



An extensive survey on the use of supervised machine learning techniques in the past two decades for prediction of drug side effects

Pranab Das¹ · Dilwar Hussain Mazumder¹

Accepted: 1 February 2023 / Published online: 15 February 2023
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

Abstract

Approved drugs for sale must be effective and safe, implying that the drug's advantages outweigh its known harmful side effects. Side effects (SE) of drugs are one of the common reasons for drug failure that may halt the whole drug discovery pipeline. The side effects might vary from minor concerns like a runny nose to potentially life-threatening issues like liver damage, heart attack, and death. Therefore, predicting the side effects of the drug is vital in drug development, discovery, and design. Supervised machine learning-based side effects prediction task has recently received much attention since it reduces time, chemical waste, design complexity, risk of failure, and cost. The advancement of supervised learning approaches for predicting side effects have emerged as essential computational tools. Supervised machine learning technique provides early information on drug side effects to develop an effective drug based on drug properties. Still, there are several challenges to predicting drug side effects. Thus, a near-exhaustive survey is carried out in this paper on the use of supervised machine learning approaches employed in drug side effects prediction tasks in the past two decades. In addition, this paper also summarized the drug descriptor required for the side effects prediction task, commonly utilized drug properties sources, computational models, and their performances. Finally, the research gap, open problems, and challenges for the further supervised learning-based side effects prediction task have been discussed.

Keywords Drug side effects · Drug properties · Supervised learning · Deep learning · Machine learning

✉ Dilwar Hussain Mazumder
dilwar2k4@yahoo.co.in

Pranab Das
dpranavdas@gmail.com

¹ Department of Computer Science and Engineering, National Institute of Technology Nagaland, Chumukedima, Dimapur, Nagaland 797103, India

1 Introduction

Side Effects (SE) of a drug can be understood as unpleasant, undesirable, adverse, unexpected hazardous reactions related to the response of the drug in organs and tissues (Naranjo et al. 1981; Edwards and Aronson 2000; Linden 2013). Adverse drug response is the main reason for the side effects of drugs; many side effects are not detected in the drug development, discovery, and design process (Katragadda et al. 2015; Liu et al. 2017), and identifying the hazardous reactions of the drug is challenging. Therefore, predicting the risky responses of drugs during the development process remains an essential task for drug safety and commercial success. Side effects may injure body organs, and patients may die due to risky drug reactions. It is essential to ensure that drugs are efficacious for treatment. Thus, the harmful responses of drugs are one of the primary causes of drug failure, and most medicines are withdrawn from the market due to hazardous side effects.

The drug development and design process needs a step-by-step clinical trial which is a tedious, complicated, challenging, and costly process and requires technological expertise, research skills, human resources, chemical compound, and billions of dollars (Willmann et al. 2008; Trouiller et al. 2017; Grygorenko et al. 2020; Sarma et al. 2021; Askar et al. 2022). Drug side effects prediction is one of the significant phases in drug discovery, development, and design (Bresso et al. 2013; Zhou and Zhong 2017). The computation models are gaining more popularity and attention to minimize monetary cost and time (Abd Elaziz and Yousri 2021; Dara et al. 2021; Das and Hussain Mazumder 2021; Das and Mazumder 2023; Das et al. 2022b). Supervised deep learning and machine learning are the two most popular computational approaches to predicting drug side effects. These approaches can learn drug side effects from past experiences of similar drugs.

Supervised machine learning plays a vital role in side effects prediction tasks. Supervised machine learning creates a model that predicts the outcome based on the available data (Lighthart et al. 2021; Chan et al. 2022; Mosqueira-Rey et al. 2022). It is possible to train a machine learning model under supervision by giving it both the correct input and output data (Muhammad et al. 2021; Himeur et al. 2022; Izadi et al. 2022; Pun et al. 2022). Finding a mapping function to link the input with the output is the goal of a supervised learning algorithm (Barrionuevo et al. 2021). Models are trained in supervised learning using the labeled dataset, which allows the model to learn about each input data. After the training, the model makes predictions about the outcome and evaluates using test data. In the supervised machine learning technique, the drug properties provided to the model act as the supervisor who teaches the model correctly predict the side effects. To predict the side effects of the drug, supervised machine learning techniques utilize drug properties and well-labeled side effects (training dataset) to learn the model and achieve the targeted side effects (output label). The training drug properties and well-labeled side effects enable the model to learn from the training dataset. Well-labeled side effects refer to the drug already assigned the appropriate side effects.

There are several survey papers on artificial intelligence in drug discovery (Chan et al. 2019; Kim et al. 2020), a computational tool for representing drug descriptors (Dong et al. 2018), application of computational approaches in drug design and screening (Lin et al. 2020), drug target identification (Dai and Zhao 2015; Jung and Cho 2020; Bagherian et al. 2021), summarization of biomedical datasets (Stein 2003; Bonner et al. 2021), molecular communication in drug discovery (Chahibi 2017), discovering

physicochemical profile in drug development (Avdeef and Testa 2002), and data-driven methods to predict adverse drug reactions (Ho et al. 2016). However, there are no specific survey papers on the supervised computational learning approach to predict drug side effects.

This research study's fundamental motivation is to summarize the side effects prediction studies for 18 years (2004–2022). The key motivation and objective of this research study are as follows:

- Side effects of the drug are common over a million serious injuries each year. Drugs must be effective and safe before they come to market, which indicates that their benefits must outweigh any known adverse side effects.
- Drug reactions are hazardous to health, and a person taking a drug may suffer from its side effects which can cause serious health complications, including liver damage, heart attack, and death.
- Studies regarding the harmful side effects are essential for drug discovery, development, and design to enhance public health concerns and drug safety.
- Also, drug development, design, and discovery processes are complicated, expensive, and time needed. So, there is a need to study the side effects of drugs efficiently to avoid the maximum cost and time.
- Drug side effects are one of the most frequent causes of drug failure and can stop the entire drug development process.
- Identifying side effects for a novel drug takes many clinical trial phases. Supervised machine learning-based side effects prediction task has recently gained much attention since they speed up the drug development process and reduces time, chemical waste, cost, design complexity, risk of failure, and human resource.
- This paper summarizes many research articles on drug side effects prediction. The survey is based on the number of samples and side effects used for the experiment, drug properties with their sources, supervised computational approach with their performances, and problems the researcher solves for the side effects prediction task.
- Also, this paper explored the evolution of supervised machine learning and its applications in side effects prediction tasks.
- The research gaps and open problems for further side effects prediction studies are also summarised for developing an effective drug.

Several side effects are unknown to the drug manufacturing organization, pharmacist, patient, and doctors. Due to harmful side effects, some patient loses their life too. Therefore, an urgent literature survey is needed for the side effects prediction study, which may help to provide beneficial guidance to build more efficient models.

The organization of this research is as follows: drug descriptors and supervised machine learning methods used for side effects prediction has been demonstrated in Sect. 2. Section 3 provided the literature survey of the supervised learning methods to predict drug side effects. Further, the discussion on the challenges and research gaps in side effects prediction have been presented in Sect. 4. At last, summarize the conclusion of the research work in Sect. 5.

2 Drug descriptors and supervised learning methods used for side effects prediction

This section presents a list and a brief description of drug properties and the supervised computational learning methods employed to predict drug side effects.

2.1 Drug descriptors used for side effects prediction

In this section, drug descriptors are presented, which have been utilized as an input feature of a supervised computational approach to predict the potential side effects of drugs. A variety of drug descriptors are available for side effects prediction. Some are clinical data (clinical trial data, health surveys, Electronic Health Records (EHR), administrative data, claim data, and disease-patient registries), and some are non-clinical information about drugs (pharmacokinetics, chemical, toxicological, safety, biological, pharmaceutical, pharmacological, and other pre-clinical data) (Lee et al. 2008; Andrade et al. 2016). Both clinical and non-clinical data are widely used for side effects prediction using computational approaches. Further, this drug information can be categorized into chemical (Che), Phenotypic (Phe), Biological(Bio), and other drug descriptors, illustrated in Fig. 1.

2.2 Supervised learning methods used for side effects prediction

This section describes the supervised computational approaches used for potential side effects prediction tasks. The traditional approach to identifying drug side effects requires several clinical trial phases and needs monitoring drug safety after being released in the market (Yao et al. 2013). However, these procedures require time, biomedicine experts, human resources, and huge capital investment. Therefore, supervised computational learning methods are essential to reduce time, chemical waste, design complexity, risk of failure, and cost. Various computational models have been used to overcome these issues to predict side effects which are shown in Fig. 2.

The multi-label task can be addressed in various ways, including adapted algorithm, problem transformation methods (classifier chains, label power-set, and binary relevance), ensemble approaches, and multi-output deep neural

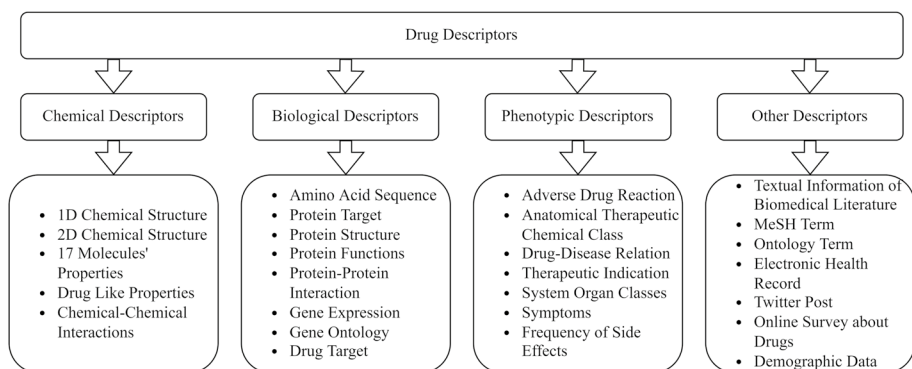


Fig. 1 Classification of drug descriptors

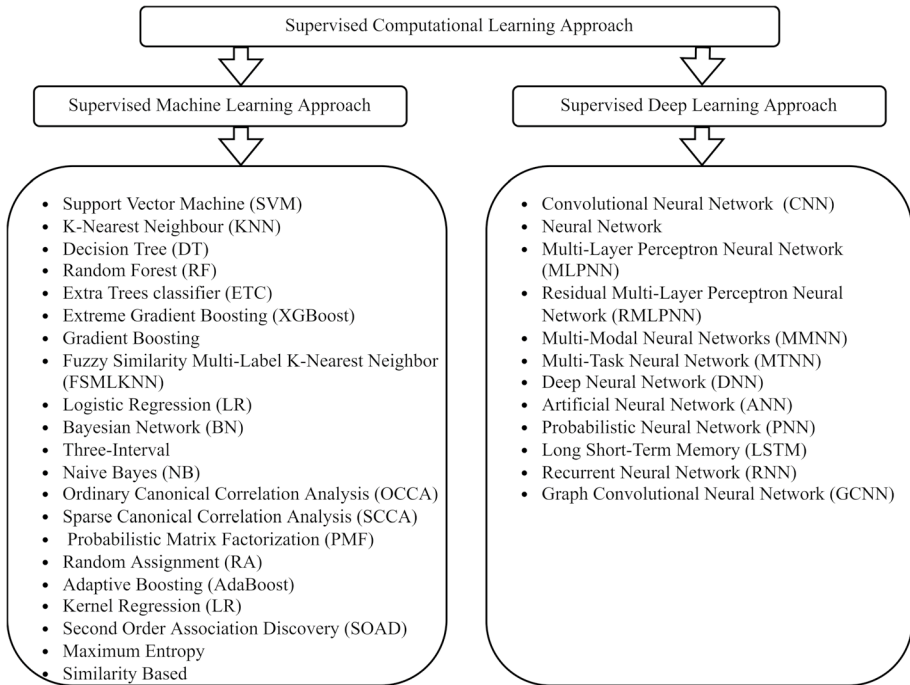


Fig. 2 Supervised computational learning approaches used for side effects prediction

framework (Tsoumakas et al. 2006; Rokach et al. 2014; Liu et al. 2022). Let us consider $\text{Drug}=\{\text{Drug}_1, \text{Drug}_2, \text{Drug}_3, \dots, \text{Drug}_n\}$ to be the set of the drugs, where n denotes the total number of drugs, $\text{Drug_Descriptors}=\{DD_1, DD_2, DD_3, \dots, DD_p\}$, where p indicates the number of drugs descriptors (input feature), and $\text{Side_Effects}=\{SE_1, SE_2, SE_3, \dots, SE_q\}$ used to indicate a set of side effects (output labels), where q denotes the total number of drug side effects. A drug can have more than one side effect, so the drug side effects prediction activity is a multi-label prediction task (Qu et al. 2012; Afdhal et al. 2020). The multi-label representation of the side effects prediction task is illustrated in Fig. 3, where zero denotes the non-appearance of a side effect, and one represents the appearance of the side effects.

