Integrative Machine learning for Drugs Side Effect Prediction

Laka Sridar
dept. AIDS
KL University
Vijayawada,
2100080250ai.ds@gmail.com

Shyam Raj
dept. AIDS
KL University
Vijayawada, India
2100080056ai.ds@gmail.com

Vujjuru Surya Charan Teja
dept. AIDS
KL University
Vijayawada, India
charanteja.sai.2003@gmail.com

***Abstract*—Predicting and understanding the potential side effects of pharmaceutical drugs is a formidable challenge in the realm of medical science. The conventional process of assessing drug safety, primarily reliant on manual clinical testing [3] and post-market surveillance [4], [5], is not only arduous but also time-consuming, making it an impediment to rapid drug development and patient well-being. This paper introduces an innovative approach that harnesses the power of advanced machine learning techniques to address this challenge [6], [7], [8], [10], [9].**

Our research delves into the intricate web of drug interactions within the human body and the complex factors contributing to adverse reactions. It is a realm where the exact mechanisms that trigger side effects often remain elusive, and the prevalence of rare and severe reactions complicates the task. Our approach encompasses the integration of diverse data sources, including drug characteristics, generic names, molecular structures (SMILES) [11], [12], and drug category [13]. Thereby providing a comprehensive understanding of the complex relationships between these factors and adverse reactions, unifying these disparate pieces of information, we aim to unveil hidden patterns and relationships that can significantly enhance the accuracy of drug side effect predictions.

In this study, we employ the Random Forest classification model, known for its robustness and interpretability, to make predictions that not only incorporate the broad spectrum of drug-related factors [14] but also ensure the inclusion of rare and severe side effects in the assessment [15]. To address the challenge of handling non-numerical data, we employ methodologies to convert features such as SMILES structures, Drug generic names, and others into meaningful numerical descriptors, enabling their seamless integration into the predictive model.

Our research does not only seek to enhance drug safety evaluation but also endeavors to bridge the chasm between the unpredictability of drug side effects and the need for a more efficient and informed drug development process. With the potential to revolutionize drug safety practices, this research has far-reaching implications in the domain of patient care and drug industry decision-making, promising a safer and more efficient future for the pharmaceutical research industry.

***Index Terms*—SMILES (Simplified Molecular Input Line Entry System), Indication or Drug Target Diseases, Drug Side Effect, Random Forest, decision tree.**

I. Introduction

The chemical composition of a medication significantly affects both how well it works and the likelihood of causing unwanted side effects. In the intricate dance between the field of pharmacology and the human body, a drug’s structure plays a vital role in how it interacts with molecules, triggers biological responses, and shares similarities with other chemical structures. [1], These factors directly impact the safety and effectiveness of pharmaceuticals. Therefore, understanding the structural foundations of these drugs is essential in the world of drug research as it unveils the intricate connections between a drug’s properties, its intended targets, and the occurrence of side effects [2].

The anticipation of potential side effects associated with drugs is a fundamental aspect of pharmaceutical research [29], with far-reaching consequences for patient well-being, drug development, and the broader healthcare landscape. Being able to predict adverse reactions to pharmaceuticals before they are released to the public is not only a matter of economic significance but, more crucially, a matter of safeguarding human health. Historically, assessing drug safety has depended on time-consuming clinical trials and meticulous post-market surveillance. Yet, this manual process often falls short in fully uncovering and predicting the complete range of potential side effects. This has, in the past, presented a significant obstacle to speeding up drug development and quickly delivering new treatments to those in need.

In this ever-evolving field of pharmacology, machine learning-based predictive methods have emerged as powerful tools to address the multifaceted challenge of anticipating drug side effects. These methods offer a way to navigate the complexities of how drugs interact within the human body, especially when the exact mechanisms behind side effects remain mysterious. Moreover, these approaches enable the thorough examination of rare and severe side effects, which has traditionally been challenging using manual techniques.

