generated the results within 11 hr. Moreover, Smith et al. [25] developed an AST platform, based on microscopy (MAST) which used deep neural networks for classification and could determine the AST after incubation of 2 hr. Further, Lechowicz et al. [26] used the combination of deep neural networks and infrared spectroscopy to classify 109 uropathogenic *E. coli* strains against cephalothin. The method generated quick results within 30 min. which are much faster than conventional AST methods (24 hr.).

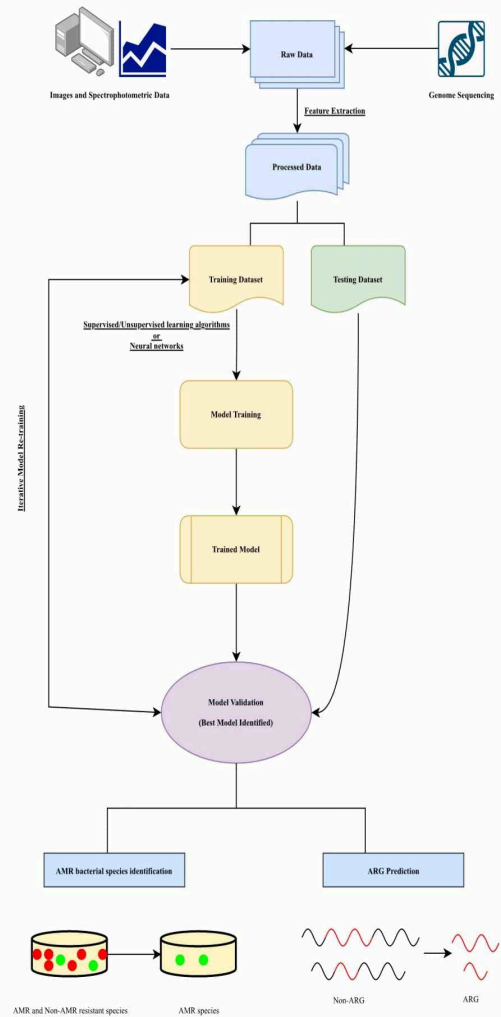


Fig. 3. Overview of the workflow for antibiotic resistant genes prediction and AMR species identification using machine learning and deep neural networks.

### C. AMR Genes Prediction

The most common way of developing antibiotic resistance is the transfer of antibiotic resistance genes (ARG) among the bacterial species. Current methods are incapable which rely on inefficient phenotypic data. Therefore, it becomes an important step to predict these ARGs accurately for a better understanding of their transmission from the environment. Most bioinformatics tools have been based on either assembly-based methods or read-based methods to identify or annotate resistance genes [Table 2] With ever-increasing genomic data, ML models have been built to learn the statistical patterns of ARGs and may potentially identify novel ones as they detect certain features rather than using sequence similarity [Fig. 3]. Some ML methods have reported models to identify novel ARGs from pan-genome and metagenomic data [27,28]. Although the results were satisfactory, these models used limited features and did not include any feature selection method to remove redundant data. Subsequently, Li et al. [29]

developed HMD-ARG (Hierarchical Multi-task Deep learning framework for prediction of the ARG) coupled with deep neural networks. The framework provides detailed information on annotated ARGs based on their biochemical properties and covers three significant aspects: resistance class, gene mobility, and mechanism. In another study, Chowdhury et al. [30] proposed a model called PARGT (Prediction of Antimicrobial Resistance via Game Theory), which can identify ARGs from bacterial species. The model utilized the supervised ML algorithm. These two methods [29,30] validated their results for feature selection. However, in the future, these methodologies should be made compatible with current sequencing technologies, which work on short reads like nanopore sequencing rather than on assembled sequences.

