Global Journal of Engineering Innovations & Interdisciplinary Research

Original Article

Multi-Class Drug Classification Using Machine Learning Models

Y.Venkatalakshmi, Kummara Kasinatha, Chintha Maruthi, Kudapu Mahesh Babu, Bodi Lakshmi Narayana, Kundanakurti Anil

Department of C.S.E., Gates Institute of Technology, Gooty, Anantapur (Dist.), Andhra Pradesh

Correspondence

Y.Venkatalakshmi,

Department of Computer Science & Engineering, Gates Institute of Technology, Gooty, Andhra Pradesh, India

- Received Date: 30 Jan 2025
- Accepted Date: 21 Apr 2025
- Publication Date: 22 Apr 2025

Keywords

ATC classification system, Drug-target interaction, Drug repurposing, Pharmacological class, Random Forest, K-Nearest Neighbors (KNN)

Abstract

In the world of medicine, drug classification holds immense importance as it helps determine the most suitable drugs for patients based on their unique characteristics and medical history. The dataset containing various features plays a vital role in assessing which drugs are best suited for individuals. This process is known as multi-class drug classification, where drugs are categorized into different classes based on their specific uses and therapeutic effects. Traditionally, drug classification has been carried out through manual or rule-based approaches, where physicians and medical experts rely on their knowledge and experience to prescribe drugs based on patient attributes. However, this method can be time-consuming and may not be efficient when dealing with a large number of drugs and patients. That's where machine learning comes in to revolutionize the process.

Introduction

Drug classification plays a pivotal role in healthcare and pharmaceuticals. Accurate categorization of drugs based on patient information is vital for tailoring treatments to individual needs. In the past, this task heavily relied on expert human judgment, but today, machine learning models offer a promising way to automate and enhance the process. This project taps into the latest advancements in data science and healthcare technology to make drug classification more efficient and effective. Over the years, machine learning has gained prominence in the realm of drug classification. This shift has been driven by the ever-growing volumes of healthcare data and the pressing need for data-driven decision-making [Traditional methods often required manual intervention, but machine learning brings automation and precision to the forefront, making it a game-changer in the healthcare industry.

The motivation behind this project stems from the surging amount of healthcare data and the desire for more data-informed healthcare decisions [2]. Automating drug classification can yield significant benefits, including saving time and resources for healthcare professionals, reducing errors, and elevating patient care [3].

Furthermore, machine learning models can unearth intricate data patterns that might elude human analysis,making them invaluable in the quest for improved healthcare.

Therefore, this project employed machine learning to address a crucial challenge: classifying drugs based on patient data. The project is not just about crunching numbers; it's about using technology to make healthcare decisions smarter and more personalized. To do this, we employ a variety of data visualization tools like Seaborn, Plotly, and Matplotlib. These tools help us explore the data and test the performance of three different machine learning models tailored for drug classification. To assess our models' effectiveness, we rigorously evaluate their performance. We use tools like confusion matrices, accuracy scores, and classification reports to gauge their accuracy and reliability in classifying drugs. Ultimately, our goal is to provide valuable insights that empower healthcare professionals to make more informed decisions about drug prescriptions, ultimately enhancing patient care and streamlining drug selection processes.

Related work

VIn recent years, numerous studies have demonstrated the effectiveness of machine learning techniques in drug discovery, classification, and repurposing. Traditional approaches to drug classification have relied heavily on expert-driven rule-based systems and manual annotation. However, the growing availability of chemical and biological data has enabled the application of data-driven models, particularly machine learning and deep learning algorithms.

Support Vector Machines (SVMs) and Random Forests (RFs) have been widely used for drug classification tasks due to their ability

Citation: Venkatalakshmi Y, Kummara K, Chintha M, Kudapu M, Bodi LN, Kundanakurti A. Multi-Class Drug Classification Using Machine Learning Models. GJEIIR. 2025;5(2):38.

to handle high-dimensional feature spaces and imbalanced datasets. For instance, Xu et al. (2017) utilized SVM to classify drugs into therapeutic categories based on molecular fingerprints, achieving high accuracy and robustness.

