EasyChair Preprint

№ 14534

Predicting Drug-Drug Interactions with Machine-Learning and ATC-SMILES Combined Representation

Yasmin Radwan, Karam Abdelghany Gouda,
Ibrahim Zaghloul Abdelbaky and Mona Mohamed Arafa

August 26, 2024

Predicting Drug-Drug Interactions with Machine-Learning and ATC-SMILES Combined Representation

Yasmin Atef Radwan*, Karam Abdelghany Gouda†, Ibrahim Zaghloul Abdelbaky†, and Mona Mohamed Arafa†
*Information Systems Department, Higher Institute of Computer Science & Information Technology, El-Shorouk Academy, Egypt
Email: Yasmin.Atef.Radwan@gmail.com
†Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, Egypt
Email: {karam.gouda, ibrahim.abdelbaky, mona.abdelmonem}@fci.bu.edu.eg

***Abstract*—Polypharmacy is a potential strategy for managing such intricate disorders, encompassing conditions like cancer, diabetes, and age-related issues in older individuals. Nonetheless, when a medication is combined with one or more drugs that either enhance, diminish, or counteract its intended effects, it can lead to undesired adverse reactions. In severe cases, these interactions can cause serious morbidity and increased mortality rates globally. In this study, we collected a Drug-Drug interaction dataset from the DrugBank database. Various chemical features were then extracted from the Simplified Molecular-Input Line-Entry System of interacting drug pairs. Our emphasis was on representing the Molecular Access System fingerprints of these drug pairs. Molecular Access System fingerprints signify the presence or absence of specific substructures in the molecule and were generated using the RDKit Open-Source Cheminformatics Software. Furthermore, we incorporated Anatomical Therapeutic Chemical classifications into our analysis. Finally, we employed various machine learning algorithms, using K-Nearest Neighbor, Random Forest, Logistic Regression, Support Vector Machine, and XGBoost for learning the extracted features and predict large-scale Drug-Drug interactions among various drug pairs. Among these models, the XGBoost model exhibited superior performance across most measurement metrics.**

***Index Terms*—Drug-Drug interactions, Simplified Molecular-Input Line-Entry System, Anatomical Therapeutic Chemical features, Molecular Access System**

I. Introduction

Drug-drug interactions (DDIs) are a critical health and safety concern that receives much attention from both academia and business [1]. DDIs may Enhance the likelihood of unexpected negative consequences and unidentified toxicity, bringing humans at threat. [2]. It is expensive and takes a while to recognize DDIs and Adverse Reactions(ARs) among several medication pairings, both in vivo and in vitro [3]. Most diseases in patients are brought on by intricate biological mechanisms that cannot be cured by one particular medication and often require multiple medications. It is unrealistic and difficult to locate every potential DDI for medicine in the initial manufacturing stages [4].

DDIs are usually assessed throughout medicinal chemistry, but most interactions remain unnoticed and produce many medications and pairings [5]. DDIs are more common among older people who have to receive ongoing therapy for one or more chronic conditions. Many ADRs are not discovered during clinical investigations before the medicine is given government approval. Analyzing DDIs by Researchers can anticipate potential, previously unidentified DDIs and investigate their correlations to pharmacodynamics and pharmacokinetic drug properties [6].

The possibility of interaction among prescribed drugs rapidly grows as the number of approved medicines grows. In general, several drugs are given to aged people and cancer survivors, putting them at a significant risk for harmful DDIs [7]. Detecting and identifying DDIs not only helps clinicians avoid chronic but will also encourage the co-prescription of safe drugs for more effective therapies. When a medicine is already on the market or in a clinic, most DDIs are unexpectedly discovered, so it is urgent before medicine is put on the market that people should be aware of any potential dangers [1].

Different types of drug interactions include pharmaceutical interactions, pharmacodynamics (PD) interactions, and pharmacokinetics (PK) interactions. Pharmaceutical interactions occur due to chemical reactions caused by improper dispensing before drugs are administered. For example, mixing tetracycline and calcium salt injection can lead to precipitation due to chelate formation in specific conditions. PD interactions happen when two drugs share the same receptor, altering each other's pharmacological effects. For instance, the combination of atropine and tubocurarine blocks the action of acetylcholine by binding to receptors. PK interactions take place since multiple medications are utilized together, influencing how they're absorbed, dispersed, metabolized, and eradicated in the human organism. These interactions can either improve the effectiveness of the drugs or lead to adverse reactions. Unlike PD interactions PK interactions primarily affect the blood concentration of the drugs involved. For instance, combining warfarin with nonsteroidal anti-inflammatory drugs can result in PK interactions [8].

The Simplified Molecular Input Line Entry System (SMILES) is a method used in chemistry to represent the

structure of a chemical compound in a compact, textual format using a standardized set of ASCII characters. This method allows for the concise representation of molecular structures, enabling efficient storage, transmission, and processing of chemical data. It provides a compact and human-readable way to encode molecular structures, making it easier to store, search, and manipulate chemical information computationally. SMILES incorporates certain rules and conventions to ensure consistency and accuracy in representing molecular structures. SMILES provides a standardized and versatile method for encoding molecular structures, facilitating their representation and analysis in various computational applications within the field of chemistry. Molecular Access System (MACCS) keys are a set of structural keys used for chemical similarity searching and structure-activity relationship analysis. They're used in cheminformatics to encode molecular structures into a series of binary fingerprints.