TABLE II. AI-BASED AMR PREDICTION TOOLS

Tool	Type of AI method	Description
ABCRpred	Supervised ML (Random Forest)	Predict ceftazidime resistance/susceptibility of beta-lactamase protein sequence
GenTB	Supervised ML (Random Forest)	Predict antibiotic resistance against <i>Mycobacterium tuberculosis</i>
VAMPPr	Supervised ML (Decision Tree)	Utilize next-generation sequencing data to determine antibiotic resistance
DeepARG	Deep learning (Artificial Neural Networks)	Utilize metagenomic data to characterize ARG

### III. CHALLENGES AND OPPORTUNITIES

Despite achieving remarkable results in applying ML techniques to combat antibiotic resistance, it has some unavoidable shortcomings. One of the fundamental challenges that should get addressed is inefficient data. The success of these techniques relies on the comprehensive quality of the databases with extensive clinical data that has been fed to train the ML models. Lack of universal standardization and sporadic data updates in AMR databases restrict ML models from training efficiently in the scenarios of AMR prediction and classification outlined above. Therefore, some ML models gave a below-par performance. These databases should be aptly curated to ascertain accurate information on the genotype-phenotype of microbial species, as errors in the training dataset will cause inaccurate ML models. Most AMR-predictive models are binary classifiers that can only predict whether AMPs are active against a specific microbe. There seems to be no information about the extent of their antimicrobial activity in those studies. In addition, a lack of interpretability arises from the "black box" approach of ML algorithms extract necessary information in such a way that makes the data interpretation difficult for the analysts and researchers. To make things easier for clinicians and lab workers, the interface of ML algorithms and models should get designed in such a user-friendly manner so that it would be easier for them to understand the mechanism and its output. Open sourcing of high-quality ML algorithms and models can result in faster adoption. In addition, laboratory professionals

can independently perform scientific research without any intervention. Also, some studies based on the diagnosis of AMR did not report the validation of the ML models, which means that the dataset (not included in the training) was not adequately employed to test the efficiency of ML models. This "external" validation ensures that the model is unbiased and not influenced by other biological factors. Further, validation in studies ensures the predicted results are significant and can be applied to the clinical setting. To achieve this, diagnostic tools like MALDI-TOF and Raman spectroscopy should get standardized across such clinical labs and, associated locally designed ML models should get generalized to the dataset of other geographical regions, especially in low and middle-income countries. Other general challenges concern while implementing ML in clinical settings are data privacy and adaptability. The former issue involves the sharing of sensitive information between the researchers and health care personnel. The developers can manipulate the data to mislead ML models for the benefit of studies, which in fact, raises ethical concerns regarding patient data confidentiality, autonomy, and informed consent. The latter issue revolves around laboratory professionals who eventually deal with ML models in the future are needed to embrace this emerging and revolutionary technology and should be actively involved in the development and implementation. Otherwise, without their cooperation, the developers or the owners may misinterpret the results to favor their study. In conclusion, due to the significant reduction in time and effort required in comparison to previous methods, AI has been an enormous help in the identification of AMR and novel antibiotics.

### IV. CONCLUSION

The use of ML to tackle AMR is in its initial stages, but due to the growing availability of genetic datasets, the most imminent applications of ML to AMR are expected to be laboratory-based, such as AST phenotypic prediction. Despite the obvious advantages of ML to improve the overall productivity of laboratory process and optimize diagnostic methods, application into everyday practice remains challenging due to concerns regarding model interpretability and data quality. There is a need to enhance knowledge among laboratory professionals about the wide ML ecosystem. It is important to understand that ML will not operate promptly, but rather will serve as a vital supporting tool. Ultimately, ML has considerably improved AMR identification, antibiotic development, and discovery procedures by significantly reducing resources, time, and effort compared to conventional methods.

