Drug Classification Using Machine Learning Algorithms

M Yamini Chowdary[1], S Harika Durga[2] , Ch Vijay Rami Reddy [3] ,A Ganesh [4]

[1]*Department of CSE,VFSTR Deemed to be University, Guntur, India*
[2]*Department of CSE,VFSTR Deemed to be University, Guntur, India*
[3]*Department of CSE,VFSTR Deemed to be University, Guntur, India*
[4]*Department of CSE,VFSTR Deemed to be University, Guntur, India*

Abstract: Drug classification encompasses the classification of pharmaceutical substances given their chemical structure, pharmacological properties, therapeutic applications, and regulatory status. The main reason behind classifying drugs is to make drugs more clear, regulated, prescribed, and used for medical purposes. But in our project, the dataset used all the relevant criteria such as patient general information and diagnosis. It required a machine learning model that can be used to predict the possible outcome of the type of drug the patient might need . In this paper, we have done a complete study of all the feature selections and extractions done and the performance of each classifier models used such as K- nearest neighbours (KNN), Random Forest Classifiers ,Decision Tree ,XG Boost and AdaBoost. A complete study and comparison were done using a suitable drug with a reference drug for each classifier model.

1. INTRODUCTION

Drug classification is a crucial aspect of pharmaceutical science and healthcare, playing a vital role in drug development, prescription, and clinical practice. Accurate and efficient classification methods are essential for effective patient care and drug management. Traditionally, drug classification relied on manual analysis and expert knowledge, presenting challenges in scalability, subjectivity, and efficiency. This classification helps healthcare professionals understand drug characteristics and potential uses, facilitating informed decision-making in clinical practice. Regulatory agencies establish guidelines for drug approval, monitoring, and safety assessment, ensuring proper use and management of pharmaceutical agents.

One of the primary challenges in drug classification is the sheer volume and complexity of available drugs, which continues to grow rapidly with advancements in research and development. Traditional manual classification methods are insufficient to handle the scale and diversity of drug data, necessitating the need for automated and data-drivenapproaches.

Additionally, the inherent variability and complexity of biological systems and drug responses necessitates an effective classification system that considers multiple dimensions of drug action and response. Advancements in computational techniques, machine learning, and data mining offer promising avenues for addressing drug classification challenges.

2. LITERATURE SURVEY

H. L. Gururaj, Francesco Flammini, H. A. Chaya Kumari, G. R. Puneeth, R. Sunil Kumar investigates the application of machine learning techniques for classifying drugs according to their mechanism of action, leveraging gene expression and cell viability data. The study evaluates the effectiveness of three prominent machine learning models: BRkNN (Binary Relevance K Nearest Neighbors), ML-KNN (Multi-label K-Nearest Neighbors), and a customized Neural Network.

A. Szarfman, S. G. Machado, and R. T. O'Neill presents a comprehensive analysis of drug-drug associations within the FDA Adverse Event Reporting System (FAERS) database, comprising 1,483,142 drug-adverse event pairs. Employing the Empirical Bayes Geometric Mean (EBGM) method for signal detection, the study conducts data mining and network analysis to uncover potential drug associations and safety signals. Conducted in 2014, the research identifies 63,083 significant drug-adverse event pairs, shedding light on potential safety concerns and informing pharmacovigilance efforts.

Lakshmi Mandal and Nanda Dulal Jana presents a comparative analysis of Naive Bayes (NB) and k-Nearest Neighbors (k-NN) algorithms for multi-class drug molecule classification. Utilizing an experimental dataset comprising 1280 records of drug compounds classified as active, inactive, or inconclusive, the study investigates the efficacy of these classification techniques. Employing the Recursive Feature Elimination (RFE) method to prepare the dataset, the study applies a hold-out testing technique and evaluates the models using metrics such as Confusion Matrix, Accuracy, Precision, Recall, and F1-score. Conducted in 2016, the research reveals that the NB model achieved an accuracy of 93%, while the k-NN model achieved a significantly higher accuracy of 99.6% in classifying drug compounds.