Anatomical Therapeutic Chemicals (ATC-Code) is structured into seven fragments, with each fragment representing a specific level of classification and denoted by a letter as follows: The Simplified Molecular Input Line Entry System (SMILES) is a method used in chemistry to represent the structure of a chemical compound in a compact, textual format using a standardized set of ASCII characters, enabling efficient storage, transmission, and processing of chemical data. ATC classification system is a widely used system for categorizing pharmaceutical drugs, with main drug classes including Alimentary Tract and Metabolism, Blood and Blood-Forming Organs, Cardiovascular System, Dermatologicals, Genitourinary System and Sex Hormones, Systemic Hormonal Preparations, Anti-infectives for Systemic Uses, Antineoplastic and Immunomodulating Agents, Musculoskeletal System, Nervous System, Antiparasitic Products, Insecticides, Repellents, Respiratory System, Sensory Organs, and Various [9].

This study presents an enhanced method for predicting DDIs by utilizing a suggested combination of features and Machine learning techniques. Our feature set is extracted from two drug representation schemes: MACCS fingerprints and ATC-Code. The paper is structured as follows: first, Section II provides a literature survey that addresses the issue of drug-drug interactions (DDIs); next, Section III outlines the materials and methods used in the study, including a description of the techniques employed for feature extraction, representation, and the application of various machine learning techniques to the embedded drug features; the methodology is then presented in Section IV; the experimental results are reported in Section V; and finally, the paper concludes in Section VI with a summary of the overall methodology.

II. LITERATURE SURVEY

When it comes to predicting drug-drug interactions (DDIs), there are multiple primary approaches, each employing a different set of methodologies and techniques. These approaches can be categorized as follows:

A. Literature-Based Approach

Text-mining tools leverage natural language processing techniques to identify meaningful associations among medications. These tools utilize text-mining methodologies to explore and compile documented DDIs sourced from diverse databases such as procurement claims, the FDA Adverse Event Report, and electronic medical records [2]. An alternative approach centers on identifying and categorizing pharmacologic substances, including drug names, brand names, group names, and active substances not approved for human use. This method extracts DDIs from sources like DrugBank and MedLine abstracts corpus using non-linear kernel techniques [10]. The researchers have evaluated the efficiency of various machine learning classifiers, including Logistic Regression (LR), Support Vector Machines (SVM), and discriminatory analysis, to identify relevant abstracts and PubMed articles that confirm the existence of chemical drug-drug interactions (DDIs) [11]. Moreover, their methodology enables the connection of causal processes to potential DDIs. In this methodology, the researchers leverage a parsing tree structure to extract various types of interactions between drugs. Building upon this, they then apply a set of logical rules to predict potential interactions between novel drugs and existing drugs based on the identified interaction patterns [12].

B. Similarity-Based Approach

The similarity-based approach is a widely used framework that aims to measure the degree of similarity or distance between data points, with the assumption that similar data points will have similar outcomes. In the context of DDIs, this approach operates on the idea that medications with similar chemical properties or structures are likely to have similar interaction patterns with other drugs. By considering the chemical similarities between drugs, this method can be used to identify potential DDIs and better understand the effects of drug combinations [6]. To measure the similarity between medications, common substructures are potentially used instead of entire chemical structures [4]. If medications both A and B interact to generate a particular response, then medications similar to drug A (or drug B) are probable to generate an identical impact as drug B (or drug A). In terms of medication similarity, interactions among novel drugs can be predicted by combining similar properties among various drugs [13].

C. Classification-Based Approach

In the traditional approach, the problem of predicting drug-drug interactions (DDIs) is framed as a binary classification task, where the objective is to classify each drug pair as either "interacting" or "non-interacting". To enhance the accuracy and reliability of predicting drug pairs using both molecular and pharmacological features, researchers have developed a probability ensemble method [14]. Additionally, the conventional approach of leveraging similarity-based and classification-based techniques can be employed to predict

unknown drug-drug interactions. However, when solely relying on these approaches, the features of drugs and their interactions may not effectively collaborate with the known interactions, leading to inaccurate predictions. Therefore, more advanced computational techniques are needed to improve the prediction of unknown drug interactions [15].

D. Graph-Based Approach

The primary goal of graph embedding techniques is to represent a graph, with all its structural details, in a compact low-dimensional vector form. This allows the rich information captured in the graph structure to be effectively encoded and utilized for various downstream applications. The graph-based encoding of the drugs and their relationships appears to capture important information that enables more accurate DDI predictions compared to approaches that do not leverage the graph structure [15]. Deep learning techniques have proven effective in extracting drug features from datasets and conducting self-training through multiple layers of the neural network to predict previously unidentified DDIs. It proposed a DNN-based approach that constructs an architecture utilizing various types of drug data. By encoding SMILES as low-dimensional vectors using one-hot encoding and incorporating topological features acquired from a knowledge graph (using GNN, they achieved an accurate prediction of unidentified DDIs [16].