3 Literature survey of the supervised learning methods to predict drug side effects

This section describes the literature research on drug side effects prediction, which consists of the number of samples (drugs) used to perform the task corresponding to the total number of output labels (side effects), utilization of drug properties with their resources, and computational supervised learning model and their performance.

In their work, Yap et al. (2004) predict Torsade De Pointes (TDP) side effects of drugs by applying probabilistic neural network (PNN), K nearest neighbors (KNN), support vector machine (SVM), and decision tree (DT) from the linear solving energy relationship

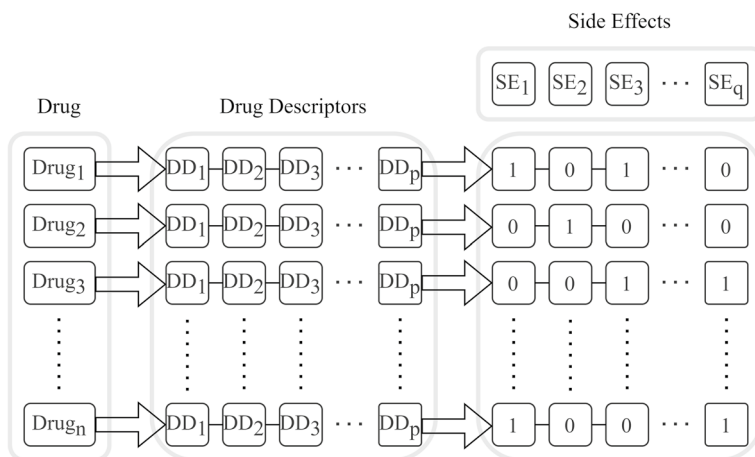


Fig. 3 Representation of multi-label side effects predictions task

(LSER) properties (hydrogen-bond term, activity term, solvent-solute interactions, and a polar term) (Abraham 1993). The authors collected data from human studies, and their dataset consists of 67 TDP⁺ and 243 TDP⁻ agents. Furthermore, their experiment outcome suggests that SVM performs the best compared to the other classifiers and achieved the highest accuracy of 91%. Scheiber et al. (2009) predict side effects by associating chemical features with side effects. The authors downloaded drugs and side effects from the PharmaPendium database (Rees et al. 2020), which consists of 4210 side effects and 1842 drugs. The Authors proposed a Laplacian-based Naive Bayes (NB) classifier to associate the chemical feature with side effects. In Pouliot et al. (2011), the authors predict system organ classes specific side effects by employing the BioAssay data, which are collected from the PubChem (Kim et al. 2019) and Canadian Adverse Drug Reactions (ADR) pharmacovigilance database (Hashimoto et al. 2009), respectively. Their collected dataset consists of 485 drugs and multiple side effects from the collection of 14,98,570 side effects. Further authors applied the logistic regression (LR) model on the dataset and achieved the highest AUC score of 92%. A machine learning methodology is presented to predict side effects from the chemical structure by Pauwels et al. (2011). They collected side effects from the SIDER (Kuhn et al. 2010) and chemical structure from PubChem (Kim et al. 2019). Their dataset consists of 1,385 side effects and 888 drugs. They applied ordinary canonical correlation analysis (OCCA), random assignment (RA), SVM, sparse canonical correlation analysis (SCCA), and the nearest neighbor classifier to predict side effects. Further, it is noticed that SCCA performed the best and obtained the highest AUC score of 89.32%. Cami et al. (2011) predict side effects events using the LR model. They extracted drug side effects associations from the Lexicom website (<http://www.lexi.com>), where taxonomic data are collected from World Health Organization (WHO) (WHO et al. 2020) and biological data from DrugBank (Wishart et al. 2018). Their experiment consists of 809 drugs and 852 side effects. The LR model achieved the highest AUROC score of 87%. LR and SVM classifiers are applied by Huang et al. (2011) to identify the cardiotoxicity side effects of drugs. The authors obtained drug targets from the DrugBank (Wishart et al. 2018), protein-protein interactions from Human Protein-Protein Interactions (HAPPI) (Chen et al. 2009), and side effects from the SIDER (Kuhn et al. 2010). The dataset

consists of 887 drugs corresponding to 1447 side effects. To reduce the high dimension of data, the authors used Fisher's exact test (FET) and Wilcoxon Rank-SUM Test (WRST) approach and handled the class imbalance with the help of the sample balancing method. Further, the authors obtained that the SVM performed the best and achieved the highest accuracy value of 67.50%. Yamanishi et al. (2012) predict side effects by integrating biological and chemical properties. Their dataset consists of 658 drugs and 969 side effects, which are collected from SIDER (Kuhn et al. 2010), chemical structures are collected from the PubChem (Kim et al. 2019), and biological properties (drug-protein interaction) were collected from DrugBank (Wishart et al. 2018). The authors applied the canonical correlation analysis (CCA), kernel regression, and multiple kernel regression model. They found that the kernel regression model performed well when applied to the integration of chemical and biological properties and achieved a 20.89% AUPR score.

In Liu et al. (2012a), the authors integrated phenotypic, chemical, and biological drug descriptors to determine the hazardous reactions of drugs. Their dataset consists of 832 drugs corresponding to 1385 side effects. The authors obtained phenotypic properties of drugs from SIDER (Kuhn et al. 2010), chemical information from PubChem (Kim et al. 2019), and biological properties from KEGG (Kanehisa et al. 2002), and DrugBank (Wishart et al. 2018). The authors applied the machine learning classifiers (SVM, KNN, LR, random forest (RF), and NB) on the three-level (Che+Bio+Phe), two-level (Che+Bio and Che+Phe), and individual drug descriptors, and their experiment revealed that the SVM classifiers achieved the highest accuracy score of 96.78% than the other approaches. A novel sparse canonical correlation analysis (SCCA) approach is introduced by Mizutani et al. (2012) to predict drug side effects from the chemical structure and drug-protein interactions. They collected chemical structures from PubChem (Kim et al. 2019), drug-protein interaction from the DrugBank (Wishart et al. 2018), and KEGG (Kanehisa et al. 2002), and side effects from the SIDER (Kuhn et al. 2010). The final dataset consists of 658 drugs and 1339 side effects. The authors compared their proposed model performance with Ordinary Canonical Correlation Analysis (OCCA). They found that SCCA performed the best compared to the OCCA and achieved the highest AUC Score of 88.95% on drug-protein interaction information. In their work, Liu et al. (2012b) predict side effects from the medical record provided by the Stanford Clinical Data Warehouse (STRIDE) (Lowe et al. 2009). The dataset consists of 1,550 drug-side effects pairs, and they trained the SVM classifier to identify side effects that achieved the highest AUC score of 85%. Further, Zhang et al. (2013) explored the potential association between Therapeutic Indication (TI) and side effects. The authors examine the prediction power of therapeutic indication information, drug Chemical Structure (CS), and target Protein (P) to predict side effects. The authors collected drug structure from PubChem (Kim et al. 2019), protein target from UniProt (Consortium 2015), and DrugBank (Wishart et al. 2018), therapeutic indication (drug-disease relation) from the National Drug File Reference Terminology (NDFRT), which is a portion of Unified Medical Language System (UMLS) (Bodenreider 2004), and side effects from SIDER (Kuhn et al. 2010). Their dataset comprises 1,385 side effects and 1,447 drugs. The authors applied the machine learning approach (SVM, LR, RF, and NB) to three-level combinations (CS+P+TI), two-level (CS+P, CS+TI, TI+P), and individual datasets. Among them, LR achieved the highest AUC in the combination of protein and therapeutic indication, and it achieved the highest ROC-AUC of 71.03%. An ensemble method is proposed by Jahid and Ruan (2013) to predict side effects. Their dataset consists of 888 drugs and 1385 side effects that are collected from the SIDER (Kuhn et al. 2010) and chemical structure collected from PubChem (Kim et al. 2019). The authors compared the proposed model with SCCA and found that their model achieved the highest ROC-AUC

score of 84%. The authors also predict unseen side effects of 2882 drugs, which are available in DrugBank (Wishart et al. 2018). Jiang and Zheng (2013) predict potential side effects from the Twitter post by employing supervised learning algorithms, including NB, Maximum Entropy (ME), and SVM. The authors collected 6829 tweets against 5 drugs. Among them, the maximum entropy classifier achieved the highest F-measure score of 84.8%. In Huang et al. (2013), Huang et al. applied SVM on integrated drug properties to predict drug side effects. The authors obtained chemical substructure from PubChem (Kim et al. 2019), protein-protein interaction from HAPPI (Chen et al. 2009), drug target from DrugBank (Wishart et al. 2018), and side effects from SIDER (Kuhn et al. 2010), which comprise 1447 side effects and 887 drugs. The authors compared SVM model performance on individual, two-level, and three-level properties and found that SVM performed better when combining protein-protein interaction, drug target, and chemical structure and achieved the highest AUC score of 70%. LaBute et al. (2014) predict side effects using the L1 regularized LR model from the drug-protein target. The authors collected side effects from the SIDER (Kuhn et al. 2010), drug-protein targets from the UniProt (Consortium 2015), and DrugBank (Wishart et al. 2018). The final dataset consists of 560 drugs and 85 side effects. Their proposed methodology achieved the highest AUC of 74%. By employing NB and SVM, Ginn et al. (2014) predict side effects from the Twitter post. They collected 10,822 tweets against 76 drugs. The SVM model performs well compared to the NB and achieved the highest accuracy of 76.6%. A general weighted profile (GWP) method is presented by Kuang et al. (2014) to predict drug side effects. The authors collected drugs from the FAERS (Vermeer et al. 2013), DrugBank (Wishart et al. 2018), KEGG (Kanehisa et al. 2002), SIDER (Kuhn et al. 2010), and drug-side effects associations from the FAERS and SIDER, which consists of 404 drugs and 461 side effects. They applied the GWP, nearest neighbor, link prediction, and regularized least-squares classification approach to predict side effects. The authors found that the proposed GWP approach achieved the highest AUPR score of 26.90%.

