Section: Research Paper
ISSN 2063-5346

Prediction of Antibacterial Properties of Piperazine Molecules using Machine Learning approach

Manju N, Aakarsh Siwani A, Nagendra Prasad H S, Shruthi K, Ananya M V, and Nidhi S C
JSS Science and Technology University, Mysuru, India

Corresponding Author: manjun007@jssstuniv.in

Abstract: The use of machine learning in the design of novel molecules with diverse biological properties has gained significant attention in recent years. Multidrug-resistant bacterial infections continue to pose a major challenge, leading to substantial healthcare costs, prolonged hospital stays and significant loss of lives. The urgent need to discover new antibiotics has prompted the exploration of alternative approaches and machine learning has emerged as a promising tool in the pharmaceutical and biotechnology sectors. While various computational methods and tools have been developed and are currently employed, there is still ample room for improvement and increased accessibility to these technologies in different stages of the drug discovery process. This work aims to address these gaps by refining computational methods, enhancing tools and fostering wider utilization of machine learning in drug discovery, thereby contributing to the development of effective antibiotics and tackling the issue of antibiotic resistance.
Keywords: Machine Learning, Decision Tree, Random Forest, Gaussian Naive Bayes, Anti-bacterial properties.

1 INTRODUCTION

The field of drug development and research is a difficult, time-consuming process that is essential to enhancing human health and wellbeing. The traditional approach to drug discovery involves synthesizing and testing a vast number of chemical compounds for their potential therapeutic activity. However, this process is time-consuming, expensive, and often results in a high rate of failure. In recent years, computational models considered promising tools in the realm of drug development, together with machine learning techniques. By speeding the identification and development of novel therapeutic compounds, models have the potential to completely transform the process of discovering drugs. By leveraging computational power and analyzing large datasets, these models can provide valuable insights and predictions about the biological activity and properties of potential drug candidates.

One of the significant advantages of computational models is their ability to save resources and research time. Instead of relying solely on costly experimental assays, researchers can use computational models to screen and prioritize compounds based on their predicted activity, allowing for a more targeted and efficient experimental validation process. This can lead to significant cost savings and reduce the time required to bring a new drug to market.

Furthermore, computational models enable the analysis of vast amounts of data simultaneously, facilitating the identification of relationships and patterns that may not be apparent through traditional experimental approaches alone. By integrating diverse data sources, such as chemical structures, genomic information, and clinical data, researchers can discover insightful information on the mechanisms of action, possible side effects, and interactions of therapeutic candidates. Foreseeing the biological activity of antimicrobial agents, cheminformatic models have been created and their limitations have prevented substantial changes to the drug discovery people. However, recent improvements in machine learning algorithms, particularly neural network-based molecular representations, have shown promise in improving the accuracy and applicability of these models.

The work explores recent developments in machine learning applications for drug discovery and development. By embracing these technological advancements, the opportunity to reshape the paradigm of discovery of drugs and bring about transformative advancements in the field of medicine. Overall, the

integration of computational models and machine learning algorithms in drug discovery holds tremendous potential to enhance the efficiency, effectiveness, and success rates of the drug development process. By combining the power of artificial intelligence with traditional experimental approaches, we can accelerate the discovery of new therapeutic agents, ultimately benefiting patients worldwide.

In the search for new medicines, computer models may prove to be a highly helpful tool. First, these models have the potential to reduce resource and research time costs. Additionally, by building associations between the data, it is feasible to simultaneously evaluate hundreds of data points and draw insightful conclusions. Many chemoinformatic models have been created to find antimicrobial drugs that work against various microorganisms however, they are only capable of predicting their biological activity in a specific strain under specific circumstances [1]. Examples of contemporary machine learning applications breakthroughs are included in review, with descriptions of how they might affect various steps in the drug discovery and development flow diagram. [2]. Analytical exploration is not a novel concept in drug design. Models for the prediction of molecular properties have been established during decades of chemoinformatic development [3]. The accuracy of these models hasn't, however, been high enough to significantly alter the conventional discovery of new drugs. The use of recent algorithms improves in modeling molecular representations based on neural networks, we are starting to have chance to change the approach to developing drugs.

When medications are tested on animals using the in-vivo testing method, more animals have reportedly died, according to a report compiled by the National Center for Biotechnology Information (NCBI). We worked on the in-silico technique, which is a computational approach to determine if medicinal compounds are active or not using numerous machine learning algorithms, to stop or reduce animal death.

- It reduces time complexity and use of resources.
- It is a new way to check the Antibacterial activity of molecules.
- The Efficiency of the drug is known using Machine learning to save man power.

1.1 Importance of Machine Learning

The discovery of new medicines to combat the rising problem of resistance of antibiotic is a time-consuming process that can span several decades. While machine learning methods have shown promise in predicting molecular properties and aiding in the development of antibiotics, current solutions often rely on large datasets and structural similarities to existing antibiotics. There is a need to overcome challenges in modeling unconventional antibiotic classes, which are gaining increased research focus. The goal of this work is to simplify and streamline the process by using machine learning models to generate scores for the antibacterial activity of molecules. By addressing these challenges, the work aims to contribute to the development of new medicines and effectively tackle the issue of antibiotic resistance.

1.2 Challenges

The Challenges faced during the modelling of Anti-Bacterial activity with Machine Learning algorithms are:

- Need for extraction of parameters of the molecules from SwissADME manually.
- The criteria of spectrum, selectivity, functionality, and essentiality should be met by antimicrobial targets.
- Desirable properties of antimicrobial targets.
- Provide adequate spectrum and selectivity.
- Designing assays and high-throughput screens requires an understanding of the function.
- All drug discovery faces numerous difficulties in candidate selection and subsequent development of antibiotics. They are barely mentioned in this study because they have been tackled, dealt with, and overcame by more conventional medicinal chemistry magic for many generations of successful antibiotics and other human health medications.
- The error provided in the label in the dataset had to be removed since string values cannot be used for modeling.