## Reference

- [1] O'Neill, "Antimicrobial Resistance; Tackling a Crisis for the Health and Wealth of Nations," *J. Review on Antimicrobial Resistance*, 2014.
- [2] J. C. Gertrudes, V. G. Maltarollo, R. A. Silva, P. R. Oliveira, K. M. Honório, and A. B. F. da Silva, "Machine learning techniques and drug design," *Curr. Med. Chem.*, vol. 19, no. 25, pp. 4289–4297, 2012.
- [3] T. Ching *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *J. R. Soc. Interface*, vol. 15, no. 141, p. 20170387, 2018.
- [4] J. M. Stokes *et al.*, "A deep learning approach to antibiotic discovery," *Cell*, vol. 180, no. 4, pp. 688–702.e13, 2020.
- [5] E. Y. Lee, M. W. Lee, B. M. Fulan, A. L. Ferguson, and G. C. L. Wong, "What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning?," *Interface Focus*, vol. 7, no. 6, p. 20160153, 2017.
- [6] A. A. Bahar and D. Ren, "Antimicrobial peptides," *Pharmaceuticals (Basel)*, vol. 6, no. 12, pp. 1543–1575, 2013.
- [7] A. Capecci, X. Cai, H. Personne, T. Köhler, C. van Delden, and J.-L. Reymond, "Machine learning designs non-hemolytic antimicrobial peptides," *Chem. Sci.*, vol. 12, no. 26, pp. 9221–9232, 2021.
- [8] K. Boone, C. Wisdom, K. Camarda, P. Spencer, and C. Tamerler, "Combining genetic algorithm with machine learning strategies for designing potent antimicrobial peptides," *BMC Bioinformatics*, vol. 22, no. 1, p. 239, 2021.
- [9] S. Mohanty, M. Harun Ai Rashid, M. Mridul, C. Mohanty, and S. Swayamsiddha, "Application of Artificial Intelligence in COVID-19 drug repurposing," *Diabetes Metab. Syndr.*, vol. 14, no. 5, pp. 1027–1031, 2020.
- [10] J. Kowalewski and A. Ray, "Predicting novel drugs for SARS-CoV-2 using machine learning from a >10 million chemical space," *Heliyon*, vol. 6, no. 8, p. e04639, 2020.
- [11] N. Singhal, M. Kumar, P. K. Kanaujia, and J. S. Viridi, "MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis," *Front. Microbiol.*, vol. 6, p. 791, 2015.
- [12] C. V. Weis, C. R. Jutzeler, and K. Borgwardt, "Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review," *Clin. Microbiol. Infect.*, vol. 26, no. 10, pp. 1310–1317, 2020.
- [13] T.-S. Huang, S. S.-J. Lee, C.-C. Lee, and F.-C. Chang, "Detection of carbapenem-resistant *Klebsiella pneumoniae* on the basis of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using supervised machine learning approach," *PLoS One*, vol. 15, no. 2, p. e0228459, 2020.
- [14] C. A. Mather, B. J. Werth, S. Sivagnanam, D. J. SenGupta, and S. M. Butler-Wu, "Rapid detection of vancomycin-intermediate staphylococcus aureus by matrix-assisted laser desorption ionization-time of flight mass spectrometry," *J. Clin. Microbiol.*, vol. 54, no. 4, pp. 883–890, 2016.
- [15] K. Sogawa *et al.*, "Rapid Discrimination between Methicillin-Sensitive and Methicillin-Resistant *Staphylococcus aureus* Using MALDI-TOF Mass Spectrometry," *Biocontrol Sci.*, vol. 22, no. 3, pp. 163–169, 2017.
- [16] H.-Y. Wang *et al.*, "Rapid classification of group B Streptococcus serotypes based on matrix-assisted laser desorption ionization-time of flight mass spectrometry and machine learning techniques," *BMC Bioinformatics*, vol. 20, no. Suppl 19, p. 703, 2019.
