International Journal of Technology and Emerging Sciences (IJTES)

www.mapscipub.com

Volume 05 || Issue 02 || July 2025 || pp. 31-36 E-ISSN: 2583-1925

# Harnessing Machine Learning for Drug Concentration Prediction: A Comprehensive Study

**J. VidhyaJanani[1], Sandeep. M[2], Anandhakumar D[2] , R. Dhinakar[2], Sharn.P[2], Sakthipriyadharshini. S[2]**

*[1]Assistant Professor, Department of Computer Science & Engineering,*
*Paavai College of Engineering, Namakkal*
*[2]Students, Department of Computer Science &Engineering(Cyber Security),*
*Paavai College of Engineering, Namakkal*

---

**Abstract -** *Accurate prediction of drug concentrations within the human body is critical for optimizing therapeutic efficacy, minimizing toxicity, and supporting personalized medicine. Traditional pharmacokinetic models, while grounded in well-established mathematical formulations, often fall short in capturing complex individual variability and nonlinear drug behavior. In recent years, machine learning (ML) has emerged as a powerful alternative, offering data-driven approaches capable of modeling intricate patterns from clinical, demographic, and biochemical data. This study provides a comprehensive overview of how machine learning techniques are being harnessed to predict drug concentrations across diverse therapeutic contexts. We explore the strengths and limitations of various ML models, including regression algorithms, decision trees, ensemble methods, and deep learning architectures, highlighting their performance in pharmacokinetic modeling. Additionally, we discuss critical challenges such as data scarcity, feature selection, model interpretability, and generalizability across populations. Case studies and recent advances illustrate real-world applications and the transformative potential of ML in precision dosing. By evaluating current methodologies and addressing prevailing obstacles, this work aims to guide researchers and clinicians in the effective integration of ML into pharmacokinetic workflows, ultimately contributing to safer and more personalized drug therapy.*

***Key Words***: Machine Learning, Drug-related side effects, Prediction

## 1. INTRODUCTION

Accurate prediction of drug concentrations within the human body is a fundamental aspect of pharmacokinetics and plays a vital role in ensuring the safety and efficacy of therapeutic treatments. Traditional approaches to drug concentration modeling—such as compartmental models and nonlinear mixed-effect modeling—have long been the standard tools in clinical pharmacology. While these methods are grounded in physiological and biochemical theory, they often require extensive domain expertise, assumptions about underlying kinetics, and large, high-quality datasets to yield robust results. As the complexity of patient data and drug interactions increases, there is a growing demand for adaptive, data-driven approaches that can handle high-dimensional, nonlinear relationships with improved predictive performance. In recent years, machine learning (ML) has emerged as a transformative tool across various domains of healthcare, including drug discovery, diagnostics, and personalized medicine. In the context of pharmacokinetics, ML offers the potential to model intricate relationships between patient characteristics, dosing regimens, and drug response, without being constrained by rigid model assumptions. Techniques such as decision trees, support vector machines, neural networks, and ensemble methods have been explored for predicting plasma drug concentrations, identifying patient-specific factors influencing pharmacokinetics, and optimizing dosing strategies. Moreover, the advent of deep learning and access to large-scale electronic health records (EHRs) and pharmacogenomic data have further propelled the interest in AI-driven therapeutic monitoring.

Despite these advancements, several challenges persist. ML models in pharmacokinetics often suffer from issues related to data sparsity, interpretability, and generalizability. Integrating heterogeneous data sources—such as genomic profiles, real-time drug level monitoring, and lifestyle variables—poses additional complexity. Furthermore, the regulatory and ethical implications of using black-box ML models in clinical decision-making demand rigorous validation, transparency, and explainability. This study presents a comprehensive overview of the application of machine learning

techniques for drug concentration prediction. We explore the current methodologies, their practical applications, performance metrics, and the key challenges limiting their widespread adoption. By bridging insights from both machine learning and pharmacology, this paper aims to contribute to the development of more accurate, personalized, and clinically reliable drug dosing models that align with the goals of precision medicine.