Our research aims to contribute to this ongoing shift in the field by introducing an innovative approach that utilizes a drug’s structure, encoded in SMILES (Simplified Molecular Input Line Entry System) notation, as a crucial factor in predicting drug side effects. The importance of drug structure in pharmacology is undeniable, as it underpins the interactions between molecules and biological responses that determine the safety and effectiveness of pharmaceuticals. Therefore, investigating the inherent relationship between drug structure and side effects holds great promise for revolutionizing drug research.

While machine learning methods have been employed in drug side effect prediction previously, our approach offers a distinctive contribution. We integrate a multitude of data sources, encompassing drug characteristics, generic names, molecular structures, drug target disease (indication) and drug category. By amalgamating these distinct dimensions, we aim to illuminate hidden patterns and relationships that have thus far eluded the grasp of traditional methodologies.

To do this, we use a Random Forest classification model, which is well-known for its reliability and ease of interpretation. This model helps us tap into the predictive power hidden within the various factors we've gathered. Importantly, it allows us to tackle the challenge of handling non-numerical data like SMILES structures and converting them into meaningful numerical descriptions.

As we embark on this research journey, our objectives In our research journey, we have two main objectives. First, we aim to improve the accuracy and efficiency of predicting drug side effects, with a particular emphasis on rare and severe ones. Second, we want to enhance our understanding of the complex mechanisms behind drug side effects, paving the way for safer and more effective drug development.

II. METHODOLOGIES

A. Problem Statement and formulation

The task at hand involves the prediction of potential side effects associated with various drugs. This problem is approached as a multi-label classification task, where each drug is associated with a binary vector of side effects. For a given drug i, the target label is represented as a binary vector $\mathbf{y_i}$, with d denoting the number of distinct side effects. Specifically, $y_{i,j}$ equals 1 when the drug i is linked to side effect j, and $y_{i,j}$ equals 0 when it is not.

The dataset consists of n drug samples, each characterized by a pair of features $\mathbf{x_i}$ that describe the drug's attributes, and an associated side effect vector $\mathbf{y_i}$ which classifies the presence or absence of side effects.

The primary objective is to develop a predictive model capable of accurately categorizing drugs based on their side effect profiles and other 'k' dependencies for prediction. This model will leverage the features $\mathbf{x_i}$ to classify each drug i into one or more side effect categories ($y_{i,j}$).

The successful solution to this problem will have significant implications for pharmaceutical research and healthcare, enabling the identification of potential side effects and improving the safety and efficacy of drug treatments.

B. Material and Datasets

C. Side Effect Extraction Resource

- **Source:** The SIDER (Side Effect Resource) 4.1 database [17] is a widely recognized and authoritative resource for collecting information on drug side effects. It aggregates data from various pharmaceutical databases, scientific literature, and clinical trials.
- **Data Extraction:** From the SIDER 4.1 database [17], we extracted specific fields, including "DRUG ATC CODE," "DRUG INDICATION," "DRUG NAME," and "DRUG SIDE EFFECT."
- **Data Records:** Our extraction process resulted in a comprehensive dataset comprising 22,115,782 data records. These records encompass a diverse range of drugs, each linked to their respective ATC code [19], drug indication, name, and associated side effects See Table I for details.
- **Uniqueness:** Within this datasets, we identified 5,820 unique side effects and 1,507 distinct drugs. These unique side effects are integral to understanding the potential adverse reactions associated with various pharmaceutical compounds See Table I for details.

TABLE I
EXTRACTED SIDE EFFECT

Number of Records	Unique Drugs	Unique Side Effects
22115782	1065	5820

D. Collection of Drug Category and Drug SMILE

- **Source:** The Pharmacogenomics Knowledge Base (PharmGKB) [16] is a reputable resource dedicated to the study of how genetic variations influence drug responses. It provides information on drug-gene interactions, pharmacokinetics, and pharmacodynamics.
- **Data Extraction:** From the PharmGKB database [16], we extracted two essential components: "SMILES" (Simplified Molecular Input Line Entry System) [20] representations of drug structures and the categorization or "type" of each drug.
- **Data Records:** The integration of the PharmGKB [16] data into our datasets resulted in a total of 6,506,423 records. These records encompass a wide array of drugs, with each drug linked to its chemical structure represented in SMILES notation and categorized by typeSee Table II for details.
- **Uniqueness:** Within this datasets, we identified 622 unique drugs and 5,057 unique side effects. The inclusion of SMILES representations enhances our ability to analyze the structural aspects of drugs, a critical component in understanding drug reactions and side effects See Table II for details.

