International Journal of Applied Pharmaceutics

ISSN- 0975-7058 Vol 16, Issue 2, 2024

Original Article

NEURAL NETWORK-BASED ADVERSE DRUG REACTION PREDICTION USING MOLECULAR SUBSTRUCTURE ANALYSES

SHIKSHA ALOK DUBEY[1*], PRASHANT S. KHARKAR[2], ANALA A. PANDIT[3]

[1,3]Department of Computer Application, Veermata Jijabai Technological Institute (VJTI), Matunga, Mumbai-400019, India. [2]Department of Pharmaceutical Sciences, Institute of Chemical Technology, Matunga, Mumbai-400019, India
*Corresponding author: Shiksha Alok Dubey; *Email: sadubey_p18@mc.vjti.ac.in

Received: 21 Nov 2023, Revised and Accepted: 08 Jan 2024

ABSTRACT

Objective: This study aims to enhance early detection and prediction by exploiting drug molecular substructures, overcoming challenges posed by limited authentic patient data in the medical domain.

Methods: The study implemented a neural network approach to optimize molecular fingerprint algorithms and employed various machine learning algorithms for predictions. Additionally, the study identified and extracted substructures associated with severe Adverse Drug Reactions (ADRs), validating their presence within drug structures through a comparison with a random set of drug structures. Predictions were made for specific molecular structures, and results were validated using clinical evidence from the literature.

Results: Optimized molecular fingerprint algorithms and diverse machine-learning models yielded promising outcomes. The Area Under Curve (AUC) value for the fingerprint dataset was obtained at approximately 65%, and integrating it with patient data significantly improved the performance by about 30%. Substructure analysis pinpointed key components linked to severe ADRs, reinforcing the predictive prowess of the model. Predictions for specific molecular structures were corroborated using clinical evidence from the literature, fortifying the credibility of the proposed approach.

Conclusion: In conclusion, this research effectively tackles challenges in the early detection and prediction of ADRs by leveraging machine learning algorithms, focusing on drug molecular substructures. The optimized model, incorporating both fingerprint and patient datasets, demonstrated significant improvements in predictive performance. Identifying and validating substructures linked to severe ADRs contribute to the model's reliability. The study's findings are vital for advancing drug safety and laying the groundwork for further strides in predictive modeling within the medical domain.

Keywords: ADR, Machine learning, Neural networks, Substructures, Fingerprints, AUC

DOI: https://dx.doi.org/10.22159/ijap.2024v16i2.49936 Journal homepage: https://innovareacademics.in/journals/index.php/ijap

INTRODUCTION

Drug development is a demanding and time-consuming process, typically spanning a period of approximately 10 to 12 y and involving substantial financial investments, all with no guaranteed outcomes. Clinical trials, which include post-marketing surveillance (phase IV), pose numerous challenges and are often burdensome. As a result, there is constant pressure to reduce the study population size [1] in which the experimental drug is tested. However, due to the controlled and specialized nature of clinical development, the patients involved may not fully represent the original population regarding genetic and physiological makeup. Consequently, there is a possibility of patients experiencing adverse reactions to the drug once it is approved and used by a diverse range of individuals with varying physiological characteristics and clinical disease presentations. Despite these challenges, as mentioned earlier, the primary objective of the healthcare and pharmaceutical industry remains the minimization of ADRs and the assurance of overall drug safety and efficacy in patients.

The World Health Organization (WHO) has defined ADRs as "noxious and unintended responses to drugs occurring at doses normally used in humans for prophylaxis, diagnosis, or therapy of diseases, or for the modification of physiological functions" [2]. In essence, ADRs refer to unexpected drug effects that often lead to hospitalization and fatalities within the patient population. These reactions can be attributed to various factors, including patient-related, drug-related, and social environment-related parameters [3]. Key patient-related factors include age and gender, whereas significant drug-related factors encompass drug dosage and drug-drug interactions, which warrant careful examination to assess the impact of ADRs on human health. Additionally, social environment-related factors such as smoking and alcoholism indirectly contribute to several ADRs. Early detection and prediction of such ADRs during the drug development cycle are crucial for enhancing patient healthcare and overall drug safety.

Adverse reactions can range from mild to severe, and in some cases, they can even be life-threatening. Common ADRs encompass symptoms like nausea, vomiting, diarrhea, dizziness, headache, rash, and fatigue [4]. However, it is important to note that adverse reactions can also lead to serious health complications such as organ damage, allergic reactions, and even mortality. While prescribing the drug, physicians should always be aware of the adverse effect of phenytoin and other many other drugs [5]. Drug structures play a significant role in the occurrence of ADRs. The molecular composition and structural characteristics of a drug can influence its interactions with biological targets in the body, leading to desired therapeutic effects as well as potential adverse reactions [6]. Understanding the relationship between drug structures and ADRs is crucial for drug design, optimization, and safety assessment. Computational methods and structure-activity relationship studies can aid in predicting potential ADRs based on structural features, facilitating the identification and modification of drug candidates to mitigate or minimize the occurrence of adverse reactions.

The present research aimed at relating chemical structures, in particular specific substructures, to the occurrence of ADRs, supported by clinical evidence, using neural network-based machine learning algorithms. Further, the models were validated for predictions of ADRs based on specific substructures present. The method described herein can be used for ADR predictions early on in preclinical and clinical candidates, which may help reduce the attrition in late-stage clinical trials or even during the post-marketing surveillance phase.

MATERIALS AND METHODS

Hardware and software

All the studies described herein were performed on HP™ machine (12th Gen Intel(R) Core(TM) i5-1235U 1.30 GHz; 64-bit operating system, x64-based processor) running Windows 11 operating system, with internal memory up to 16 GB. The programming language used for implementation was Python 3.0. Experiments were conducted in the Google Colab programming framework. It allows researchers to write and execute arbitrary Python code through the browser and is especially well suited to ML, data analysis, and education.

Datasets

The following datasets were used extensively throughout the studies.