Drug-related side effects include undesirable, unpleasant, unexpected, and adverse hazardous reactions in organs and tissues [1]. Some market-approved drugs may cause unacceptable side effects, endangering human health and raising concerns among pharmaceutical companies [2]. Ensuring drug efficacy is crucial since unfavorable drug responses are the main cause of drug failure, often leading to side effects and drug withdrawal [2,3]. However, the traditional method of identifying side effects through solid clinical trials is time-consuming and expensive, making it unsuitable for large-scale tests [4,5]. As a result, there is a critical need to develop rapid and cost-effective methods for predicting drug-related side effects [6].

The ability to predict drug-related side effects presents itself as an indispensable facet of contemporary pharmaceutical research and development [7]. By enabling the early and accurate identification of potential side effects, such methodologies have the potential to revolutionize the drug development landscape, which can lead to significant time and resource efficiencies [8]. This transformative capacity facilitates the prioritization of drug candidates with favorable safety profiles while concurrently enabling the exclusion of those exhibiting a high propensity to induce adverse events. Ultimately, the development of robust drug side effect prediction methodologies paves the way for the introduction of safer and more efficacious medications, thereby fostering improved patient outcomes and propelling advancements in personalized medicine [9,10].

The development of advanced computational algorithms provides strong technical support for addressing a wide range of medical challenges [11]. Specifically, numerous computational methods have been developed for predicting drug-related side effects, with a strong emphasis on machine learning-based approaches [13]. These methods delve into current information on drug-related side effects to create patterns that allow for the prediction of side effects for various drugs [14]. To our knowledge, none of the studies specifically focused on predicting drug-related side effects using drugs chemical, phenotypic, or biological features and machine learning techniques. Therefore, the aim of the current study was to review studies in which machine-learning techniques were used to predict drug-related side effects based on chemical, biological, or phenotypic features.

## 2. MATERIALS & METHODS

This scoping review was conducted according to Arksey and O'Malley's framework in 2023 [6]. Before conducting the research, ethics approval was obtained from the ethics committee of Iran University of Medical Sciences (IR.IUMS.REC.1401.1007).

### *2.1. Stage 1: Identifying Research Questions*

A comprehensive understanding of machine learning techniques is essential to predict drug-related side effects based on chemical, biological, or phenotypic features for improving personalized medicine and safe medication prescriptions. Therefore, the research questions were as follows:

- What were the machine learning techniques used for predicting drug-related side effects?
- What were the main features used for predicting drug-related side effects?

### *2.2. Stage 2: Identifying Relevant Studies*

The related articles were searched in different databases, including Web of Science, PubMed, Ovid, Scopus, ProQuest, IEEE Xplore, and the Cochrane Library. The search strategy included three main concepts: namely, "drug-related side effect", "machine learning", and "prediction". The MeSH terms, synonyms, and other related keywords were also included in the search strategies. The citations and reference lists of the retrieved papers were also checked to ensure that all relevant studies were included.

### *2.3. Stage 3: Study Selection*

In this study, the original research papers published in English between 2013 and 2023 with a focus on predicting drug-related side effects using chemical, biological, or phenotypical features were included. However, for papers that were published in languages other than English, there

was no access to their full texts, review papers, letters to the editor, and papers that did not primarily focus on machine learning techniques were excluded.

The retrieved papers were entered into the Endnote software version 19, and after removing duplicates, the remaining articles were assessed in terms of the title and abstract relevancy to the study objective. After removing the irrelevant articles, the full texts of the remaining ones were examined by two authors (E.T. and H.A.) separately, and any disagreements were resolved by the third author (A.F.S.).

### *2.4. Stage 4: Charting the Data*

We used a data extraction form to collect the required data. This form contained the author's name, publication year, country, study objective, selected features and data sources, algorithms, evaluation metrics, and main results. In this study, conducting a meta-analysis was not feasible due to the inherent heterogeneity of the study design and methodologies. As a result, the findings were organized and reported narratively. Regarding the evaluation metrics, including precision, accuracy, recall, F1 score, area under the curve (AUC), and area under the precision–recall curve (AUPR), the average was calculated and reported.