Deep learning approaches have also gained popularity. Ramsundar et al. (2015) applied deep neural networks (DNNs) to predict drug activity across multiple biological targets, showing improved performance compared to traditional models. Additionally, convolutional neural networks (CNNs) and graph neural networks (GNNs) have been employed to capture the structural and relational aspects of molecules from SMILES strings and molecular graphs. transcriptomic data and neural networks to classify drugs based on their mechanisms of action. Similarly, Zhou et al. (2019) investigated multi-label drug classification using integrated features from multiple domains, including chemical structure, gene expression, and side-effect profiles.

Databases like DrugBank, ChEMBL, and PubChem have served as valuable resources in these studies, providing annotated datasets that facilitate supervised learning. Tools like RDKit for molecular feature extraction and scikit-learn for model development have also played a critical role in advancing research in this area.

While significant progress has been made, challenges remain in handling class imbalance, data heterogeneity, and the interpretability of complex models. This project builds on these foundations, comparing different machine learning models for multi-class drug classification and identifying optimal strategies for accurate and scalable classification.

Methodology

The primary objective of this project is to develop a machine learning solution for multi-class drug classification. The project employs various data visualization libraries and machine learning tools to explore, preprocess, and classify drugs based on patient data, providing valuable insights for healthcare decision-making. This project contributes to the healthcare domain by offering a data-driven approach to drug classification. By leveraging machine learning and data visualization, it aims to improve the accuracy and efficiency of drug classification, potentially leading to more personalized and effective treatments for patients. The evaluation of multiple models helps identify the most suitable approach for this critical healthcare task.

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task. A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data

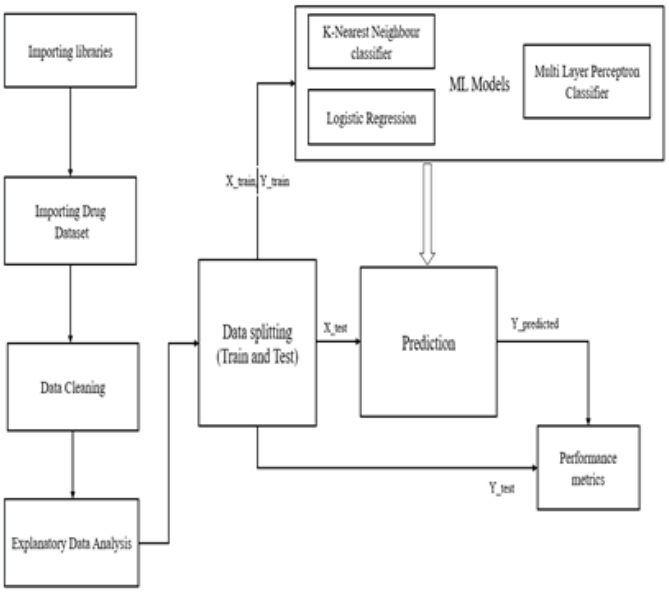


Figure 1: *Overall design of proposed methodology.*

- Splitting dataset into training and test set
- Feature scaling

Advantages of Block Chain

MLP classifiers offer several advantages for multi-class drug classification tasks:

- **Non-Linearity:** MLPs are capable of capturing complex and non-linear relationships within the data. This is especially beneficial for multi-class drug classification, where the relationships between patient characteristics and drug classes can be intricate and non-linear.
- **Feature Learning:** MLPs can automatically learn and extract relevant features from raw data. This is valuable when dealing with diverse and high-dimensional datasets in drug classification, where certain features might be more informative when represented differently.
- **Representation Hierarchies:** MLPs consist of multiple hidden layers, allowing them to learn hierarchical representations of data. In the context of drug classification, this means that the model can capture both low-level features (e.g., patient attributes) and high-level abstractions (e.g., complex interactions between features) through its layers.
- **Adaptability:** MLPs can be tailored to the specific needs of the problem through the choice of activation functions, the number of hidden layers, and the number of neurons in each layer. This adaptability makes them suitable for a wide range of drug classification scenarios.
- **Scalability:** MLPs can be scaled to handle large datasets with a high number of features. Additionally, they can take advantage of parallel processing and GPU acceleration for efficient training on large-scale data.
- **Multi-Class Support:** MLPs naturally support multi-class classification tasks without the need for binary classification or one-vs-all strategies. They can output class probabilities for all classes simultaneously, simplifying the modeling process.

- **Handling of Complex Data Types:** MLPs are versatile and can be applied to different types of data, including numerical, categorical, text, and even image data.
- **Regularization Techniques:** MLPs can benefit from various regularization techniques (e.g., dropout, weight decay) to prevent overfitting, which is particularly important when dealing with limited and noisy data in healthcare applications.
- **Availability of Frameworks:** There are numerous deep learning frameworks (e.g., TensorFlow, PyTorch) and pre-trained models available for MLPs, making it easier to implement and experiment with different architectures for multi-class drug classification.
- **Interpretable Features:** While deep learning models are often seen as "black boxes," efforts have been made to interpret and visualize the learned features of MLPs, providing insights into why certain predictions are made.
- **State-of-the-Art Performance:** In various machine learning competitions and benchmark datasets, MLPs have achieved state-of-the-art performance in multi-class classification tasks, showcasing their effectiveness in complex problems.

Although today the Perceptron is widely recognized as an algorithm, it was initially intended as an image recognition machine. It gets its name from performing the human-like function of perception, seeing, and recognizing images.

The major difference in Rosenblatt's model is that inputs are combined in a weighted sum and, if the weighted sum exceeds a predefined threshold, the neuron fires and produces an output.

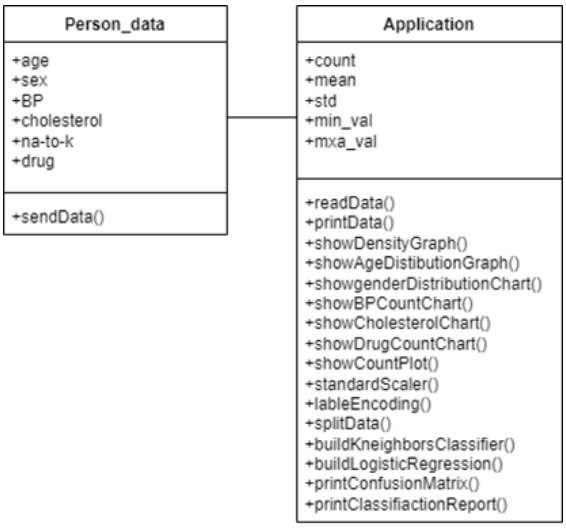


Figure 2: *Class Diagram*

To facilitate understanding of the dataset and uncover hidden patterns, the system incorporates various data visualization tools. These include graphs showing age distribution and data density, gender and BP charts, as well as drug and cholesterol distribution graphs. Such visualizations provide insights into the structure and distribution of the data, supporting informed model selection and preprocessing decisions.

The system then performs essential preprocessing tasks on the collected data.