TABLE II
EXTRACTING FROM PHARMGKB DATABASE

Records	Drugs	Side Effects	Structure	Category
6506423	612	5057	619	619

The integration of these two diverse datasets forms the backbone of our research. By merging data from the authoritative SIDER 4.1 database [17], which focuses on side effects and drug characteristics, with information from the

PharmGKB database [16], which provides valuable insights into drug structures and categorization, we have created a comprehensive and multidimensional dataset. This dataset, consisting of a vast number of records, unique drugs, and distinct side effects, enables us to explore and predict drug side effects with depth and precision See Table III for details. The amalgamation of these datasets empowers our research to provide a more holistic understanding of the complex relationships between drug properties, chemical structures, and adverse reactions [18], ultimately contributing to the advancement of drug safety evaluation and patient care.

III. STATISTICAL ANALYSIS AND VISUALIZATION

The word cloud visualization, presented in Fig. 1, offers a striking portrayal of the most frequently mentioned drug names within the datasets. Each drug name is represented by a word, with the size of the word directly corresponding to its frequency of occurrence. Larger words within the cloud denote drugs that are most frequently administered or cited. This visualization provides an immediate and intuitive overview of the dominant drugs within the datasets, which is pivotal in discerning their potential implications in relation to side effects.

Fig. 2 and 3 showcase a bar chart depicting the most prevalent drug names and drug indication within the dataset. This chart is limited to the top N drug names, each accompanied by the corresponding count of occurrences. The visualization efficiently spotlights the drugs that are most commonly prescribed or referenced, enabling an in-depth analysis of frequently administered drugs in the context of side effects.

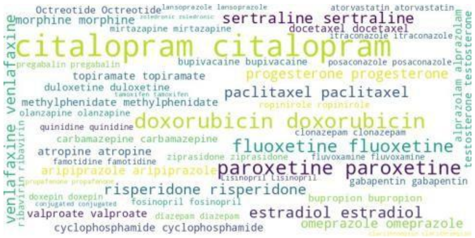


Fig. 1. World cloud Visualization

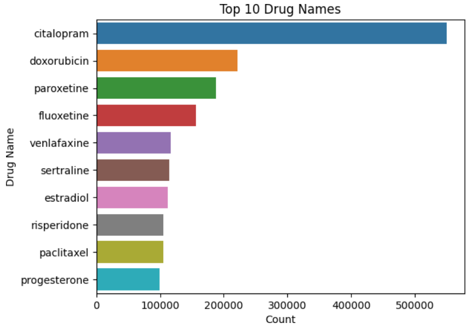


Fig. 2. Drug Count Chart

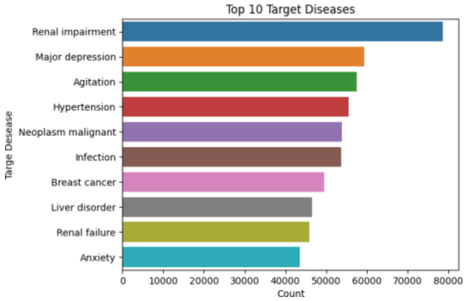


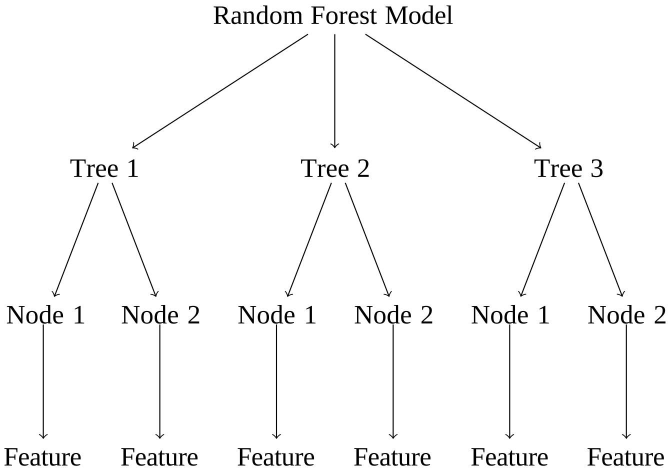
Fig. 3. Indication Count Chart

A. *Abbreviations and Acronyms*

- **SMILES:** Simplified Molecular Input Line Entry System.
- **ATC CODE:** Anatomical Therapeutic Chemical.
- **PharmGKB:** Pharmacogenomics Knowledge Base.