Section: Research Paper
ISSN 2063-5346

1.3 Process of Machine Learning Model

Process of Modelling of Anti-Bacterial activity with Machine Learning is to Develop an in-silico computational method using machine learning models to assess the activity of chemical compounds, with the goal of replacing animal testing and reducing costs while maintaining accuracy. Train machine learning models on the physicochemical properties of drug molecules to create computational models capable of predicting the activity of the molecules. Aim for high accuracy in the predictions to ensure reliable results. Evaluate the performance and accuracy of the developed machine learning models by employing a confusion matrix that includes known toxins and commonly used constituents of health products. This evaluation will demonstrate the effectiveness and reliability of the proposed models in accurately classifying chemical compounds as active or inactive. Continuously refine and optimize the ML models to improve their accuracy and performance, ensuring their suitability for use in assessing chemical activity.

2. LITERATURE SURVEY

There is a rising need to find new antibiotics since antibiotic-resistant bacteria are emerging quickly. Halicin is a drug repurposing hub molecule that is structurally different from conventional antibiotics and exhibits bactericidal activity against a wide phylogenetic spectrum of pathogens, including Mycobacterium tuberculosis and carbapenem-resistant Enterobacteriaceae. This was discovered through predictions on multiple chemical libraries [4]. An examination of the structural connections between these substances, ZINC15 molecules with prediction scores greater than 0.9, the primary training set, the Drug Repurposing Hub, and the WuXi anti-tuberculosis library. Intriguingly, analysis showed [5] that the molecules in the WuXi anti-tuberculosis library mostly occupied a different chemical space from substances with antibacterial activity, which is consistent with the findings that even the highest predicted concentrations of these couldn't stop the growth of E. coli. Use of SVMs in a QSAR study of transcription factors activator protein (AP)-1 and nuclear factor (NF)-kB by ethyl 2-[(3-methyl-2,5-dioxo(3-pyrrolidinyl)) amino] -4-(trifluoromethyl) pyrimidine-5-carboxylate derivatives [6].

Work from [7] indicates that the urease enzyme is crucial for the infamously cancerous H. pylori bacterium to colonize the human stomach [8]. An acid neutral urease that is constitutively expressed by the pathogen is crucial for infection establishment. However, current data point to an increase in strains that are resistant to clarithromycin and metronidazole. Perhaps as a result of this, a small percentage of infected people received regular triple therapy but infection persisted. To treat these people, alternative therapeutic approaches are needed. H. pylori urease is a crucial component of the formation of infection, making it a prospective therapeutic target.

The activity's numerical quantities were dissociated [9] as they were discovered. Following unidentified descriptor filtering, models have been created. For all models, the cross-validated balanced accuracy results ranged from 73% to 83%. The test sets' balanced accuracy for the same models was also calculated, and significant values in the range of 77% to 89% were discovered. The consensus model was calculated to determine which of the compounds to be synthesized had the best anti-urease activity, as specified for the highest productivity.

Each data set originally contained millions of pharmacophore models, and the data set generated the most pharmacophore models. The largest conformational flexibility is observed for the molecules in the data set, which is consistent with this observation. Then, significance analysis was done on the pharmacophore models that were kept. The threshold values were chosen at 100 evenly spaced intervals from 0 to the point where the ranking scores and reference scores differed from each other by the most. The number of falsely significant pharmacophore models declines while the number of actually significant models essentially holds constant when the threshold value rises in a bottom-up fashion. Therefore, as the number of mistakenly significant pharmacophore models declines to zero, the ideal threshold values (*) for each data set may be found. The best subsets of the pharmacophore fingerprint bits were subsequently discovered for four data sets. The fingerprint bits that were important for classification made up a very modest percentage of the total [10]. Ronak Y. Patel [11] founded it in 2012 with the intention of predicting bioactive conformers using a supervised learning technique known as multiple-instance learning. A molecule is considered inactive if none of its conformers are accountable

for the observed bioactivity. Instead, a single molecule, treated as a bag of conformers, is biologically active if and only if at least one of its conformers is viewed as an instance.

3 Methodology

Figure 1 shows the block diagram of proposed method which depicts the process of the activities carried out in our experimentation.

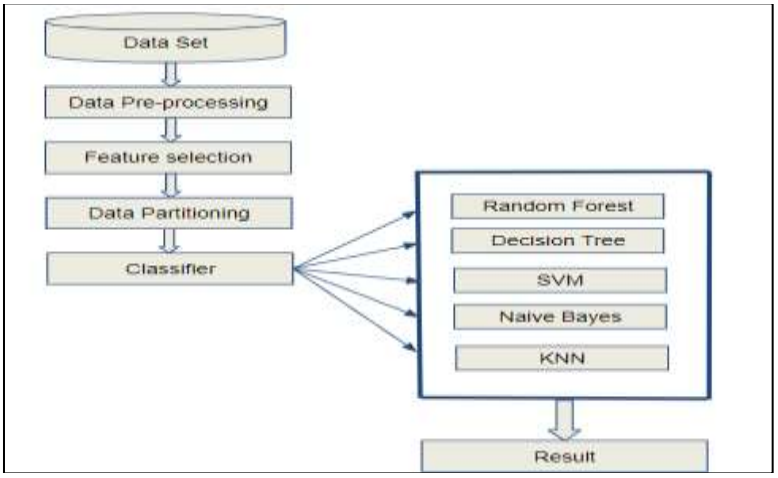

Figure 1: Block Diagram of proposed method

We employ machine learning technology to forecast drug activity, and we created a prediction pipeline that makes use of machine learning for activity prediction possible. This prediction pipeline was created for the dataset and applies machine learning techniques to it in order to draw a conclusion based on the findings that were predicted.