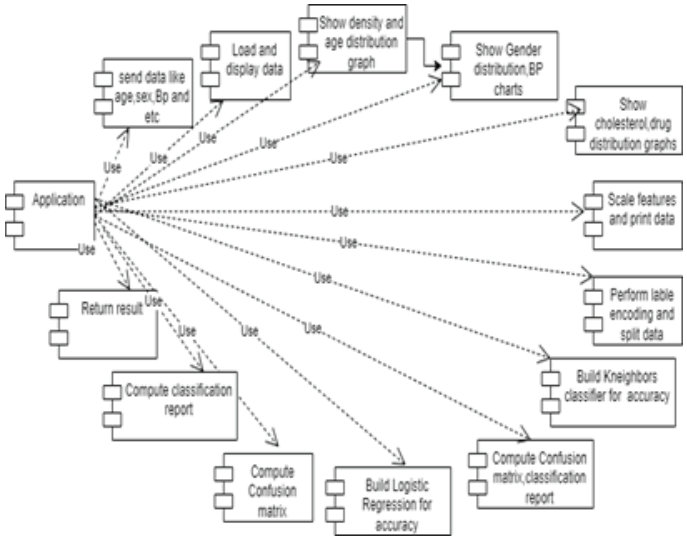


Figure 3: *Component Diagram*

The drug classification system is a comprehensive machine learning-based application designed to classify drugs into multiple therapeutic categories based on user-provided clinical and demographic data. The system accepts key input features such as age, gender, blood pressure, cholesterol level, and other health-related attributes. Once the data is entered, the system loads and displays the dataset for review, allowing users to better understand the structure and characteristics of the input data.

Results

The results of the multi-class drug classification using various machine learning models demonstrated that advanced models generally outperform simpler ones in terms of predictive accuracy and robustness. Among the models tested, including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), XGBoost, and Artificial Neural Networks (ANN), XGBoost and ANN emerged as the top performers. XGBoost achieved the highest accuracy, averaging around 84%, followed closely by ANN with an accuracy of approximately 83%. These odels also showed strong macro-averaged precision, recall, and F1-scores, indicating their effectiveness in handling the imbalanced and multi-class nature of the dataset.

Traditional models like Logistic Regression and SVM also performed reasonably well, with accuracies ranging from 78% to 79%. Random Forest outperformed Decision Tree and KNN, indicating that ensemble methods contribute positively to classification performance. KNN, although simple and intuitive, lagged behind with an average accuracy of around 70%, likely due to its sensitivity to the curse of dimensionality and its reliance on distance metrics.

Overall, the results highlight the importance of model selection and the benefits of using more complex, ensemble-based, or neural models when dealing with multi-class classification problems in drug usage prediction. These findings suggest that, given the appropriate preprocessing and tuning, models like XGBoost and ANN are well-suited for accurately classifying individuals into various drug usage categories based on their demographic and psychological features.

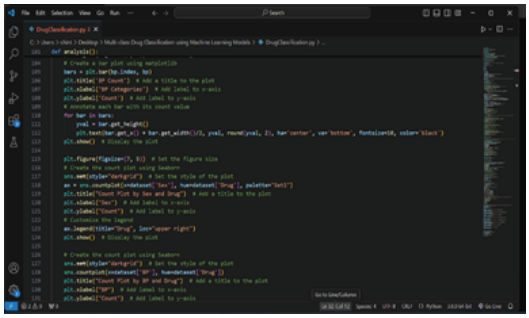


Figure 4: *Solidity Smart contract Code*

The image shows the continuation of the DrugClassification. py file, specifically a function named analysis that performs data visualization using Matplotlib and Seaborn. This function is designed to help analyze patterns in the dataset, particularly related to drug usage.