SIDER dataset

It is a database with marketed drugs and ADRs. The version of the SIDER dataset in DeepChem has grouped drug side effects into 27 system organ classes following MedDRA classifications measured for 1427 approved drugs [7]. It is one of the most popular datasets used in ADR detection and prediction-based research studies. It has been used in almost 60% of the research work done up till now [8]. A pictorial representation of the dataset is shown in table 1.

Table 1: Sample dataset

Smiles	Hepatobiliary disorders
C(CNCCNCCNCCN)N	1
CC(C)(C)C1=CC(=C(C=C1NC(=O)C2=CNC3=CC=CC=C3C2=O)O)C(C)(C)C	0
CC[C@]12CC(=C)[C@H]3[C@H]([C@@H]1CC[C@]2(C#C)O)CCC4=CCCC[C@H]34	0
CCC12CC(=C)C3C(C1CC[C@]2(C#C)O)CCC4=CC(=O)CCC34	1
CCCCCC(C=CC1C(CC(=O)C1CC=CCCCC(=O)O)O)O	0

The drug SMILES are converted into fingerprints for the application of ML algorithms. A detailed introduction to fingerprint algorithms is described in the next section.

FAERS dataset

This is a primary data source [9]. The data is collected and stored through authentic processes and validated. This dataset is presented both in ASCII and CSV format. Around 3 million records were collected from the FAERS dataset dated from 2019 to 2020 end in ASCII format. Once downloaded and extracted, the overall dataset is visualized in fig. 1.

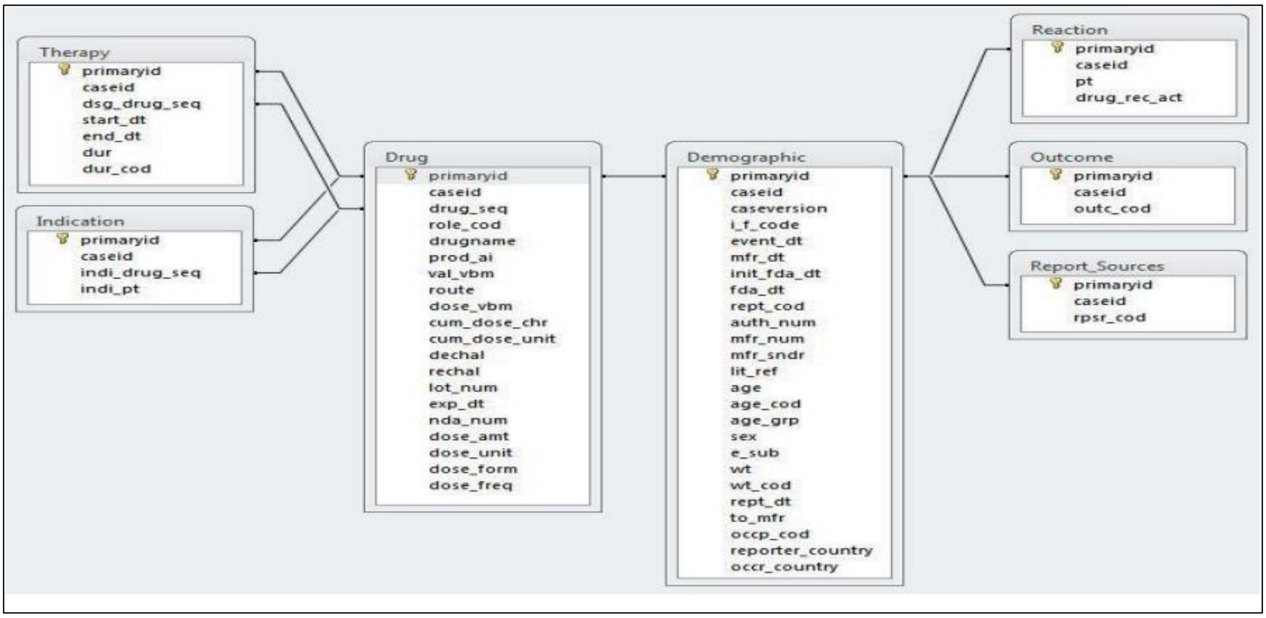

Fig. 1: Overview of the FAERS dataset [10]

Table 2: Description of each table of the FAERS dataset

Name	Description
Demographic	It contains patient demographic and administrative information, a single record for each event report.
Drug	It contains drug/biological information for as many medications as were reported for the event (1 or more per event).
Indication	It contains all Medical Dictionary for Regulatory Activities (MedDRA) terms coded for the indications for use (diagnoses) for the reported drugs (0 or more per drug per event).
Reaction	It contains all MedDRA terms coded for the adverse event (1 or more).
Outcome	It contains patient outcomes for the event (0 or more).
Therapy	It contains drug therapy start dates and end dates for the reported drugs (0 or more per drug per event)
Report Sources	It contains report sources for the event (0 or more).

The FAERS dataset is segregated across multiple tables that need to be integrated using primary ID and case ID. The drug names of the drug database should be converted into smile structure format using the Chemical Identifier Resolver (CIR) [11] from the RDKit package. Further, the smile structures were manually checked for consistency with the drug structure itself.

ChEMBL database

It is a manually curated database of bioactive molecules with drug-like properties [12]. It brings together chemical, bioactivity, and genomic data to aid the translation of genomic information into effective new drugs. It includes information about how small molecules interact with their protein targets, how these compounds affect cells and whole organisms, and information on Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET). ChEMBL holds two-dimensional structures, calculated molecular properties (e. g., logP, molecular weight, Lipinski 'Rule of Five' parameters), and bioactivity data (such as binding constants and pharmacology). The bioactivity data is tagged to show links between molecular targets and published essays. A diagrammatic representation of the ChEMBL dataset is shown in fig. 2.

These datasets form the basis of the present research work. The SMILES 1D representation of drug structures is converted into some fixed-size bit vector, i.e., fingerprint for the application of the ML algorithms.

Fingerprint algorithms

Molecular graph theory and fingerprints have a long history of applications in drug discovery and development [13]. Some predefined molecular fingerprints already in use for the drug structures are listed in table 3.

