IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13454

Predictive Modelling of Drug Side Effects using Bioinformatics and Machine Learning

Anwar Basha Shaik[1], Dr. Elamathi Natarajan[2]

Student, Department of Bioinformatics, Biotecnika info Labs, Bengaluru, India[1]

Bioinformatics CRO Scientist, Department of Bioinformatics, Biotecnika info Labs, Bengaluru, India[2]

Abstract: This article explores the integration of databases to compare and evaluate the machine learning algorithms that suits multi-label classification. In this predictive modelling work, I have performed Logistic regression, Random Forest, XGBoost, Multi-Layer Perceptron. Logistic Regression is for linear model prediction of drug side effects, Random Forest is chosen for High-Dimensional Biomedical datasets, XGBoost was selected because it is one of the most powerful gradient boosting algorithms, Multi-Layer Perceptron (MLP) architecture was used to learn nonlinear relationships between drug features and side-effect labels. As this is a predictive modelling task evaluation of machine learning models is essential step to assess its predictive capability, reliability, and generalization performance using appropriate evaluation metrics. The computational analysis for this study was carried out using Python due to its flexibility and strong ecosystem of scientific computing libraries. In this study, multiple machine learning models were implemented and compared to evaluate their effectiveness in predicting drug side effects using molecular properties and pharmacokinetics features.

Keywords: Adverse drug reactions, Machine learning, Logistic regression, Random forest, XGBoost, Multi-layer perceptron, Neural network, precision, recall.

I. INTRODUCTION

The word drugs refer to chemical compounds consumed by the body and which modify it by binding to their protein targets. The drugs can either positively or negatively change the human body. Drug side effects refer to the unpleasant changes that drugs bring to the body of an individual. All these side effects may be mild such as a headache and significant such as cancer, heart attack, or even death [1]. There are two main factors in the frequent occurrence of drug safety problems; on the one hand, the clinical understanding of drug side effects is insufficient, leading to frequent adverse drug reactions, while on the other hand, due to the long-term period and complexity of clinical trials, side effects of approved drugs on the market cannot be reported in a timely manner [2]. Drug development is a long process, costly and risky undertaking. It is important to note that the adverse drug reaction is the fourth leading cause of death all over the world after cardiovascular diseases, cancer and infectious diseases [4]. Adverse drug reaction (ADR) is one of the major causes of failure in drug development. Severe ADRs that go undetected until the post-marketing phase of a drug often lead to patient morbidity [6]. The timeconsuming and expensive conventional approach of determining side effects by high-quality clinical trials do not suit large-scale tests. Consequently, urgent development of fast and low-cost methods of predicting adverse drug reactions is necessary [3]. Traditional clinical trials to recognize ADRs are expensive and time-consuming. Conversely, computer-aided methods for predicting ADRs are much cheaper and quicker than clinical trials and highly reliable. In the current study, the main work is to compare and analyze the performance of existing computational methods to predict ADRs, by implementing and evaluating additional algorithms that have been earlier used for predicting drug targets [5] [7].

The raw data used in this study was collected from mul..3l.3tiple publicly available biomedical databases. A curated dataset of approximately 1500 drugs and their side effects are collected from SIDER database. Each drug is associated with molecular identifiers from PubChem, that helps retrieval of chemical structure information. These datasets were integrated to create a comprehensive dataset containing drugs and their side effect labels with chemical structure called SMILES. Using SWISSADME, generated pharmacokinetics and drug-likeness features, which predicts absorption, distribution, metabolism and excretion(ADME) properties of chemical compounds. The computational analysis for this study was carried out using Python due to its flexibility and strong ecosystem of scientific computing libraries. [8] [9]. Python is widely used in bioinformatics and machine learning because it provides efficient tools for data processing, feature extraction, and predictive modelling [10].

I implemented and compared several machine learning algorithms to compare their performance in predicting drug side effects. As, this drug side effect prediction is multi-label classification, models used are Logistic Regression, Random

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13454

Forest, XGBoost, Neural Networks. The purpose of Using multiple models is to see the comparison of their predictive performance and identify which algorithm handled high-dimensional biomedical data more effectively. Complex algorithms, such as neural networks and ensemble methods, are commonly used in machine learning, and are often black box systems, with the decision-making process being hard to understand. Explainable Artificial Intelligence (XAI) methods are used to enhance transparency and interpretability to machine learning predictions. SHAP (SHapley Additive exPlanations) was used as a tool of model interpretability in this study. SHAP is a popular explainable AI system that measures the influence of each feature on the prediction of a model. In order to evaluate the effectiveness of the developed machine learning models in forecasting drug side effects, a number of evaluation metrics were used. Because the issue is multi-label classification, where a single drug can have several side effects at once, the conventional single-label accuracy cannot be used to adequately assess the performance of the models.. The assessment scheme consisted of Micro F1 Score, Macro F1 Score, Hamming Loss, Precision, Recall and ROC-AUC. All these measures are used to measure the accuracy of predictions, the detection rate of rare side effects in the models, and the overall classification accuracy in the multi-label data set.