E. Ensemble-Based Approach

An ensemble-based approach is utilized for drug-drug interaction prediction by integrating predictions obtained from various independent models to enhance the overall accuracy as well as the resilience of the predictions. For multiple-label DDI prediction, three key processes are included. The first step is the creation of a knowledge graph using the four knowledge graphs that are established in Bio2RDF (DrugBank, KEGG, PharmaGKB, and Comparative Toxicogenomics Database). Secondly, in addition to the drug KG, biological DDI text which is composed of DDI documents from DrugBank and MEDLINE, Abstracts were embedded into a Low-Dimensional vector. Third, DDI prediction is effectively computed using the learned embedding as a link prediction methodology [17].

III. MATERIALS AND METHODS

A. Dataset

DrugBank is the primary data source used in this study, as our dataset was obtained from DrugBank (version 5.0, released on August 1, 2017). DrugBank 5.0 offers extensive data on a broad spectrum of drugs involving both approved medications and exploratory compounds, as well as the most recent drug-related data. It additionally provides detailed information concerning the binding proteins, enzymes, and receptors in the body, associated pathways and biological functions, and additional molecular targets with which drugs interact. The DrugBank dataset, which includes more than 4,100 drug entries, provides comprehensive information on drug interactions and includes data on pharmacokinetic and pharmacodynamic interactions. Additionally, the dataset includes 365,984 direct interactions among these drugs [18].

B. Feature Extraction and Representation

We use two schemes for drug representations to build up our feature set: MACCS fingerprints and ATC-Code. In MACCS fingerprints, drugs are represented based on the presence or absence of particular sub-structures in the chemical molecule. We calculated MACCS fingerprints using the RDKit package in Python [19]. The inputs to the RDKit are SMILES strings of drugs extracted from PubChem [20]. This step results in 166-dimensional MACCS fingerprints.

ATC-Code is a one-of-a-kind code assigned to each medicine based on the organ or system it works on and how it works. ATC-Code was extracted from the WHO Collaborating Centre for Drug Statistics Methodology (WHOCC). ATC-Code is structured into seven fragments; the first fragment, known as the anatomical main group, represents the broadest category. It classifies drugs based on their primary anatomical or therapeutic area of action. The second fragment, the therapeutic subgroup, is represented by a numeric value. This fragment provides more specific information about the pharmacological or therapeutic properties of the drug. The third fragment, the pharmacological subgroup, is represented by an alphanumeric code and describes the drug's pharmacological characteristics. The fourth fragment, the chemical subgroup, is also represented by an alphanumeric code and provides information on the drug's chemical structure or chemical class. The fifth fragment, the chemical substance, designates a specific drug or active ingredient. It is represented by an alphanumeric code, often derived from the drug's generic name. The sixth fragment, the formulation level, indicates the drug's formulation or presentation. The seventh and final fragment is used for additional classification purposes or to provide more specific information about the particular drug formulation.

Using one-hot encoding, a method for encoding categorical variables into binary vectors, the ATC-Code can be expressed. A category is any ATC-Code fragment represented by a letter or numerical code. In this vector, each position represents a unique category, and the value 1 indicates the presence of that category in the ATC-Code. The encodings of chemical features described by the MACCS fingerprints and the ATC-Code one hot encoding features are then concatenated to construct the final feature vector for the modeling phase [21].

C. Data Preparation

Many drugs in commonly utilized pharmaceutical databases do not have ATC-Codes. In this study, we chose drugs that have SMILES and ATC-Codes, but some drugs do not have an ATC-Code, so we eliminated such records from the entire dataset, ending up with 207,096 Drug-Drug interactions with both SMILES and ATC-Codes. The statistics of our dataset are as follows: it consists of 207,096 drug-drug interactions (DDIs) with both SMILES and ATC-Codes available. Out of these interactions, there are 127,220 positive labels indicating

the presence of DDIs, while there are 26,009 negative labels indicating the absence of DDIs. Data imbalance is a common issue in modeling problems, where the imbalance in the input classes makes predictive categorization in machine learning difficult.

D. Balancing Dataset Classes

Data imbalance occurs when the proportion of classes or categories inside a dataset is unequal or skewed. This signifies that several classes had more or substantially fewer instances than others. This imbalance can cause problems with various data-driven activities, such as categorization, prediction, and model-based machine learning training. To address the class imbalance issue, Over-sampling was employed as a technique to balance the positive and negative classes. The Synthetic Minority Oversampling Technique (SMOTE) technique tackles class imbalance by generating synthetic examples for the minority class, thereby enhancing the performance of machine learning algorithms when dealing with imbalanced datasets [8]. SMOTE technique randomly chooses a sample from the minority class and determines its nearest neighbors. It then generates synthetic examples by filling in the gaps along the line segments connecting the selected sample to its neighbors. This is done by interpolating the feature values based on the existing samples. In simpler terms, SMOTE creates new minority class examples by blending the characteristics of existing samples. This approach allows for a more equitable representation of positive and negative instances, potentially leading to improved outcomes in predicting DDIs accurately [22].