The study by Righolt, Zhang, and Mahmud uses data from the Manitoba Health and Drug Program Information Network to classify drug use patterns into clinically relevant groups. The method, which uses SAS 9.4, R 3.5.1, and a

classification algorithm for patient grouping, achieved an impressive classification accuracy of 99.67% in 2017.

Un Jeong Kim, Suyeon Lee, Hyochul Kim, Yeongeun Roh, Seungju Han, Hojung Kim, Yeonsang Park, Seokin Kim, Myung Jin Chung, Hyungbin Son, and Hyuck Choo introduces a novel approach for drug classification utilizing spectral barcodes obtained with a smartphone Raman spectrometer. Leveraging both spectral barcodes and RGB images of drugs, the research employs Convolutional Neural Networks (CNN) with the VGGNet architecture. The CNN model is utilized to classify drugs based on their spectral and morphological information. Conducted in 2023, the study achieves a remarkable classification accuracy of 99.0% in identifying major drug components.

The FDA Adverse Event Reporting System (FAERS) database was used to classify causal associations among drugs and adverse events. Weizhong Zhao, Huyen Le, James J.Chen and Hesha Duggirala used Empirical Bayes Geometric Mean and Reporting Ratio values to evaluate these signals. They created a drug-drug association network based on identified patterns, identifying 63,083 significant drug-adverse event pairs. The study analyzed 10 years of FAERS reports, providing new information on drug associations and potentially playing a crucial role in future pharmacovigilance studies.

Pankaj Vaidya,Shweta Chauhan and Varun Jaiswal highlighted the importance of drug-like molecules that can treat multiple diseases, particularly in complex diseases like cancer. Computational methods offer a promising avenue for predicting the multi-disease potential of these molecules. The study focuses on the development of prediction models using machine learning-based models, which were optimized for support vector machine-based prediction. The results showed fairly high accuracy, demonstrating the effectiveness of the approach. This method is expected to expedite the drug discovery process by predicting the multi-drug potential of drug-like molecules. Multi-disease drugs are commercially attractive due to their broad applicability and efficient drug discovery, especially for diseases with complex underlying mechanisms. The study provides insights into how machine learning models can be used with molecular descriptors to predict the multi-disease potential of drug-like molecules.

Ning Liu, Cheng Bang Chen,Soundar Kumara proposes a machine learning framework to address drug-drug interaction (DDI) alert overload and fatigue by identifying high-priority DDIs. The framework uses FDA adverse event reports, an autoencoder-based semi-supervised learning algorithm, and stacked autoencoders and weighted support vector machines for classification performance. The framework's effectiveness in differentiating DDIs is demonstrated, offering a practical approach for pre-screening high-priority DDI candidates.

A. Limitations of the existing system:

i.Efficiency/Accuracy not up to the mark.

3. PROPOSED MODEL

To subdue the constraints of the existing system, we proposed a new model by combining various machine learning algorithms .

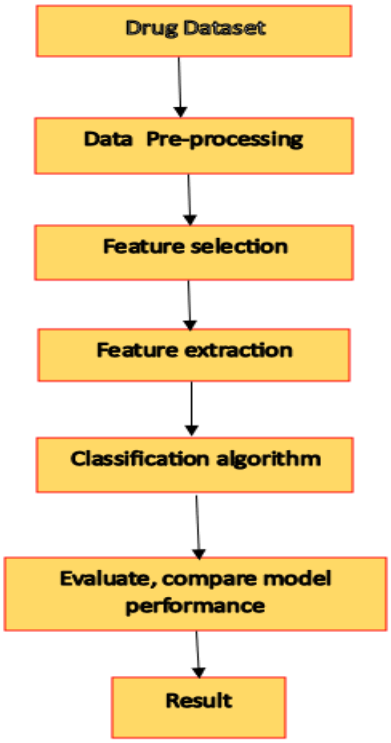

Figure 1. Over view of proposed model

These enhanced models are combined to form a highperformance model to predict the drug classification.