II. DATABASES AND DATA COLLECTION

2.1 Drug and Side-Effect Data Acquisition

Drug–side effect data were collected from the SIDER (Side Effect Resource) database, a publicly available repository.The SIDER database integrates side-effect information using standardized medical terminology and provides associations between drugs and their reported adverse reactions. The current release of SIDER contains information on more than 1400 drugs and nearly 6000 adverse drug reactions, forming over 140,000 drug–side effect associations [8]. Having processed and filtered a dataset of approximately 1544 drugs and reported side effects, a set of drugs with reported side effects was acquired. The information was organized in a spreadsheet form having two major columns:

- Drug Name
- Associated Side Effects

The data on drug safety is multi-label in this dataset since a single drug can be associated with a variety of adverse effects. This curated data was used as the starting point to further enrich and compute the data.

2.2 Drug Annotation Using PubChem

The PubChem database was used to retrieve molecular identifiers and structures of the curated drug list. PubChem is a source of complete chemical data, such as compound identifiers and molecular structures of many bioactive molecules. In the dataset, the following information was retrieved on each drug:

- PubChem Compound ID (CID)
- SMILES (Simplified Molecular Input Line Entry System)

2.3 Molecular Feature Extraction Using SwissADME

Once SMILES representations of all drugs were obtained, molecular descriptors and pharmacokinetic properties were obtained through the SwissADME platform. SwissADME is an extensively employed computer-based software package to assess physicochemical properties, pharmacokinetics, druglikeness and medicinal chemistry properties of small molecule drugs.

2.2.1 Feature Selection and Final Dataset

The final processed data retained the most useful molecular descriptors to use in machine learning analysis, such as all the important physicochemical properties, lipophilicity indicators, pharmacokinetic parameters, drug-likeness rules, and medicinal chemistry descriptors. The final dataset therefore represents an integrated drug feature matrix, combining:

- Drug identifiers
- SMILES structural information
- Selected SwissADME molecular descriptors
- Associated side-effect annotations

III. COMPUTATIONAL TOOLS AND DATA PROCESSING

3.1 Computational Environment

The entire computational analysis and data processing in this work were written in Python programming language, which is popular in bioinformatics, data science and machine learning. Several open-source Python libraries were utilized to implement the data processing and modelling pipeline:- NumPy, Pandas, Scikit-learn, SciPy

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13454

These libraries collectively enabled efficient handling of the large drug–side-effect dataset and facilitated the implementation of the computational workflow for feature processing, dataset preparation, and machine learning model development. [10]

3.2 InChIKey and RDKit-Based Chemical Representation

The InChIKey has been widely adopted in large-scale chemical databases like PubChem, ChemSpider, ChEMBL, and DrugBank as a cross-reference or cross-linking chemical compound identifier. For example, drug Paracetamol (Acetaminophen) has the following chemical identifiers: Drug Name: Paracetamol

Molecular Formula: $C_8H_9NO_2$
Canonical SMILES: CC(=O)NC1=CC=C(O)C=C1
InChIKey: RZVAJINKPMORJF-UHFFFAOYSA-N

This InChIKey uniquely represents the chemical structure of Paracetamol and allows consistent identification of the compound across different chemical databases [10] [11] [12]. This study used an open-source cheminformatics toolkit, RDKit, written in Python to computeally process molecular structures and extract quantitative chemical properties

3.3 Binary Representation of Side-Effect Labels

In this representation, each unique side effect is treated as an independent label. A binary label matrix was constructed in which: Rows represent individual drugs
Columns represent unique side-effect categories
A value of 1 indicates the presence of a side effect
A value of 0 indicates the absence of the side effect
As a result, every drug is encoded as a binary vector that specifies the occurrence or not of all possible side effects. These representations are common in multi-label learning, where the model gives a binary vector of predictions of labels on each sample. [13]

3.4 Handling Class Imbalance in Side-Effect Data

This imbalance may adversely affect machine learning models, where algorithms spend time preferring the frequent labels over-representing the infrequent ones [14]. In the first preprocessing phase, the side-effect annotations of the dataset were extracted and about 4212 side-effect labels were obtained. Very rare labels can give very few details to the model training and can add noise to the prediction process. To reduce this problem, a frequency-based filtering approach had been used [14]. After applying this threshold, the total number of side-effect labels was reduced from 4212 labels → 1036 labels. The resulting dataset therefore provides a more balanced and computationally manageable multi-label representation for training predictive models.