## 3. RESULTS & DISCUSSION

The study findings revealed that the selected features across various studies could be classified into four main categories, including general, chemical, biological, and phenotypical features. Different models employed one or more of these categories in predicting drug-related side effects. Furthermore, the data sources utilized for feature extraction displayed a degree of variability. DrugBank, Liu's dataset, and SIDER 4 were consistently employed for extracting features across all categories. Bio2RDF v2 was utilized for all categories except for the general category, and Mizutani's dataset was utilized across all categories except for the phenotypical category. The subsequent sections entail the features and data sources encompassed within each category.

In total, 1698 papers were retrieved from databases. After removing duplicates (n = 809), the remaining papers (n = 889) were examined in terms of their titles and abstracts, and irrelevant papers were excluded (n = 827). Among the remaining papers (n = 62), the full texts of three papers were not retrieved. As a result, the full texts of 59 papers were reviewed. Finally, 22 papers were selected to be included in the study [18,] A total of 37 papers were removed as either they were not related to machine learning algorithms, or they did not include the expected features. The process of selecting the articles is illustrated in Figure 1.

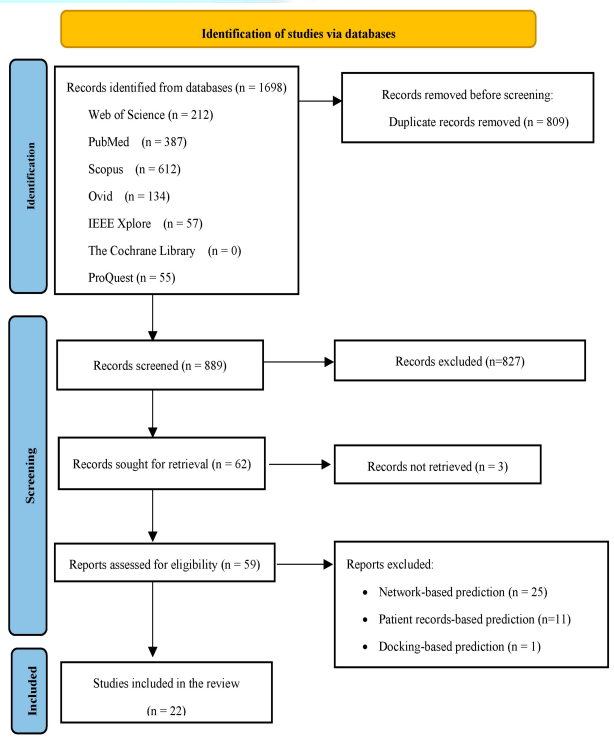


Fig1:Paper selection Process review

This scoping review investigated the use of machine learning techniques for the prediction of drug-related side effects. Based on the findings, general features were mainly extracted from SIDER, Pauwel's dataset, Mizutani's dataset, Liu's dataset, and DrugBank. Chemical features predominantly were obtained from PubChem, Molecular Operating Environment, and DrugBank using fingerprint analysis software. DrugBank, Liu's dataset, and Pauwels' dataset were used to provide biological features, and SIDER 4, Liu's dataset, SIDER, DrugBank, and Bio2RDF v2 provided therapeutic indications and phenotypes.

According to the current review findings, when chemical and biological features were combined, the prediction outcomes were impressive. Moreover, ensemble methods

showed the best results in terms of precision and AURP metrics. SVM exhibited superior performance in accuracy and recall measures, and decision trees excelled in F1 score metrics. In addition, clustering methods demonstrate proficiency in AUC assessment.

The results showed that careful selection of features from relevant databases or datasets is crucial in predicting drug-related side effects. In the present study, features were classified into four primary groups. This classification scheme is aligned with the findings reported by Das and Mazumder's study [1]. Likewise, the review conducted by Sachdev and Gupta on computational techniques for identifying drug-related side effects introduced some features and datasets [13]; however, the focus was not primarily on machine learning techniques, resulting in a limited range of features compared to the current study.

Various studies highlighted the importance of specific features in predicting drug-related side effects, such as chemical fingerprints from SMILES strings and target protein associations from DrugBank, indicating the necessity for a combination of chemical and biological data for accurate predictions. However, biases exist within data sources like SIDER, which may skew towards common side effects [10], and limitations in PubChem exclude information on biologic drugs, urging integration with databases capturing biologic complexities [11]. Feature engineering techniques, like fingerprint generation algorithms and text-mining, aid in translating raw data into interpretable formats [12], while network-based approaches offer promise in modeling complex relationships between chemical structures, biological targets, and side effects [13]. Despite the potential of emerging data sources such as electronic health records and genomics data for personalized prediction, challenges like data standardization and interoperability persist [14], highlighting the need for standardized efforts and common ontologies to facilitate comprehensive dataset creation for machine learning models in side effect prediction.

According to the findings of this review, the integration of chemical and biological features showcased proficiency in precision, F1 score, AUC, and AUPR metrics. In the research conducted by Mizutani et al., canonical correlation analysis and sparse canonical correlation analysis were used, which provided valuable insights into the significance of feature selection. Their study highlighted the superiority of employing the targeted protein-based approach as a biological feature for the prediction of drug-related side effects [19]. Moreover, the research conducted by Liu et al. evaluated various machine-learning algorithms by different features and demonstrated the exceptional performance of SVM when combining chemical, biological, and phenotypic features [17].

Random Forest emerged as the most common algorithm used across the included studies, followed by KNN and SVM. However, there are discrepancies regarding the most frequently used algorithms within this research domain. Das and Mazumder reported that SVM and logistic regression are commonly used for predicting drug-related side effects [1]. In contrast, Sachdev and Gupta emphasized the efficacy of multi-label KNN learning, SVM, and random forest [13]. Random Forest interpretability and resistance to overfitting are among the advantages of this algorithm; however, it may struggle with high-dimensional data. Techniques like Mean Decrease in Impurity (MDI) could enhance its efficacy. KNN is valued for simplicity but requires careful parameter selection, while SVM handles high-dimensional data well but can be computationally expensive. Beyond these, gradient-boosting machines and deep learning architectures offer promising alternatives and are adept at capturing complex relationships in drug data [6].

This study highlighted the significance of different feature combinations in predicting outcomes. Similarly, Das and Mazumder focused on four distinct features, namely, chemical, biological, phenotypic, and other drug descriptors [1]. Other studies concentrated on patient-centric data sources such as prospective data collection and derived data from Electronic Health Records (EHRs) and social media platforms to enrich their predictive capabilities. For example, Zhao et al. used EHR data to predict drug-related side effects. They applied multiple supervised algorithms to analyze patient data, including demographics, lab results, and medication history, achieving significant accuracy with the Random Forest algorithm in identifying potential drug-related side effects before they manifested clinically. Ietswaart et al. used data from the FDA's Adverse Event Reporting System (FAERS) to train a Random Forest model. This model was able to detect subtle patterns and correlations within the vast datasets, effectively predicting the side effects of new and existing drugs.

It is essential to distinguish between studies that used patient-centric data and those that focused on drug features, as their objectives vary significantly. Patient-centric studies primarily aim to predict the overall incidence of specific drug-related side effects, diagnose individuals experiencing side effects, or prognosticate patients at high risk of drug side effects. Conversely, studies included in this review predominantly focused on predicting drug-related side effects based on drug features prior to their manifestation in patients. For instance, Kim et al. reviewed existing statistical and machine-learning methods to detect drug-related side effects in humans. La et al. integrated theoretical biological data into machine-learning models to predict Active Pharmaceutical Ingredient (API) side effects, validating their approach against real-world clinical outcomes. This underscores the multifaceted nature of data used in predicting drug-related side effects, reflecting the inherent challenges in directly comparing machine learning techniques used across these two distinct groups of studies.

The results showed that Random Forest had superior performance compared to other machine learning algorithms included in this study. However, the prominent algorithm in Das and Mazumder's study was SVM [1], and multi-label KNN learning prevailed in Sachdev and Gupta's research [13]. Random Forest's prominence in drug-related side effect prediction arises from its adeptness at handling high-dimensional data and its robustness to imbalanced class distributions commonly found in such datasets [9]. Ensemble methods like Random Forest often outperform single-learner methods like SVM due to their ability to leverage multiple learners for greater generalizability, although SVMs may excel in specific scenarios, particularly with limited dataset sizes. However, a deeper analysis beyond average performance metrics is essential to unveil algorithm-specific nuances and assess generalizability across independent datasets. Combining chemical and biological features enhances performance, but further exploration into specific types of features and feature selection techniques is warranted.

Overall, a comprehensive examination of multiple studies reveals common trends and variations in the selection of features, databases, and algorithms for predicting drug-related side effects. The diversity of machine learning approaches highlighted the complex nature of this task, and the emphasis on using different evaluation metrics underscores the significance of thorough evaluation to guarantee the reliability and effectiveness of predictive models in the pharmaceutical research domain.

## 4. CONCLUSION

In conclusion, this scoping review comprehensively analyzed the use of machine learning techniques for predicting drug-related side effects. The findings underscore the critical role of selecting features from diverse databases encompassing chemical, biological, and phenotypic data for robust prediction. Ensemble methods, particularly Random Forest, emerged as superior algorithms across a spectrum of evaluation metrics, including AUC, precision, recall, F1 score, and AUPR. To predict drug-related side effects, the integration of chemical and biological features enhanced performance. These findings suggested that machine learning algorithms are useful for various applications in the pharmaceutical domain, including drug development through early prediction of side effects and optimizing clinical trial designs via patient stratification based on the predicted risk of side effects. Future research should delve into exploring specific feature types, refining feature selection techniques, and investigating the potential of graph-based methods to predict even more accurate drug-related side effects.

***Disclosure:*** *The authors certify that all content, data, and organizational references in this manuscript have been included with proper authorization and consent. They accept full legal and ethical responsibility for the accuracy, originality, and integrity of the work. The journal, publisher, and editorial board assume no liability for any disputes arising from the submitted material.*

## REFERENCES


[1]. Das, P.; Mazumder, D.H. An extensive survey on the use of supervised machine learning techniques in the past two decades for prediction of drug side effects. Artif. Intell. Rev. 2023, 56, 9809–9836.

[2]. Downing, N.S.; Shah, N.D.; Aminawung, J.A.; Pease, A.M.; Zeitoun, J.-D.; Krumholz, H.M.; Ross, J.S. Postmarket safety events among novel therapeutics approved by the US food and drug administration between 2001 and 2010. JAMA 2017, 317, 1854–1863.

[3]. Craveiro, S.N.; Lopes, S.B.; Tomás, L.; Almeida, F.S. Drug withdrawal due to safety: A review of the data supporting withdrawal decision. Curr. Drug Saf. 2020, 15, 4–12.

[4]. Subbiah, V. The next generation of evidence-based medicine. Nat. Med. 2023, 29, 49–58.

[5]. Lavertu, A.; Vora, B.; Giacomini, K.M.; Altman, R.; Rensi, S. A new era in pharmacovigilance: Toward real-

world data and digital monitoring. Clin. Pharmacol. Ther. 2021, 109, 1197–1202.
[6]. Vora, L.K.; Gholap, A.D.; Jetha, K.; Thakur, R.R.; Solanki, H.K.; Chavda, V.P. Artificial intelligence in pharmaceutical technology and drug delivery design. Pharmaceutics 2023, 15, 1916.
[7]. Yang, S.; Kar, S. Application of artificial intelligence and machine learning in early detection of adverse drug reactions (ADRs) and drug-induced toxicity. Artif. Intell. Chem. 2023, 1, 100011.
[8]. Biala, G.; Kedzierska, E.; Kruk-Slomka, M.; Orzelska-Gorka, J.; Hmaidan, S.; Skrok, A.; Kaminski, J.; Havrankova, E.; Nadaska, D.; Malik, I. Research in the field of drug design and development. Pharmaceuticals 2023, 16, 1283.
[9]. Han, R.; Yoon, H.; Kim, G.; Lee, H.; Lee, Y. Revolutionizing medicinal chemistry: The application of artificial intelligence (AI) in early drug discovery. Pharmaceuticals 2023, 16, 1259.
[10]. Johnson, K.B.; Wei, W.Q.; Weeraratne, D.; Frisse, M.E.; Misulis, K.; Rhee, K.; Zhao, J.; Snowdon, J.L. Precision medicine, AI, and the future of personalized health care. Clin. Transl. Sci. 2021, 14, 86–93.
[11]. Singh, S.; Kumar, R.; Payra, S.; Singh, S.K. Artificial intelligence and machine learning in pharmacological research: Bridging the gap between data and drug discovery. Cureus 2023, 15, e44359
[12]. Alowais, S.A.; Alghamdi, S.S.; Alsuhebany, N.; Alqahtani, T.; Alshaya, A.I.; Almohareb, S.N.; Aldairem, A.; Alrashed, M.; Bin Saleh, K.; Badreldin, H.A.; et al. Revolutionizing healthcare: The role of artificial intelligence in clinical practice. BMC Med. Educ. 2023, 23, 689.
[13]. Sachdev, K.; Gupta, M.K. A comprehensive review of computational techniques for the prediction of drug side effects. Drug Dev. Res. 2020, 81, 650–670.
[14]. Ho, T.B.; Le, L.; Thai, D.T.; Taewijit, S. Data-driven approach to detect and predict adverse drug reactions. Curr. Pharm. Des. 2016, 22, 3498–3526.
[15]. Deimazar, G.; Sheikhtaheri, A. Machine learning models to detect and predict patient safety events using electronic health records: A systematic review. Int. J. Med. Inform. 2023, 180, 105246.
[16]. Rajpoot, K.; Desai, N.; Koppisetti, H.; Tekade, M.; Sharma, M.C.; Behera, S.K.; Tekade, R.K. In silico methods for the prediction of drug toxicity. In Pharmacokinetics and Toxicokinetic Considerations; Tekade, R.K., Ed.; Academic Press: New York, NY, USA, 2022; Volume 2, pp. 357–383.
[17]. Liu, M.; Wu, Y.; Chen, Y.; Sun, J.; Zhao, Z.; Chen, X.W.; Matheny, M.E.; Xu, H. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. J. Am. Med. Inform. Assoc. JAMIA 2012, 19, e28–e35.
[18]. Pauwels, E.; Stoven, V.; Yamanishi, Y. Predicting drug side-effect profiles: A chemical fragment-based approach. BMC Bioinf. 2011, 12, 169
[19]. Mizutani, S.; Pauwels, E.; Stoven, V.; Goto, S.; Yamanishi, Y. Relating drug-protein interaction network with drug side effects. Bioinformatics 2012, 28, i522–i528.
[20]. Amaro, R.E.; Mulholland, A.J. Multiscale methods in drug design bridge chemical and biological complexity in the search for cures. Nat. Rev. Chem. 2018, 2, 0148.
[21]. Duran-Frigola, M.; Aloy, P. Analysis of chemical and biological features yields mechanistic insights into drug side effects. Chem. Biol. 2013, 20, 594–603
[22]. Boland, M.R.; Jacunski, A.; Lorberbaum, T.; Romano, J.D.; Moskovitch, R.; Tatonetti, N.P. Systems biology approaches for identifying adverse drug reactions and elucidating their underlying biological mechanisms. Wiley Interdiscip. Rev. Syst. Biol. Med. 2016, 8, 104–122.
[23]. Yoo, S.; Noh, K.; Shin, M.; Park, J.; Lee, K.-H.; Nam, H.; Lee, D. In silico profiling of systemic effects of drugs to predict unexpected interactions. Sci. Rep. 2018, 8, 1612.
[24]. Zitnik, M.; Nguyen, F.; Wang, B.; Leskovec, J.; Goldenberg, A.; Hoffman, M.M. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Int. J. Inf. Fusion 2019, 50, 71–91
[25]. Marques, L.; Costa, B.; Pereira, M.; Silva, A.; Santos, J.; Saldanha, L.; Silva, I.; Magalhães, P.; Schmidt, S.; Vale, N. Advancing precision medicine: A review of innovative In Silico approaches for drug development, clinical pharmacology and personalized healthcare. Pharmaceutics 2024, 16, 332.