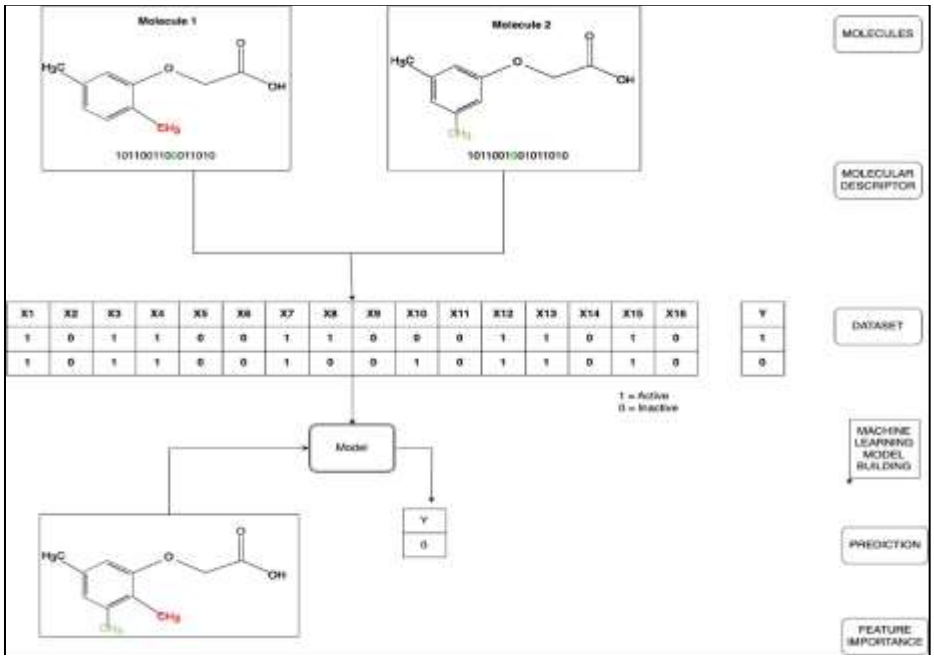

Figure 2: Drug activity prediction using Machine Learning Technology

Drug activity prediction using Machine Learning as shown in figure 2 is a process that begins with the selection of molecular descriptors that characterize chemical compounds. These descriptors are used to

create a dataset containing information about various compounds and their corresponding activities or properties. Machine learning models are then built using this dataset, with the goal of learning patterns and relationships between molecular features and drug activities. Once the models are trained, they can be employed to predict the activity or behavior of new, untested compounds, aiding in drug discovery, toxicity assessment, and therapeutic development. This process plays a vital role in streamlining and accelerating the drug development pipeline by prioritizing promising compounds for further experimentation and reducing the need for time-consuming and costly laboratory tests.

3.1 DATASET

In the dataset, 86 pharmacological compounds are present. There are 47 active drug molecules and 38 inactive drug molecules in the collection, which has 86 drug molecules in total. Each molecule has seven characteristics, including physical traits as shown in figure 3. Molecular Descriptors are another name for these physical characteristics. Consequently, Molecular Descriptors categorize pharmacological molecules according to their physical characteristics. The dataset's activity is separated into binary classes. The first class is active, whereas the second is inactive. There are two levels of drug molecule activity: '0' and '1'. Out of these, '0' is utilized for inactive molecules and '1' is used for inactive medication molecules. In the dataset, we have taken into account 4 different split types among these pharmacological compounds. They are:

- 80% data has been set for training dataset and remaining 20% as a testing dataset.
- 75% data has been set for training dataset and remaining 25% as a testing dataset.
- 70% data has been set for training dataset and remaining 30% as a testing dataset.
- 60% data has been set for training dataset and remaining 40% as a testing dataset.

cLogP	cLogS	H-Acceptors	H-Donors	Total Surface Area	Polar Surface Area	Druglikeness	label
2.3877	-2.245	5	1	290.37	75.85	6.2911	63
2.9937	-2.981	5	1	305.79	75.85	6.2831	61
2.3177	-2.263	6	1	312.63	85.08	6.0736	53
2.042	-1.949	6	2	296.72	96.08	6.1535	47
2.9937	-2.981	5	1	305.79	75.85	6.2831	35
2.7316	-2.589	5	1	302.63	75.85	6.1811	42
2.9937	-2.981	5	1	305.79	75.85	6.2831	45
3.5997	-3.717	5	1	321.21	75.85	6.2831	50
2.4885	-2.559	5	1	296.72	75.85	4.9511	52
3.1129	-3.079	5	1	309	75.85	4.5011	38
2.8248	-3.261	5	1	316.65	75.85	6.6638	41
1.2826	-2.705	6	1	322.45	128.34	5.1404	43
1.991	-2.161	6	1	275.41	60.83	7.9514	57
2.3349	-2.505	6	1	287.67	60.83	7.865	37
1.921	-2.179	7	1	297.67	70.06	7.8466	46
1.6453	-1.865	7	2	281.76	81.06	7.9133	48
2.597	-2.897	6	1	290.83	60.83	7.9533	57
2.597	-2.897	6	1	290.83	60.83	7.9533	55
2.597	-2.897	6	1	290.83	60.83	7.9533	69
3.203	-3.633	6	1	306.25	60.83	7.9533	70
2.0918	-2.475	6	1	281.76	60.83	6.6114	61

Figure 3: Dataset description

3.2 Dataset Feature Selection

Correlated variables, molecular descriptors with numerous zero values, missing values, and noise are all removed from the dataset via feature selection. The process of feature selection helps the model perform better by removing unnecessary and redundant features while also extracting the most important features. According to the discussion, our dataset comprises 7 features, many of which are numerous and require a model's execution to take a long time. We are aware that the selection of features is an important phase in predictive modeling. The steps are as follows:

Step 1: By generating shuffled copies of each feature (referred to as shadow features), it adds unpredictability to the provided data set.

Step 2: It applies a feature importance measure (the default is Mean Decrease Accuracy) to assess the relevance of each feature, with higher being more significant, then trains a random forest classifier on the expanded data set.