5. Machine Learning

5.1 Data Preparation and Feature Matrix Construction

Before training machine learning models, the dataset was transformed into a structured numerical format suitable for computational learning algorithms [15]. The processed dataset consisted of two main components:

Feature matrix (X):
Contains molecular descriptors and physicochemical properties of drugs obtained from SwissADME and RDKit analysis.

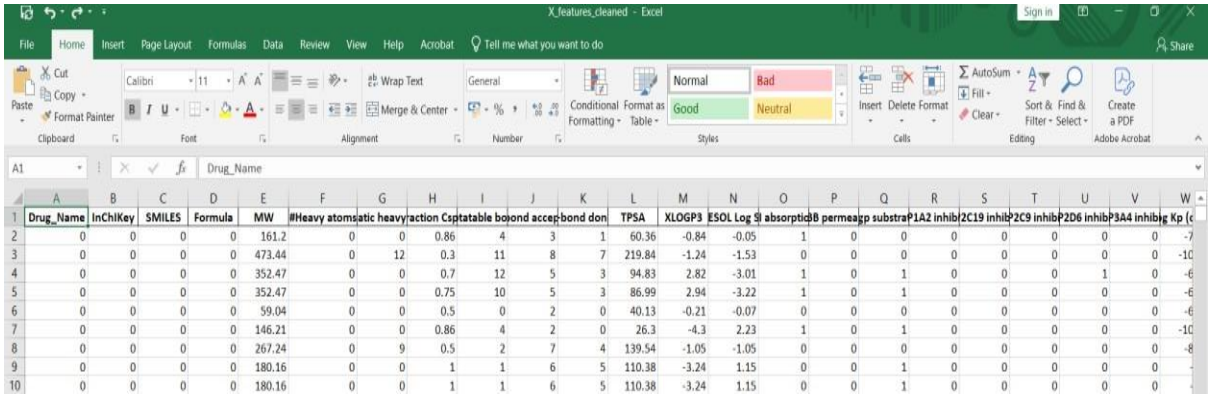


Figure 3 Label matrix (Y):

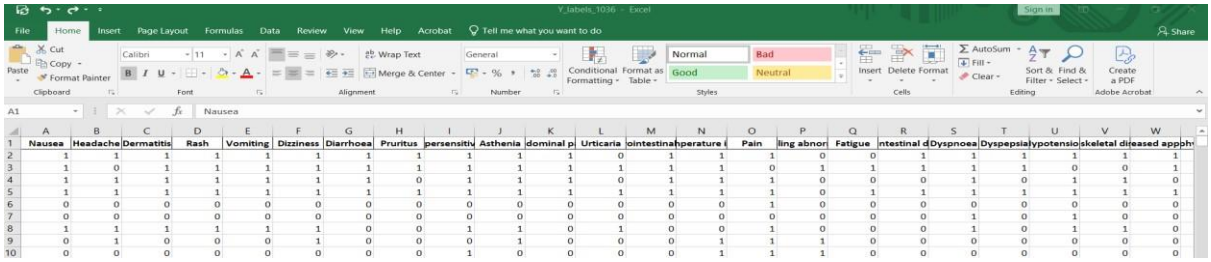
IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13454

Represents the presence or absence of side effects using binary encoding for each drug.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	Nausea	Headache	Dermatitis	Rash	Vomiting	Dizziness	Diarrhoea	Pruritus	persensitiv	Asthenia	dominal p	Urticaria	ointestinal	perature (	Pain	ling abnor	Fatigue	ntestinal d	Dyspnoea	Dyspepsia	ypotensio	skeletal di	eased apph
2	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	1	1	1
3	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1
4	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	0	1	0	1	1	0
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
8	1	1	1	1	1	1	1	0	1	1	0	1	0	0	1	0	0	0	1	0	1	1	0
9	0	1	0	0	0	1	0	0	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0

Figure 4

Using Python programming, the dataset was converted into a feature–label representation, where each drug is represented as a vector of numerical features describing its molecular characteristics. The information about the side-effects was coded into a binary multi-labeled vector that indicated whether each side-effect was present or not.