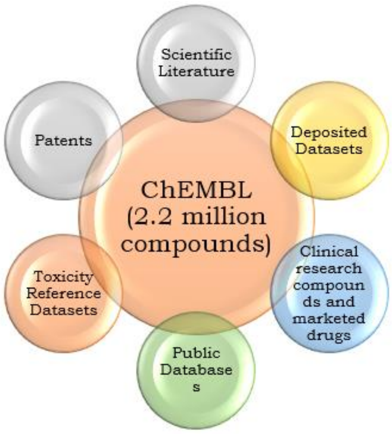
Scientific Literature
Deposited Datasets
Patents
ChEMBL (2.2 million compounds)
Clinical research compoun ds and marketed drugs
Toxicity Reference Datasets
Public Database s

Fig. 2: Data included in the ChEMBL database

Table 3: A representative list of fingerprint algorithms

Algorithm	Brief description
Pubchem Fingerprinter	These fingerprints are used by PubChem for similarity neighboring and similarity searching. A substructure is a fragment of a chemical structure. A fingerprint is an ordered list of binaries (1/0) bits [14].
AtomPairs2D Fingerprinter	The fingerprints are generated by connecting atoms within a molecule and creating a two-dimensional graph. This type of fingerprinting allows scientists to identify compounds from their chemical structure [15].
Estate Fingerprinter	It is an AI algorithm designed to automatically identify molecules within a substance. It works by comparing a molecule's structure to a database of known compounds, allowing scientists to quickly and accurately identify the molecules present in a given sample [16].
Extended Fingerprinter	It works by comparing a substance's molecular structure to a known database of compounds. The fingerprints generated by this technology are also able to identify the conformational and stereochemical properties of molecules [17].
GraphOnly Fingerprinter [18]	It works by analyzing a substance's molecular structures and creating a graph of atoms. This graph can then be compared to a database of known compounds, allowing scientists to quickly and accurately identify the substances in question.
KlekotaRoth Fingerprinter [18]	It works by analyzing a substance's molecular structure and using a statistical algorithm to compare the result with a known database of compounds. This type of fingerprinting can identify subtle differences between molecules and thus can be used to quickly and accurately identify new or unknown compounds.
Molecular ACCess System (MACCS)	This type of fingerprinting uses 166 specific bits to represent particular chemical features. For measuring molecular similarity, 166-bit 2D structure fingerprints are provided by MACCS keys. The binary bit is either 0 (or off) or 1 (or on) to represent it. MACCS provides more than 9.3x1049 distinguishable fingerprint vectors [19].
Substructure	This is a type of molecular or compound fingerprinting technology based on the concept of a "substructure". It looks at smaller molecular structures within molecules and uses them to identify different compounds. A substructure is made up of individual atoms that are connected in a certain pattern [18]
Circular Fingerprint	This type of fingerprinting works by analyzing a substance's molecular structure and then creating a circular graph of its atoms. The representation of molecular structures by atom neighborhoods--has been applied to a wide range of applications, such as similarity searching and the prediction of absorption, distribution, metabolism, excretion, and toxicity properties [20].
Morgan fingerprint	The fingerprint is a reimplementation of the Extended Connectivity Fingerprint (ECFP). It goes through each atom of the molecule and obtains all possible paths through this atom with a specific radius. Then, each unique path is hashed into a number with a maximum based on the bit number [21]. The higher the radius, the bigger fragments are encoded.

Molecular fingerprints are limited by their ability to accurately represent the chemical structure and properties of molecules. Additionally, molecular fingerprints can be prone to false positives and false negatives due to their size. This can lead to inaccurate results in certain applications. The primary drawback of the discussed current molecular structure fingerprints is their general-purpose use. This involves encoding structures into large-sized bit-vectors and encoding all possible substructures, resulting in redundancy. To counter the limitation of the existing molecular fingerprinting algorithms, the research further discusses the application of neural networks to drug molecular structures.

Neural fingerprint methodology

A replacement for the molecular fingerprint of drug structure is to apply a neural network to drug structures [22]. Neural network fingerprints are machine learning techniques that can be used to identify and classify molecules. They rely on deep neural networks, which take in a molecule's structure and output a fingerprint that captures the essential features of that molecule. The steps involved in the algorithm designed for neural network fingerprint are given in fig. 3.

The drug molecular structure is considered as input to the neural network. The radius of the molecule and the input and output weights are the initial hyperparameters set for the model. A bit array vector for storing the fingerprints is initialized. For each atom in the molecule, the neighboring atoms are identified and summed up.

A smoothing function is applied to obtain approximate values. The fingerprint obtained for each atom is added to the fingerprint vector. After performing the above process for all the atoms in the molecule, the entire fingerprint vector is returned. These neural graph fingerprints offer several advantages over fixed fingerprints.

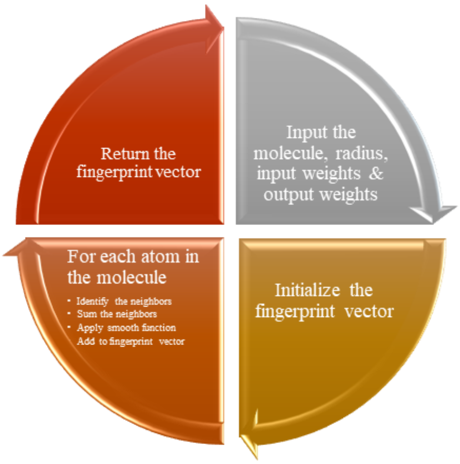
Return the fingerprint vector
Input the molecule, radius, input weights & output weights
Initialize the fingerprint vector
For each atom in the molecule
Identify the neighbors
Sum the neighbors
Apply smooth function
Add to fingerprint vector

Fig. 3: Steps involved in the neural fingerprint algorithm

Predictive performance

ML fingerprints can provide better performance at prediction tasks than the predefined fixed fingerprint technique by using the available data at hand. The prediction performance of the neural graph fingerprint technique is comparatively better at solubility, drug efficacy, and organic photovoltaic efficiency datasets than the existing molecular fingerprinting technique.

Parsimony

To encode all possible substructures without any overlap, the standard fingerprints need to be very large. The fingerprint vector size can go up to 43,000 even after eliminating the rarely occurring features [23]. Only the relevant features are encoded by the differentiable fingerprint, thus reducing the downstream computation and regularization requirements.

Interpretability

The problem with the existing fingerprint was that it encoded each fragment separately without identifying the similarity between them. Compared to neural graph fingerprint, it encodes distinct features separately and identifies the overlap, making the fragment representation more meaningful.

After identifying the optimum fingerprint algorithm, the next step was to identify the critical substructures of the drug structure responsible for causing an ADR. The fingerprint similarity was done for all the adjacent molecular structures in the dataset, but a common substructure for all structures in the given dataset was not able to be obtained. Therefore, the 'Maximum Common Substructure' algorithm was applied for the positive drug structure samples.

Maximum common substructure (MCS) [24] refers to a set of atoms or molecules that are shared between two or more molecules. These atoms and molecules will form the same structural arrangement despite differences in their functional groups. MCS can be used to identify similarities, generate leads for drug discovery, and determine structure-activity relationships. Fig. 4 shows the MCS between two representative molecules.

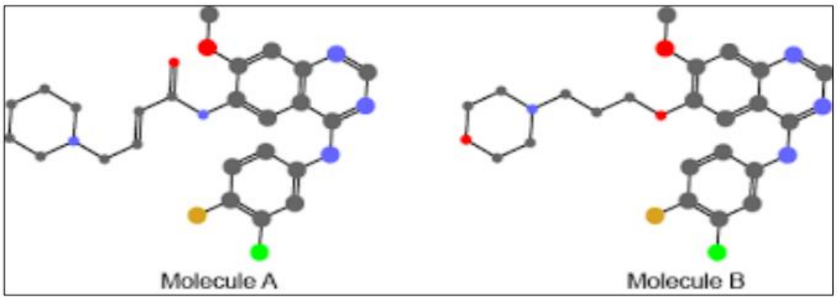

Fig. 4: Maximum common substructure between two representative molecules

The common substructure was identified for the given set of drug structures. Next, the drug substructure was compared with a random set of drug structures to identify its presence and predict its possibility for ADR association.

RESULTS AND DISCUSSION

Over the last decade, substantial research has been carried out in the field of ADR detection and prediction. Initially, the ADRs were detected based on their temporal association with drugs, as discussed by Shanmugapriya *et al.* in their research. Signal detection techniques [25] were also applied to Spontaneous Reporting System (SRS) databases to identify the true signal among all the reported ADR instances. After successfully performing ADR detection on both reported ADRs and medical reports of patients, further research efforts were applied to successfully predict the occurrence of severe and harmful ADRs soon.

The majority of research carried out in the domain of ADR prediction is based on the SRS dataset and electronic health records. The limitations of the SRS dataset are under-reporting [26], data duplication, and data quality issues, while the drawback of electronic health records is their unavailability [27]. Therefore, a methodology needs to be developed to counter the issue of data quality as well as its unavailability. In 2012, a research study was performed by Liu *et al.* [28] to predict ADRs by integrating the drug's biological, chemical, and phenotypic properties. Although the performance metrics reported by the research were above 90% in terms of accuracy for all ADRs the drawback of this research work was the model's interpretability for acceptance in the medical domain.

The integration among different datasets was done through the network and knowledge-graph representation techniques [29]. The inference of these research studies showed that to some extent, the molecular structure of drugs was associated with the ADRs caused due to it. This concept was mainly discussed in the research study done by Dey, *et al.* [30] in 2018 where the prediction of ADR was done using the molecular structure of drugs. For prediction algorithms to be applied to drug molecular structures these structures should be converted into fixed-size bit vector arrays. The entire process of conversion is performed using fingerprinting techniques [30]. The performance assessment of different fingerprint techniques was also done as part of their research work. Although the research study tackles the issue of model interpretability, the prediction model does not account for the severity of adverse drug reactions. The molecular structure can further be partitioned into several substructures. The analysis of these substructures can be associated with the prediction of ADR in an early stage of the drug development life cycle.

Processes related to substructures can be identifying its presence in the given drug structure, substructure-substructure similarity matching [31], and extracting MCS from a set of molecular structures, which is based on the mathematical concept of maximum common subgraphs derived by Cao *et al.* [32] in his research work. The application of drug substructures is not only limited to drug discovery [33] as well as drug repositioning [34] for different diseases but also associating the side effects of drugs with similar drug substructures [35]. Therefore, the authors of this research study have made an attempt to address the issue of prediction of severe ADRs based on their drug substructure analysis and develop an interpretable ML model for acceptance in the medical domain.

This fingerprint algorithm is tested on a sample dataset referenced in table 1 and the prediction results were evaluated based on training accuracy and test accuracy. The results are shown in the following table 4.

Table 4: Results obtained from model development and validation using fingerprint algorithms

Fingerprint algorithm	Accuracy of the training dataset	Accuracy of the test dataset
Pubchem fingerprinter	0.9667	0.5333
AtomPairs2D fingerprinter	0.9194	0.4615
Estate fingerprinter	0.8675	0.4596
Extended fingerprinter	0.9877	0.5035
GraphOnly fingerprinter	0.9649	0.4650
KlekotaRoth fingerprinter	0.9675	0.5298
MACCS fingerprinter	0.9675	0.5018
Substructure fingerprinter	0.8781	0.5088
Circular fingerprinter	0.9947	0.6573
Morgan fingerprinter	0.7212	0.7118
Neural fingerprinter	0.7953	0.7318

After analyzing the results from table 4, it was evident that the outcomes of the Morgan fingerprint and Neural fingerprints were comparable. However, the performance of the Neural fingerprint algorithm was superior. While both algorithms shared a similar initial framework, the Neural fingerprint algorithm incorporated a neural network, which contributed to its enhanced performance. In the Neural fingerprint algorithm, a summation operation was conducted for each atom in the molecule instead of concatenation. A smooth function was also applied to the final layer, contrasting with the hash function utilized in the Morgan fingerprint algorithm. Furthermore, the Neural fingerprint algorithm added the fingerprint to the fingerprint vector instead of indexing, as is done in the Morgan fingerprint.

Based on the obtained results for both algorithms, it was evident that the Neural fingerprint algorithm exhibited optimal performance compared to other Molecular fingerprints [36].

Substructure analysis of molecular structures

The subsequent step involved applying the MCS algorithm to the positive drug structure observations. The extracted common substructure was then compared with the entire dataset of samples. It was observed that the extracted common substructure existed in approximately 90% of the drug structures within the sample dataset. This outcome could be attributed to the fact that the comparison was performed using the same dataset from which the common substructure was extracted. To address this limitation, a random dataset was obtained from the ChEMBL database [12], as described earlier. This dataset encompasses the molecular structures of approximately 14,000 drugs. The common substructure derived from the sample dataset was subsequently compared with this new dataset. Around 100 data samples were extracted for this comparison, and upon evaluation, it was found that five drug structures returned true values, indicating the presence of the common substructure. The results of this comparison can be seen in fig. 5.

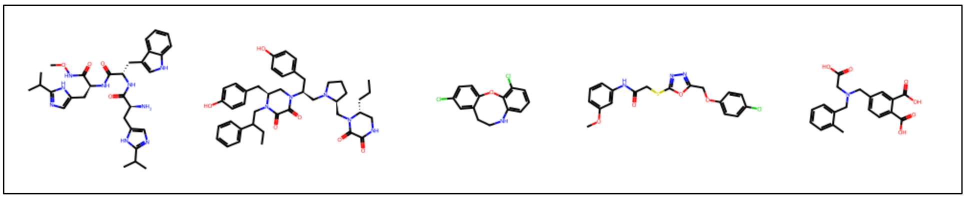

Fig. 5: Molecular structures obtained as hits from a pilot study

The obtained drug structures were initially compared with the drugs in the sample dataset to identify any common records. However, no such records were found. Subsequently, the drug structures were transformed into fingerprints, and predictions were made for all five molecular structures. Among the tested structures, four were predicted to be true for causing the specified ADR (hepatobiliary disorder). This pilot study highlighted the significant role that drug substructures play in the occurrence of ADRs. It emphasized the importance of early detection and prediction of ADRs based on the analysis of drug structures themselves. By identifying the specific substructures associated with ADRs, this approach lays the foundation for proactive measures in drug safety assessment.

Generalizing the results of the pilot study on the real-world dataset:- The FAERS dataset described earlier was pre-processed using the steps shown in fig. 6. As illustrated in fig. 6, the main steps of the process included converting drug names into SMILES format for neural fingerprint techniques, extracting external drug characteristics, and encoding patient-related data for the application of machine learning algorithms. The drug dataset was then integrated with the reaction dataset using primary ID and case ID as key identifiers. Subsequently, the most frequently occurring adverse drug reactions were identified and presented in table 5.

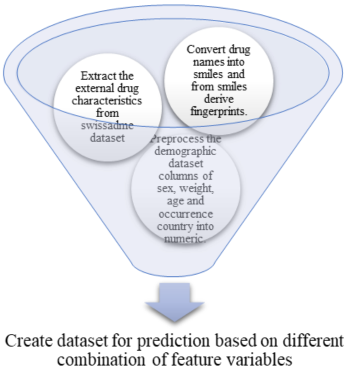
Extract the external drug characteristics from swissadme dataset
Convert drug names into smiles and from smiles derive fingerprints.
Preprocess the demographic dataset columns of sex, weight, age and occurrence country into numeric.
Create dataset for prediction based on different combination of feature variables

Fig. 6: Pre-processing of the FAERS dataset

Table 5: Top-10 most occurring ADRs

S. No.	ADR	Occurrence
1	Aplastic anemia	347
2	Mucosal inflammation	272
3	Nausea	146
4	Hypogonadism	142
5	Pancreatitis acute	142
6	Pain	137
7	Vomiting	125
8	Dry mouth	102
9	Somnolence	96
10	Sepsis	96

ADR prediction using ML approaches: -The prediction of adverse reactions to drugs was performed by incorporating different compositions of feature variables. Strategy 1. To predict ADRs using only neural fingerprints of drugs:- As seen in table 6, the ML algorithms were used to predict the probability of the occurrence of different ADRs based only on the drug fingerprints.

Table 6: Fingerprint-based prediction based on AUC

Adverse drug reactions										
AUC	**Aplastic anemia**	**Pain**	**Nausea**	**Mucosal inflammation**	**Hypogonadism**	**Pancreatitis acute**	**Vomiting**	**Dry mouth**	**Somnolence**	**Sepsis**
Random forest model	0.56	0.68	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
Support vector machine (SVM)	0.51	0.67	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
Logistic regression	0.51	0.67	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
ANN	0.53	0.50	0.50	0.54	0.50	0.50	0.50	0.50	0.50	0.50

Instead of using accuracy as the evaluation metric for the model, AUC was employed. It is a preferred metric as it ensures that the performance of the classification model remains independent of the threshold value chosen. The achieved performance for the drug structure fingerprints exceeded 65%. This indicated that the model performed well when relying solely on fingerprints. However, there is potential for further improvement by incorporating additional feature variables alongside the fingerprints.

Strategy 2. To predict ADRs using fingerprints as well as other characteristics of drugs: -The external features of drugs, namely target inhibitors and toxicity, are known to have an impact on the occurrence of ADRs. To account for these factors, the target inhibitors and toxicity information were concatenated with the drug SMILES fingerprints. Subsequently, predictions were made using this combined dataset for the 10 most frequently observed ADRs. The results obtained from this analysis are presented in table 7.

Table 7: Prediction based on fingerprint and other drug characteristics

	Adverse drug reactions									
AUC	**Aplastic anemia**	**Pain**	**Nausea**	**Mucosal inflammation**	**Hypogonadism**	**Pancreatitis acute**	**Vomiting**	**Dry mouth**	**Somnolence**	**Sepsis**
Random forest model	0.56	0.68	0.55	0.56	0.55	0.55	0.55	0.55	0.55	0.55
SVM	0.56	0.71	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
Logistic regression	0.56	0.69	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56
ANN	0.55	0.66	0.55	0.55	0.50	0.55	0.51	0.55	0.52	0.52

Based on the findings presented in table 7, it can be deduced that the incorporation of drug characteristics, in combination with fingerprints, results in a noticeable improvement of approximately 5% in the AUC compared to our initial prediction model. This indicates that the inclusion of drug characteristics has enhanced the classifier's ability to classify adverse drug reactions. Notably, the SVM algorithm demonstrates the most significant improvement among the applied methods.

Strategy 3. To perform prediction using fingerprints, drug characteristics, and patient data: -

To assess the real-world impact, a dataset was constructed by incorporating patient data, including age, weight, gender, and demographic details, alongside drug and adverse reaction information. The occurrence of adverse drug reactions within the patient population was analyzed using this dataset. Finally, predictions for ADRs were made by combining fingerprints, drug characteristics, and patient data. The results of these predictions are presented in table 8.

Table 8 demonstrated a substantial enhancement in the performance of the prediction model when the patient dataset is combined with the drug data. This indicates that the inclusion of patient data, in conjunction with drug data, effectively predicts the occurrence of adverse drug reactions, resulting in a noteworthy increase of 30% in AUC. Notably, certain adverse reactions, such as aplastic anemia, hypogonadism, and acute pancreatitis, exhibit a considerable boost in AUC.

To evaluate the severity of the ADRs, their impact on human health was assessed, as depicted in fig. 7. The analysis took into account the range of effects, ranging from mild symptoms such as cold and cough to more severe consequences that necessitate hospitalization or even result in fatalities.

Extracting the most commonly occurring substructure for most severe ADRs

To identify the most common substructure associated with the given ADRs, it was compared with the original dataset to check for repetition. To broaden the comparison, a random dataset comprising approximately 8 million drug structure records in SMILES format was downloaded from ChEMBL [23]. The provided substructure was then compared with all the records in the ChEMBL dataset, yielding the following results:

MCS for aplastic anemia: -. The structures shown in fig. 8 were found to be true, containing the highlighted MCS.

Table 8: Prediction based on drug and patient data

Adverse drug reactions										
AUC	Aplastic anemia	Pain	Nausea	Mucosal inflammation	Hypogonadism	Pancreatitis acute	Vomiting	Dry mouth	Somnolence	Sepsis
Random forest model	0.89	0.71	0.64	0.85	0.95	0.95	0.75	0.75	0.79	0.79
SVM	0.92	0.79	0.75	0.91	0.91	0.92	0.77	0.85	0.82	0.9
Logistic regression	0.92	0.84	0.78	0.93	0.92	0.92	0.78	0.87	0.83	0.87
ANN	0.54	0.65	0.65	0.82	0.59	0.54	0.5	0.5	0.61	0.7

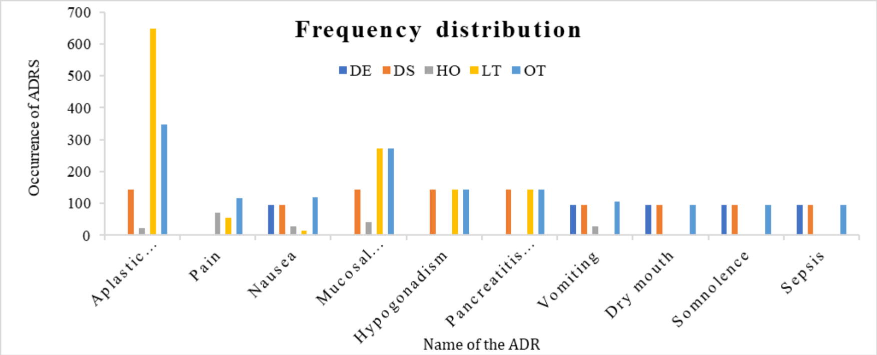

Fig. 7: Bar chart representing frequency distribution of ADR severity; after analyzing the impact of adverse reactions of drugs on patients' health, the three most severe ADRs were aplastic anemia, mucosal inflammation, and vomiting

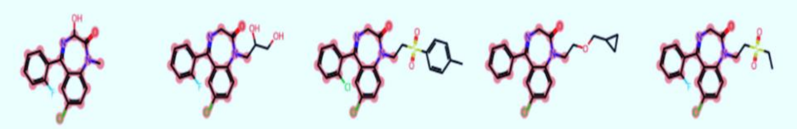

Fig. 8: Highlighted MCS containing hits for aplastic anemia ADR

The predicted model was utilized to perform predictions for all of these structures using their fingerprints, and all of them returned true values, indicating that these structures had the potential to cause aplastic anemia. The accuracy of the predictions was further validated using literature evidence [37]. This validation supported the reliability of the prediction model in identifying structures that were associated with the occurrence of aplastic anemia.

MCS for mucosal inflammation*:* -For this ADR, three molecular structures returned to be true of containing the highlighted MCS as depicted in fig. 9.

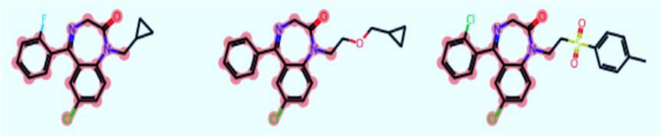

Fig. 9: Highlighted MCS containing hits for mucosal inflammation ADR

The structures for which the common substructure returned a true value were converted into fingerprints, and predictions were made based on these fingerprints. The predictions resulted in positive values for all the structures, indicating that these structures were also associated with mucosal inflammation. To validate these predicted outcomes, literature evidence was referenced [38]. This validation strengthened the reliability of the prediction model in identifying structures that contribute to the occurrence of mucosal inflammation.

MCS for vomiting: -For this ADR, three molecular structures returned to be true as depicted in fig. 10.

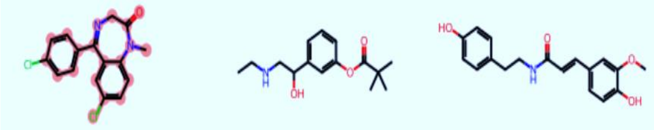

Fig. 10: Hits for vomiting ADR

Upon applying the prediction model to these structures, it was observed that only two of them returned a positive prediction, while the third structure yielded a negative prediction, as reported [39]. The model correctly predicted the first and third structures as positive, indicating their association with the specified outcome. However, the second structure was predicted to be negative, suggesting it might not be linked to the outcome. Further analysis revealed that the third structure could be realigned or reconfigured to form an MCS, which would then align with the positive prediction. This discrepancy highlighted the complexity and intricacies involved in predicting the relationship between structures and specific outcomes. It emphasized the importance of continuous validation and refinement of prediction models based on real-world evidence and alignment with MCS.

In summary, this research study aimed to detect and predict ADRs early by utilizing drug and patient characteristics, employing the drug's molecular structural fingerprint technique. One major challenge in medical research is obtaining patient data, which was addressed in this study. The prediction model based on drug fingerprints achieved performance with an AUC metric of above 65%. By incorporating patient data alongside drug characteristics, the algorithms showed a significant improvement in AUC, increasing it by 30%, and indicating its effectiveness as a classifier. The extracted substructure was also identified and visually presented in fig. 8, 9, and 10, demonstrating the interpretability of the entire process. Furthermore, the prediction model successfully predicted ADRs for unknown drugs and validated its predictions using literature evidence. Overall, this research study effectively achieved its objectives and contributed to the advancement of early detection and prediction of ADRs.

CONCLUSION

The pilot study revealed that specific drug components significantly influence ADRs. This insight was applied to the pre-processed FAERS dataset using drug fingerprints and patient data. Machine learning showed drug fingerprints alone achieved over 65% prediction accuracy, addressing data limitations. Incorporating patient data improved overall prediction by approximately 30%. The study also focused on model interpretability using outcome visualization techniques. In conclusion, our research successfully tackled early ADR detection, data limitations, and model interpretability challenges, offering a valuable framework for future studies in the field.

FUNDING

No funding was received for the presented work

AUTHORS CONTRIBUTIONS

SD and AP conceived the idea. PK fine-tuned the idea. All authors contributed to the manuscript preparation.

CONFLICT OF INTERESTS

The authors declare no conflict of interest

REFERENCES

1. Pushparajah DS, Geissler J, Eupati WN. Collaboration between patients, academia, and industry to champion the informed patient in the research and development of medicines. J Dev Sci. 2016 Nov 17;1(1):74. doi: 10.18063/jmds.v1i1.122.
2. The ICH Expert Working Group. Post-approval safety data management: definitions and standards for expedited reporting. ICH harmonised tripartite; 2003. Available from: http://www.fda.gov/cber/gdlns/ichexrep.htm. [Last accessed on 03 Feb 2024]
3. Alomar MJ. Factors affecting the development of adverse drug reactions (Review article). Saudi Pharm J. 2014;22(2):83-94. doi: 10.1016/j.jsps.2013.02.003, PMID 24648818.
4. Cyriac ST, Iype DS. Neuropsychiatric adverse effects of antibacterial agents. Int J Pharm Pharm Sci. 2021;13(12):1-8. doi: 10.22159/ijpps.2021v13i12.42482.
5. Kaur A, Singh J. A case report on well-known but has to be reported adverse effect of phenytoin-induced rash and assessment of its severity based on adverse drug reaction reporting scales. Asian J Pharm Clin Res. 2022;15(12):1-2. doi: 10.22159/ajpcr.2022.v15i12.46006.
6. Chagas CM, Moss S, Alisaraie L. Drug metabolites and their effects on the development of adverse reactions: revisiting lipinski's rule of five. Int J Pharm. 2018;549(1-2):133-49. doi: 10.1016/j.ijpharm.2018.07.046, PMID 30040971.
7. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016;44(D1):D1075-9. doi: 10.1093/nar/gkv1075, PMID 26481350.
8. Pandit AA, Dubey SA. A comprehensive review on Adverse Drug Reactions (ADRs) Detection and Prediction Models. In: 13th International Conference on Computational Intelligence and Communication Networks (CICN). Vol. 2021. IEEE Publications; 2021 Sep. p. 123-7. doi: 10.1109/CICN51697.2021.9574639.
9. Peng L, Xiao K, Ottaviani S, Stebbing J, Wang YJ. A real-world disproportionality analysis of FDA adverse event reporting system (FAERS) events for baricitinib. Expert Opin Drug Saf. 2020 Nov;19(11):1505-11. doi: 10.1080/14740338.2020.1799975, PMID 32693646.
10. Chiappini S, Vickers-Smith R, Guirguis A, Corkery JM, Martinotti G, Harris DR. Pharmacovigilance signals of the opioid epidemic over 10 years: data mining methods in the analysis of pharmacovigilance datasets collecting adverse drug reactions (ADRs) Reported to EudraVigilance (EV) and the FDA Adverse Event Reporting System (FAERS). Pharmaceuticals (Basel). 2022 Jun;15(6):675. doi: 10.3390/ph15060675, PMID 35745593.
11. Wohlgemuth G, Haldiya PK, Willighagen E, Kind T, Fiehn O. The chemical translation service-a web-based tool to improve standardization of metabolomic reports. Bioinformatics. 2010 Oct 15;26(20):2647-8. doi: 10.1093/bioinformatics/btq476, PMID 20829444.
12. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D. The ChEMBL database in 2017. Nucleic Acids Res. 2017 Jan 4;45(D1):D945-54. doi: 10.1093/nar/gkw1074, PMID 27899562.
13. Gonzalez Diaz H, Vilar S, Santana L, Uriarte E. Medicinal chemistry and bioinformatics-current trends in drugs discovery with networks topological indices. Curr Top Med Chem. 2007;7(10):1015-29. doi: 10.2174/156802607780906771, PMID 17508935.
14. Xie XQS. Exploiting PubChem for virtual screening. Expert Opin Drug Discov. 2010 Dec;5(12):1205-20. doi: 10.1517/17460441.2010.524924, PMID 21691435.
15. Awale M, Reymond JL. Atom pair 2D-fingerprints perceive 3D-molecular shape and pharmacophores for very fast virtual screening of ZINC and GDB-17. J Chem Inf Model. 2014 Jul 28;54(7):1892-907. doi: 10.1021/ci500232g, PMID 24988038.
16. Warr WA. Twenty-five years of progress in cheminformatics. In: American Chemical Society National Meeting; 2005. p. 229.
17. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010 May 24;50(5):742-54. doi: 10.1021/ci100050t, PMID 20426451.
18. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011 Jun 30;32(7):1466-74. doi: 10.1002/jcc.21707, PMID 21425294.
19. Fernandez-de Gortari E, Garcia Jacas CR, Martinez 0Mayorga K, Medina Franco JL. Database fingerprint (DFP): an approach to represent molecular databases. J Cheminform. 2017;9:9. doi: 10.1186/s13321-017-0195-1, PMID 28224019.
20. Glem RC, Bender A, Arnby CH, Carlsson L, Boyer S, Smith J. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. I Drugs. 2006 Mar;9(3):199-204. PMID 16523386.
21. Withnall M, Lindelöf E, Engkvist O, Chen H. Building attention and edge message passing neural networks for bioactivity and physical-chemical property prediction. J Cheminform. 2020;12(1):1. doi: 10.1186/s13321-019-0407-y, PMID 33430988.
22. Meyer JG, Liu S, Miller IJ, Coon JJ, Gitter A. Learning drug functions from chemical structures with convolutional neural networks and random forests. J Chem Inf Model. 2019 Oct 28;59(10):4438-49. doi: 10.1021/acs.jcim.9b00236, PMID 31518132.
23. Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru Guzik A. Convolutional networks on graphs for learning molecular fingerprints. Adv Neural Inf Process Syst. 2015;28.

24. Cao Y, Jiang T, Girke T. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. Bioinformatics. 2008 Jul 1;24(13):i366-74. doi: 10.1093/bioinformatics/btn186, PMID 18586736.
25. Shanmugapriya K. N-Unexpected temporal association rule for diagnosing adverse drug reaction from health database. Int Proc Comput Sci Inf Technol (IPCSIT). 2011;18.
26. Babu LP, Robin N, Babu JV, Jose J, George S. Adverse drug reactions among drug-resistant tuberculosis treatment: an observational cohort study. Int J Pharm Pharm Sci. 2021;13(9):50-5. doi: 10.22159/ijpps.2021v13i9.42460.
27. Kasliwal R. Spontaneous reporting in pharmacovigilance: strengths, weaknesses and recent methods of analysis. J Clin Prev Cardiol. 2012;1:20-3.
28. Liu M, Wu Y, Chen Y, Sun J, Zhao Z, Chen XW. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. J Am Med Inform Assoc. 2012 Jan 1;19(e1):e28-35. doi: 10.1136/amiajnl-2011-000699, PMID 22718037.
29. Grangel Gonzalez I. A knowledge graph-based integration approach for industry 4.0: Universitats-und Landesbibliothek Bonn; 2019.
30. Dey S, Luo H, Fokoue A, Hu J, Zhang P. Predicting adverse drug reactions through interpretable deep learning framework. BMC Bioinformatics. 2018 Nov 29;19Suppl 21:476. doi: 10.1186/s12859-018-2544-0, PMID 30591036.
31. Yang Z, Zhong W, Lv Q, Yu-Chian Chen C. Learning size-adaptive molecular substructures for explainable drug–drug interaction prediction by substructure-aware graph neural network. Chem Sci. 2022;13(29):8693-703. doi: 10.1039/d2sc02023h, PMID 35974769.
32. Cao Y, Jiang T, Girke T. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. Bioinformatics. 2008 Jul 1;24(13):i366-74. doi: 10.1093/bioinformatics/btn186, PMID 18586736.
33. Merlot C, Domine D, Cleva C, Church DJ. Chemical substructures in drug discovery. Drug Discov Today. 2003;8(13):594-602. doi: 10.1016/s1359-6446(03)02740-5, PMID 12850335.
34. Yang J, Zhang D, Liu L, Li G, Cai Y, Zhang Y. Computational drug repositioning based on the relationships between substructure–indication. Brief Bioinform. 2021 Jul 1;22(4):bbaa348. doi: 10.1093/bib/bbaa348, PMID 33313675.
35. Zhou B, Zhao X, Lu J, Sun Z, Liu M, Zhou Y. Relating substructures and side effects of drugs with chemical-chemical interactions. Comb Chem High Throughput Screen. 2020;23(4):285-94. doi: 10.2174/1386207322666190702102752, PMID 31267865.
36. Hentabli H, Bengherbia B, Saeed F, Salim N, Nafea I, Toubal A. Convolutional neural network model based on 2D fingerprint for bioactivity prediction. Int J Mol Sci. 2022 Nov;23(21):13230. doi: 10.3390/ijms232113230, PMID 36362018.
37. Issaragrisil S, Kaufman DW, Anderson T, Chansung K, Thamprasit T, Sirijirachai J. Low drug attributability of aplastic anemia in Thailand. The Aplastic Anemia Study Group. Blood. 1997 Jun 1;89(11):4034-9. PMID 9166842.
38. Jääskeläinen SK, Woda A. Burning mouth syndrome. Cephalalgia. 2017 Jun;37(7):627-47. doi: 10.1177/0333102417694883, PMID 28569120.
39. Bover Cid S, Latorre Moratalla ML, Veciana Nogues MT, Vidal Carou MD. Processing contaminants: biogenic amines encyclopedia of food safety; 2014. p. 381-91.