In their work, Zhang et al. (2015) proposed an essential Feature Selection-Based Multi-Label KNN (FS-MLKNN) approach to predict side effects from the drug targets and chemical information which are collected from DrugBank (Wishart et al. 2018) and PubChem (Kim et al. 2019), respectively. Their dataset consists of 1080 drugs and 2260 side effects which are collected from SIDER (Kuhn et al. 2010). They also used the benchmark datasets (Mizutani's 2012, Liu et al. 2012a, and Pauwel's 2011 datasets) to compare their proposed model with Mizutani's, Liu's, and Pauwel's methods. Their proposed model performed well on all three benchmark datasets and achieved an AUPR score of 42.86%, 40.04%, and 48.02% on Pauwel's, Mizutani's, and Liu's datasets, respectively. Kanji et al. (2015) identify side effects of drugs from the 3D chemical properties, 2D chemical properties, and drug target information. The authors collected 996 drugs, and 4192 side effects from the SIDER (Kuhn et al. 2010), and drug target information from the DrugBank (Wishart et al. 2018). Further, they generated 3D and 2D chemical properties using the molecular properties module of Discovery Studio 4.0. The authors applied an ordinary canonical correlation (OCC) approach to their collected properties and found that their proposed model OCC achieved the highest AUC score of 92% on 3D chemical properties. A boltzmann machine-based approach (BMBA), Integrated Neighborhood-Based Approach (INBA), and an ensemble method are employed by Zhang et al. (2016a) to predict side effects from biological and chemical information. The authors used Liu et al. 2012a, Mizutani's 2012, and Pauwel's 2011 data as benchmark datasets. The authors found that their ensemble approach performed the best and achieved the highest accuracy of 96.20%. In Niu et al. (2015), a drug side effects prediction (DSEP) neural network approach is proposed by

Niu et al. to predict the drug side effects. Their dataset consists of 4,192 side effects and 996 drugs, which are collected from SIDER (Kuhn et al. 2010). The authors used chemical descriptors, drug targets, and chemical substructures as input features for their model, which are collected from KEGG (Kanehisa et al. 2002), DrugBank (Wishart et al. 2018), and Pubchem (Kim et al. 2019), respectively. Their proposed model achieved the highest AUC of 89.27%.

To predict side effects, a variational Bayesian ensemble learning approach based on SVM, KNN, LR, and extreme LR is provided by Ngufor et al. (2015). The authors collected side effects from SIDER (Kuhn et al. 2010), which consists of 1450 side effects and 888 drugs. They collected biological and chemical properties from DrugBank (Wishart et al. 2018), and demographic data (age, gender, occur country, weight, route of administration, dose amount, and reporter country) were collected from the FAERS database (Vermeer et al. 2013). Their proposed methodology achieved the highest AUC score of 80% compared to the majority voting, sum rule, and stacking ensemble method. In Jamal et al. (2017), the authors aim to identify neurological side effects from the combination of phenotypic (Pheno), Biological (Bio), and Chemical (Ch) information of compounds. The authors reduced the dimension of the input vector by applying a feature extraction approach named the relief-based method and addressed the class imbalance issue using the synthetic minority oversampling technique (SMOTE) technique. The authors downloaded the phenotypic properties (side effects and therapeutic indication) from SIDER (Kuhn et al. 2010), biological properties (target, enzyme, and transporters) from DrugBank (Wishart et al. 2018), and Chemical information from PubChem (Kim et al. 2019). Their dataset consists of 22 neurological side effects and 913 drugs. Further, the authors applied a support vector machine on the three-level (Phe+Bio+Che), two-level (Phe+Bio, Phe+Che, and Bio+Che), and their model revealed that SVM on the combination of three drug descriptors performs well and it achieved the accuracy of 94.18%. Wang et al. (2016) integrated gene information and chemical structure and applied extra tree classifier (ETC), LR, SVM, and RF for side effects prediction. The authors collected gene information from the LINCS L1000 (Duan et al. 2016), which consists of more than 20,000 small molecules and chemical structures from PubChem (Kim et al. 2019). Further, they collected side effects from the SIDER (Kuhn et al. 2010), OFFSIDE (Tatonetti et al. 2012a), and OMOP (Ryan et al. 2012), which consists of '4192 ADR, and 996 drugs', '1,322 drugs, and 10,097 side effects', 60, 40, 62, 78 drugs for liver failure, gastrointestinal ulcer, kidney failure, and myocardial infarction respectively. Furthermore, the authors found that ETC performed better than others and achieved the highest accuracy (more than 90%) by integrating gene ontology and chemical structure. In Zhang et al. (2016b), the authors integrated different drug information and presented a novel LNSM (Linear Neighborhood Similarity Method) to predict side effects. The authors used Pauwels's 2011, Liu et al. 2012a, and Mizutan's 2012 datasets. Further, it extended the LNSM and proposed two models named as CMI (Cost Minimization Integration) approach and SMI (Similarity Interaction approach). They applied their proposed model, which performed the best compared to the other classifiers (FS-MLKNN, Pauwels's 2011, Liu et al. 2012a, and Mizutan's 2012 methods), and they found that the LNSM-CMI method achieved the highest AUC score of 90.91%. Raja et al. (2017) aim to enhance the efficiency of the machine learning approach to identify drug-drug interaction and side effects from drug-gene interactions. The authors retrieved more than 5,00,000 drug-gene interactions from the comparative toxicogenomics database (CTD) (Davis et al. 2015) and 8176 drugs from MEDLINE (Conn et al. 2003), and DrugBank (Wishart et al. 2018). They handled the class imbalance issue by employing SMOTE and applied RF, KNN, random tree (RT), DT, and bayesian network (BN) classifiers to

predict side effects. The authors found that the RT classifier achieved the highest F-score of 90%. In their work (Hu et al. 2017), the authors presented a novel second order association discovery (SOAD) approach to identify side effects from the drug substructure. They used Pauwels's 2011 dataset as a benchmark dataset and collected drug substructure from the KEGG (Kanehisa et al. 2002), PubChem (Kim et al. 2019), and DrugBank (Wishart et al. 2018), as well as side effects from the SIDER (Kuhn et al. 2010). Their dataset consists of 888 drugs and 1385 side effects. The authors applied SOAD, SCCA, SVM, NB, and OCCA to the dataset. Their experiment outcome suggests that the SOAD method has the best predictive ability and achieved the highest AUROC score of 88.10%. A Linear Neighborhood Similarity Measure (LNSM) method is presented by Zhang et al. (2017) to predict side effects. The authors collected chemical structure from PubChem (Kim et al. 2019), biological information from DrugBank (Wishart et al. 2018), and KEGG (Kanehisa et al. 2002). They also utilized Pauwels's 2011, Liu et al. 2012a, and Mizutani's 2012 benchmark datasets, and their dataset consists of 1,080 drugs and 2260 side effects. The authors compared their proposed model with Liu's method (Liu et al. 2012a), FS-MLKNN method, and found that their proposed model achieved the highest AUC score of 94.8%. In a different work, a novel three-interval classification method is introduced by Lee et al. (2017) to identify drug side effects by integrating multiple drug properties. The authors collected substructure of drugs information from DrugBank (Wishart et al. 2018), and PubChem (Kim et al. 2019), protein information from DrugBank (Wishart et al. 2018), and UniProt (Consortium 2015), side effects from the SIDER (Kuhn et al. 2010), and the relationship between therapeutic indication and drug-disease are collected from the NDFRT (Bodenreider 2004). Their dataset consists of 1002 drugs and 3903 side effects. The authors compared their three-interval approach with RF, NB, and KNN, and the proposed model performs well compared to other classifiers, which achieved the highest accuracy of 95.20%. In Muñoz et al. (2019), the authors predict side effects by employing Knowledge Graph Similarity PROPagation (KG-SIM-PROP), LR, multi-layer perceptron (MLP), RF, DT, KNN, LNSM-CMI, LNSM-SMI, FS-MLKNN, and Liu's method (Liu et al. 2012a). They found that FS-MLKNN achieved the highest ROC-AUC value of 90.34%. The authors obtained drug types, chemical information from DrugBank (Wishart et al. 2018), side effects from SIDER (Kuhn et al. 2010), pathways, genes, and drugs from KEGG (Kanehisa et al. 2002). Their dataset consists of 1,080 drugs and 5,579 side effects and also utilized Liu et al. 2012a, Bio2RDF (Belleau et al. 2008), and Aeolus (Banda et al. 2016) benchmark datasets.

Dey et al. (2018) describe the importance of combining deep learning frameworks and machine learning models to identify chemical substructure relations with side effects by generating distinct fingerprints from the compound structure with the help of convolutional deep learning methodology. The authors collected chemical substructure from PubChem (Kim et al. 2019) and side effects from the SIDER (Kuhn et al. 2010), and their dataset consists of 1420 drugs and 6123 side effects. Their proposed model achieved the highest accuracy of 97.70% on skin striae side effects. In their work, Zheng et al. (2018) analyzed and identified side effects of drugs by applying Highly Credible Negative Samples (HCNS) from pathways, target protein, chemical substructure, substituents, and gene-disease relationship. Their dataset consists of 1048 drugs and 1276 side effects. The authors collected target protein chemicals, substructure, and substituents from the DrugBank (Wishart et al. 2018). The gene-disease relationship is obtained from the comparative toxicogenomics database (CTD) (Davis et al. 2015), and the side effects-drug association is collected from the Tatonetti Lab (Tatonetti et al. 2012b). The authors applied RF, KNN, LR, and SVM with HCNS and found that LR with HCNS achieved the highest accuracy of 98%. A graph

side effects (GraphSE) prediction model is presented by Hu et al. (2018) to associate chemical substructure and side effects. The authors used Pauwels's 2011 dataset, which included 1,385 side effects and 888 drugs, and Liu's dataset (Liu et al. 2012a) comprised 1385 side effects and 832 drugs. They collected 881 chemical structures from KEGG (Kanehisa et al. 2002), DrugBank (Wishart et al. 2018), and PubChem (Kim et al. 2019) for the test. The authors applied SVM, GraphSE, OCCA, NB, GraphSE-RanksClass, and SCCA and found that their proposed GraphSE model performs the best compared to the other classifiers. It achieved the ROC of 88.70% in Liu's dataset and 89.20% in Pauwel's data set. A similarity-based approach is provided by Zhao et al. (2018) to identify the harmful response of drugs from the heterogeneous drug information (fingerprint similarity, structure similarity, anatomical therapeutic chemical (ATC) code, literature information, and target protein). They collected fingerprints from the DrugBank (Wishart et al. 2018), drug structure from the KEGG (Kanehisa et al. 2002), ATC code and literature information from the STITCH (Szkarczyk et al. 2016), target protein from the DrugBank (Wishart et al. 2018), and side effects from SIDER (Kuhn et al. 2010). Their dataset consists of 888 drugs and 1,385 side effects. Further, they found that their proposed method performed well than the SVM, dagging, and nearest neighbors method and achieved the highest AUC of 84.92%. Liu et al. (2018) predict side effects of osteoarthritis disease from the electronic medical records, which are collected from the osteoarthritis initiative dataset (Razmjoo et al. 2021). The authors applied SVM, XGBoost, gradient boosting decision tree (GBDT), DT, and LR and found that XGBoost performed well than the others and achieved a ROC-AUC value of 92%. In Zitnik et al. (2018), the authors predict polypharmacy drug pairs side effects with the help of a graph convolutional network named Decagon. Their dataset comprises 5,868 side effects and 1556 drugs obtained from the SIDER (Kuhn et al. 2010). The authors extracted protein-protein interactions from Menche et al. (2015), Chatr-Aryamontri et al. (2015), and polypharmacy side effects are extracted from TWOSIDES (Tatonetti et al. 2012a). Their presented method achieved an AUROC score of 87.20%.

A methodology is proposed by Islam et al. (2018) to detect death side effects from the electronic health record. The authors obtained data from the Food and Drug Administration (FDA) (Edwards et al. 2013). Their dataset includes demographic data as input, including preferred term, dechal, route, patient weight, sex, and age; output labels include minor injury, major injury, and death. They employed MLP, RF, and SVM and obtained that RF classifiers achieved the highest death prediction accuracy score of 91.4% compared to others. Hu et al. (2018) provided a deep heterogeneous information embedding approach to predict side effects. The authors obtained chemical structure from PubChem (Kim et al. 2019), target protein information and treatment disease from Drugbank (Wishart et al. 2018), drug-drug interaction from TWOSIDES (Tatonetti et al. 2012a), and protein-protein interaction from human protein reference database (HPRD) (<http://www.hprd.org>). Their final dataset consists of 548 drugs and 1318 side effects, which are collected from the SIDER (Kuhn et al. 2010) and OFFSIDES (Tatonetti et al. 2012a). Their proposed model achieved a ROC-AUC score of 84.07% than graph convolutional neural network (GraphCNN). In their work, Uner et al. (2019) evaluate the deep neural network model performance to analyze the hazardous reactions of the drug by merging chemical structure (CS), gene ontology (GO), gene expression META information of drugs, and gene expression signature (GEX). The authors analyzed the performance of 'Simplified Molecular-Input Line-Entry System-convolutional neural network (CNN)' (SMILES-CNN), MLP, multi-task neural network, multi-model neural network, and residual MLP neural network on the different levels of drug description combination such as three-level (META+CS+GEX), two-level (GO+CS, GEX+CS), and individual level. Their

SMILES-CNN model on chemical descriptors exhibits better results compared to the other methods and drug descriptors and their combination, and it achieved the highest micro AUC of 88.50%. The authors collected gene information from the LINCS project (Duan et al. 2016), and side-effects from SIDER (Kuhn et al. 2010), and the SMILES are collected from PubChem (Kim et al. 2019). Their dataset consists of 1,052 side effects and 791 drugs. Odeh and Taweel (2019) aims to extract side effects from Twitter posts by applying the deep learning method. The authors collected 7574 Twitter posts for 74 drugs to check whether the Twitter posts have side effects or not. They used convolutional neural network (CNN) with attention (CNNA), recurrent convolutional neural network (RCNN), CNN, majority vote RCNN, ADR-Classifer, 'CNN-and-Google-News' and found that their proposed model achieved the highest precision score of 77.93% than the other approaches. Jamal et al. (2019) predict cardiovascular drug side effects by combining phenotypic, chemical, and biological information. The authors obtained biological and chemical information from the DrugBank (Wishart et al. 2018), and phenotypic properties from the SIDER (Kuhn et al. 2010), and the dataset consists of 970 drugs and 36 cardiovascular side effects. The authors combined drug descriptors into three-level combinations (Ch+Pheno+Bio), two-level (Che+Phe, Che+Bio, and Phe+Bio), and individual drug properties. Further, the authors applied sequential minimization optimization (SMO) and RF classifier to predict the side effects of drugs. Remove useless filter (RUF) is used to reduce the high dimension of data and address the class imbalance issue with the help of SMOTE. The authors obtained that their experiment in both models achieved the highest accuracy score of 93.83% on phenotypic drug properties compared to the other drug properties and their combinations.

A deep learning model is presented by Wang et al. (2019) to analyze the side effects from the different drug descriptors. The authors collected the textual information of biomedical literature from MEDLINE (Conn et al. 2003), 17 drug-like properties from PubChem (Kim et al. 2019), and biological properties (carriers, enzymes, target, and transporters) from DrugBank (Wishart et al. 2018) and integrated them into a single dataset. Further, the MLP model with two hidden layers is applied to the integrated drug properties. Their proposed model achieved the highest AUC of 84.40% and performed well compared to the other classifier models, such as Gaussian NB, Linear SVM, and probability matrix factorization (PMF). In Swathi et al. (2020), the authors analyze the side effects from the medical health forums. The authors obtained the data (satisfaction rating, age, gender, comments, information on health, effectiveness, disease, symptoms, and number of dosages) from WebMB.com. The authors applied linear SVC, LR, SVM, DT, NB, and RF and obtained that the RF classifier have the best predictive ability and achieved the highest accuracy score of 62.40%. Liang et al. (2020) used an advanced negative instance selection approach to analyze side effects. The authors collected drug fingerprints, drug literature, and ATC code from the STITCH (Szkłarczyk et al. 2016), drug structure from the KEGG (Kanehisa et al. 2002), and drug target prediction from the DrugBank (Wishart et al. 2018). Their dataset consists of 1385 side effects and 888 drugs which are downloaded from the SIDER website (Kuhn et al. 2010). The authors applied artificial neural network (ANN), SVM, and RF to predict side effects and found that the RF model achieved the highest accuracy of 97.50%. They applied the random walk with restart approach to select the high-quality negative instances. In Zhou et al. (2020), the authors predict side effects from protein targets, treatments, transporter, enzyme, pathways, and chemical structure of drugs using boosted random forest classifier. The authors collected drug indication data from <https://clinicaltrials.gov/>, SIDER (Kuhn et al. 2010), and therapeutic targeted database (Li et al. 2018). Their dataset consists of 4251 side effects and 1426 drugs, which are collected

from the SIDER (Kuhn et al. 2010). The authors used PharmacotherapyDB (<https://think-lab.github.io/d/182/>), DrugCentral (<https://drugcentral.org/>), and ClinicalTrialsSlim (<https://clinicaltrials.gov/>) datasets to compare their models. Their proposed model achieved the highest precision score of 78% to predict side effects and performed the best compared to the KNN and MLP classifiers. Further, Shankar et al. (2021) identify side effects of drug-pair from the gene expression and chemical structure by employing an artificial neural network model. The authors collected gene expression from LINC L1000 (Duan et al. 2016), chemical substructure from DrugBank (Wishart et al. 2018), and side effects obtained from TWOSIDES (Tatonetti et al. 2012a). Their database comprises 34,549 drug-pair and 243 side effects. Their model achieved the highest accuracy score of 82%. A methodology is presented by Ietswaart et al. (2020) to associate drug and side effects from pharmacological assays by employing an RF classifier. The authors collected side effects from SIDER (Kuhn et al. 2010), and FAERS (Vermeer et al. 2013), and pharmacovigilance assays were obtained from the Novartis Institutes of Biomedical Research (NIBR) (Tuntland et al. 2014). Their database comprises 2134 drugs and 40 side effects. Their model achieved the highest accuracy score of 98%. In Das et al. (2021), a multi-label machine learning methodology is presented to detect the side effects. The authors checked whether drug functions are utilized to predict harmful side effects or not. The authors extracted 10,819 drugs corresponding to the 12 drug functions from PubChem (Kim et al. 2019) and 6,123 side effects collected from the SIDER (Kuhn et al. 2010). After mapping the drug with side effects, the final dataset consists of 670 drugs corresponding to those 6123 side effects. They applied the binary relevance method to solve the problem of multi-label task with five multi-label supported supervised machine learning approaches, namely RF, MLP, KNN, ETC, and DT. The authors found that ETC achieved the highest accuracy of 99.95%.

In their work, Hatmal et al. (2021) identify and analyze COVID-19 vaccine side effects. The authors collected data through an online survey of 2213 participants. Further, the authors applied the supervised machine learning approach (RF, K-star, XGboost, and MLP) to predict side effects. They found that the RF classifier achieved the highest accuracy of 80% compared to others. In a different work, Hatmal et al. (2022) predict side effects of the coronavirus disease vaccine from a survey of 10,064 participants. Their survey consists of age category, infected with COVID-19 before vaccination, is suffering from any other diseases?, gender, is a smoker?, educational level, the participant's symptoms, and type of the vaccine (AstraZeneca, Johnson & Johnson, Sputnik V, Covaxin and Moderna). The authors applied gradient boost, MLP, RF, and XGBoost methods and found that the Xgboost classifier performed the best, and achieved an accuracy value in the range of 74–58%. Güneş et al. (2021) predict side effects of antidepressant drugs from chemical and biological properties. Their dataset consists of 329 side effects and 27 drugs. The authors collected biological properties (transporter, enzyme, and drug target) from the DrugBank (Wishart et al. 2018), side effects are obtained from Drug.com, and SIDER (Kuhn et al. 2010), and drug structure from PubChem (Kim et al. 2019). The authors applied SVM, KNN, and MLP classifiers on a single level and integrated drug properties; among them, MLP classifiers perform well on integrated chemical and biological properties, which achieved the highest AUC of 69.50%. In Das et al. (2022a), A deep neural network (DNN) is presented to identify side effects from the 17 molecules, drug functions, and chemical 1D structure. The authors collected drug properties from PubChem (Kim et al. 2019) and side effects from the SIDER (Kuhn et al. 2010). Their dataset consists of 1430 drugs and 6123 side effects. Further, they integrated the three-level combination (17 molecules + drug functions + chemical 1D structure), two-level combination (17 molecules + drug functions, 17 molecules + chemical 1D structure, drug functions + chemical 1D structure), and individual

drug properties with side effects. Furthermore, they applied a DNN to predict side effects. The experimental outcomes showed that the combination of drug functions and chemical 1D structure performs better than the other drug properties combinations. The DNN on the combination of drug functions and chemical 1D structure achieved the highest ROC-AUC score of 99.99%. Zhao et al. (2022) introduce a similarity-based deep neural network methodology to identify the frequency of drug side effects. The authors used the Galeano et al. (2020) dataset. Further, they divided side effects frequency into five classes, namely very frequent, frequent, infrequent, rare, and very rare. The authors collected drug information and drug-drug association score, chemical substructure from STITCH database (Szkarczyk et al. 2016), drug target from DrugBank (Wishart et al. 2018), and side effects and their frequency are obtained from SIDER (Kuhn et al. 2010). Their database comprises 757 drugs, 994 side effects, and 37,366 frequency items. Their proposed model performs well compared to Galeano's et al. (Galeano et al. 2020) method and achieved the highest AUROC score of 94.52%.

The summarization of related work is shown in Fig. 4 and Table 1, which comprises of Reference (Ref.) and the tasks solved by the researcher to predict the side effects of drugs from several drug descriptors by applying computational approaches.

It can be observed from the Fig. 4, the literature survey of supervised machine learning models to predict drug side effects have been conducted based on the tasks, which have been addressed by the researcher for a duration of 18 years (2004–2022). The literature survey includes the data sources of drugs, drug properties, and side effects that are utilized to conduct the drug side effects prediction tasks. The information related to the drug is available in several data sources, including PubChem, SIDER, DrugBank, WHO, HAP, KEGG, STRIDE, UniProt, NDFRT, Twitter, FAERS, LINC L1000, OMOP, MEDLINE, WebMD, and DrugCentral, etc (Kanehisa et al. 2002; Conn et al. 2003; Bodenreider 2004; Chen et al. 2009; Lowe et al. 2009; Kuhn et al. 2010; Ryan et al. 2012; Vermeer et al. 2013; Consortium 2015; Duan et al. 2016; Wishart et al. 2018; Kim et al. 2019; WHO et al. 2020).

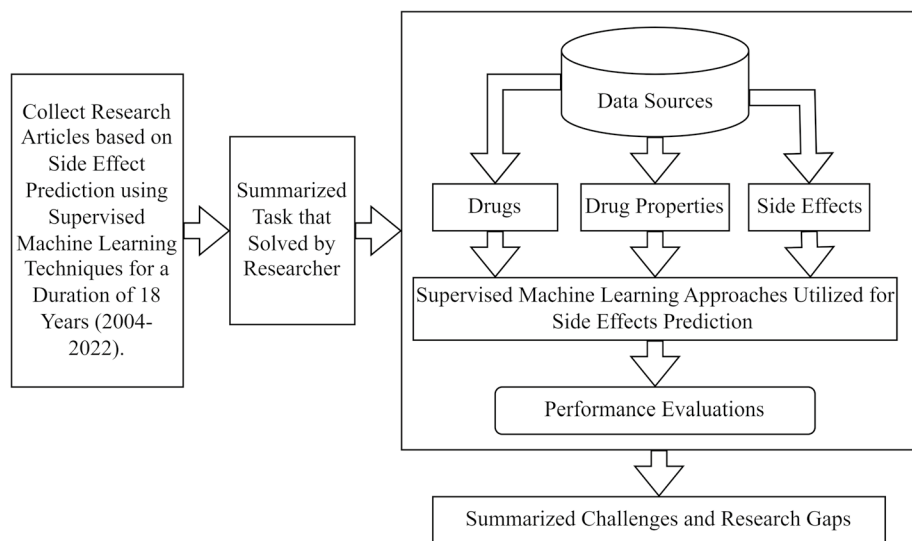


Fig. 4 Summary of the literature survey based on the supervised machine learning techniques in the past two decades for the prediction of drug side effects

Table 1 Summary of computational approaches, side effects prediction tasks, drug properties, and their sources

Ref.	Task	Drug properties	Data source(s)	Approach(es)
Yap et al. (2004)	Predict torsade de pointes (TDP) side effects of drug	Hydrogen-bond term, a cavity term, solvent-solute interactions, and a polar term	University of Arizona	PNN, DT, SVM, KNN
Scheiber et al. (2009)	Associating chemical features with side effects	Cheminformatics	PharmaPendium	Laplacian-based NB classifier
Pouliot et al. (2011)	Utilized BioAssay information to predict side effects	BioAssay information	PubChem and Canadian ADR pharmacovigilance database	LR
Pauwels et al. (2011)	Predict potential side effects from the chemical structure	Chemical structure	SIDER, PubChem	OCCA, RA, SVM, SCCA, KNN
Cami et al. (2011)	Identify side effects events using the LR model	Physiochemical, 16 drug molecule properties, taxonomic data, biological properties	DrugBank, PubChem, WHO	LR
Huang et al. (2011)	Identify cardiotoxicity side effects of drugs	Gene ontology, protein-protein interaction, drug-target information	SIDER, DrugBank, HAPPI	LR, SVM
Yamanishi et al. (2012)	Predict side effects by integrating biological and chemical properties	Chemical structure, biological properties (drug-protein interaction)	SIDER, DrugBank, PubChem	CCA, kernel regression, multiple kernel regression model
Liu et al. (2012a)	Investigate side effects from the integrated drug descriptors	Chemical information, therapeutic indication, biological properties	PubChem, KEGG, DrugBank, SIDER	SVM, KNN, LR, RF, NB
Mizutani et al. (2012)	Introduce a novel SCCA approach to identify the side effects from the chemical structure and drug-protein interactions	Chemical structure, drug-protein interactions	PubChem, DrugBank, SIDER, KEGG	OCCA, SCCA
Liu et al. (2012b)	Predict side effects from the medical record	Medical record	STRIDE	SVM
Zhang et al. (2013)	Explored the potential association between therapeutic indication and side effects	Chemical information, target protein, therapeutic indication	SIDER, DrugBank, PubChem, UniProt, NDFRT	SVM, LR, RF, NB
Jahid and Ruan (2013)	An ensemble method is proposed to predict side effects	Chemical structure	SIDER, PubChem, DrugBank	Ensemble, SVM, LR, SCCA

Table 1 (continued)

Ref.	Task	Drug properties	Data source(s)	Approach(es)
Jiang and Zheng (2013)	Predict potential side effects from the Twitter data	Tweet of the drug user	Twitter	ME, SVM, NB
Huang et al. (2013)	Apply SVM on integrated drug properties to identify side effects	Drug target, chemical structure, and protein-protein interactions	PubChem, DrugBank, SIDER	SVM
LaBute et al. (2014)	Predict side effects using the L1 regularized LR model from the drug-protein target	Drug-protein target	DrugBank, UniProt, SIDER	L1 regularized LR model
Ginn et al. (2014)	Predict side effects from the Twitter dataset	Tweet of the drug user	Twitter	NB, SVM
Kuang et al. (2014)	Proposed a general weighted profile approach to detect side effects	Drug-side effects association	FAERS, DrugBank, KEGG, SIDER	GWP, nearest neighbor, link prediction, regularized least-squares classification approach
Zhang et al. (2015)	Proposed an essential FS-MLKNN model to predict side effects	Chemical information, drug target	PubChem, DrugBank	FS-MLKNN, Mizutani's, Liu's, Pauwel's methods
Kanji et al. (2015)	Identify side effects of drugs from the 3D chemical properties, 2D chemical properties, and drug target information	3D chemical properties, 2D chemical properties, drug target information	SIDER, DrugBank	OCC
Zhang et al. (2016a)	Machine-based approach (BMBA), integrated neighborhood-based approach (INBA), and their ensemble method are presented to predict side effects	Chemical and biological information	Liu's, Pauwel's datasets, Mizutani's	Machine-Based Approach (BMBA), Integrated INBA, ensemble methods, Liu's, Pauwel's, Mizutani's methods
Niu et al. (2015)	A drug side effects prediction model is proposed	Chemical descriptors, drug target, chemical substructure	KEGG, DrugBank, Pubchem, SIDER	Neural Network (NN), kernel regression, SCCA, SVM
Ngufo et al. (2015)	A variational Bayesian ensemble learning approach is provided to predict drug side effects	Chemical, biological, and demographic data	SIDER, DrugBank, FAERS	Variational Bayesian, majority voting, sum rule, and stacking ensemble learning approach, SVM, KNN, LR, extreme LR

Table 1 (continued)

Ref.	Task	Drug properties	Data source(s)	Approach(es)
Jamal et al. (2017)	Identified neurological drug side effects	Chemical information, therapeutic indication, and biological properties	PubChem, DrugBank, SIDER	SVM
Wang et al. (2016)	Predict side effects by integrating chemical structure and gene information	Gene expression, gene ontology, chemical structure	LINCS L1000, SIDER, OMOP, FAERS	ETC, LR, SVM, RF
Zhang et al. (2016b)	Integrated different data sources of drugs and presented a novel linear neighborhood similarity (LNSM) method to predict side effects	Biological information (target, pathways, enzyme, treatment, and transporter), chemical substructure	SIDER, PubChem, DrugBank, KEGG, Mizutani's dataset, Pauwels's, Liu's, dataset	FS-MLKNN, Pauwels's, Liu's, Mizutani's methods
Raja et al. (2017)	Aim to enhance the performance of machine learning to predict side effects from the drug-gene interactions	Drug-gene interaction, drug-drug interaction	MEDLINE, DrugBank, CTD	KNN, RT, DT, BN, RF
Hu et al. (2017)	Presented a novel second-order association discovery approach to identify side effects	Drug substructure	Pauwels's dataset, KEGG, PubChem, DrugBank	SOAD, SCCA, SVM, NB, OCCA
Zhang et al. (2017)	Proposed a linear neighborhood similarity measure approach to predict side effects	Chemical structure, protein information (transporter, pathways, target, enzyme, indication, carrier)	PubChem, DrugBank, KEGG, Pauwels's, Liu's, Mizutani's benchmark datasets	LNSM, Liu's method, FS-MLKNN method
Lee et al. (2017)	Introduce a novel three-interval classification method to identify drug side effects	Chemical substructure, protein information, therapeutic indication, drug-diseases treatment relation	NDFERT, PubChem, DrugBank, UniProt, SIDER	RF, NB, KNN, three-interval
Muñoz et al. (2019)	Predict side effects from distinct drug descriptors	Drug types, chemical information, pathways, and genes	DrugBank, SIDER, KEGG	LR, MLP, RF, DT, KNN, KG-SIM-PROP, LNSM-CMI, LNSM-SMI, FS-MLKNN, Liu's method

Table 1 (continued)

Ref.	Task	Drug properties	Data source(s)	Approach(es)
Dey et al. (2018)	Describe the importance of the integration of deep learning and machine learning to identify side effects	Chemical fingerprint	PubChem, SIDER	Convolution deep learning, L2 norm regularized LR
Zheng et al. (2018)	Analysis and identify side effects of drugs from the combination of different pharmacologic properties	Pathways, target protein, chemical structure, substituents, gene-disease relationship	DrugBank, Toxicogenomics Database	RF, KNN, LR, SVM, HCNS
Hu et al. (2018)	Presented a GraphSE model to associate chemical substructure and side effects	Chemical substructure	PubChem, DrugBank, KEGG	SVM, SCCA, GraphSE, OCCA, NB, GraphSE-RanksClass
Zhao et al. (2018)	A similarity-based approach is proposed to predict side effects	Fingerprint, drug structure, ATC code, literature information, target protein	KEGG, DrugBank, STITCH, SIDER	SVM, dagging, similarity-based, nearest neighbors method
Liu et al. (2018)	Predict side effects of osteoarthritis drugs	Electronic medical records	Osteoarthritis initiative dataset	XGBoost, DT, SVM, GBDT, LR
Zitnik et al. (2018)	Predict polypharmacy side effects of drug pair	Protein-protein interactions, drug-drug interactions	OFFSIDES, TWOSIDES, SIDER, Menche's, Chattr's dataset	Decagon
Islam et al. (2018)	Proposed a methodology to detect death side effects from the electronic health record	Electronic health record	FDA	Multi-layer perceptron neural network, random forest, support vector machine
Hu et al. (2018)	Proposed a deep heterogeneous information embedding method to predict side effects	Chemical structure, target protein information, treatment disease, drug-drug interaction, protein-protein interaction	FAERS, TWOSIDES, PubChem, HPRD, OFFSIDES, SIDER, DrugBank	GraphCNN, deep heterogeneous information embedding method
Uner et al. (2019)	Evaluate the deep learning models performance to identify side effects of drugs	Gene ontology, gene expression, chemical structure, META information of drugs	PubChem, SIDER, LINC1000	Residual multi-layer perceptron neural network, SMILES-CNN, MLP, multi-model neural network, multi-tasking neural network

Table 1 (continued)

Ref.	Task	Drug properties	Data source(s)	Approach(es)
Odeh and Taweel (2019)	Aim to extract side effects from the Twitter post by applying the deep learning method	Twitter posts	Twitter	Convolutional neural network, Convolutional Neural Network with Attention, Recurrent Convolutional Neural Network, majority vote CRNN, ADR-Classifer, CNN-and-Google News
Jamal et al. (2019)	Identify potential cardiovascular side effects	Chemical, biological, therapeutic indication properties	DrugBank, SIDER	Sequential minimization optimization approach, random forest
Wang et al. (2019)	Presented a DNN model to detect the side effects by integrating different drug descriptors	17 drug-like properties, biomedical literature, biological properties	PubChem, DrugBank, MEDLINE, SIDER	Gaussian NB, linear SVM, probability matrix factorization
Swathi et al. (2020)	Identifies the side effects of drugs from the medical health forums	Health forums	WebMD	Linear SVC, LR, SVM, DT, NB, RF
Liang et al. (2020)	Predict side effects of drugs	Fingerprint, drug literature, ATC code, drug structure, drug target	STITCH, KEGG, DrugBank, SIDER	SVM, RF, ANN
Zhou et al. (2020)	Predict side effects from drug indication using boosted random forest classifier	Protein target, treatments, transporter, enzyme, pathways, chemical structure	SIDER, PharmacotherapyDB, DrugCentral, ClinicalTrialslim	RF, KNN, MLP
Shankar et al. (2021)	Predict the side effects of drug-pair	Gene expression, chemical sub-structure	LINCS L1000, DrugBank, TWOSIDES	ANN
Ietswaart et al. (2020)	Proposed a methodology to associate drug and side effects	Pharmacovigilance assays	SIDER, FAERS, NIBR	RF
Das et al. (2021)	Checked whether drug functions are used to predict harmful reactions of drugs or not	Drug functions	PubChem, SIDER	RF, MLP, KNN, ETC, DT
Hatmal et al. (2021)	Identify COVID-19 vaccine side effects	Textual information of Covid-19 vaccine	Online survey	RF, K-star, XGboost, MLP

Table 1 (continued)

Ref.	Task	Drug properties	Data source(s)	Approach(es)
Hatmal et al. (2022)	Predict side effects of the coronavirus disease vaccine from the survey	Survey of the covid vaccine	Online survey	Gradient boost, AdaBoost, SVM, KNN, and K-star, RF, Probabilistic Neural Network (PNN)
Güneş et al. (2021)	Predict side effects of antidepressant drug	Chemical structure, transporter, enzyme, and drug target	Drug.com, PubChem, DrugBank, SIDER	SVM, KNN, MLP
Das et al. (2022a)	A DNN model is presented to predict side effects	17 molecules, drug functions, and chemical ID structure	SIDER, PubChem	DNN
Zhao et al. (2022)	Introduce a similarity-based DNN method to identify the frequency of side effects	Drug-drug association score, chemical substructure, and drug target	STITCH, DrugBank, SIDER, Galeano's dataset	Similarity-based deep neural network MGPred, Galeano's method

Further, the supervised computational models and their performance are briefly discussed. Furthermore summarizes the challenges, research gaps, future scope of the research, and suggestions for further research, which helps to address the existing challenges.

4 Discussion on the challenges and research gaps in side effects prediction

This section briefly discusses the challenges, issues, and research gaps of the side effects prediction task. Most side effects prediction researchers proposed various supervised computation approaches to solve the different challenges of side effects prediction. However, there is still a need for some improvement to achieve optimized side effects of drugs. There are various issues, challenges, and research gaps, including the following:

- Although there are different existing multi-label supervised computational approaches to analyze the side effects, an urgent efficient model is still required to obtain an optimized result.
- The most common and significant challenge in the multi-label side effects prediction task is dealing with the class imbalance problem, which can affect the predicted model's performance.
- The feature selection method is also crucial in the side effects prediction task. An effective feature selection method is also needed to deal with the curse of data dimensionality.
- To obtain drug-drug similarity, an efficient similarity measure technique is required.
- An advanced computational approach may embed the drug properties effectively, so an advanced deep learning or machine learning-based method is required to represent the drug properties in machine-understandable format.
- Drug side effects prediction may benefit from the combination of distinct drug descriptors, but mapping and combining several drug descriptors is a complex procedure.
- One of the significant challenges is gathering drug properties from various sources. The researcher still faced challenges in determining which drug properties sources are best for them to obtain an effective outcome.
- There are still several unused drug properties and resources to analyze the side effects, which can be utilized further to identify side effects.
- Obtaining side effects of the modern drug is challenging because most pharmaceutical manufacturing organizations do not give the details of novel medicines.
- Pharmaceutical organizations are concerned about drug failure because it losses billions of dollars and is the uncertainty of drug success.
- Slower growth of the computational model in the area of drug discovery, design, and development.
- Poor clinical trial phase on the animal model; therefore, an advanced computational model is required for drug screening.
- Difficult to balance adverse drug reactions, toxicity, and biological complexity in a chemical compound.

The supervised computational learning approach can be employed to learn side effects from distinct drug descriptors. The learned approach can be utilized to develop new drugs by labeling unseen side effects of novel drugs. Furthermore, it enhances the success rate of

drug development, design, and discovery. The supervised computational learning approach brings new life into the drug discovery procedure by analyzing the potential side effects. The pharmaceutical drug discovery, development, and design process is complicated, tedious, costly, chemical wastage, and requires billions of dollars. One of the main causes of drug failure in the market is the harmful side effects of drugs that can halt the entire drug discovery process. Every year, most drug discovery researchers research on the side effects of identification and propose various supervised learning approaches to reduce the cost, chemical wastage, time, and complexity of drug discovery, development, and design and overcome issues encountered during drug discovery. However, side effects prediction studies still face complex challenges, such as gathering drug information from various sources and determining which is best suited to predict drug side effects. Furthermore, adjusting the supervised computational approach parameters is also complex to obtain optimization results. There are only some specific instructions for the parameter settings of computational approaches, but the models are still ineffective in getting an optimized result. Supervised computational learning approaches for identifying side effects are still in their early stages. In the near future, the supervised learning approach is expected to solve all the problems associated with the side effects prediction task. In the current work, a large number of side effects prediction articles based on the supervised learning approach are summarized from 2004–2022. Each side effects prediction research article is summarized based on the problem that the researcher solves. The number of samples and side effects used for the experiment, which type of drug properties are utilized with their corresponding sources, and the supervised computational approach and their performance have been summarized.

5 Conclusion

This work have reviewed and summarized the side effects prediction articles in the last two decades. The computational method, especially the supervised learning approach, can predict side effects from the several drug descriptors that give some prior knowledge about that drug, which may reduce the cost, risk of failure, design complexity, chemical wastage, and time. However, some complex challenges still exist that affect the computation model performance to predict side effects. Only some drug side effects pairs are known in the drug side effects prediction task. In fact, the number of drugs with side effects is significantly lesser than the number of available drugs. So, a semi-supervised method is necessary to give the labels to drugs with no side effects labels. Therefore the main future scope of the side effects prediction task is labeling side effects to the drugs by employing some clustering methods for further classifying side effects with the help of the supervised machine learning method. The drug side effects prediction task belongs to the multi-label prediction task. So handling class imbalance, the curse of dimensionality for multi-label classification tasks, is vital for improving the supervised machine learning approaches performance. However, most drug properties are required to convert computer-understandable format, including 1D chemical structure (simplified molecular input line entry system string), 2D chemical structure, amino acid sequence, protein function, drug target, and drug ATC code. Therefore, an advanced machine learning or deep learning-based method is required to represent the drug properties in machine-understandable format, which may embed the drug properties effectively. Drug side effects prediction task may aid by integrating distinct drug properties. Several combinations of drug properties remain to predict drug side effects,

which may give adequate outcomes. Such as ATC + 17 molecular properties, chemical 1D structure + amino acid sequence, 17 molecular properties + amino acid sequence, system organ class + drug-like properties, 17 molecular properties + protein-protein interaction, drug-disease relation + drug-like properties, and chemical 2D structure + Chemical 3D conformer, Etc. Increased research community attention to these issues will help to solve some of them in the coming days. The supervised learning approach is expected to cover all aspects of drug discovery, design, and development in the near future. This work focused on the supervised computational approach utilized in side effects prediction task and their performances. The supervised learning approaches are rapidly increasing in several fields of bioinformatics, significantly predicting side effects. However, the development of supervised learning approaches is still in the early stage. In conclusion, supervised learning approaches in side effects prediction research will likely emerge soon. Utilizing available drug descriptors with appropriate supervised learning methods remains a significant challenge. Drug side effects prediction is essential, and the exhibited side effects are challenging problems to concentrate on the effective drug.

Funding There has been no significant financial support for this work that could have influenced its outcome.

Data availability Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Declarations

Conflict of interest The authors declare that no conflicts of interest are associated with this publication.

References

- Abd Elaziz M, Yousri D (2021) Automatic selection of heavy-tailed distributions-based synergy henry gas solubility and harris hawk optimizer for feature selection: case study drug design and discovery. *Artif Intell Rev* 54(6):4685–4730
- Abraham MH (1993) Scales of solute hydrogen-bonding: their construction and application to physico-chemical and biochemical processes. *Chem Soc Rev* 22(2):73–83
- Afdhal D, Ananta KW, Hartono WS (2020) Adverse drug reactions prediction using multi-label linear discriminant analysis and multi-label learning. In: 2020 International conference on advanced computer science and information systems (ICACSIS), pp. 69–76. IEEE
- Andrade E, Bento A, Cavalli J, Oliveira S, Schwanke R, Siqueira J, Freitas C, Marcon R, Calixto J (2016) Non-clinical studies in the process of new drug development-part ii: Good laboratory practice, metabolism, pharmacokinetics, safety and dose translation to clinical studies. *Braz J Med Biol Res* 49:e5646
- Askr H, Elgeldawi E, Aboul Ella H, Elshaier YA, Gomaa MM, Hassanien AE (2022) Deep learning in drug discovery: an integrative review and future challenges. *Artif Intell Rev* 2022:1–63
- Avdeef A, Testa B (2002) Physicochemical profiling in drug research: a brief survey of the state-of-the-art of experimental techniques. *Cell Mol Life Sci CMLS* 59(10):1681–1689
- Bagherian M, Sabeti E, Wang K, Sartor MA, Nikolovska-Coleska Z, Najarian K (2021) Machine learning approaches and databases for prediction of drug-target interaction: a survey paper. *Brief Bioinfo* 22(1):247–269
- Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH (2016) A curated and standardized adverse drug event resource to accelerate drug safety research. *Sci Data* 3(1):1–11
- Barrionuevo GO, Ramos-Grez JA, Walczak M, Betancourt CA (2021) Comparative evaluation of supervised machine learning algorithms in the prediction of the relative density of 316L stainless steel fabricated by selective laser melting. *Int J Adv Manuf Technol* 113(1):419–433
- Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J (2008) Bio2rdf: towards a mashup to build bioinformatics knowledge systems. *J Biomed Info* 41(5):706–716

- Bodenreider O (2004) The unified medical language system (umls): integrating biomedical terminology. *Nucl Acids Res* 32(suppl-1):267–270
- Bonner S, Barrett IP, Ye C, Swiers R, Engkvist O, Bender A, Hoyt CT, Hamilton W (2021) A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *arXiv preprint arXiv:2102.10062*
- Bresso E, Grisoni R, Marchetti G, Karaboga AS, Souchet M, Devignes M-D, Smaïl-Tabbone M (2013) Integrative relational machine-learning for understanding drug side-effect profiles. *BMC Bioinfo* 14(1):1–11
- Cami A, Arnold A, Manzi S, Reis B (2011) Predicting adverse drug events using pharmacological network models. *Sci Trans Med* 3(114):114–127
- Chahibi Y (2017) Molecular communication for drug delivery systems: a survey. *Nano Commun Netw* 11:90–102
- Chan JY-L, Bea KT, Leow SMH, Phoong SW, Cheng WK (2022) State of the art: a review of sentiment analysis based on sequential transfer learning. *Artif Intell Rev* 1–32
- Chan HS, Shan H, Dahoun T, Vogel H, Yuan S (2019) Advancing drug discovery via artificial intelligence. *Trends Pharmacol Sci* 40(8):592–604
- Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L et al (2015) The biogrid interaction database: 2015 update. *Nucl Acids Res* 43(D1):470–478
- Chen JY, Mamidipalli S, Huan T (2009) Happi: an online database of comprehensive human annotated and predicted protein interactions. *BMC Genom* 10(1):1–11
- Conn VS, Isaramalai SA, Rath S, Jantarakupt P, Wadhawan R, Dash Y (2003) Beyond medline for literature searches. *J Nurs Scholarsh* 35(2):177–182
- Consortium U (2015) Uniprot: a hub for protein information. *Nucl Acids Res* 43(D1):204–212
- Dai Y-F, Zhao X-M (2015) A survey on the computational approaches to identify drug targets in the post-genomic era. *BioMed Res Int* 2015:1–10
- Dara S, Dhamercherla S, JadaV SS, Babu C, Ahsan MJ (2021) Machine learning in drug discovery: a review. *Artif Intell Rev* 2021:1–53
- Das P, Hussain Mazumder D (2021) Predicting anatomical therapeutic chemical drug classes from 17 molecules' properties of drugs by multi-label binary relevance approach with mlsmote. In: 2021 5th International conference on computational biology and bioinformatics, pp. 1–7
- Das P, Mazumder DH (2023) Predicting drug functions from adverse drug reactions by multi-label deep neural network. In: Multimodal AI in healthcare, pp. 215–226. Springer
- Das P, Pal V et al (2022a) Integrative analysis of chemical properties and functions of drugs for adverse drug reaction prediction based on multi-label deep neural network. *J Integr Bioinfo* 19(3):20220007
- Das P, Sangma JW, Pal V, et al. (2021) Predicting adverse drug reactions from drug functions by binary relevance multi-label classification and mlsmote. In: International conference on practical applications of computational biology & bioinformatics, pp. 165–173. Springer
- Das P, Thakran Y, Anal SN, Pal V, Yadav A (2022b) Brmcf: Binary relevance and mlsmote based computational framework to predict drug functions from chemical and biological properties of drugs. *IEEE/ACM transactions on computational biology and bioinformatics*, IEEE
- Davis AP, Grondin CJ, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, Wiegers TC, Mattingly CJ (2015) The comparative toxicogenomics database's 10th year anniversary: update 2015. *Nucl Acids Res* 43(D1):914–920
- Dey S, Luo H, Fokoue A, Hu J, Zhang P (2018) Predicting adverse drug reactions through interpretable deep learning framework. *BMC Bioinfo* 19(21):1–13
- Dong J, Yao ZJ, Zhang L, Luo F, Lin Q, Lu AP, Chen AF, Cao DS (2018) Pybiomed: a python library for various molecular representations of chemicals, proteins and dnas and their interactions. *J Cheminfo* 10(1):1–11
- Duan Q, Reid SP, Clark NR, Wang Z, Fernandez NF, Rouillard AD, Readhead B, Tritsch SR, Hodos R, Hafner M et al (2016) L1000cds2: Lincs 1000 characteristic direction signatures search engine. *NPJ Syst Biol Appl* 2(1):1–12
- Edwards BJ, Bunta AD, Lane J, Odvina C, Rao DS, Raisch DW, McKoy JM, Omar I, Belknap SM, Garg V et al (2013) Bisphosphonates and nonhealing femoral fractures: analysis of the fda adverse event reporting system (faers) and international safety efforts: a systematic review from the research on adverse drug events and reports (radar) project. *J Bone Joint Surg.* 95(4):297
- Edwards IR, Aronson JK (2000) Adverse drug reactions: definitions, diagnosis, and management. *The Lancet* 356(9237):1255–1259
- Galeano D, Li S, Gerstein M, Paccanaro A (2020) Predicting the frequencies of drug side effects. *Nat Commun* 11(1):1–14

- Ginn R, Pimpalkhute P, Nikfarjam A, Patki A, O'Connor K, Sarker A, Smith K, Gonzalez G (2014) Mining twitter for adverse drug reaction mentions: a corpus and classification benchmark. In: Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing, pp. 1–8. Citeseer
- Grygorenko OO, Volochnyuk DM, Ryabukhin SV, Judd DB (2020) The symbiotic relationship between drug discovery and organic chemistry. *Chem-A Euro J* 26(6):1196–1237
- Güneş SS, Yeşil Ç, Gurdal EE, Korkmaz EE, Yarım M, Aydın A, Sipahi H (2021) Primum non nocere: In silico prediction of adverse drug reactions of antidepressant drugs. *Comput Toxicol* 18:100165
- Hashimoto S, Ball N, Tremlett H (2009) Progressive lipoatrophy after cessation of glatiramer acetate injections: a case report. *Multi Scler J* 15(4):521
- Hatmal MM, Al-Hatamleh MA, Olaimat AN, Hatmal M, Alhaj-Qasem DM, Olaimat TM, Mohamud R (2021) Side effects and perceptions following covid-19 vaccination in Jordan: a randomized, cross-sectional study implementing machine learning for predicting severity of side effects. *Vaccines* 9(6):556
- Hatmal MM, Al-Hatamleh MA, Olaimat AN, Mohamud R, Fawaz M, Kateeb ET, Alkhairi OK, Tayyem R, Lounis M, Al-Raei M et al (2022) Reported adverse effects and attitudes among Arab populations following covid-19 vaccination: a large-scale multinational study implementing machine learning tools in predicting post-vaccination adverse effects based on predisposing factors. *Vaccines* 10(3):366
- Himeur Y, Elnour M, Fadli F, Meskin N, Petri I, Rezgui Y, Bensaali F, Amira A (2022) AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives. *Artif Intell Rev* 2022: 1–93
- Ho TB, Le L, Thai DT, Taewijit S (2016) Data-driven approach to detect and predict adverse drug reactions. *Curr Pharm Des* 22(23):3498–3526
- Hu B, Wang H, Wang L, Yuan W (2018) Adverse drug reaction predictions using stacking deep heterogeneous information network embedding approach. *Molecules* 23(12):3193
- Huang LC, Wu X, Chen JY (2011) Predicting adverse side effects of drugs. *BMC Genom* 12(5):1–10
- Huang L-C, Wu X, Chen JY (2013) Predicting adverse drug reaction profiles by integrating protein interaction networks with drug structures. *Proteomics* 13(2):313–324
- Hu P, Chan KC, Hu L, Leung H (2017) Discovering second-order sub-structure associations in drug molecules for side-effect prediction. In: 2017 IEEE International conference on bioinformatics and biomedicine (BIBM), pp. 2250–2253. IEEE
- Hu P, You ZH, He T, Li S, Gu S, Chan KC (2018) Learning latent patterns in molecular data for explainable drug side effects prediction. In: 2018 IEEE International conference on bioinformatics and biomedicine (BIBM), pp. 1163–1169. IEEE
- Ietswaart R, Arat S, Chen AX, Farahmand S, Kim B, DuMouchel W, Armstrong D, Fekete A, Sutherland JJ, Urban L (2020) Machine learning guided association of adverse drug reactions with in vitro target-based pharmacology. *EBioMedicine* 57:102837
- Islam T, Hussain N, Islam S, Chakrabarty A (2018) Detecting adverse drug reaction with data mining and predicting its severity with machine learning. In: 2018 IEEE region 10 humanitarian technology conference (R10-HTC), pp. 1–5. IEEE
- Izadi S, Sutton D, Hamarneh G (2022) Image denoising in the deep learning era. *Artif Intell Rev* 2022:1–46
- Jahid MJ, Ruan J (2013) An ensemble approach for drug side effect prediction. In: 2013 IEEE international conference on bioinformatics and biomedicine, pp. 440–445. IEEE
- Jamal S, Goyal S, Shanker A, Grover A (2017) Predicting neurological adverse drug reactions based on biological, chemical and phenotypic properties of drugs using machine learning models. *Sci Rep* 7(1):1–12
- Jamal S, Ali W, Nagpal P, Grover S, Grover A (2019) Computational models for the prediction of adverse cardiovascular drug reactions. *J Trans Med* 17(1):1–13
- Jiang K, Zheng Y (2013) Mining twitter data for potential drug effects. In: International conference on advanced data mining and applications, pp. 434–443. Springer
- Jung LS, Cho Y-R (2020) Survey of network-based approaches of drug-target interaction prediction. In: 2020 IEEE International conference on bioinformatics and biomedicine (BIBM), pp. 1793–1796. IEEE
- Kanehisa M, et al (2002) The KEGG database. In: Novartis foundation symposium, pp. 91–100. Wiley
- Kanji R, Sharma A, Bagler G (2015) Phenotypic side effects prediction by optimizing correlation with chemical and target profiles of drugs. *Mol Biosyst* 11(11):2900–2906
- Katragadda S, Karnati H, Pusala M, Raghavan V, Benton R (2015) Detecting adverse drug effects using link classification on twitter data. In: 2015 IEEE International conference on bioinformatics and biomedicine (BIBM), pp. 675–679. IEEE

- Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B et al (2019) Pubchem 2019 update: improved access to chemical data. *Nucl Acids Res* 47(D1):1102–1109
- Kim H, Kim E, Lee I, Bae B, Park M, Nam H (2020) Artificial intelligence in drug discovery: a comprehensive review of data-driven and machine learning approaches. *Biotechnol Bioprocess Eng* 25(6):895–930
- Kuang Q, Wang M, Li R, Dong Y, Li Y, Li M (2014) A systematic investigation of computation models for predicting adverse drug reactions (adrs). *PloS one* 9(9):105889
- Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P (2010) A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 6(1):343
- LaBute MX, Zhang X, Lenderman J, Bennion BJ, Wong SE, Lightstone FC (2014) Adverse drug reaction prediction using scores produced by large-scale drug-protein target docking on high-performance computing machines. *PloS one* 9(9):106298
- Lee K, Bacchetti P, Sim I (2008) Publication of clinical trials supporting successful new drug applications: a literature analysis. *PLoS Med* 5(9):191
- Lee WP, Huang JY, Chang HH, Lee KT, Lai CT (2017) Predicting drug side effects using data analytics and the integration of multiple data sources. *IEEE Access* 5:20449–20462
- Li YH, Yu CY, Li XX, Zhang P, Tang J, Yang Q, Fu T, Zhang X, Cui X, Tu G et al (2018) Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucl Acids Res* 46(D1):1121–1127
- Liang H, Chen L, Zhao X, Zhang X (2020) Prediction of drug side effects with a refined negative sample selection strategy. *Comput Math Methods Med* 2020:1–10
- Ligthart A, Catal C, Tekinerdogan B (2021) Systematic reviews in sentiment analysis: a tertiary study. *Artif Intell Rev* 54(7):4997–5053
- Lin X, Li X, Lin X (2020) A review on applications of computational methods in drug screening and design. *Molecules* 25(6):1375
- Linden M (2013) How to define, find and classify side effects in psychotherapy: from unwanted events to adverse treatment reactions. *Clini Psychol Psychother* 20(4):286–296
- Liu M, Wu Y, Chen Y, Sun J, Zhao Z, Chen XW, Matheny ME, Xu H (2012a) Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *J Am Med Assoc* 19(e1):28–35
- Liu Y, LePendu P, Iyer S, Shah NH (2012b) Using temporal patterns in medical records to discern adverse drug events from indications. *AMIA Summits Trans Sci Proc* 2012:47
- Liu R, AbdulHameed MDM, Kumar K, Yu X, Wallqvist A, Reifman J (2017) Data-driven prediction of adverse drug reactions induced by drug-drug interactions. *BMC Pharmacol Toxicol* 18(1):1–18
- Liu L, Yu Y, Fei Z, Li M, Wu FX, Li HD, Pan Y, Wang J (2018) An interpretable boosting model to predict side effects of analgesics for osteoarthritis. *BMC Syst Biol* 12(6):29–38
- Liu S, Song X, Ma Z, Ganaa ED, Shen X (2022) More: Multi-output residual embedding for multi-label classification. *Pattern Recognit* 126:108584
- Lowe HJ, Ferris TA, Hernandez PM, Weber SC (2009) Stride—an integrated standards-based translational research informatics platform. In: *AMIA annual symposium proceedings*, vol. 2009, p. 391. American Medical Informatics Association
- Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási AL (2015) Uncovering disease-disease relationships through the incomplete interactome. *Science* 347(6224):1257601
- Mizutani S, Pauwels E, Stoven V, Goto S, Yamanishi Y (2012) Relating drug-protein interaction network with drug side effects. *Bioinformatics* 28(18):522–528
- Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á (2022) Human-in-the-loop machine learning: a state of the art. *Artif Intell Rev* 1–50
- Muhammad L, Algehyne EA, Usman SS, Ahmad A, Chakraborty C, Mohammed IA (2021) Supervised machine learning models for prediction of covid-19 infection using epidemiology dataset. *SN Comput Sci* 2(1):1–13
- Muñoz E, Nováček V, Vandenbussche PY (2019) Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models. *Brief Bioinfo* 20(1):190–202
- Naranjo CA, Busto U, Sellers EM, Sandor P, Ruiz I, Roberts E, Janecek E, Domecq C, Greenblatt D (1981) A method for estimating the probability of adverse drug reactions. *Clin Pharmacol Ther* 30(2):239–245
- Ngufor C, Wojtusiak J, Pathak J (2015) A systematic prediction of adverse drug reactions using pre-clinical drug characteristics and spontaneous reports. In: *2015 International conference on healthcare informatics*, pp. 76–81. IEEE
- Niu SY, Xin MY, Luo J, Liu MY, Jiang ZR (2015) Dsep: a tool implementing novel method to predict side effects of drugs. *J Comput Biol* 22(12):1108–1117

- Odeh F, Taweel A (2019) A deep learning approach to extracting adverse drug reactions. In: 2019 IEEE/ACS 16th International conference on computer systems and applications (AICCSA), pp. 1–6. IEEE
- Pauwels E, Stoven V, Yamanishi Y (2011) Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinfo* 12(1):1–13
- Pouliot Y, Chiang AP, Butte AJ (2011) Predicting adverse drug reactions using publicly available pubchem bioassay data. *Clin Pharmacol Ther* 90(1):90–99
- Pun CS, Lee SX, Xia K (2022) Persistent-homology-based machine learning: a survey and a comparative study. *Artif Intell Rev* 2022: 1–45
- Qu G, Wu H, Hartrick CT, Niu J (2012) Local analgesia adverse effects prediction using multi-label classification. *Neurocomputing* 92:18–27
- Raja K, Patrick M, Elder JT, Tsoi LC (2017) Machine learning workflow to enhance predictions of adverse drug reactions (adrs) through drug-gene interactions: application to drugs for cutaneous diseases. *Sci Rep* 7(1):1–11
- Razmjoo A, Caliva F, Lee J, Liu F, Joseph GB, Link TM, Majumdar S, Pedito V (2021) T2 analysis of the entire osteoarthritis initiative dataset. *J Orthop Res* 39(1):74–85
- Rees KE, Chyou TY, Nishtala PS (2020) A disproportionality analysis of the adverse drug events associated with lurasidone in paediatric patients using the us fda adverse event reporting system (faers). *Drug Saf* 43(6):607–609
- Rokach L, Schclar A, Itach E (2014) Ensemble methods for multi-label classification. *Exp Syst Appl* 41(16):7507–7523
- Ryan PB, Madigan D, Stang PE, Marc Overhage J, Racoosin JA, Hartzema AG (2012) Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the observational medical outcomes partnership. *Stat Med* 31(30):4401–4415
- Sarma H, Upadhyaya M, Gogoi B, Phukan M, Kashyap P, Das B, Devi R, Sharma HK (2021) Cardiovascular drugs: an insight of in silico drug design tools. *J Pharm Innov* 2021:1–26
- Scheiber J, Jenkins JL, Sukuru SCK, Bender A, Mikhailov D, Milik M, Azzaoui K, Whitebread S, Hamon J, Urban L et al (2009) Mapping adverse drug reactions in chemical space. *J Med Chem* 52(9):3103–3107
- Shankar S, Bhandari I, Okou DT, Srinivasa G, Athri P (2021) Predicting adverse drug reactions of two-drug combinations using structural and transcriptomic drug representations to train an artificial neural network. *Chem Biol Drug Des* 97(3):665–673
- Stein LD (2003) Integrating biological databases. *Nat Rev Gene* 4(5):337–345
- Swathi DN, et al. (2020) Predicting drug side-effects from open source health forums using supervised classifier approach. In: 2020 5th International conference on communication and electronics systems (ICCES), pp. 796–800. IEEE
- Szklarczyk D, Santos A, Von Mering C, Jensen LJ, Bork P, Kuhn M (2016) Stitch 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucl Acids Res* 44(D1):380–384
- Tatonetti NP, Ye PP, Daneshjou R, Altman RB (2012a) Data-driven prediction of drug effects and interactions. *Sci Trans Med* 4(125):125–3112531
- Tatonetti N, Ye P, Daneshjou R, Altman R (2012b) Data-driven prediction of drug effects and interactions. *Sci. Transl. Med.* 4:125ra131
- Trouiller P, Olliaro P, Torreele E, Orbinski J, Laing R, Ford N (2017) Drug development for neglected diseases: a deficient market and a public-health policy failure. *Global Health* 267–273
- Tsoumakas G, Katakis I, Vlahavas I (2006) A review of multi-label classification methods. In: Proceedings of the 2nd ADBIS workshop on data mining and knowledge discovery (ADMKD 2006), pp. 99–109
- Tuntland T, Ethell B, Kosaka T, Blasco F, Zang RX, Jain M, Gould T, Hoffmaster K (2014) Implementation of pharmacokinetic and pharmacodynamic strategies in early research phases of drug discovery and development at novartis institute of biomedical research. *Front Pharmacol* 5:174
- Uner OC, Cinbis RG, Tastan O, Cicek AE (2019) Deepside: a deep learning framework for drug side effect prediction. *Biorxiv* 2019:843029
- Vermeer NS, Straus SM, Mantel-Teeuwisse AK, Domergue F, Egberts TC, Leufkens HG, De Bruin ML (2013) Traceability of biopharmaceuticals in spontaneous reporting systems: a cross-sectional study in the fda adverse event reporting system (faers) and eudravigilance databases. *Drug Saf* 36(8):617–625
- Wang Z, Clark NR, Ma'ayan A (2016) Drug-induced adverse events prediction with the lincs 1000 data. *Bioinformatics* 32(15):2338–2345
- Wang CS, Lin PJ, Cheng CL, Tai SH, Yang YHK, Chiang JH et al (2019) Detecting potential adverse drug reactions using a deep neural network model. *J Med Internet Res* 21(2):11016
- WHO CO, et al (2020) World health organization. Responding to Community Spread of COVID-19. Reference WHO/COVID-19/Community_Transmission/2020.1

- Willmann JK, Van Bruggen N, Dinkelborg LM, Gambhir SS (2008) Molecular imaging in drug development. *Nat Rev Drug Discov* 7(7):591–607
- Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z et al (2018) Drugbank 5.0: a major update to the drugbank database for 2018. *Nucl Acids Res* 46(D1):1074–1082
- Yamanishi Y, Pauwels E, Kotera M (2012) Drug side-effect prediction based on the integration of chemical and biological spaces. *J Chem Info Model* 52(12):3284–3292
- Yao B, Zhu L, Jiang Q, Xia HA (2013) Safety monitoring in clinical trials. *Pharm* 5(1):94–106
- Yap C, Cai C, Xue Y, Chen Y (2004) Prediction of torsade-causing potential of drugs by support vector machine approach. *Toxicol Sci* 79(1):170–177
- Zhang W, Liu F, Luo L, Zhang J (2015) Predicting drug side effects by multi-label learning and ensemble learning. *BMC Bioinfo* 16(1):1–11
- Zhang W, Zou H, Luo L, Liu Q, Wu W, Xiao W (2016a) Predicting potential side effects of drugs by recommender methods and ensemble learning. *Neurocomputing* 173:979–987
- Zhang W, Yue X, Liu F, Chen Y, Tu S, Zhang X (2017) A unified frame of predicting side effects of drugs by using linear neighborhood similarity. *BMC Syst Biol* 11(6):23–34
- Zhang W, Chen Y, Tu S, Liu F, Qu Q (2016b) Drug side effect prediction through linear neighborhoods and multiple data source integration. In: 2016 IEEE international conference on bioinformatics and biomedicine (BIBM), pp. 427–434. IEEE
- Zhang P, Wang F, Hu J, Sorrentino R (2013) Exploring the relationship between drug side-effects and therapeutic indications. In: AMIA annual symposium proceedings, vol. 2013, p. 1568. American Medical Informatics Association
- Zhao X, Chen L, Lu J (2018) A similarity-based method for prediction of drug side effects with heterogeneous information. *Math Biosci* 306:136–144
- Zhao H, Wang S, Zheng K, Zhao Q, Zhu F, Wang J (2022) A similarity-based deep learning approach for determining the frequencies of drug side effects. *Brief Bioinfo* 23(1):449
- Zheng Y, Peng H, Zhang X, Zhao Z, Yin J, Li J (2018) Predicting adverse drug reactions of combined medication from heterogeneous pharmacologic databases. *BMC Bioinfo* 19(19):49–59
- Zhou SF, Zhong WZ (2017) Drug design and discovery: principles and applications. *MDPI* 22:279
- Zhou H, Cao H, Matyunina L, Shelby M, Cassels L, McDonald JF, Skolnick J (2020) Medicascy: a machine learning approach for predicting small-molecule drug side effects, indications, efficacy, and modes of action. *Mol Pharm* 17(5):1558–1574
- Zitnik M, Agrawal M, Leskovec J (2018) Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 34(13):457–466

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.