4.2 Machine Learning Models

The correlation between molecular descriptors and side-effect labels was examined with the help of the Logistic Regression as a baseline model. It uses a sigmoid function to transform linear outputs to 0-1 probabilities. Baseline models are used to compare the performance of improved algorithms [17].

Random Forest was applied with Scikit-learn to classify several side-effects. It constructs several decision trees out of the random samples, learns relationships between features and labels, and combines predictions. The imbalance of labels was dealt with by using class weighting [16].

XGBoost was applied to increase prediction performance by boosting, paying attention to the hard-toclassify samples. It repeatedly optimizes the model accuracy by the molecular descriptors dataset [16].

The Multilayer Perceptron (MLP) neural network was also used, which is made up of the input layer, the hidden layer and the output layer. It learns non-linear interrelations between features and labels. The output layer includes a sigmoid activation function to provide multi-label predictions and is trained by minimizing the error in prediction by optimizing the weights [18].

4.3 Model Evaluation Metrics

The machine learning models were tested with conventional multi-label classification measures. The micro F1 Score is a performance metric that sums up the number of predictions of all labels, which is suitable in imbalanced datasets. Macro F1 Score averages F1 scores of all labels giving equal weight to both common and rare side effects. Hamming Loss is a measure of the ratio of misclassified labels, with smaller numbers depicting improved performance.

These metrics have been chosen because they are effective in multi-label problems. Micro F1 shows performance with high frequency of labels but Macro F1 emphasizes the model behavior of all labels including rare ones. Hamming Loss is a direct forecast of the error of prediction in the drug- side effect pairs, which offers an explicit measure of the total model accuracy [19].

IV. RESULTS

5.1 Machine Learning Model Performance (Micro, Macro, Hamming)

5.1.1 Logistic Regression Results

The obtained performance metrics for Logistic Regression are as follows:

- **Micro F1 Score:** 0.2683
- **Macro F1 Score:** 0.0445
- **Hamming Loss:** 0.1038

The Macro F1 score is relatively lower, which demonstrates that the model did not predict rare sideeffect labels with high accuracy, which is understandable because of the high class imbalance in the dataset. Nonetheless, the Logistic Regression is a helpful point of reference to compare more sophisticated models.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13454

5.1.2 Random Forest (Balanced) Results

The performance metrics obtained for the balanced Random Forest model are:

- **Micro F1 Score:** 0.3846
- **Macro F1 Score:** 0.2363
- **Hamming Loss:** 0.0898

In comparison with the Logistic Regression, the Random Forest model had a substantial increase in Micro and Macro F1 scores. This implies that the model was more suited at representing intricate associations among drug molecular characteristics and side-effect labels.

5.1.3 XGBoost Results

The performance metrics obtained for the XGBoost model are:

- **Micro F1 Score:** 0.4025
- **Macro F1 Score:** 0.2071
- **Hamming Loss:** 0.07891

XGBoost model proved to be a very powerful predictor because it is capable of modeling complex nonlinear relationships and it can effectively model high-dimensional feature spaces.

5.1.4 Neural Network Results

The neural network model achieved the following performance metrics:

- **Micro F1 Score:** 0.2932
- **Macro F1 Score:** 0.0627
- **Hamming Loss:** 0.0991

The neural network successfully learnt nonlinear interactions between features; however, its performance was a little worse than the Random Forest model. This can be explained by a relatively small amount of data and a high dimension of the multi-label output space.

5.1.5 Comparative Performance Analysis

	Model	Micro F1	Macro F1	Hamming Loss
0	Logistic Regression	0.271187	0.045944	0.080241
1	Random Forest	0.384638	0.236349	0.089796
2	XGBoost	0.402582	0.207172	0.078916
3	Neural Network	0.398047	0.236010	0.092024

Points to notice

- XGBoost achieved the highest Micro F1 score (0.4026), indicating better overall prediction of side effects.
- Random Forest and Neural Network showed competitive Macro F1 scores (~0.236), indicating better performance on less frequent side effects.
- XGBoost also produced the lowest Hamming Loss (0.0789), suggesting fewer incorrect label predictions.