- [17] S. Stöckel, J. Kirchoff, U. Neugebauer, P. Rösch, and J. Popp, "The application of Raman spectroscopy for the detection and identification of microorganisms: Raman spectroscopy for microorganism detection and identification," *J. Raman Spectrosc.*, vol. 47, no. 1, pp. 89–109, 2016.
- [18] R. Ullah, S. Khan, I. I. Chaudhary, S. Shahzad, H. Ali, and M. Bilal, "Cost effective and efficient screening of tuberculosis disease with Raman spectroscopy and machine learning algorithms," *Photodiagnosis Photodyn. Ther.*, vol. 32, no. 101963, p. 101963, 2020.
- [19] A. A. Moawad *et al.*, "A machine learning-based Raman spectroscopic assay for the identification of *Burkholderia mallei* and related species," *Molecules*, vol. 24, no. 24, p. 4516, 2019.
- [20] K. Rebřošová *et al.*, "Rapid identification of staphylococci by Raman spectroscopy," *Sci. Rep.*, vol. 7, no. 1, p. 14846, 2017.
- [21] C.-S. Ho *et al.*, "Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning," *Nat. Commun.*, vol. 10, no. 1, p. 4927, 2019.
- [22] X. Yi *et al.*, "Development of a fast Raman-assisted antibiotic susceptibility test (FRASST) for the antibiotic resistance analysis of clinical urine and blood samples," *Anal. Chem.*, vol. 93, no. 12, pp. 5098–5106, 2021.
- [23] T. J. J. Inglis, T. F. Paton, M. K. Koczyk, K. T. Mulrone, and C. F. Carson, "Same-day antimicrobial susceptibility test using acoustic-enhanced flow cytometry visualized with supervised machine learning," *J. Med. Microbiol.*, vol. 69, no. 5, pp. 657–669, 2020.
- [24] Y. Maeda *et al.*, "Colony fingerprint-based discrimination of *Staphylococcus* species with machine learning approaches," *Sensors (Basel)*, vol. 18, no. 9, p. 2789, 2018.
- [25] K. P. Smith, D. L. Richmond, T. Brennan-Krohn, H. L. Elliott, and J. E. Kirby, "Development of MAST: A microscopy-based antimicrobial susceptibility testing platform," *SLAS Technol.*, vol. 22, no. 6, pp. 662–674, 2017.
- [26] L. Lechowicz, M. Urbaniak, W. Adamus-Bialek, and W. Kaca, "The use of infrared spectroscopy and artificial neural networks for detection of uropathogenic *Escherichia coli* strains' susceptibility to cephalothin," *Acta Biochim. Pol.*, vol. 60, no. 4, pp. 713–718, 2013.
- [27] D. Moradigaravand, M. Palm, A. Farewell, V. Mustonen, J. Warringer, and L. Parts, "Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data," *PLoS Comput. Biol.*, vol. 14, no. 12, p. e1006258, 2018.
- [28] G. Arango-Argoty, E. Garner, A. Pruden, L. S. Heath, P. Vikesland, and L. Zhang, "DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data," *Microbiome*, vol. 6, no. 1, 2018.
- [29] Y. Li *et al.*, "HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes," *Microbiome*, vol. 9, no. 1, p. 40, 2021.
- [30] A. S. Chowdhury, D. R. Call, and S. L. Broschat, "PARGT: a software tool for predicting antimicrobial resistance in bacteria," *Sci. Rep.*, vol. 10, no. 1, p. 11033, 2020.





ASHISH KUKRETI &lt;ashkukreti2000@gmail.com&gt;

---

**IEEE - ICACCS 2022 - Acceptance Letter - Reg**

---

ICACCS 2022 &lt;icaccs@sece.ac.in&gt;

Mon, Feb 28, 2022 at 11:08 PM

To: ashkukreti2000@gmail.com, yashahasija06@gmail.com

**Dear Author(s),****Greetings from ICACCS 2022.****2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS).****Organized by Sri Eshwar College of Engineering, Coimbatore, Tamilnadu, India.**

We are very happy to inform you that your paper entitled “ **Machine Learning: A promising in-silico approach to curb antimicrobial resistance** ” with Paper-ID: **ICACCS\_2022\_paper\_354** reviewed by the Technical Committee and recommended for presentation in the conference. The suggestions/comments made by the Technical Committee are listed at the end of this acceptance.

Kindly make all the necessary corrections as per their remarks before submitting camera ready paper. This is mandatory to include your paper in the proceedings. Formatting your paper with the below procedures and **keeping your paper without plagiarism, will increase the chance of publishing the paper in IEEE Digital Library publications.**

The "camera-ready" PDF paper can be created with IEEE PDF Express (**Conference-ID: 54159X**), (<https://ieee-pdf-express.org/>). All papers should strictly comply with IEEE Conference paper format attached with this mail and send the soft copy of camera ready paper both in PDF and WORD format to [icaccs2022@gmail.com](mailto:icaccs2022@gmail.com). **Kindly indicate your paper ID all the time in the subject line** whenever you send mail to us.

**Kindly upload your Camera Ready Documents in [icaccs.sece.ac.in](http://icaccs.sece.ac.in) - > Camera Ready Submission link.**

**The following guidelines are to be adhered to complete the registration process.**

1. Before sending the PDF file generated through PDF express please ensure the format generated is in proper format and according to IEEE standard, otherwise try till proper format achieved. (Procedure is attached with this mail)

**Note : IEEE Copyright notice will be updated by the conference publication Team, Authors no need to add the same**

2. The Conference registration kindly visit our conference website [icaccs.sece.ac.in](http://icaccs.sece.ac.in)

3. If the no. of pages of your paper exceeds 6 (SIX), Rs.500 will be charged for an extra additional page, which should be included in your payment.

4. A Scanned copy of completed and signed **IEEE copyright form must be sent to [icaccs2022@gmail.com](mailto:icaccs2022@gmail.com)** along with your Camera ready paper (**required for IEEE Digital Library Publication**).

5. All the accepted papers which are registered for **IEEE ICACCS 2022 will be published in the IEEE conference proceedings with an ISBN number (assigned by IEEE, USA)** and also submitted for inclusion in **IEEE Digital Library, with author's concern.**

You have to complete the registration process as per the steps given in our website [icaccs.sece.ac.in](http://icaccs.sece.ac.in)

**Review Reports**

<https://ijaccs.com/ICACCS2022/index.html>

**Registration Link**

<http://icaccs.sece.ac.in/Registration>

**Camera Ready Link**

<https://www.emailmeform.com/builder/form/MkflI3sm6ORWz>

**General Editor Comments**

- **Kindly verify the attached Plagiarism Report and improve your paper during Camera ready Submission.**
- **We can accept plagiarism upto 15% during final submission. ( only 3% Single source plagiarism recommended)**
  - **(Note: With reference 20%, without reference to 15% will be accepted)**

**Editor Comments for final Manuscript preparation :**

1. Highlight the major contributions
2. Sections must be organized under the introduction.
3. Complete proofreading recommended
4. Image quality needs to improve.
5. Check the title and special characters
6. Remove unwanted/unused citations from your manuscript.
7. Add 6 to 10 keywords and organize sections under section 1.

Follow the proper IEEE Paper format.

**\*\*Detailed reviewers comments are available with the author terminal system/review report system.**

**NOTE:**

- (1) If you have more than one paper, you need to repeat the registration procedure.
- (2) If one of the authors registers for the paper presentation and if any of the other **co-authors** wish to participate, they should pay the participation fees.
- (3) The paid registration fee is non-refundable.
- (4) One regular registration is within Six Pages including all figures, tables, and references. Extra pages will be charged.
- (5) Filled registration form is one of the mandate requirements from the authors.

**Authors' Registration Fee includes**

1. Technical Sessions (Hybrid/Virtual/Online)
2. 10 Minutes oral presentation (Q&A included)
3. Conference Soft Proceedings. ( Book format)
4. Softcopy of the Certificates.