IV. Random Forest Model

A Random Forest is an ensemble machine learning model that combines multiple decision tree classifiers to make predictions. Each decision tree in the forest contributes to the final prediction, and the ensemble approach enhances the model's robustness and generalization [21], [22], [23], [27].



A. *Random Forest Model Interpretation for Drug Side Effect Prediction*

In our Drug Side Effect Prediction model, the Random Forest algorithm employs a tree-based structure to make predictions based on the attributes of our datasets. Each tree within the model consists of nodes and features, which are directly related to the columns of our datasets.

ATC Code:

- Within our Random Forest model, a node may split the data based on the ATC Code. This code represents the Anatomical Therapeutic Chemical Classification, a system for classifying drugs based on their therapeutic and chemical properties.

TABLE III
DATA SUMMARY STATISTICS

	ATC Code	Drug Name	Indication	Side Effect	Type	SMILES
count	6506423	6506423	6506423	6506423	6506423	6506423
unique	622	612	2156	5057	4	619
top	N06AB04	citalopram	Renal impairment	Dizziness	Drug	CN(C)CCCC1(C2=C(CO1)C=C(C=C2)C#N)C3=CC=C(C=C3)F
freq	275310	550620	78667	44429	6023642	275310

Drug Name:

- In another part of the tree, the model might utilize the Drug Name column to make distinctions. The drug's name can provide valuable information about potential side effects.

Indication:

- Indication is another feature that plays a role in our model's predictions. It relates to the specific medical condition or purpose for which a drug is prescribed.

Side Effect:

- The Side Effect column is of utmost importance in our model, as it directly relates to the prediction task. It's highly likely that nodes in the tree will make decisions based on the presence or absence of specific side effects.

Type:

- The Type column can represent various drug types or categories, and it may guide the model in making predictions based on the type of drug.

SMILES:

- The Simplified Molecular Input Line Entry System (SMILES) notation can capture the chemical structure of drugs. Nodes in the tree may use this information to differentiate between different drugs and their potential side effects. Each node in the tree represents a decision point, where the model chooses how to split the data based on one of these features. This decision process continues down the tree, with each node guiding the model closer to a prediction regarding drug side effects.

The Random Forest model combines the predictions of multiple trees, enhancing the overall accuracy and robustness of our drug side effect predictions. By analyzing the structure of these trees, we gain insights into how the model leverages the attributes of our datasets to make informed decisions about potential drug side effects.

It's important to note that the exact formulas and decision boundaries in the nodes are determined by the training process of the Random Forest algorithm, and the interpretation provided here is based on the general principles of how Random Forest models work.

- **Bootstrapping:** The process of creating subsets of the training data for each tree is called bootstrapping. It involves sampling with replacement, which results in each subset having some duplicate data points [28].
- **Voting or Averaging:** For classification tasks, each tree predicts the class label, and the final prediction is determined by majority voting among the trees. For regression tasks, the final prediction is the average of the predictions from all trees.
- **Feature Importance:** Random Forest provides a measure of feature importance, indicating which features contribute most to the model's predictions. This is valuable for understanding the model's decision-making process.

C. Model Evaluation

Model performance evaluation is a critical aspect of assessing the effectiveness of machine learning models, Several metrics have been used for our side effect prediction.

Accuracy: Accuracy is a widely used metric to measure the overall correctness of a classification model. It calculates the ratio of correctly predicted instances to the total number of instances in the datasets.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision: Precision measures the accuracy of positive predictions. It quantifies the proportion of true positive predictions out of all positive predictions made by the model.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall (Sensitivity or True Positive Rate): Recall assesses the model's ability to identify all relevant instances in the datasets. It calculates the proportion of true positive predictions out of all actual positive instances.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-Score:

$$F1 - Score = \frac{2 \cdot (Precision \cdot Recall)}{Precision + Recall} \tag{4}$$

B. Key characteristics of a Random Forest include:

- **Decision Trees:** The Random Forest consists of multiple decision trees. Each decision tree is trained on a random subset of the training data, and at each node, it selects the best split among a random subset of features. This introduces diversity among the trees [24], [25], [26].
- **TP (True Positives):** The number of correctly predicted positive instances.
- **TN (True Negatives):** The number of correctly predicted negative instances.
- **FP (False Positives):** The number of instances predicted as positive but are actually negative.
-
- **FN (False Negatives):** The number of instances predicted as negative but are actually positive.

In the context of our Drug Side Effect Prediction model, these metrics collectively underscore its strong performance. It excels in correctly predicting the presence or absence of drug side effects and effectively capturing the actual positive cases. Such performance is invaluable for real-world applications where the accurate identification of side effects is paramount for patient safety and treatment efficacy.

It is crucial to consider the specific goals and requirements of our model's application. Depending on the domain and the relative significance of false positives and false negatives, we may fine-tune the model to achieve the desired balance between precision and recall. Overall, with an accuracy of 85% and robust precision, recall, and F1-Score values, our Drug Side Effect Prediction model proves to be highly effective in its task.

V. RESULTS

Our model **Accuracy**, which measures the overall correctness of our model's predictions, stands at an impressive 85%. This indicates that our model correctly predicts drug side effects in 85% of cases, demonstrating its reliability.

Our model's **precision** is 84%, signifying that when it predicts a drug to have a specific side effect, it is accurate 84% of the time. This metric is vital as it reflects the model's ability to minimize false positive predictions.

The **recall**, or sensitivity, score of 88% showcases the model's capability to identify 88% of actual drug side effects. In other words, it effectively captures most of the true positive cases.

The F1-Score, at 86%, provides a balanced assessment of our model's performance. It demonstrates the model's ability to strike a harmonious balance between making accurate positive predictions (precision) and capturing a significant proportion of positive cases (recall).

REFERENCES

[1] Wang, Xiaoyan and Hripcsak, George and Markatou, Marianthi and Friedman, Carol, "Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study," BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, vol. 16, pp. 328–337, April 1955.

[2] Irwin D. Kuntz, Structure-Based Strategies for Drug Design and Discovery, American Association for the Advancement of Science, 1892, pp.1078–1082.

[3] Sorrentino R., "Exploring the relationship between drug side-effects and therapeutic indications.,"AMIA Annu Symp Proc. 2013;2013:1568- 1577., 2013 Nov 16.

[4] Soohyeon Lee and Sangjoon Shin and Lynn Howie and Hyunjin Jung and Steven Yoo and David Hong, "740 A first-in-human trial of hSTC810 (anti-BTN1A1 Ab), a novel immune checkpoint with a mutually exclusive expression with PD-1/PD-L1, in patients with relapsed/refractory solid tumors," BMJ Specialist Journals, vol. 2, pp. 740–741, August 1987 Journal for ImmunoTherapy of Cancer, p. A773–A773, 2022.

[5] Lu, S., Pan, H., Wu, L. et al., Efficacy, safety and pharmacokinetics of Unecritinib (TQ-B3101) for patients with ROS1 positive advanced non-small cell lung cancer Mill Valley, a Phase I/II Trial. Sig Transduct Target Ther 8, 249., 2023.

[6] Amiri, Marjan and Michel, Martin C, "Expectations and satisfaction of academic investigators in nonclinical collaboration with the pharmaceutical industry," Naunyn-Schmiedeberg's Archives of Pharmacology, vol. 338, pp. 613–622, 2015.

[7] Yildirim, Oktay and Gottwald, Matthias and Schu¨ler, Peter and Michel, Martin C, "Opportunities and challenges for drug development: public–private partnerships, adaptive designs and big data," Frontiers Media SA, vol. 7, pp. 461, 2016.

[8] Uddin, Shahid. "Chapter 25: Peptide Drug/Device Combinations." Development of Biopharmaceutical Drug-Device Products (2020): 613-637.

[9] De Rycker, Manu, et al. "Challenges and recent progress in drug discovery for tropical diseases." Nature 559.7715 (2018): 498-506.

[10] VC Guido, Rafael, Glaucius Oliva, and Adriano D Andricopulo. "Modern drug discovery technologies: opportunities and challenges in lead discovery." Combinatorial chemistry & high throughput screening 14.10 (2011): 830-839.

[11] Tatonetti, Nicholas P., Tianyun Liu, and Russ B. Altman. "Predicting drug side-effects by chemical systems biology." Genome biology 10.9 (2009): 1–4.

[12] Pauwels, Edouard, Ve´ronique Stoven, and Yoshihiro Yamanishi. "Predicting drug side-effect profiles: a chemical fragment-based approach." BMC bioinformatics 12.1 (2011): 1-13.

[13] Pauwels, Edouard, Ve´ronique Stoven, and Yoshihiro Yamanishi. "Predicting drug side-effect profiles: a chemical fragment-based approach." BMC bioinformatics 12.1 (2011): 1-13.

[14] Pauwels, Edouard, Ve´ronique Stoven, and Yoshihiro Yamanishi. "Predicting drug side-effect profiles: a chemical fragment-based approach." BMC bioinformatics 12.1 (2011): 1-13.

[15] Kapsiani, Sofia, and Brendan J. Howlin. "Random forest classification for predicting lifespan-extending chemical compounds." Scientific reports 11.1 (2021): 13812.

[16] Thorn, Caroline F., Teri E. Klein, and Russ B. Altman. "PharmGKB: the pharmacogenomics knowledge base." Pharmacogenomics: Methods and Protocols (2013): 311-320.

[17] Kuhn, Michael, et al. "The SIDER database of drugs and side effects." Nucleic acids research 44.D1 (2016): D1075-D1079.

[18] Wishart, David S., et al. "DrugBank 5.0: a major update to the DrugBank database for 2018." Nucleic acids research 46.D1 (2018): D1074-D1082.

[19] Miller, G. C., and H. Britt. "A new drug classification for computer systems: the ATC extension code." International journal of bio-medical computing 40.2 (1995): 121-124.

[20] Kalyaanamoorthy, Subha, and Yi-Ping Phoebe Chen. "Structure-based drug design to augment hit discovery." Drug discovery today 16.17-18 (2011): 831-839.

[21] Paul, Angshuman, et al. "Improved random forest for classification." IEEE Transactions on Image Processing 27.8 (2018): 4012-4024.

[22] Liaw, Andy, and Matthew Wiener. ”Classification and regression by randomForest.” R news 2.3 (2002): 18-22.
[23] Denisko, Danielle, and Michael M. Hoffman. ”Classification and interaction in random forests.” Proceedings of the National Academy of Sciences 115.8 (2018): 1690-1692.
[24] Liu, Yingchun. ”Random forest algorithm in big data environment.” Computer modelling & new technologies 18.12A (2014): 147-151.
[25] Ali, Jehad, et al. ”Random forests and decision trees.” International Journal of Computer Science Issues (IJCSI) 9.5 (2012): 272.
[26] Ali, Jehad, et al. ”Random forests and decision trees.” International Journal of Computer Science Issues (IJCSI) 9.5 (2012): 272.
[27] Jain, Nishant, and Prasanta K. Jana. ”LRF: A logically randomized forest algorithm for classification and regression problems.” Expert Systems with Applications 213 (2023): 119225.
[28] Shah, Kanish, et al. ”A comparative analysis of logistic regression, random forest and KNN models for the text classification.” Augmented Human Research 5 (2020): 1-16.
[29] REDDY ENUMULA, Raveendra, and Rama KRISHNA RAO. ”Alzheimer’s disease prediction and classification using CT images through machine learning.” Bratislava Medical Journal/Bratislavske Lekarske Listy 124.5 (2023).