Step 3: It constantly eliminates characteristics that are regarded to be of extremely low relevance and determines whether a real feature has a higher importance than the best of its shadow features at each iteration.

3.3 Machine Learning Algorithms

Following are the machine learning algorithms used in our experimentation.

3.1.1 Decision Tree Classifier

The decision tree is the most efficient and popular method for categorization and prediction. It has a tree structure resembling a flowchart, each internal node corresponding to a test on an attribute, each branch a test result, with a class label for each leaf node (terminal node). Using decision trees has the following advantages: Clear rules can be produced through decision trees. Which fields are most important for classification or prediction is made abundantly clear by them. Decision trees can handle both continuous and categorical inputs and can do classification without requiring a lot of processing [12,13].

3.1.2 Random Forest Classifier

A forest is built using a supervised learning algorithm, which also makes it appear random. An ensemble of Decision Trees, frequently trained via the "bagging" approach, makes up the "forest" that it constructs. By selecting a random sample from the training data set as the input, a unique decision tree is created. Only a random sample of predictors is selected at each node of the tree to calculate the split point. The essential advantage of using random forest is that it can be used for both classification and regression issues, which are the main components of most contemporary machine learning systems [14,15,16].

Several decision trees are constructed in a random forest classification utilizing several randomly chosen subsets of the data and attributes. Each decision tree assumes the role of an expert in selecting the appropriate category for the data. Predictions are based on the most typical result after each decision tree's forecast has been calculated.

3.1.3 Support Vector Machine (SVM)

Vector machines are built upon support decision planes, which define decision boundaries. A decision plane is a diagram that makes distinctions between a group of objects with different class memberships. In general, the greater the margin, the smaller the generalization error of the classifier, therefore it makes sense that the hyperplane with the greatest distance from the closest training data points of any class (referred to as the functional margin) achieves a respectable separation [17].

3.1.4 K Nearest Neighbor

The popular machine learning technique known as K-Nearest Neighbors is utilized for both classification and regression problems. It is a non-parametric and instance-based learning method, which means that it bases its predictions not on generalizations about the distribution of the underlying data, but rather on the immediate surroundings of the data points. When a new data point needs to be classified or its target value needs to be predicted, the algorithm searches for the K nearest neighbors of that data point from the training set based on a distance metric. The distance can be calculated using various mathematical formulas depending on the type of data being used. The majority rule among an object's k nearest neighbors, where k is an integer, is used to classify it in this instance-based learning method. In KNN, the labels from the closest nodes are transmitted to the query using a majority-voting method. The data including labeled and unlabeled nodes are represented in a high-dimensional feature space. The number of nearest neighbors taking part in the voting process is indicated by the value k in this case [18].

3.1.2 Gaussian Naive Bayes Classifier

A probabilistic strategy that bases membership predictions on feature independence and the Bayes rule. Naive Bayes classifiers are Bayes' rule-based probabilistic models. Using the prior probability distribution reflecting the relative quantities of labels in training sets, they calculate the probability that a specific piece of data would be accurately attributed to a specific label. The probability attached to each label is conditionally independent when numerous labels are offered.

Section: Research Paper
ISSN 2063-5346

4. RESULTS AND DISCUSSION

In the following section, tables and graphs show the results obtained using various machine learning approaches.

4.1 DECISION TREE CLASSIFIER

Table 1: Results obtained Decision Tree

Split Ratio	Accuracy	Precision	Recall	F1 score
80:20	0.95	0.95	0.95	0.95
75:25	0.96	0.96	0.96	0.96
70:30	0.93	0.93	0.93	0.93
60:40	0.87	0.88	0.88	0.87

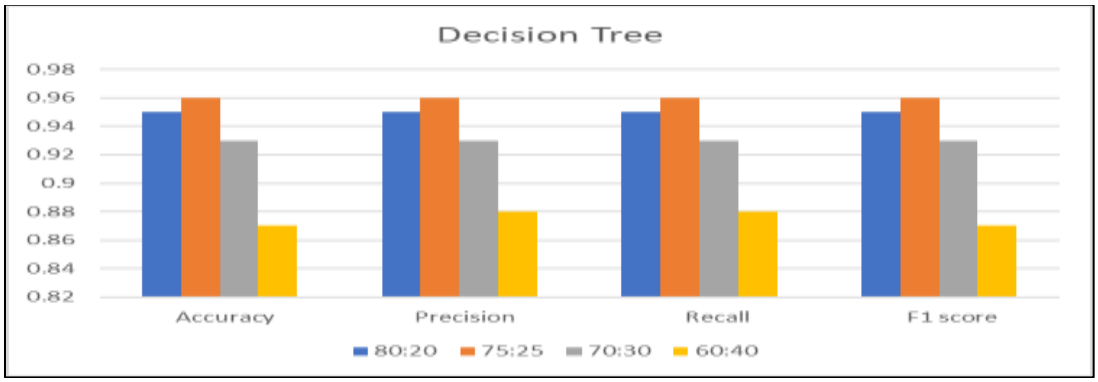

Figure 4: Decision Tree Graph

The Decision Tree classifier produced the highest accuracy in a split ratio of 75:25 with 0.96 accuracy. This is obtained due to the tree constructed in the particular split ratio by the attribute selection measures like Information Gain and Gini Index.