**Note:**

- **IEEE Xplore accepts upto 20% plagiarism during your final submission, authors are requested to ensure the same during your camera ready submission. IEEE Recommend Cross Check / Turnitin for plagiarism verification.**
- **Registration Mapped with the first author mentioned in the paper. Authors need to pay the first author fee (Either Student or Faculty) during registration.**
- **Foreign (Indian) authors need to pay USD payments mentioned in the registration form.**






**Any Queries Contact/ WhatsApp: 9486137910 or Email to: [icaccs@sece.ac.in](mailto:icaccs@sece.ac.in) / [anandakumar.h@sece.ac.in](mailto:anandakumar.h@sece.ac.in)**

Thanks and Regards

**ICACCS 2022  
Organizing Committee**

---

**5 attachments**

-  **IEEE\_Paper\_Format\_ICACCS 2022.doc**  
59K
-  **54159X.PDF**  
100K
-  **ICACCS 2022 - IEEE - Copyright form.pdf**  
106K
-  **Acceptance Letter - ICACCS 2022.pdf**  
255K
-  **ICACCS\_2022\_paper\_354.pdf**  
1884K





### Acceptance Letter

To

Dr./Ms./Mr. Ashish Kukreti and Yasha Hasija

Paper ID: ICACCS\_2022\_paper\_354

Dear Sir/Madam,

**Sub: Acceptance Letter – IEEE 2022 8<sup>th</sup> International Conference on Advanced Computing and Communication Systems (ICACCS) 25<sup>th</sup> – 26<sup>th</sup> March 2022 Technically Sponsored by IEEE and IEEE Madras Section.**

The organizing Committee is pleased to inform you that the peer- reviewed and refereed conference paper titled as “**Machine Learning: A promising in-silico approach to curb antimicrobial resistance**”, has been conditionally accepted for Hybrid (Oral/Virtual) presentation at the ICACCS 2022 conference on 25<sup>th</sup> – 26<sup>th</sup> March 2022.

We would like to kindly invite you to register for the conference on or before 07.03.2022 and present the paper at the conference venue in Coimbatore. On behalf of the organizing committee, I would like to congratulate you.

**Note: Authors can present their paper through virtual / video conferencing.**

**Dr. H. Anandakumar**  
Conference Chair – ICACCS 2022

Sri Eshwar College of Engineering (Autonomous), Coimbatore, Tamil Nadu, India



# Your Registration for ICACCS 2022!



**Organizer:**  
**icaccs 2022**

**Hi Ashish Kukreti,**

Thank you for purchasing 1 ticket(s) for **ICACCS 2022**.

Your Registration details are as follows:


Booking Id : **7669386**

Booking Date : **03/15/2022 10:41(IST)**


<b>NAME</b>	<b>TYPE</b>	<b>PRICE</b>
Ashish Kukreti	Indian Authors-Students(Non-IEEE Members)	INR 6500.00



**Event Details :**

 Organizer : **icaccs 2022**

 Venue : **Sri Eshwar College of Engineering, Coimbatore, India**

 Event Link : **icaccs2022**

(\*Note : Please bring the printout of this email to the event *OR* show this on your smart phone at event venue)



2022

Event by & at :



**Sri Eshwar**  
College of Engineering  
An Autonomous Institution

Coimbatore, India



8<sup>th</sup>

2022

International Conference on

**Advanced Computing &  
Communication Systems**

TECHNICAL SPONSORS



# Certificate of Presentation

Certify that

**Ashish Kukreti**

**Delhi Technological University, Delhi, India.**

has presented a paper in the International Conference on  
**Advanced Computing & Communication Systems - ICACCS 2022**  
on 25<sup>th</sup> & 26<sup>th</sup> March 2022 at Sri Eshwar College of Engineering,  
Coimbatore, TamilNadu, India.