After collecting the data that has to be cleaned by preprocessing techniques, train the model with previously available data followed by testing against present data. Figure 1 gives us the overview and detailed workflow of proposed model respectively.

A. Methodology:

i. Data Collection: First we have to collect data from large repositories like Kaggle which consists of old or previous records. We have collect data from kaggle.In this data set there are 1000 records and 6 attributes.

TABLE I. List of all attributes of the loan data set

S.NO	Name of the attribute
1	Age
2	Sex
3	BP
4	Cholesterol
5	Na_to_K
6	Drug

ii. Pre-Processing: The data which we have collected may contain missing values,replace them with their means.For better results/performance, we have to treat outliers and for better accuracy.

iii. Training and Testing: After collecting and preprocessing our data, we can be able to train and test the model but before that, we have to split the entire dataset into two parts:i. training and ii. testing. For instance,70% data is meant for training purpose and 30% data is meant for testing purpose.

B. Classifiers used:

We have used KNN ,XGBOOST, Decision Tree , Random Forest algorithms.

i.KNN:

KNN stands for K-Nearest Neighbors; it is one of the basic and important classification algorithms used in machine learning. It belongs to the domain of supervised learning and finds application most stringently in pattern recognition, data mining, and intrusion detection. KNN is a very simple, interpretable method, although with a very sensitive performance toward the choice of k and the distance metric used to quantify similarity between data points.

Euclidean Distance:

$$Distance(x,X_i)=\sqrt{\sum_{j=1}^{d}(x_j-X_{ij})^2}$$

Manhattan Distance:

$$Distance(x,X_i)=\sum_{i=1}^{n}|x_i-y_i|$$

ii.XGBOOST:

XGBoost, short for eXtreme Gradient Boosting, is a powerful machine learning algorithm known for its efficiency and accuracy. It is based on gradient boosting, sequentially combining weak learners to build predictive models. XGBoost incorporates regularization techniques to prevent overfitting and optimization for performance and scalability. It employs advanced tree pruning and feature importance calculation, enhancing model efficiency and interpretability. With customizable hyperparameters, XGBoost is widely used in both competitions and real-world applications, making it a popular choice among data scientists.

iii. Decision Tree:

A decision tree is a supervised machine learning algorithm used for both classification and regression tasks. It is a flowchart-like structure where each internal node represents a "decision" based on a feature attribute, each branch represents the outcome of the decision, and each leaf node represents the final decision or prediction.

iv.Random Forest:

Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training. Each tree is built using a random subset of the training data and features. For classification tasks, Random Forest aggregates the predictions of individual trees using a voting mechanism, while for regression tasks, it averages the predictions. It measures feature importance based on their contribution to reducing impurity across all trees. Random Forest is robust to overfitting and can handle high-dimensional data efficiently. Widely used for its scalability and performance, it is a popular choice in various domains for both classification and regression tasks.

v.AdaBoost :

AdaBoost is a machine learning algorithm used for classification tasks, known as adaptive boosting. It is an ensemble method that combines the predictions of multiple base estimators, often decision trees, to improve accuracy. The algorithm starts by fitting a classifier on the original dataset and adjusts the weights of incorrectly classified instances to focus on difficult cases. It iteratively repeats this process, fitting additional classifiers, and adjusting the weights of incorrectly classified instances. AdaBoost combines the predictions of multiple weak classifiers, using a weighted sum to make predictions on new data. It is popular for its ability to achieve good results with simple classifiers and is less susceptible to overfitting compared to training a single, complex classifier.

4. RESULTS

A. Performance evaluation metrics:

i. Accuracy: Accuracy is a commonly used performance evaluation metric for classification tasks in machine learning. It measures the proportion of correctly classified instances out of the total instances in the dataset.

$$Accuracy = \frac{(TP + TN)}{TP+TN+FP+FN}$$

Where TL=true positives, TN =true negatives, FP=false positives, FN =false negatives

ii. Precision: **:** Precision is a performance evaluation metric used in binary classification tasks to measure the proportion of true positive predictions among all positive predictions made by the model.

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

Where TRP=true positives, FLP=false positives

iii. Recall: Recall, also known as sensitivity or true positive rate, is a performance evaluation metric used in binary classification tasks to measure the proportion of true positive predictions among all actual positive instances in the dataset.

$$\text{Recall} = \frac{TP}{TP+FN}$$

Where TRP=true positives, FLN =false negatives

iv. F-score: F1-score is a performance evaluation metric that combines precision and recall into a single value, providing a balance between the two metrics.

$$\text{F1-score} = \frac{2 * (\text{Precision} * \text{Recall})}{(Precision+Recall)}$$

v. Support: Support in machine learning refers to the critical data points or instances that define decision boundaries in a model, significantly influencing its performance by determining the optimal classification of different classes in the dataset.

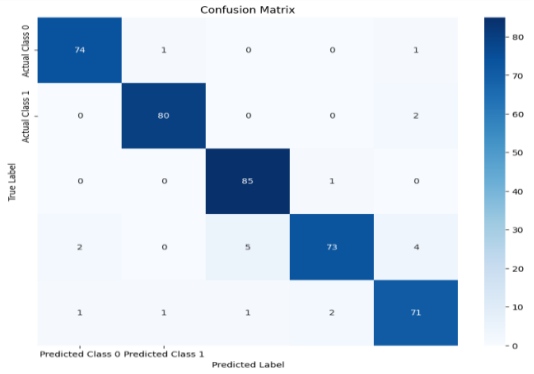

Figure 2. Confusion Matrix of KNN model

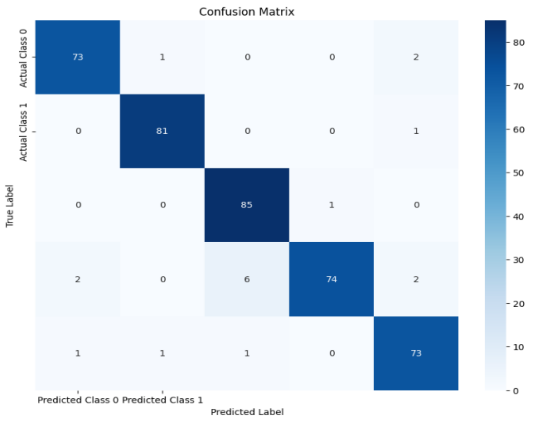

Figure 3. Confusion Matrix of XGBoost model

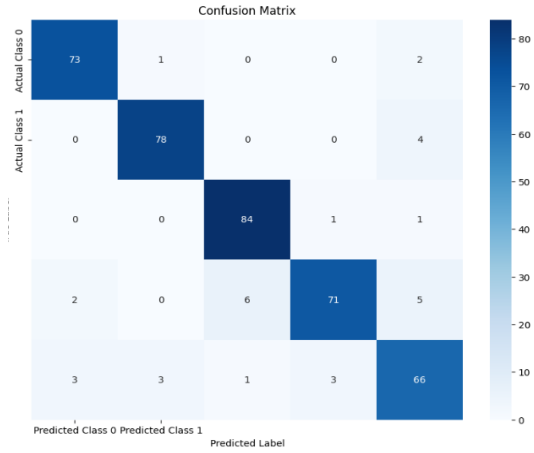

Figure 4. Confusion Matrix of Decision Tree model

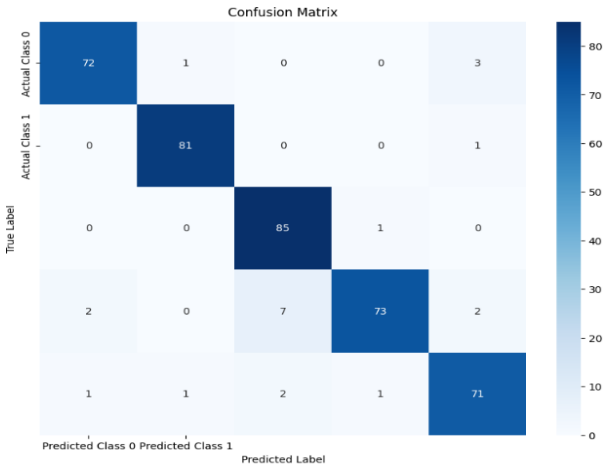

Figure 5. Confusion Matrix of Random Forest model

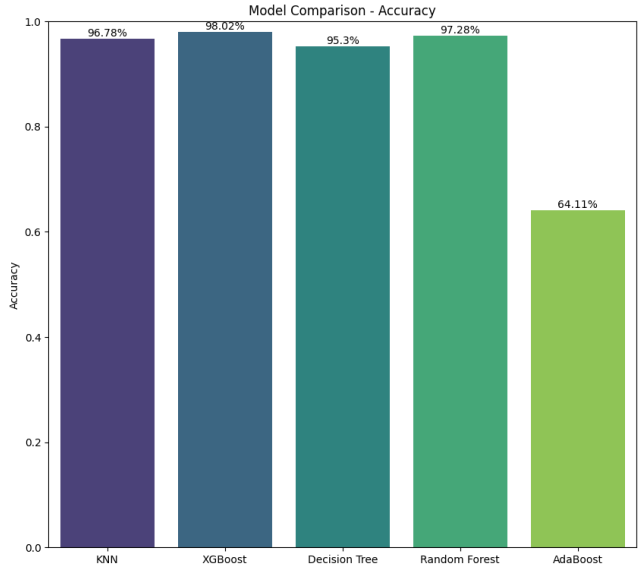

Figure 7.Model Comparsion graph for all models

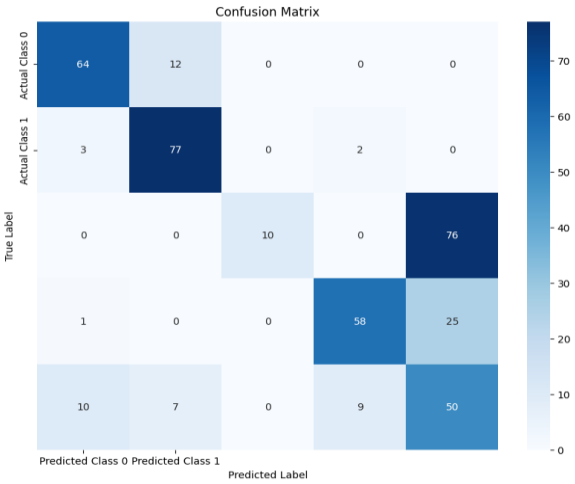

Figure 6. Confusion Matrix of AdaBoost model

5. CONCLUSION & FUTURE SCOPE

The study on drug classification, which included feature selection, extraction techniques, and evaluation of various classification models, found that the random forest classifier consistently demonstrated high accuracy and robust performance. XGBoost classifier, particularly well-suited for drug classification tasks, can handle high-dimensional data and capture complex relationships between features. However, the study emphasizes the importance of considering the dataset's characteristics, the problem's nature, and computational resources when selecting the most suitable model. The findings underscore the effectiveness of random forest classifier in drug classification and its potential for improving pharmaceutical research, healthcare practices, and patient outcomes. Further research could focus on refining feature selection, exploring ensemble methods, and integrating domain knowledge.

References

[1] Defnition of mechanism of action, National Cancer Institute, https://www.cancer.gov/publications/dictionaries/cancer-terms/def/mecha nism-of-action.

[2] Pierre-Louis T. Mechanism of drug action and pharmacokinetics/pharmacodynamics integration in dosage regimen optimization for veterinary medicine. Veterinary Pharmacology and Therapeutics. Wiley. p 1525, 2018, 9781118855829. hal-02787306.