4.2 GAUSSIAN NAIVE BAYES CLASSIFIER

Table 2: Results obtained using Gaussian Naive Bayes Classifier

Split Ratio	Accuracy	Precision	Recall	F1 score
80:20	0.80	0.82	0.80	0.80
75:25	0.80	0.80	0.80	0.80
70:30	0.833	0.84	0.83	0.83
60:40	0.85	0.85	0.85	0.85

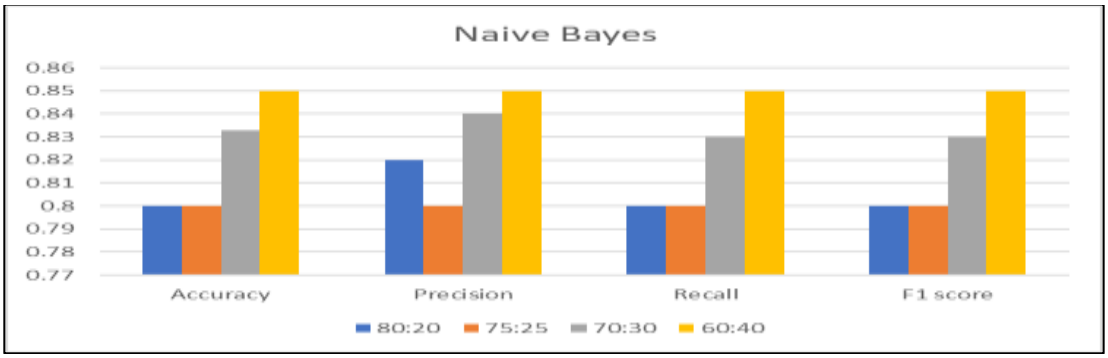

Figure 4: Gaussian Naive Bayes Graph

The Gaussian Naive Bayes classifier produced the highest accuracy in a split ratio of 60:40 due to less number of the molecules taken into consideration. The more trained molecules created more complexity in the classification. When splitting a dataset into training and test sets, the choice of the split ratio can have an impact on the accuracy of the model. In our case, we are using a 60:40 split ratio, with 60% of the data for training and 40% for testing.

4.3 SUPPORT VECTOR MACHINE

Table 3: Results obtained using Support Vector Machine

Split Ratio	Accuracy	Precision	Recall	F1 score
80:20	0.95	0.92	1.0	0.96
75:25	0.96	0.92	1.0	0.96
70:30	0.933	0.87	1.0	0.93
60:40	0.975	1.0	0.95	0.97

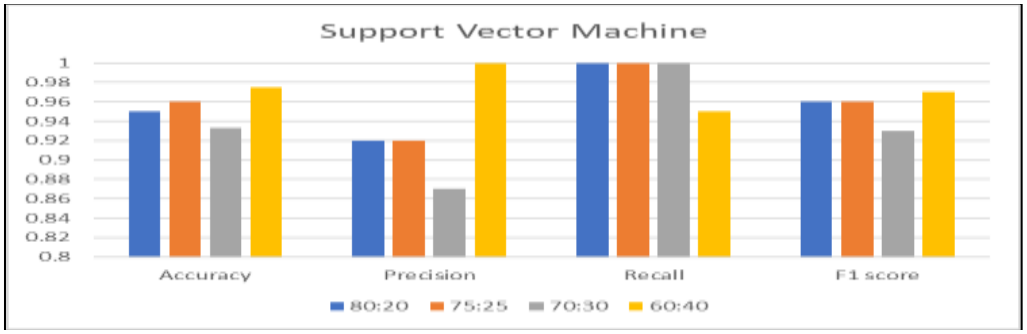

Figure 5: Support Vector Machine Graph

SVM classifiers produced the highest accuracy of 0.97 in a split ratio of 60:40 due to less number of the molecules taken into consideration for training. The more trained molecules created more complexity in the classification. When splitting a dataset into training and test sets, the choice of the split ratio can have an impact on the accuracy of the model. In our case, we are using a 60:40 split ratio, with 60% of the data for training and 40% for testing.

Section: Research Paper
ISSN 2063-5346

4.4 K NEAREST NEIGHBOR

Here **p =1** which means **Manhattan distance** and **p = 2** is **Euclidean distance**

Table 4: Results obtained using K Nearest Neighbour

Accuracy 80:20	K neighbors=3	K neighbors=4	K neighbors=5
p=1	0.85	0.65	0.80
p=2	0.90	0.65	0.75

Accuracy 75:25	K neighbors=3	K neighbors=4	K neighbors=5
p=1	0.84	0.76	0.84
p=2	0.88	0.72	0.84

Accuracy 70:30	K neighbors=3	K neighbors=4	K neighbors=5
p=1	0.93	0.80	0.86
p=2	0.83	0.76	0.86

Accuracy 60:40	K neighbors=3	K neighbors=4	K neighbors=5
p=1	0.85	0.72	0.80
p=2	0.82	0.67	0.85