E. Feature Selection

In the process of building a predictive model, it is essential to carefully choose the most important features from a given set. Recursive Feature Selection (RFE) is a commonly used technique in machine learning for this purpose. Its goal is to identify the most relevant features in a dataset [13]. In our study, we utilized the RFE-DT algorithm, which can be divided into four stages. Initially, a Decision Tree (DT) is trained using the training set. Then, the features are ranked based on the weights derived from the resulting classifier. Subsequently, the features with the smallest weights are removed. Finally, the process is repeated on the training set with the remaining features [23]. To optimize the modeling process and improve overall performance, we utilized the RFE-DT algorithm to extract a subset of features from the initial pool of 768 features. This feature selection technique aimed to identify the most relevant and informative features that would contribute significantly to the predictive modeling task. By narrowing down the feature set, we aimed to streamline the analysis and enhance the efficiency and accuracy of our models.

IV. METHODOLOGY

The DDI prediction process in our study revolves around two input features: SMILES and ATC-Code. These features represent the molecular structure and therapeutic classification of the two interacting drugs, respectively. The goal is to generate a binary output prediction indicating whether there is an interaction between the drug pairs or not.

As shown in the Fig. 1, Our Proposed Contribution includes the following steps:

- Extracting the Simplified Molecular Input Line-Entry System of drug pairs from PubChem.
- SMILES is a string-based depiction of a chemical compound's molecular structure. It is tight and understandable by humans which enables the special encoding of molecular structures. SMILES syntax enables the representation of molecular structures distinctly.
- Calculating MACCS fingerprints (166 bits): Each bit position represents the presence or absence of structural fragments encoded by RDKit. The bit vector is a string of one (1) and zero (0) characters, with each character representing the state of a single bit. The figure above shows how to use RDKit to represent a molecular structure to a MACCS fingerprint.
- Extracting the anatomical therapeutic chemical code, which was extracted from the WHO Collaborating Centre for Drug Statistics Methodology.
- Representing ATC-Codes of Drug pairs as low-dimensional vectors via one-hot encoding. The methodology of one-hot encoding has been employed for expressing categorical parameters just like binary vectors.
- One-Hot-Encode function accepts the category just like an input and creates a binary vector of zeros. If a particular category is found in the terms, it recognizes its associated index and initiates the vector's value at that position to 1.
- The encodings of chemical features represented by the MACCS fingerprints in addition to ATC-Code one hot encoding features are then concatenated to construct the final feature vector for the modeling phase.
- Applying Recursive Feature Selection with Decision Tree to the training dataset.
- Developing various machine learning models that predict Drug-Drug interactions.

A. Development of Machine Learning Models

In our study, we employed various machine learning methods to predict DDIs, including Support Vector Machines [24], k-Nearest Neighbors [4], Logistic Regression [16], Random Forest [6], and Extreme Gradient Boosting [25]. In order to train and evaluate the models, the input data was divided into a training set and a test set. The training set comprised 70% of the data, which was used to train the models. The remaining 30% of the data was reserved as the test dataset, which was used to assess the performance of the trained models.

- Random Forest
  Random Forest is an ensemble-based learning methodology that is widely used for a variety of tasks, including classification, regression, and others [16]. our study utilized Random Forest for binary classification tasks,

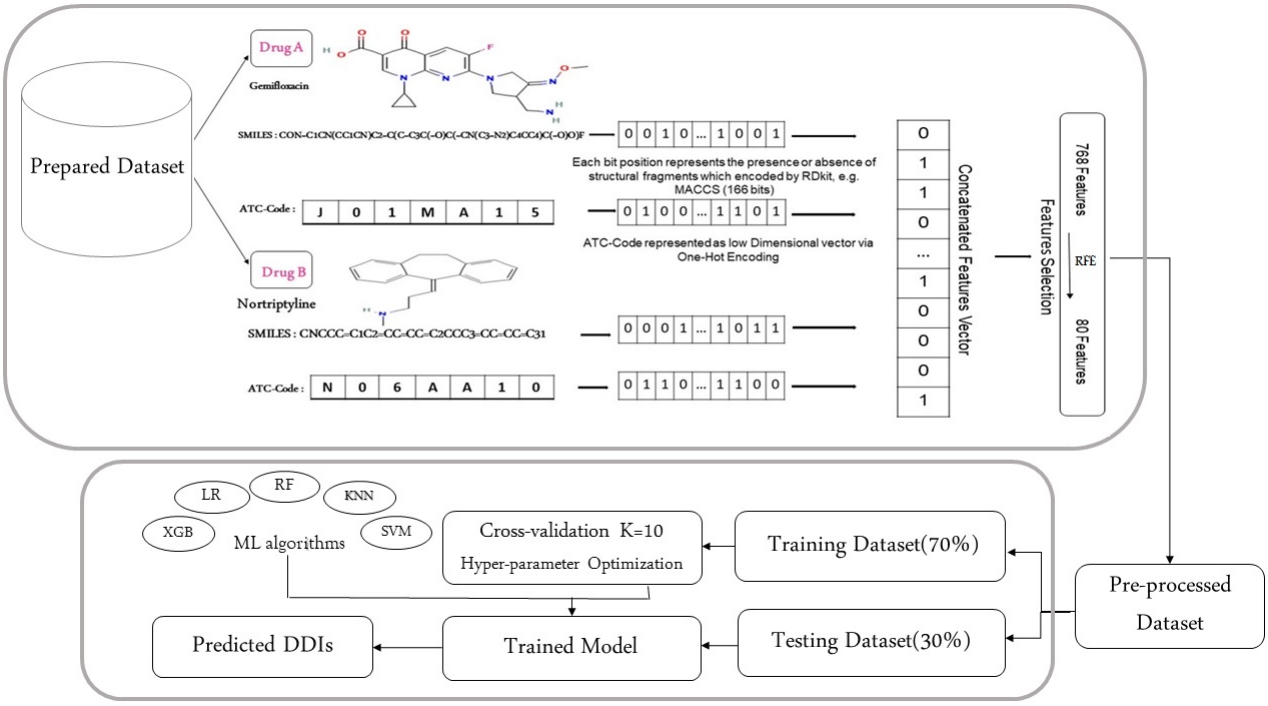

Fig. 1. Architecture of the Proposed Framework for DDIs Prediction through ML Models

specifically in predicting drug-drug interactions (DDIs). Positive and negative DDI pairs were used as inputs to construct the classification model, primarily relying on known interactions between drug pairs [15].

- K-Nearest Neighbors
  The K-Nearest Neighbors (KNN) algorithm is a supervised machine learning technique that can be used for both classification and prediction tasks. It is a versatile algorithm that classifies new data points based on their similarity to existing data [4]. The key idea behind KNN is that similar data points tend to be located in close proximity to each other. The algorithm stores the available training data and calculates the distances between the new data point and the existing data points. It then identifies the K nearest neighbors, where K is a predetermined parameter. The majority class among these K nearest neighbors is used to determine the classification of the new data point. By relying on the proximity and similarity of data points, the KNN algorithm is able to effectively classify and predict new samples based on the characteristics of the nearest known data points. [24].
- XGBoost
  In drug-drug interaction (DDI) prediction, XGBoost, along with decision trees and random forests, plays a significant role in classifying positive and negative DDI pairs. It can enhance weak models and improve predictive performance, making it valuable for DDI prediction [16].
- Support Vector Machine
  The Support Vector Machine (SVM) algorithm is a supervised machine learning technique that can be applied to both classification and regression problems. SVM assigns each training vector to a class based on its position relative to the hyperplane. The optimal hyperplane is the one that maximizes the margin between classes, even though there may be multiple hyperplanes that correctly classify all elements in the feature set [24].
- Logistic Regression
  Logistic regression is a supervised machine learning algorithm used for binary classification tasks. In our study, logistic regression was employed to predict drug interactions. By utilizing an input dataset, logistic regression calculates the probability of DDIs [16].

To identify the optimal values for the model hyperparameters, we conducted parameter tuning for each machine-learning method. The optimal parameter values for each method are presented in Table I. To ensure robust evaluation, we employed the Cross-Validation technique with k-fold validation. In this approach, we divided the preprocessed data into k equal-sized subsets, with k set to 10 in our study. This allowed us to perform training and evaluation iteratively, ensuring a comprehensive assessment of the model's performance.

B. Evaluation

To assess the effectiveness of the machine learning model in the training dataset, we utilized 10-fold cross-validation (10-CV). Equal-sized subsets of the retrieved features from known DDIs were chosen at random. Our models' performance is

TABLE I
HYPERPARAMETER SEARCH GRID AND THE OPTIMAL VALUE OF XGBOOST ALGORITHM

Algorithm	*Hyperparameter*	*Grid search range*	*Optimal value*
XGBoost	n-estimators	20,40,60,80,100,200,500,1000,1500	1500
	Max-depth	10,20,30,40,50,None	30
	gamma	0.5,1,1.5,2,5	2
	min-child-weight	10,15,20,25	20

evaluated using various performance metrics, including Accuracy, F1-Score, Precision, and Recall.

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

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

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

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

In the context of the evaluation metrics used in this study, the following definitions apply: True Positive (TP): The number of drug-drug interactions (DDIs) that were correctly predicted as positive. True Negative (TN): The number of non-interacting drug pairs that were correctly predicted as negative. False Positive (FP): The number of non-interacting drug pairs that were incorrectly predicted as positive. False Negative (FN): The number of actual DDIs that were incorrectly predicted as negative.

Recall is defined as the fraction of correctly predicted DDIs (true positives) divided by the total number of true DDIs. It measures the model's ability to correctly identify positive instances, indicating how many of the actual DDIs were predicted correctly.

Precision is defined as the fraction of correctly predicted DDIs (true positives) divided by the total number of predicted DDIs (true positives + false positives). It measures the accuracy of the positive predictions, indicating how many of the predicted DDIs were true.