Paper Title :

**Machine Learning: A promising in-silico approach to curb  
antimicrobial resistance**

**Dr. H. Anandakumar**  
Conference Chair

**Dr. R. Subha**  
Convener

**Dr. Sudha Mohanram**  
Patron





2022

Event by & at :



**Sri Eshwar**  
College of Engineering  
An Autonomous Institution

Coimbatore, India



8<sup>th</sup>

2022

International Conference on

**Advanced Computing &  
Communication Systems**

TECHNICAL SPONSORS



# Certificate of Presentation

Certify that

**Yasha Hasija**

**Delhi Technological University, Delhi, India.**

has presented a paper in the International Conference on  
**Advanced Computing & Communication Systems - ICACCS 2022**  
on 25<sup>th</sup> & 26<sup>th</sup> March 2022 at Sri Eshwar College of Engineering,  
Coimbatore, TamilNadu, India.

Paper Title :

**Machine Learning: A promising in-silico approach to curb  
antimicrobial resistance**

**Dr. H. Anandakumar**  
Conference Chair

**Dr. R. Subha**  
Convener

**Dr. Sudha Mohanram**  
Patron





## DEPARTMENT OF BIOTECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY (Formerly Delhi College of Engineering)  
Bawana Road, Delhi - 110042

### CANDIDATE'S DECLARATION

I Ashish Kukreti, Roll Number: 2K20/MSCBIO/02, student of M.Sc. Biotechnology, hereby declare that the work which is presented in the Major Project entitled — **Prediction of small antimicrobial peptides using Machine learning** in the fulfilment of the requirement for the award of the degree of Master of Science in Biotechnology and submitted to the Department of Biotechnology, Delhi Technological University, Delhi, is an authentic record of my own carried out during the period from January- May 2022, under the supervision of Prof. Yasha Hasija.

The matter presented in this report has not been submitted by me for the award for any other degree of this or any other Institute/University. The work has been accepted in SCI/SCI expanded /SSCI/Scopus Indexed Journal OR peer reviewed Scopus Index Conference with the following details:

**Title of the Paper:** Machine Learning: A promising in-silico approach to curb antimicrobial resistance

**Author Names:** Kukreti, Ashish and Hasija, Yasha

**Name of Conference:** ICACCS 2022 (2022 8th International Conference on Advanced Computing and Communication Systems)-IEEE conference

**Conference Date and Venue:** 25-26 March 2022 at Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu, India.

**Registration:** Done

**Status of Paper:** Acceptance Received

**Date of Paper Communication:** 24th February 2022

**Date of Paper Acceptance:** 28th February 2022

**Date of Paper Publication:** NA

Date: 06/05/2022

Ashish Kukreti

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

**CERTIFICATE**

I hereby certify that the Project dissertation titled "Prediction of small antimicrobial peptides using Machine learning" which is submitted by Ashish Kukreti, 2K20/MSCBIO/02, Department of Biotechnology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Science, is a record for the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date: 06/05/2022



**Prof. Yasha Hasija**

**(SUPERVISOR)**

Professor

Department of Biotechnology

Delhi Technological University



**Prof. Pravir Kumar**

**Head of Department**

Department of Biotechnology

Delhi Technological University

## Acknowledgement

I would like to express my gratitude towards my supervisor, Prof. Yasha Hasija, for giving me the opportunity to do research and providing invaluable guidance throughout this research. Her dynamism, vision, sincerity, and motivation have deeply inspired me. She has motivated to carry out the research and to present my work works as clearly as possible. It was a great privilege and honour to work and study under her guidance. I am extremely grateful for what she has offered me. Her insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I am equally grateful and wish to express my wholehearted thanks to respected lab senior Mr. Rajkumar for their kind support and help in the course of my research work.

I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future.

I would also like the institution Delhi Technological University, Delhi for giving me the opportunities throughout the tenure of study.

Finally, my thanks go to all the people who have supported me to complete the research work directly or indirectly.

  
Ashish Kukreti

2K20/MSCBIO/02