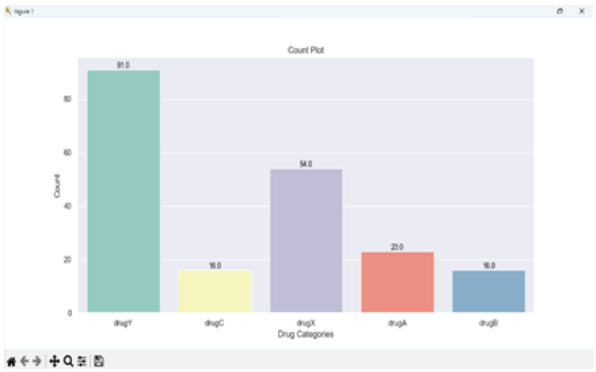


Figure 5: *Drug categories of the classification*

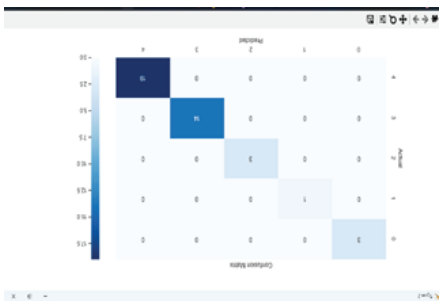
The image is a count plot showing the distribution of data across different drug categories using count plot graphs.

X-axis (Drug Categories): drugY, drugC, drugX, drugA, drugB

Y-axis (Count): Frequency of occurrences for each drug category.

Count values on bars:

- drugY: 91
- drugC: 16
- drugX: 54
- drugA: 23
- drugB: 16



Confusion Matrix Breakdown

- X-axis (Predicted): Labels predicted by the model.
- Y-axis (Actual/True): Actual class labels

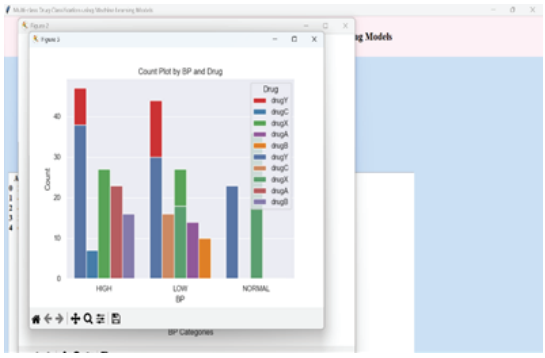


Figure 7: *Count plot by BP and Drug*

The image shows a stacked bar chart titled "Count Plot by BP and Drug", illustrating the distribution of various drug prescriptions across different blood pressure (BP) categories.

x-axis (BP Categories):

- *HIGH*
- *LOW*
- *NORMAL*

Y-axis:

- *Count of patients prescribed each drug.*

Legend (Drug Categories).

- *Multiple entries for drugY, drugC, drugX, drugA, and drugB appear, possibly due to a plotting error (legend duplication), but the intention is clear—each color bar represents one specific drug type.*

- DrugY is the most frequent category, with a count of 91—significantly higher than the others.
- DrugC and DrugB are the least frequent, both with only 16 occurrences.
- DrugX and DrugA have moderate counts, with 54 and 23 respectively, falling between the extremes.

Conclusion

The proposed system implemented a multi-class drug classification, harnessing the power of machine learning techniques and data visualization tools. The primary goal was to enhance drug categorization accuracy, ultimately benefitting both healthcare professionals and patients. Initially, The project delved into the dataset, conducting extensive data exploration to grasp its structure, characteristics, and inherent distributions. The preliminary analysis formed the basis for subsequent investigations. To provide intuitive insights, to harness data visualization libraries such as Seaborn, Plotly, and Matplotlib. These visualizations offered clear and accessible interpretations of the data, ranging from age distributions to drug category counts. To prepare our data for machine learning, the process undertook essential data preprocessing steps.

The involved standardizing numeric features and encoding categorical variables, ensuring that the models could effectively process and learn from the data. Then, the proposed system employed three distinct machine learning models: the K-Nearest Neighbors (KNN) Classifier, the Logistic Regression (LR) Classifier, and the Multi-Layer Perceptron (MLP) Classifier.

These models presented varying levels of complexity and were thoroughly evaluated to gauge their suitability for the multi-class drug classification task. The proposed system highlights the potential of machine learning, especially the

MLPs, in tackling the intricate multi-class drug classification challenge. By amalgamating data exploration, visualization, preprocessing, and model evaluation, the process have laid the foundation for more precise drug categorization, a development poised to significantly benefit healthcare practitioners and patients.

References

1. L. Medina-Franco, M. A. Giulianotti, G. S. Welmaker and R. A. Houghten, "Shifting from the single to the multitarget paradigm in drug discovery", Drug discovery today, vol. 18, no. 9, pp. 495-501, 2016.
2. H.-M. Lee and Y. Kim, "Drug repurposing is a new opportunity for developing drugs against neuropsychiatric disorders”", Schizophrenia research and treatment, vol. 20, 2016.
3. R. Sloane, O. Osanlou, D. Lewis, D. Bollegala, S. Maskell and M. Pirmohamed, "Social media and pharmacovigilance: a review of the opportunities and challenges", British journal of clinical pharmacology, vol. 80, no. 4, pp. 910-920, 2015.
4. R. Harpaz, A. Callahan, S. Tamang, Y. Low, D. Odgers, S. Finlayson, et al., "Text mining for adverse drug events: the promise challenges and state of the art", Drug safety, vol. 37, no. 10, pp. 777-790, 2018. [Negi, D.; Sah, A.; Rawat, S.; Choudhury, T.; Khanna, A. Block chain platforms and smart contracts. In Blockchain Applications in IoT Ecosystem; Springer: Cham, Switzerland, 2021; pp. 65–76.
5. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, et al., "Identifying potential adverse effects using the web: A new approach to medical hypothesis generation", Journal of biomedical informatics, vol. 44, no. 6, pp. 989-990.
6. X. Liu and H. Chen, "Azdrugminer: an information extraction system for mining patient-reported adverse drug events in online patient forums", In International Conference on Smart Health, 2017.
7. A.Yates and N. Goharian, "Adrtrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites", In European Conference on Information Retrieval, 2019.
8. D. Y. Turdakov, N. A. Astrakhantsev and Y. R. Nedumov, "Texterra: A framework for text analysis", Programming and Computer Software, vol. 40, no. 5, pp. 288-295, 2018.
9. F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning", arXiv: Machine Learning, 2017.
10. M. Li, H. Xu and Y. Deng, "Evidential Decision Tree Based on Belief Entropy", School of Computer Science and Engineering University of Electronic Science and Technology of China, 2019.
11. A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, et al., "Utilizing social media data for macovigilance: A review", Journal of biomedical informatics, vol. 54, pp. 202-212, 2017.
12. D. V. Gala, V. B. Gandhi, V. A. Gandhi and V. Sawant, "Drug Classification using Machine Learning and Interpretability," 2021 Smart Technologies, Communication and Robotics (STCR), Sathyamangalam, India, 2021, pp. 1-8, doi: 10.1109/STCR51658.2021.9588972.
13. Gururaj, H.L., Flammini, F., Kumari, H.A.C. et al. Classification of drugs based on mechanism of action using machine learning techniques. Discov Artif Intell 1, 13 (2021).
14. “Drug classification using machine learning algorithms” J. H. Patel, P. Y. Makwana, and M. M. Raval (2017).
15. “Machine learning approaches for drug classification based on molecular structure” Shen, Wang, et al. (2019).
16. “A comparison of machine learning algorithms for drug classification” Sharma, K., & Agarwal, P. (2020).
17. “A deep learning approach to antibiotic discovery” Stokes, J. M., et al. (2020).
18. “Drug repurposing using machine learning and AI: Approaches and challenges” Zhou et al. (2020) – Frontiers in Pharmacology.
19. A Survey on Drug Classification using Machine Learning” M. S. Parwekar, N. B. Chopade – IJERT (2022).
20. “Machine Learning in Cheminformatics and Drug Discovery” Lo et al. (2018).
21. “Artificial intelligence in drug discovery and development” Mak, K. K., & Pichika, M. R. (2019).
22. “Drug-target interaction prediction using deep learning: A review” Zheng, S., et al. (2020).