The F1-score is the harmonic mean of precision and recall. It provides a balanced measure of the model's performance by considering both precision and recall. The F1-score is often used to evaluate the overall effectiveness of the prediction outcomes.

V. RESULTS

We conducted experiments on training and testing datasets to evaluate the impact of these algorithms on the results. Based on the results shown in Table II, it can be observed that XGBoost achieves superior performance when applied to the selected features of the combined SMILES and ATC-Codes dataset. These results indicate that XGBoost outperforms other algorithms across various measurement metrics.

TABLE II
RESULTS OF DIFFERENT ML ALGORITHMS

Algorithm	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Logistic Regression	0.813	0.806	0.809	0.811
k-Nearest Neighbors	0.861	0.856	0.860	0.858
Support Vector Machine	0.909	0.916	0.907	0.905
Random Forest	0.916	0.920	0.919	0.908
XGBoost	**0.945**	**0.940**	**0.941**	**0.939**

A comprehensive comparison of different machine learning algorithms is presented in Fig. 2, demonstrating their performance on the combined SMILES and ATC-Code datasets. The results indicate that XGBoost outperforms the other algorithms, exhibiting superior performance.

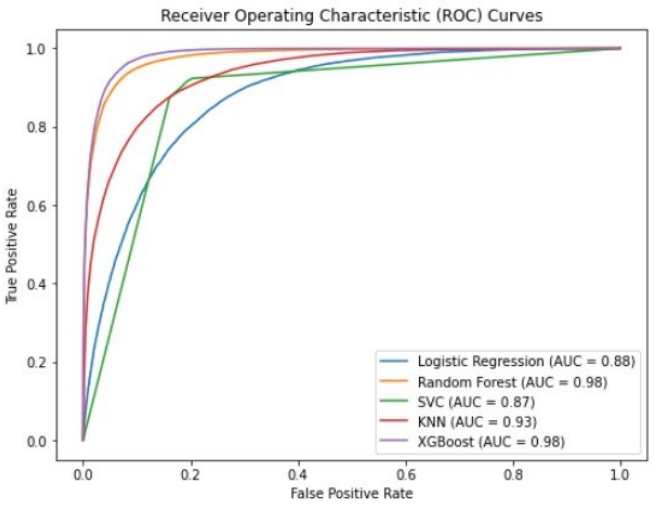

Fig. 2. XGBoosting Model Superior Compared to other ML Models

The experiment was conducted using only the features of SMILES and ATC-Code individually, as well as by combining the two features. The best outcome was achieved when the experiment was run using both features together. These combined features are presented in Table III, along with the corresponding measurement metrics.

TABLE III
RESULTS OF DIFFERENT ML ALGORITHMS

Algorithm	*Accuracy*	*Precision*	*Recall*	*F1-Score*
SMILES	0.907	0.909	0.901	0.906
ATC-Code	0.928	0.920	0.919	0.907
Our Study	**0.945**	**0.940**	**0.941**	**0.939**

We carried out a comparison between the results of the XGBoost model and other methods presented in Table IV, including Vilar's methods, the label propagation method, and Stacked RF-XGBoost. Vilar's methods utilized drug interaction profile fingerprints (IPFs) to predict DDIs [26] [27]. These methods employ similarity measurements and classify input from known DDIs to previously unknown nodes by calculating drug similarity. This approach produces weight values for

edges on the DDI network. The label propagation method focuses on predicting DDIs by using similarity measurements. It classifies input from known DDIs to unknown nodes by calculating drug similarity and determining the weight values of edges on the DDI network [1]. Stacked RF-XGBoost is a method specifically designed to predict DDIs between osteoporosis and Paget's disease [24]. This method focuses solely on the SMILES representations of drugs to predict DDIs.

TABLE IV
PERFORMANCE COMPARISON ON DRUGBANK DATASET

Method	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Vilar 1 [26]	0.719	0.253	0.495	0.334
Vilar 2 [27]	0.862	0.515	0.569	0.540
LP [1]	0.809	0.729	0.685	0.706
Stacked RF-XGBoost [24]	0.740	0.730	0.730	0.730
Our Study	**0.945**	**0.940**	**0.941**	**0.939**

The comparison results illustrate the superior performance of our model across multiple evaluation metrics including Accuracy, Precision, Recall, and F1-Score, which is depicted in Fig. 3.

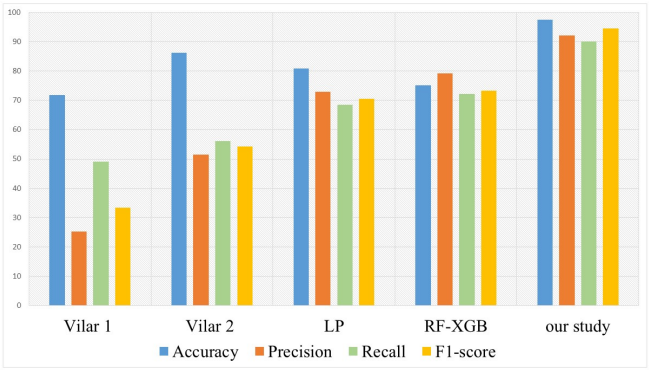

Fig. 3. Comparison Results of Different Studies on DrugBank Dataset

The DrugBank interaction checker is a valuable tool that allows users to evaluate potential interactions between multiple drugs. By utilizing a comprehensive drug database, the DrugBank interaction checker provides information on the severity of each interaction, categorizing them as minor, moderate, or severe. Additionally, it offers recommendations on how to manage these interactions effectively. Table V displays the most recent predictions of the top 20 Drug-Drug Interactions (DDIs) generated using our methodology. Out of these predictions, 15 have been validated using the DrugBank interaction checker website [18].

VI. CONCLUSION

In our study, we proposed an efficient methodology for predicting DDIs by leveraging drug information from different sources. To achieve this, we collected the Simplified Molecular Input Line Entry System representations of drug pairs from PubChem. Furthermore, we acquired the Anatomical Therapeutic Chemical from the WHO Collaborating Centre for Drug Statistics Methodology. For the SMILES of drugs, we utilized the MACCS fingerprint to create binary vectors. The MACCS fingerprint indicates the presence or absence of specific substructures in a molecule, and we employed RDKit to generate these fingerprints for the drug pairs. As for encoding the ATC-Codes, we employed One-Hot Encoding to convert them into binary vectors. The resulting binary vectors for each drug in a pair are then concatenated into a single vector, combining the representations of both drugs. For the prediction task, we employed various machine learning algorithms, including Random Forest, XGBoosting, K-Nearest Neighbor, Support Vector Machine, and Logistic Regression. These algorithms were trained on the extracted features obtained from the MACCS fingerprints and ATC-Code one hot encoding features, using cross-validation to ensure robustness. After evaluating the performance of the different models, we found that the XGBoosting model outperformed the others in terms of most measurement metrics. This indicates that the XGBoosting algorithm was particularly effective in predicting DDIs among various drug pairs.

TABLE V
NEW PREDICTED DDIS (CONFIRMED INTERACTIONS SHOWN IN BOLD)

Rank	Drug1-ID	Drug1-Name	Drug2-ID	Drug2-Name
1	**DB01013**	**Clobetasol propionate**	**DB01072**	**Atazanavir**
2	**DB00699**	**Nicergoline**	**DB00813**	**Fentanyl**
3	**DB00745**	**Modafinil**	**DB01050**	**Ibuprofen**
4	**DB00193**	**Tramadol**	**DB00423**	**Methocarbamol**
5	**DB00433**	**Prochlorperazine**	**DB00962**	**Zaleplon**
6	**DB01028**	**Methoxyflurane**	**DB01351**	**Amobarbital**
7	DB00641	Simvastatin	DB08815	Lurasidone
8	**DB00353**	**Methylergometrine**	**DB08810**	**Cinitapride**
9	DB01179	Podofilox	DB01320	Fosphenytoin
10	**DB00292**	**Etomidate**	**DB01576**	**Dextroamphetamine**
11	**DB00863**	**Ranitidine**	**DB01242**	**Clomipramine**
12	**DB06153**	**Pizotifen**	**DB06237**	**Avanafil**
13	DB04573	Estriol	DB09213	Dexibuprofen
14	DB00227	Lovastatin	DB00883	Isosorbide Dinitrate
15	**DB00496**	**Darifenacin**	**DB00611**	**Butorphanol**
16	**DB00218**	**Moxifloxacin**	**DB00687**	**Fludrocortisone**
17	**DB00502**	**Haloperidol**	**DB01623**	**Thiothixene**
18	**DB01241**	**Gemfibrozil**	**DB06403**	**Ambrisentan**
19	**DB00246**	**Ziprasidone**	**DB09038**	**Empagliflozin**
20	DB00996	Gabapentin	DB09031	Miltefosine

REFERENCES

[1] P. Zhang, F. Wang, J. Hu, and R. Sorrentino, "Label propagation prediction of drug-drug interactions based on clinical side effects," *Scientific reports*, vol. 5, no. 1, p. 12339, 2015.

[2] Y.-H. Feng and S.-W. Zhang, "Prediction of drug-drug interaction using an attention-based graph neural network on drug molecular graphs," *Molecules*, vol. 27, no. 9, p. 3004, 2022.

[3] Y.-H. Feng, S.-W. Zhang, and J.-Y. Shi, "Dpddi: a deep predictor for drug-drug interactions," *BMC bioinformatics*, vol. 21, no. 1, p. 419, 2020.

[4] B. Bumgardner, F. Tanvir, K. M. Saifuddin, and E. Akbas, "Drug-drug interaction prediction: a purely smiles based approach," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 5571–5579.

[5] F. Rafiei, H. Zeraati, K. Abbasi, P. Razzaghi, J. B. Ghasemi, M. Parsaeian, and A. Masoudi-Nejad, "Cfssynergy: combining feature-based and similarity-based methods for drug synergy prediction," *Journal of Chemical Information and Modeling*, vol. 64, no. 7, pp. 2577–2585, 2024.

[6] A. Kastrin, P. Ferk, and B. Leskošek, "Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning," *PloS one*, vol. 13, no. 5, p. e0196865, 2018.