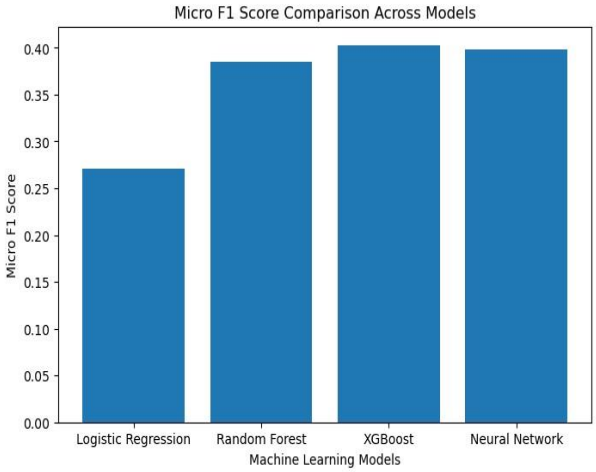


Figure 5

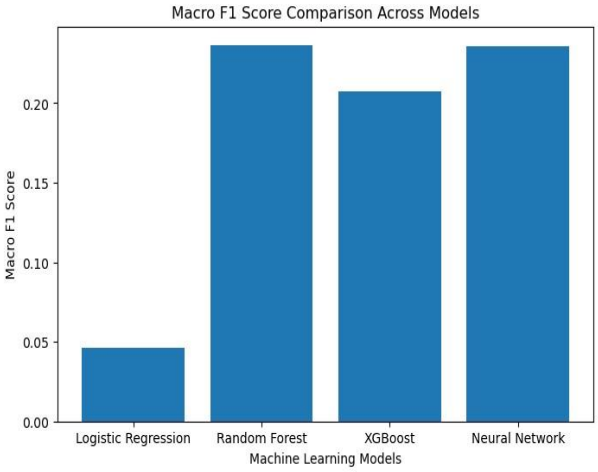


Figure 6

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13454

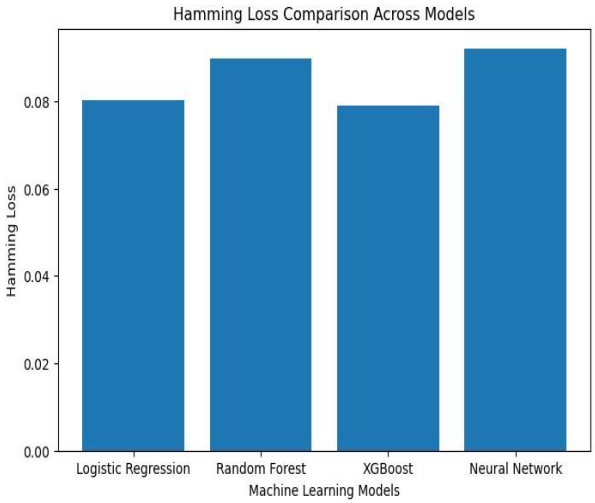


Figure 7

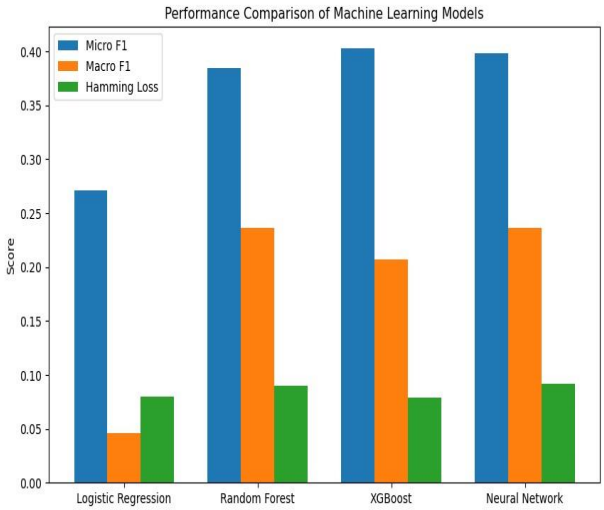


Figure 8

5.2 ROC–AUC Analysis

- Receiver Operating Characteristic – Area Under the Curve (ROC–AUC) was used to further evaluate the discrimination capability of the machine learning models in predicting drug side effects. ROC analysis compares the True Positive Rate (TPR) or the False Positive Rate (FPR) over various classification thresholds and hence gives a threshold-free assessment of the functionalities of the model [20].

- The ROC–AUC results obtained for the evaluated models are summarized below.

Model	Micro ROC–AUC	Macro ROC–AUC
Logistic Regression	0.8308	0.6617
Random Forest	**0.8492**	**0.7259**
XGBoost	0.8415	0.7052
Neural Network	0.8112	0.6826

- XGBoost was also found to be highly discriminative, whereas the Logistic Regression and the Neural Network exhibited relatively low yet decent ROC–AUC scores. Outcomes imply ensemble tree-based models are especially efficient at modeling complex interactions