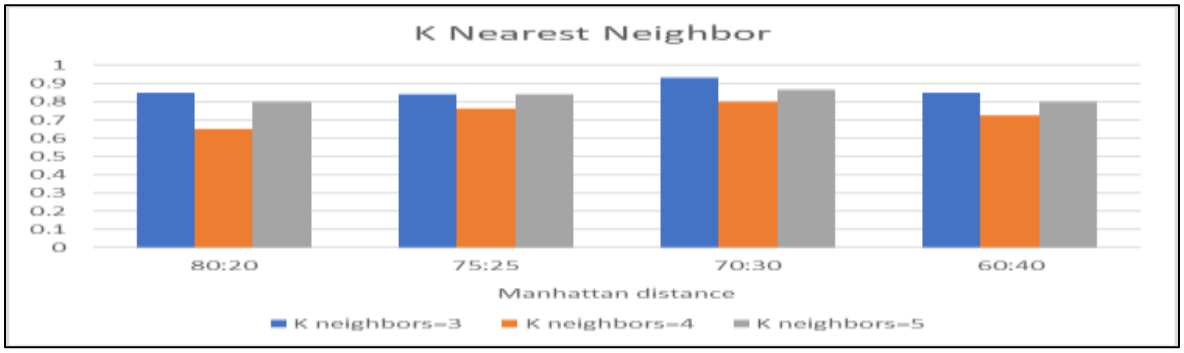

Figure 6: K Nearest Neighbor (Manhattan Distance) Graph

Section: Research Paper
ISSN 2063-5346

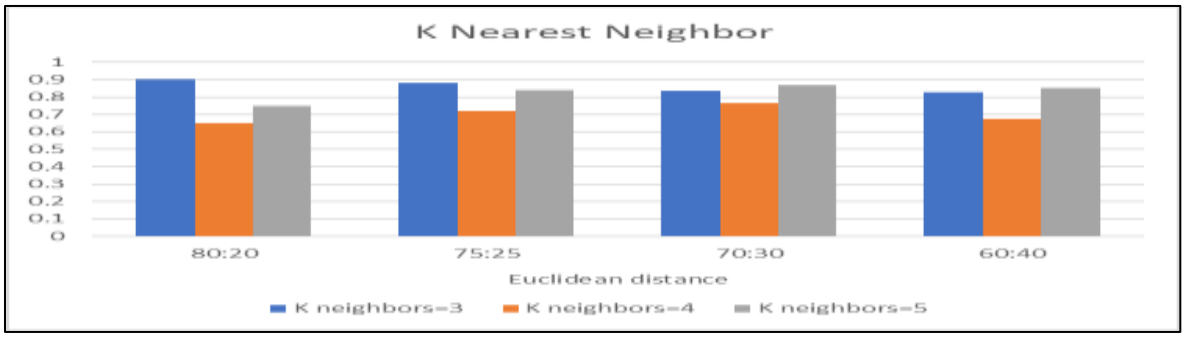

Figure 7: K Nearest Neighbor (Euclidean Distance) Graph

Euclidean distance has produced accuracy of 0.9 at 80:20 split ratio and Manhattan has produced 0.933 accuracy at 70:30 split ratio both with neighbors 3. KNN produced the highest accuracy in a split ratio of 70:30 using Manhattan distance for the n value 3. With 60% of the data allocated for training, the model has access to a substantial amount of information to learn from. With a larger training set, the model can fit more complex patterns, reducing bias.

4.5 RANDOM FOREST CLASSIFIER

Table 5: Results obtained using Random Forest Classifier (N_estimators=100)

Accuracy	max_depth=3	max_depth=4	max_depth=5
max_features=5	0.92	0.88	0.84
max_ features=7	0.92	0.92	0.92

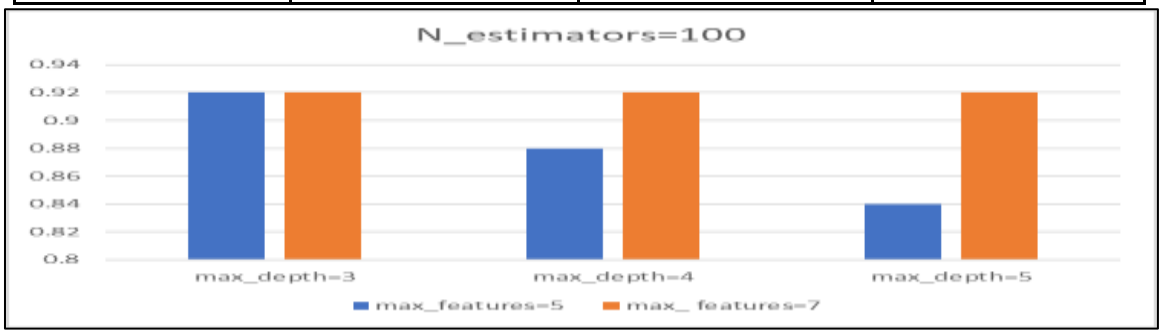

Figure 8: Random Forest Algorithm Graph (N_estimators=100)

Table 6: Results obtained uing Random Forest Classifier (N_estimators=150)

Accuracy	max_depth=3	max_depth=4	max_depth=5
max_features=5	0.88	0.88	0.92
max_features=7	0.92	0.92	0.92

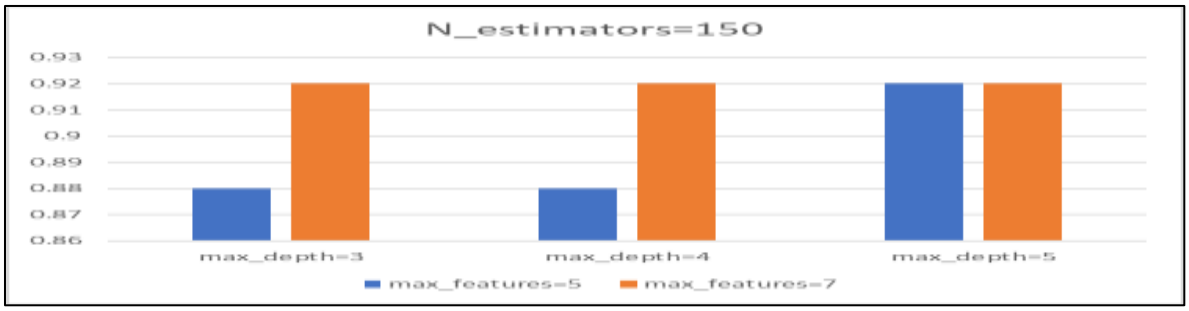

Figure 8: Random Forest Algorithm Graph (N_estimators=150)

Table 7: Results obtained Random Forest Classifier (N_estimators=200)

Accuracy	max_depth=3	max_depth=4	max_depth=5
max_features=5	0.92	0.88	0.88
max_features=7	0.92	0.92	0.92

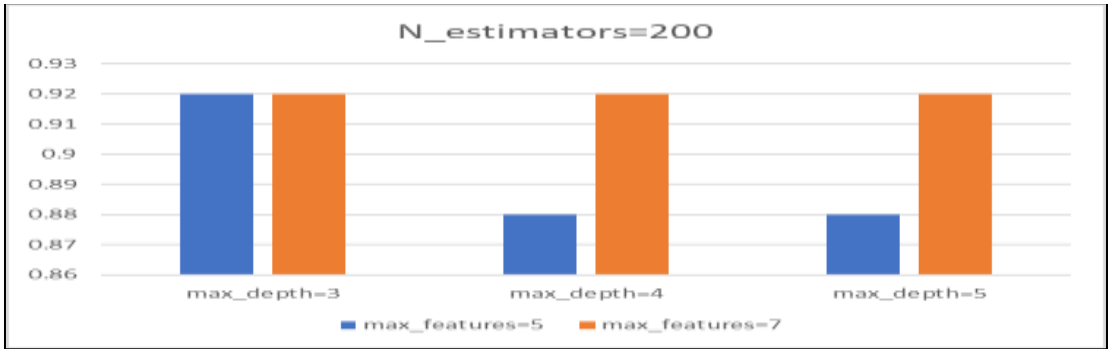

Figure 9: Random Forest Algorithm Graph (N_estimators=200)

Compared to the table 5, table 6 and table 7, the best accuracy in Random Forest Classifier for estimators = 200 and depth=3.

4.6 COMPARATIVE ANALYSIS

Table 8: Comparison of Results of all Algorithms

	SVM	RF	DT	NB	KNN
80:20	0.95	0.93	0.85	0.80	0.80
75:25	0.96	0.91	0.92	0.80	0.80
70:30	0.93	0.88	0.90	0.83	0.86
60:40	0.97	0.92	0.87	0.85	0.85

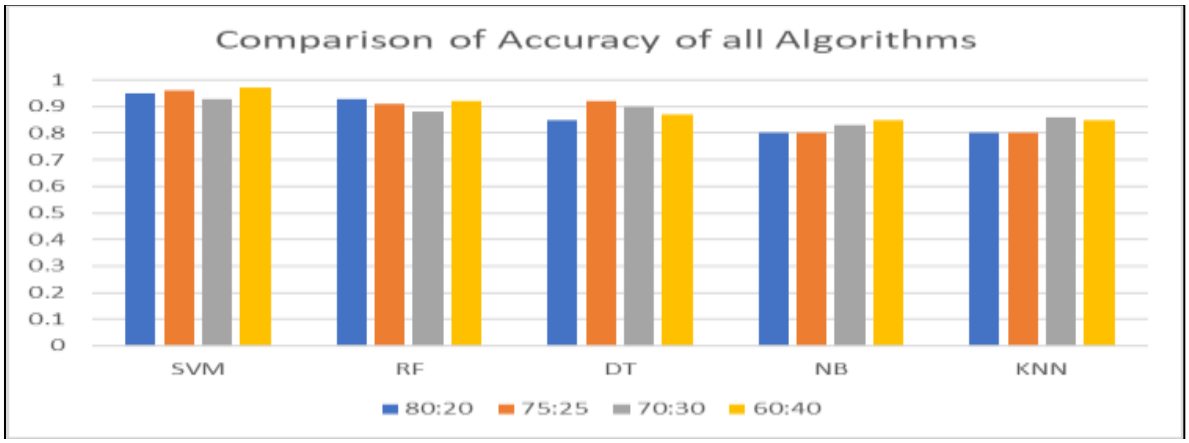

Figure 10: Comparison of Accuracy of all Algorithms

In light of the comparative study's findings and evaluation criteria, it can be said that the SVM algorithm for a 60:40 split ratio outperforms all others. The best option for this project is due to its high accuracy and excellent performance in other evaluation criteria. In terms of accuracy, precision, and recall, Support Vector Machines (SVM) clearly surpass other techniques. In comparison to other algorithms, the SVM algorithm consistently produced greater accuracy rates and showed superior performance in correctly identifying the data points. This demonstrates SVM's effectiveness in managing the dataset's complexity and its potential to produce solid and trustworthy results. As a result, SVM is the suggested option for the particular assignment because it offers better performance and prediction abilities. It allows for better differentiation between active and inactive pharmacological compounds due to its capacity to manage non-linear connections and record complicated decision boundaries. SVM's enhanced performance was also aided by its resistance to noisy and unbalanced datasets.

Upon evaluating the performance of multiple machine learning algorithms, including Decision Trees (DT), Naive Bayes (NB), Random Forest (RF), and k-Nearest Neighbors (KNN), it is observed that these algorithms did not achieve better accuracy than Support Vector Machines (SVM) in predicting the Antibacterial activity of molecules. One possible reason for the lower performance of DT, NB, RF, and KNN could be their inability to effectively capture complex relationships and decision boundaries within the dataset. DT and RF, being tree-based algorithms, may struggle with non-linear patterns, leading to less accurate predictions. NB, while simple and computationally efficient, assumes independence between features, which might not hold in the dataset, thus limiting its accuracy. KNN reliance on local neighborhoods may not be optimal for the specific dataset, resulting in suboptimal performance. In contrast, SVM demonstrated superior performance due to its ability to handle non-linear relationships and capture intricate decision boundaries. By leveraging kernel functions, SVM effectively mapped the data into higher-dimensional spaces, enabling better separation of active and inactive drug molecules. Furthermore, the regularization capability of SVM helped prevent overfitting, leading to improved generalization performance on unseen data. The optimization of hyperparameters also contributed to fine-tuning the SVM model, enhancing its accuracy.

In summary, the lower accuracy of DT, NB, RF, and KNN compared to SVM in predicting Antibacterial activity can be attributed to their limitations in capturing complex relationships, non-linear patterns, and the assumptions made by these algorithms. SVM's capability to handle these challenges, along with its regularization and hyperparameter optimization, resulted in superior accuracy and made it the preferred choice for this particular task.

5 CONCLUSION

It was discovered after evaluating the performance of several machine learning algorithms, including Decision Trees (DT), Naive Bayes (NB), Random Forest (RF), and k-Nearest Neighbors (KNN), that these algorithms did not produce predictions of molecule antibacterial activity that were more accurate than Support Vector Machines.

The superiority of SVM can be attributed to its ability to handle non-linear relationships and capture complex decision boundaries, making it more effective in distinguishing between active and inactive drug molecules. Furthermore, SVM demonstrates robustness to noisy and imbalanced datasets, which are common challenges in this domain. The regularization capability of SVM helps prevent overfitting, allowing for better generalization to unseen data. By optimizing hyperparameters, such as the choice of kernel function and regularization parameter, SVM's predictive power can be further enhanced.

Based on the findings of this study, it is recommended to use SVM as the preferred algorithm for predicting antibacterial activity. Its higher accuracy, discrimination capabilities, robustness to various dataset characteristics, and ability to handle non-linear relationships make it the most suitable choice for this specific application. This research provides valuable insights into the performance of different machine learning algorithms and highlights the significance of selecting the appropriate algorithm for accurate antibacterial activity prediction.

Overall, SVM showed that it was appropriate for the Antibacterial activity prediction challenge by outperforming other algorithms in terms of accuracy and discrimination, the use of SVM in antibacterial activity prediction shows great promise and can significantly contribute to drug discovery and development efforts. As a result, SVM is suggested as the ideal option for this particular application.

5.1 FUTURE SCOPE

Machine learning models have a bright future in predicting antibacterial activity and assisting in medication discovery.

- **Enhanced Accuracy:** Ongoing improvements to machine learning techniques like deep learning and reinforcement learning may help make predictions even more accurate and dependable. As a result, there would be a decrease in false positives and false negatives and a more accurate identification of prospective antibiotic compounds.
- **Virtual Screening and Virtual Clinical Trials:** Machine learning models can be utilized for virtual screening of large chemical libraries to identify potential drug candidates with desired properties. This can significantly reduce the time and cost associated with experimental screening. Additionally, virtual clinical trials, where models simulate the effects of drugs on a virtual patient population, can help predict drug responses, optimize dosages, and identify potential adverse effects.
- **Personalized Medicine:** Machine learning models can be used in personalized medicine, where the choice of antibiotics is based on a person's unique genetic profile and microbiological profile. Machine learning algorithms can help anticipate the effectiveness and potential negative effects of antibiotics for individualized treatment programmes by analyzing large-scale genomes and microbiome data.
- **Continuous Research and Collaboration:** The scope of machine learning in the field of drug discovery and predicting antibacterial activity is vast. Advancements in algorithms, data integration, personalized medicine, and virtual screening can revolutionize the drug discovery process, leading to the development of more effective antibiotics and personalized treatment strategies. Continued research and collaboration between experts in the fields of machine learning, biology, and medicine will be crucial in realizing this potential.

REFERENCES

[1] Speck-Planche, A.; Kleandrova, V. V.; Cordeiro, M. N. D. S. Cheminformatics for rational discovery of safe antibacterial drugs: Simultaneous predictions of biological activity against streptococci and toxicological profiles in laboratory animals. Bioorg. Med. Chem. 2013, 21, 2727−2732.

[2] Jane Panteleev, Hua Gao, Lei Jia, Recent applications of machine learning in medicinal chemistry, Bioorganic & Medicinal Chemistry Letters, Volume 28, Issue 17, 2018.

[3] Mayr et al., 2018; Wu et al., 2017

[4] Machine learning in chemoinformatics and drug discovery Yu-Chen Lo, Stefano E. Rensi, Wen Torng and Russ B. Altman
[5] A deep learning approach to antibiotic discovery.Jonathan M. Stokes, Kevin Yang
[6] Liu, H. et al. (2003) QSAR study of ethyl 2-[(3-methyl-2, 5-dioxo (3-pyrrolidinyl)) amino]-4-(trifluoromethyl) pyrimidine-5-carboxylate: an inhibitor of AP-1 and NF-kB mediated gene expression based on support vector machines. J. Chem. Inf. Comput. Sci. 43, 1288–1296
[7] Design, synthesize and anti urease activity of novel thiazole derivatives: Machine learning, molecular docking and biological investigation Arif Mermer
[8] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre and A. Jemal, Ca-Cancer J. Clin., 2018, 68, 394–424.
[9] Identification of novel bacterial urease inhibitors through molecular shape and structure based virtual screening approaches Muhammad Imran, Saba Waqar
[10] Implementation of multiple-instance learning in drug activity prediction Gang Fu1, Xiaofei Nan2, Haining Liu1, Ronak Y Patel1,4, Pankaj R Daga1, Yixin Chen2*, Dawn E Wilkins2*, Robert J Doerksen1,3*
[11] Fu, Gang & Nan, Xiaofei & Liu, Haining & Patel, Ronak & Daga, Pankaj & Chen, Yixin & Wilkins, Dawn & Doerksen, Robert. (2012). Implementation of Multiple-Instance Learning in Drug Activity Prediction. BMC bioinformatics.
[12] L. Breiman, "Random forests", *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, Oct. 2001.
[13] S. Martínez-Agüero, I. Mora-Jiménez, J. Lérida-García, J. Álvarez-Rodríguez and C. Soguero-Ruiz, "Machine learning techniques to identify antimicrobial resistance in the intensive care unit", *Entropy*, vol. 21, no. 6, pp. 603, Jun. 2019
[14] Svetnik, V. et al. (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 43, 1947– 1958 78 Noble, W.S. (2006) What is a support vector machine? Nat. Biotechnol. 24, 1565– 1567
[15] E. Elyan and M. M. Gaber, "A fine-grained random forest using class decomposition: An application to medical diagnosis", *Neural Comput. Appl.*, vol. 27, no. 8, pp. 2279-2288, Nov. 2016.
[16] E. Elyan and M. M. Gaber, "A genetic algorithm approach to optimizing random forests applied to class engineered data", *Inf. Sci.*, vol. 384, pp. 220-234, Apr. 2017
[17] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines", *IEEE Intell. Syst.*, vol. 13, no. 4, pp. 18-28, Jul./Aug. 1998.
[18] Khanfar, M.A. and Taha, M.O. (2013) Elaborate ligand-based modeling coupled with multiple linear regression and k nearest neighbor QSAR analyses unveiled new nanomolar mTOR inhibitors. J. Chem. Inf. Model. 53, 2587–2612