[7] I. Z. Abdelbaky *et al.*, "A comprehensive survey explores drug-drug interaction prediction using machine-learning techniques," *Benha Journal of Applied Sciences*, vol. 9, no. 5, pp. 13–21, 2024.

[8] Y. Zhao, J. Yin, L. Zhang, Y. Zhang, and X. Chen, "Drug–drug interaction prediction: databases, web servers and computational models," *Briefings in Bioinformatics*, vol. 25, no. 1, p. bbad445, 2024.

[9] S. Park, S. Lee, M. Pak, and S. Kim, "Dual representation learning for predicting drug-side effect frequency using protein target information," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[10] M. Dou, J. Tang, P. Tiwari, Y. Ding, and F. Guo, "Drug–drug interaction relation extraction based on deep learning: A review," *ACM Computing Surveys*, vol. 56, no. 6, pp. 1–33, 2024.

[11] M. A. Shamami, M. A. Ilani, and B. Teimourpour, "An optimized deep neural network framework for classification of drug–drug interactions," 2024.

[12] K. Wang, X. Fu, Y. Liu, W. Chen, and J. Chen, "Ptda: Improving drug-drug interaction extraction from biomedical literature based on prompt tuning and data augmentation." *IAENG International Journal of Computer Science*, vol. 51, no. 5, 2024.

[13] B. Shaker, K. M. Tran, C. Jung, and D. Na, "Introduction of advanced methods for structure-based drug discovery," *Current Bioinformatics*, vol. 16, no. 3, pp. 351–363, 2021.

[14] J.-Y. Shi, K. Gao, X.-Q. Shang, and S.-M. Yiu, "Lcm-ds: a novel approach of predicting drug-drug interactions for new drugs via dempster-shafer theory of evidence," in *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE, 2016, pp. 512–515.

[15] K. Han, P. Cao, Y. Wang, F. Xie, J. Ma, M. Yu, J. Wang, Y. Xu, Y. Zhang, and J. Wan, "A review of approaches for predicting drug–drug interactions based on machine learning," *Frontiers in pharmacology*, vol. 12, p. 814858, 2022.

[16] L. H. Dang, N. T. Dung, L. X. Quang, L. Q. Hung, N. H. Le, N. T. N. Le, N. T. Diem, N. T. T. Nga, S.-H. Hung, and N. Q. K. Le, "Machine learning-based prediction of drug-drug interactions for histamine antagonist using hybrid chemical features," *Cells*, vol. 10, no. 11, p. 3092, 2021.

[17] M. Wang, H. Wang, X. Liu, X. Ma, and B. Wang, "Drug-drug interaction predictions via knowledge graph and text embedding: instrument validation study," *JMIR Medical Informatics*, vol. 9, no. 6, p. e28277, 2021.

[18] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda *et al.*, "Drugbank 5.0: a major update to the drugbank database for 2018," *Nucleic acids research*, vol. 46, no. D1, pp. D1074–D1082, 2018.

[19] G. Landrum, "Rdkit: open-source cheminformatics http://www. rdkit. org," *Google Scholar There is no corresponding record for this reference*, vol. 3, no. 8, 2016.

[20] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu *et al.*, "Pubchem 2023 update," *Nucleic acids research*, vol. 51, no. D1, pp. D1373–D1380, 2023.

[21] D. Rahme, M. Ayoub, K. Shaito, N. Saleh, S. Assaf, and N. Lahoud, "First trend analysis of antifungals consumption in lebanon using the world health organization collaborating center for drug statistics methodology," *BMC Infectious Diseases*, vol. 22, no. 1, p. 882, 2022.

[22] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, "Predicting diabetes mellitus using smote and ensemble machine learning approach: The henry ford exercise testing (fit) project," *PloS one*, vol. 12, no. 7, p. e0179805, 2017.

[23] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An overview on the advancements of support vector machine models in healthcare applications: A review," *Information*, vol. 15, no. 4, p. 235, 2024.

[24] T. N. K. Hung, N. Q. K. Le, N. H. Le, L. Van Tuan, T. P. Nguyen, C. Thi, and J.-H. Kang, "An ai-based prediction model for drug-drug interactions in osteoporosis and paget's diseases from smiles," *Molecular informatics*, vol. 41, no. 6, p. 2100264, 2022.

[25] G. Abdurrahman and M. Sintawati, "Implementation of xgboost for classification of parkinson's disease," in *Journal of Physics: Conference Series*, vol. 1538, no. 1. IOP Publishing, 2020, p. 012024.

[26] S. Vilar, R. Harpaz, E. Uriarte, L. Santana, R. Rabadan, and C. Friedman, "Drug—drug interaction through molecular structure similarity analysis," *Journal of the American Medical Informatics Association*, vol. 19, no. 6, pp. 1066–1074, 2012.

[27] S. Vilar, E. Uriarte, L. Santana, N. P. Tatonetti, and C. Friedman, "Detection of drug-drug interactions by modeling interaction profile fingerprints," *PloS one*, vol. 8, no. 3, p. e58321, 2013.