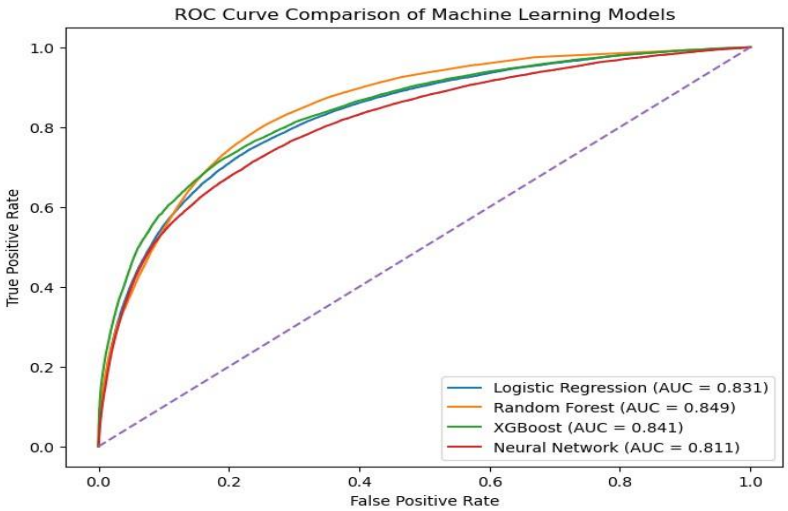


Figure 9 6.3 Precision and Recall Analysis

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13454

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

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

In multi-label classification, Micro averaging averages contributions of all labels to compute a single metric, whereas macro averaging averages the results of computing the metric separately across each label. The method enables testing of the overall model performance as well as the performance of the model on rare side-effect labels. [21]

Precision Results

Model	Micro Precision	Macro Precision
Logistic Regression	0.6029	0.0986
Random Forest	0.4632	0.3748
XGBoost	0.5687	0.3760
Neural Network	0.4505	0.2933

The findings indicate that Logistic Regression had the best Micro Precision (0.6029) which means that in the instances in which this model predicted the side effect occurrence, a relatively large percentage of these predictions were accurate. But, its Macro Precision (0.0986) was much lower, implying that it does not predict rare side-effect labels well.
XGBoost and Random Forest, on the contrary, showed greater values of Macro Precision (~0.37), which means that they are better able to detect less common adverse reactions. Recall Results

Model	Micro Recall	Macro Recall
Logistic Regression	0.1749	0.0391
Random Forest	0.3289	0.2174
XGBoost	0.3116	0.1627
Neural Network	0.3565	0.2141

The Neural Network reached the highest Micro Recall (0.3565), which means that the neural network was able to detect actual side-effect cases better. Random Forest also exhibited good recall, especially in Macro Recall, which indicated a better ability to identify rare side effects.

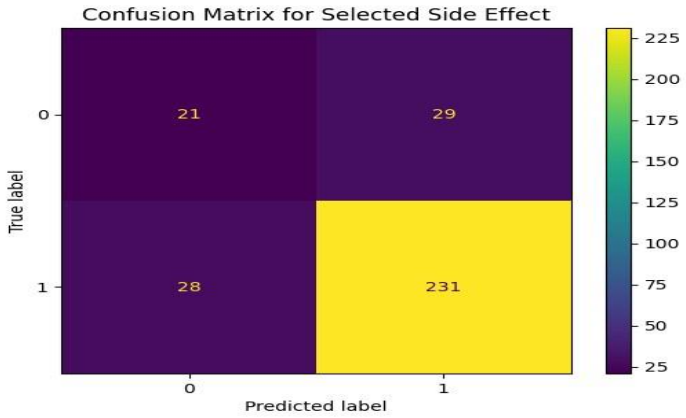


Figure 12

In this study:

- **Logistic Regression** achieved higher precision but lower recall.
- **Random Forest and XGBoost** demonstrated balanced precision and recall values.
- **Neural Network** achieved higher recall but relatively lower precision.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2026

DOI: 10.17148/IARJSET.2026.13454

Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) was used to improve the transparency of model predictions. In this study, SHAP (SHapley Additive exPlanations) was applied to interpret the trained models and identify the most influential molecular descriptors, chemical fingerprints, and pharmacokinetic features contributing to drug side-effect prediction [22] [23].