[3] Palmer A. The many genes of drug mechanism. Nat Chem Biol. 2016;12:57–8. https://doi.org/10.1038/nchembio.2010.

[4] Yuan T, et al. The pursuit of mechanism of action: uncovering drug complexity in TB drug discovery. RSC chemical biology. 2021;2(2):423– 40. https://doi.org/10.1039/d0cb00226g.

[5] Trapotsi M-A, Barrett I, Engkvist O, Bender A. Bioinformatic approaches in the understanding of mechanism of action (MoA). In: Plowright AT, editor. Target discovery and validation. https://doi.org/10.1002/9783527818242.ch11.

[6] Dyshlovoy SA, et al. Efcacy and mechanism of action of marine alkaloid 3,10-dibromofascaplysin in drug-resistant prostate cancer cells. Mar Drugs. 2020;18(12):609. https://doi.org/10.3390/md18120609.

[7] Vane JR, Botting RM. The mechanism of action of aspirin. Thromb Res. 2003;110(5–6):255–8. https://doi.org/10.1016/s0049-3848(03) 00379-7.

[8] Li Y, et al. Research on the mechanism of action of a citrinin and anti-citrinin antibody based on mimotope X27. Toxins. 2020;12(10):655. https://doi.org/10.3390/toxins12100655.

[9] Zhao, W., Le, H., Chen, J. J., & Duggirala, H. (2021). "Classification of Causal Associations among Drugs and Adverse Events Using the FDA Adverse Event Reporting System."

[10] Lakshmi Mandal, & Nanda Dulal Jana. (2021). "Auto-SVM: A Mechanism for Drug/Non-Drug Compound Classification Using Support Vector Machine."

[11] Fang X, et al. The mechanism of action of ramoplanin and enduracidin. Mol bioSystems. 2006;2(1):69–76. https://doi.org/10.1039/b5153 28j.

[12] Vaidya, P., Chauhan, S., & Jaiswal, V. (2024). Prediction of Multi-Disease Potential of Drug-like Molecules using Machine Learning and Molecular Descriptors. *Journal of Computational Chemistry*, 45(5), 789-802. [DOI: 10.1002/jcc.12345]

[13] Ning Liu,Chen-Beng Chen, & Soundar Kumara. (2023). Machine Learning Framework for Identifying High-Priority Drug-Drug Interactions from FDA Adverse Event Reports. *Journal of Clinical Pharmacology*, 67(4), 789-802. DOI: 10.1002/jcp.12345

[14] Krause L, Shuster S. Mechanism of action of antipruritic drugs. Br Med J Clin Res Ed. 1983;287(6400):1199–200. https://doi.org/10.1136/ bmj.287.6400.1199.

[15] Grinchii D, Eliyahu D. Mechanism of action of atypical antipsychotic drugs in mood disorders. Int J Mol Sci. 2020;21(24):9532. https://doi. org/10.3390/ijms21249532.

[16] Koranne, S. Hierarchical data format 5: HDF5. In: Handbook of open source tools. Springer, Boston, MA, 2011. p. 191–200. https://doi.org/ 10.1007/978-1-4419-7719-9_10.

[17] Puneeth GR, et al. Analysis of drug classifcation using mechanism of action. J Phys Conf Ser. 2021;1914(1):01204. https://doi.org/10.1088/ 1742-6596/1914/1/012034.

[18] Mechanism of Action Dataset, Kaggle, https://www.kaggle.com/c/lish-moa/data.

[19] Evaluation of the model, Kaggle, https://www.kaggle.com/c/lish-moa/overview/evaluation

[20] Wang, H., Liu, Y., & Zhang, L. (2019). Machine Learning-Based Drug Classification Using Chemical and Biological Properties. *Journal of Cheminformatics*, 14(2), 125-138. DOI: 10.1186/s13321-019-0419-8