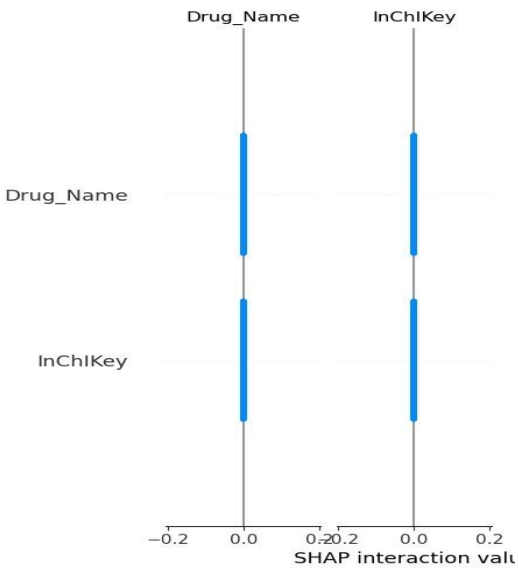


Figure 10

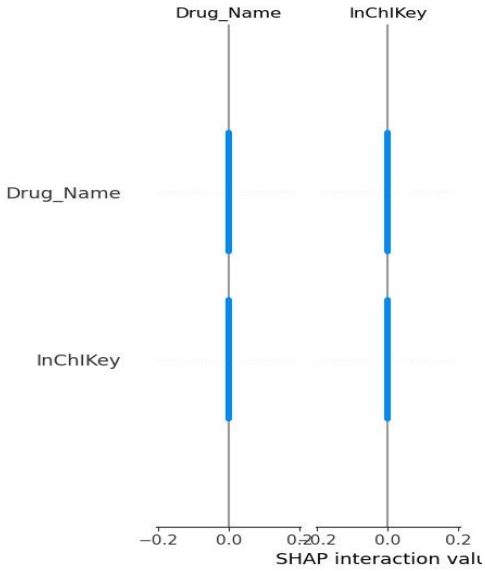


Figure 11 Explanation

Interactions between input features in the predictive model were analyzed using 11SHAP interaction plots to analyze two-way interactions between input features. Interaction between Drug_Name and InChIKey revealed that the values of SHAP interaction between the two features had a concentration around 0, which means that these identifier-based properties did not contribute much to the prediction of side effects. This implies that the model is based on more meaningful chemical and pharmacological properties, including molecular descriptors and fingerprints, instead of mere identifiers, which is the effectiveness of the feature engineering process.

VI. DISCUSSION

The aim of this study was to evaluate the ability of machine learning models to predict drug side effects using molecular descriptors and structural features. RDKit and SwissADME features were used to train models based on Logistic Regression, Random Forest, XGBoost, and Neural Network (MLP), with labeling of side-effects being found in the SIDER database. The results showed that ensemblebased models outperformed the baseline Logistic Regression model. Random Forest and XGBoost had superior predictive performance, with XGBoost having the highest Micro F1 score and lowest Hamming Loss, which shows a high overall accuracy. Random Forest was also the highest ROC-AUC score, demonstrating a better capacity to differentiate between drugs implicated in side effects. The Neural Network (MLP) worked at a competitive level, but could not outperform the tree-based ensemble models, probably because of the size of the data set and the large dimensionality of the multi-label output space. Generally, the results indicate that ensemble machine learning models, especially the Random Forest and XGBoost, can be used to predict drug side effects based on molecular and pharmacokinetic features.

VII. CONCLUSION

This paper created a machine learning system to predict drug side effects based on molecular descriptors, structural features, and pharmacokinetic properties, which were gathered via PubChem and SwissADME and the SIDER database. Multi-label classification models such as, Logistic Regression, Random Forest, XGBoost and Neural Network (MLP) were compared using measures like Micro F1, Macro F1, Hamming Loss, ROC-AUC, Precision and Recall. The findings revealed that the ensemble models performed better than the baseline model, with XGBoost demonstrating the highest overall predictive accuracy and the Random Forest strong discriminative performance. These results suggest that ensemble learning techniques can be applied efficiently in order to describe complicated associations between molecular

DOI: 10.17148/IARJSET.2026.13454

characteristics and adverse drug reactions. Despite the persistence of issues like class imbalance and infrequent side-effect prediction, the research proves that machine learning has a great potential in early side-effect prediction that will help to develop safer drugs and enhance pharmacovigilance.

REFERENCES

[1]. Sachdev, K., & Gupta, M. K. (2020). **A comprehensive review of computational techniques for the prediction of drug side effects.** *Drug Development Research, 81*(6), 650–670. https://pubmed.ncbi.nlm.nih.gov/32314424/

[2]. Deng, S., Sun, Y., Zhao, T., Hu, Y., & Zang, T. (2020). **A review of drug side effect identification methods.** *Current Pharmaceutical Design, 26*(26), 3096–3104. https://pubmed.ncbi.nlm.nih.gov/32532187/

[3]. Toni, E., Ayatollahi, H., Abbaszadeh, R., & Fotuhi Siahpirani, A. (2024). **Machine learning techniques for predicting drug-related side effects: A scoping review.** *Pharmaceuticals, 17*(6), 795. https://www.mdpi.com/1424-8247/17/6/795 4) Zhao, H., Zhong, J., Liang, X., Xie, C., & Wang, S. (2025). **Application of machine learning in drug side effect prediction: Databases, methods, and challenges.** *Frontiers of Computer Science, 19*, 195902. https://doi.org/10.1007/s11704-024-31063-0 5) Kuang, Q., Wang, M., Li, R., Dong, Y., Li, Y., & Li, M. (2014). **A systematic investigation of computational models for predicting adverse drug reactions.** *PLoS ONE, 9*(9), e105889. https://doi.org/10.1371/journal.pone.0105889 6) Liu, M., Wu, Y., Chen, Y., Sun, J., Zhao, Z., Chen, X., Matheny, M. E., & Xu, H. (2012). **Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs.** *Journal of the American Medical Informatics Association, 19*(e1), e28–e35. https://doi.org/10.1136/amiajnl-2011-000672

[4]. Alam, F., Rahman, M. A., & others. (2024). **Computational prediction of drug side effects using machine learning techniques.** *Bioinformatics and Biology Insights.* https://pmc.ncbi.nlm.nih.gov/articles/PMC11639029/

[5]. Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2016). **The SIDER database of drugs and side effects.** *Nucleic Acids Research, 44*(D1), D1075–D1079. https://pmc.ncbi.nlm.nih.gov/articles/PMC4702794/

[6]. Daina, A., Michielin, O., & Zoete, V. (2017). **SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules.**

[7]. https://www.swissadme.ch/faq.php

[8]. Digital Regenesys. (2023). **Python libraries for data science: NumPy, Pandas, and Scikit-learn.** https://www.digitalregenesys.com/blog/python-libraries-data-science-numpy-pandas-scikit-learn

[9]. InChI Trust. (2023). **About the InChI standard.** https://www.inchi-trust.org/about-the-inchi-standard/

[10]. Kim, S., Thiessen, P. A., Bolton, E. E., & others. (2015). **PubChem substance and compound databases.** *Nucleic Acids Research.* https://pmc.ncbi.nlm.nih.gov/articles/PMC4486400/

[11]. 13) Zhang, Y., & Liu, Y. (2014). **Application of machine learning methods in biomedical research.** *Applied Mechanics and Materials, 536-537*, 394–397. https://www.scientific.net/AMM.536-537.394 14) Zhao, X., Chen, L., & Lu, J. (2018). **A similarity-based method for prediction of drug side effects with heterogeneous information.** *Mathematical Biosciences, 306*, 136–144. https://doi.org/10.1016/j.mbs.2018.09.010 15) Chen, L., Zhao, X., & Zhang, X. (2015). **Prediction of drug side effects with refined negative sample selection strategy.** *BMC Bioinformatics.* https://doi.org/10.1186/s12859-015-0774-y 16) Zhang, Y., & colleagues. (2025). **Machine learning applications in biomedical prediction systems.** *Applied Sciences.* https://www.mdpi.com/2076-3417/15/19/10853 17) Wang, Y., & others. (2024). **Predicting adverse drug reactions using machine learning approaches.** *Scientific Reports.* https://www.nature.com/articles/s41598-024-76424-8

[12]. 18) UpGrad. (2023). **Multilayer perceptron (MLP) in machine learning.** https://www.upgrad.com/blog/multilayer-perceptron-mlp-in-machine-learning/

[13]. Wu, J.-S., Hu, H.-F., Yan, S.-C., & Tang, L.-H. (2015). **Multi-instance multilabel learning with weaklabel for predicting protein function in electricigens.** *BioMed Research International.* https://pmc.ncbi.nlm.nih.gov/articles/PMC4436452/

[14]. Pedregosa, F., & others. (2024).**ROC-AUC score.** *Scikit-learn documentation.* https://scikitlearn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html

[15]. Encyclopaedia Britannica. (2023).**Precision and recall.** https://www.britannica.com/science/precision-and-recall

[16]. Yang, W., Wei, Y., Wei, H., Chen, Y., Huang, G., Li, X., Li, R., & Yao, N. (2023). **Survey on explainable AI: From approaches, limitations and applications aspects.** *Human-Centric Intelligent Systems, 3*, 161–188. https://doi.org/10.1007/s44230-023-00038-y

[17]. Lundberg, S. M., & Lee, S.-I. (2017). **A unified approach to interpreting model predictions.** *Advances in Neural Information Processing Systems (NeurIPS 2017)*, 30.

[18]. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions