Performance Analysis of Machine Learning Models for Drug Toxicity Prediction

Kamal
Dept. of CSE
Baba Mastnath University,
Rohtak, India

Devender Kumar
Dept. of Computer Science and Application,
Baba Mastnath University,
Rohtak, India

Anuj Kumar Sharma
Dept. of CSE
BRCM College of Engineering,
Bahal, India

***Abstract*— Toxicology testing is a crucial stage of the medication production process. However, the cost and duration of the existing empirical procedures utilized to calculate drug toxicity indicate these are unsuitable for extensive assessments of drug toxicity mostly in initial stages of drug production. Consequently, there is a great requirement for computer systems that can forecast the potential of medication toxicity. Expensive in respect of both individual health and financial resources are unanticipated patient security incidents in clinical studies for innovative medications. The drug development business does thorough preliminary security screening in an effort to reduce these occurrences. Low bogus experimental toxicology outcomes remain a possibility despite the fact that present best procedures are effective at avoiding dangerous substances from becoming evaluated in clinical settings. The experimental domain must always be continuously improved. With increased knowledge of possible toxicity and related causes, greater medicines could be developed. Artificial intelligence and machine learning techniques offer new views on drug toxicity forecasts. In order to provide ratings to discovered characteristics depending on its choices and importance, machine learning attribute selection approaches are helpful. The effectiveness of the forecasting system is also improved via cluster assessment. The medical industry will gain immensely if we're able to predict medication toxicity, which is why machine learning is so important for understanding drug toxicity in various aspects.**

***Keywords*—Drug toxicity, prediction model, machine learning, ensemble.**

I. INTRODUCTION AND BACKGROUND

Human body is frequently subjected to a wide range of chemical compounds in the modern environment, including cosmetics, particles, and common hazardous and toxic compounds. Furthermore, we are unaware of the precise mechanisms by what substances cause adverse symptoms or, in the most severe situation, non-acute or sub-acute toxicity. Harm to organs and possibly death could result from it[1]. This really is taking place due to the toxicity of the medication. Machine learning, which is presently frequently applied to a wide range of contexts along with speech recognition, natural language processing, picture recognition, computation chemistry, and bioinformatics to advantageous achievement, can help with toxicity prediction[2]. It is primarily used during Big Data and artificial intelligence production. Toxicology calculations are essential for detecting any negative effects that chemicals may have. These compounds have an impact on both people and animals as well as plants. All medications must undergo medical testing before ever being licenced for use on humans. Unfortunately, there is considerable risk associated with drug research. Approximately two-thirds of drugs have been shown to be harmful or worthless in late patient medical tests, depending to reports. Preclinical assessments are crucial for avoiding hazardous pharmaceuticals from getting through clinical testing, that is highlighted by clinical trial uncertainty. Forecasting drug toxicity is crucial for the creation of new medications[3-6]. Even though in vivo animal testing is constrained by expense, duration, and ethical issues, animal procedures are routinely used to determine toxicity. Experts preferred analytically modelling rather more conventional techniques for predicting degrees of danger in account of such factors. Over the past ten years, the top pharmaceuticals companies in the nation have started to use a homoeopathic strategy to therapy and rehabilitative services. This change led to better advancements in diseases protection and therapy, however it simultaneously raised drug costs that

constituted a social burden. Despite becoming incredibly varied and specialised to opportunities, the expense of drug discovery and production has risen steadily and dramatically.
In attempt to build a design, machine learning techniques are utilised to train and evaluate the instances. The machine learning method produces forecasts depending on the framework as additional data is introduced into it. Inside this work, machine learning and deep learning can be used to identify the attitudes and extract particular activity[7-11]. The principle of machine learning is to educate and train machines by giving them information and distinguishing characteristics. Machines educate, modify, and extend while depending on specific programming when given new and pertinent information. If there are no records, machines might recall things less. The Machine analyses data, looks for similarities therein, trains by responsive action, or produces predictions. Machine learning techniques build a system that forecasts or judges without even any specific instructions by using training knowledge. Machine learning techniques can be used in many circumstances wherein traditional methods are impractical to deploy. Examples include consumer machine interaction, fraud detection, human voice analytics, and pharmaceutical research.

ML Algorithms

Drug manufacturing has been significantly enhanced by machine learning techniques. The biopharmaceutical business has substantially benefited from the development of medications utilising a range of machine learning approaches. In the last decade, a variety of machine learning (ML) approaches were extensively used in the monitoring of diabetes. Various approaches for forecasting the biochemical, biological, and physiological features of molecules are developed using machine learning techniques. The proposed strategy proposes an original method to assess the severity of cardiac diseases using conventional lifespan research and machine learning methodologies focused on limits. Machine learning is becoming essential for handling innovative medical treatments, medical data, and clinical history as well as for the field of healthcare[12-14]. At many stages of the drug development procedure, machine learning techniques may be useful. For example, machine learning techniques have been extensively used to find new applications for medications, assess drug effectiveness, discover drug linkages, ensure safety surveillance, and enhance chemical biocompatibility[15]. Medicinal study frequently utilised the machine learning methods Random Forest, Naive Bayesian, and Support Vector Machine.

Ensembling

It seemed like here were many of ways to make learning lethargic. Probably most widely used examples of lazy training are ensemble classifications. Unsupervised learning techniques known as ensemble processes create a number of classifications and then use the bulk of their estimations to find new collected data[16-19]. The likelihood of selecting a classifier with bad efficiency is reduced whenever the results of a variety of sensors with comparable educational capabilities have different generalization results[20-24]. It has actually been proven that a classifying ensemble consistently outperforms a keep alternative.
The ensemble methodology is described as a diverse architecture in machine learning where numerous classifiers and approaches are purposefully merged to produce a forecasting machine[25-28]. The ensemble technique also helps in correctly identifying and forecasting statistics from complex problems, reducing biases in the forecasting framework, and decreasing scatter in predicted values.

II. LITERATURE REVIEW

This area focuses on important study performed in its field of drug toxicity and machine learning by a number of scholars, which we emphasise using the literature review that is presented here.

- **Tharwat et al. 2018** High total count of blastocysts, consensual mistake is just not feasible quite so, slow and incorrect drug tonicity are problems mentioned in this paper. ML and microbially methodologies are recommended as part of a completely computerised procedure to examine the tonicity of microscope images of treated zebra fish embryos.
- **Haixin Ali 2019** this research to create a framework for the development of chemical toxic effects. Recursive function reduction and support vector machines were coupled to create a regression prototype. Utilizing Three ML approaches in classification structures, three ensemble models were constructed.
- **Sonal Mishra et al. 2015** difficulty to create a framework for predicting protein architecture that was discovered in this work. A comparison of ML-based proteins architecture forecasting frameworks.
- **Zhi-hua Zhou 2009** the issue mentioned in this paper is detection of poor students. This can be resolved by using putting together ML algorithms.
- **J Cai 2018** the issue mentioned in this research is ML, high dimensional data processing is difficult. Effective FS techniques increased accuracy and made the learning model easier to comprehend.

- **Asha S Manek 2015** Opinion mining, sentiment analysis, opinion extraction, affect analysis, and decision-making for potential customers over whether or not to purchase the goods. It is suggested to use a Gini-Index based FS approach with SVM classifiers to identify mood in a sizable dataset of movie reviews.
- **Karim 2021** the issue mentioned in this research is he effectiveness of the algorithm is constrained by using only one kind of feature description and one kind of neural network. Five distinct basic deep learning models were used to provide a deep learning architecture for the forecasting of quantitative toxicity.
- **Maini 2021** the topic mentioned in this paper is e early cardiac disease prediction is necessary. Comprehensive attempts have been undertaken to improve the efficiency of a system that was created to forecast cardiac disease.
- **Ramana et al. 2019** the necessity for early disease prediction is an issue this paper identifies. Evaluation of efficiency across several datasets.
- **Roy 2021** it takes a lot of time and effort for doctors to diagnose the invariveductal carcinoma stage of breast cancer. Created a model for computer-aided breast cancer detection using assembly.
- **Singh 2018** the problem raised in this study the use of an ensemble approach is necessary for precise forecasting. Researchers can develop a method for neuro-fuzzy assembly.
- **Smith 2016** in this research examples of noisy data with incorrect labels and outliers are evaluated. Researchers studied anomaly filtering and ML method assembly.
- **Sumonja 2019** Nutrient relationships are predicted in this paper. Newly developed expertly built series, evolutionary, and graphical characteristics have been included, and autonomous feature engineering has been used to further broaden and enhance forecasting.

III. Result and Discussion

This section concentrates on contrasting our proposed method with widely used machine learning methods.

Table 2: Correlation coefficient

Regeration Model	Correlation Coefficient
Simple linear regression	0.53
IBk	0.61
AdditiveRegression	0.59
RandomCommittee	0.71
Randomizable filterd Classifier	0.61
Decision table	0.63
M5P	0.67
Random Tree	0.58

The correlation coefficient is a numerical measure of the strength of the relationship between the relative changes of two items. The above table demonstrates how the correlation coefficient numbers for the different classifiers differ.

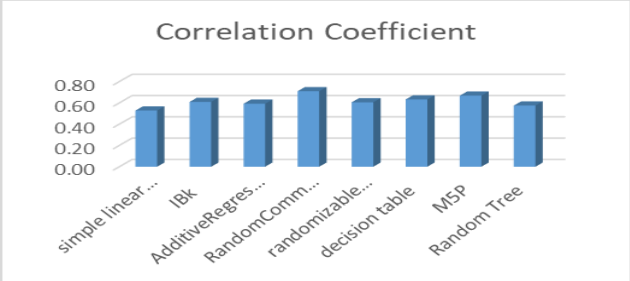

Fig 6: Correlation coefficient

We can examine various correlation coefficient numbers in the graph.

Table 3: Mean absolute error

Regeration Model	Mean Absolute Error
simple linear regression	10.71
IBk	9.72
AdditiveRegression	10.51
RandomCommittee	8.51
randomizable filterd Classifier	9.83
decision table	9.60
M5P	9.26
Random Tree	10.68

The **Mean Absolute Error** (MAE) The Mean Absolute Error is the average of all relative errors. The Mean Absolute Error determines the typical number of anomalies in a group of predictions while accounting for the origin of the errors. The measuring criterion is what affects the degree of accuracy.

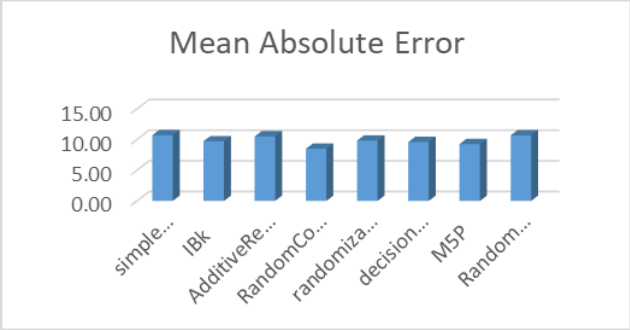

Fig 7 : Mean Absolute Error

The accuracy of the system is strongly impacted by the various Mean Absolute Errors that may be assessed in the above image.

Table 4: Root mean squared error

Regeration Model	Root Mean Squared Error
Simple linear regression	1.41
IBk	1.45
AdditiveRegression	1.35
RandomCommittee	1.18
Randomizable filterd Classifier	1.44
Decision table	1.32
M5P	1.24
Random Tree	1.55

The root mean squared error is described as the square root of the mean of the square of each inaccuracy. For quantitative predictions, RMSE is commonly used and is considered as a better all-purpose error metric. We'll find that the RMSE values for each classifier fluctuate.

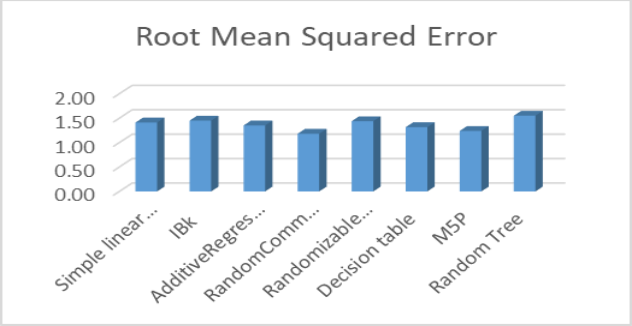

Fig 8 : Root mean squared error

As seen in the above figure, several classifiers have varying RMSE values.Table 5: Accuracy

Regeration Model	**Accuracy**
Simple linear regression	89.29
IBk	90.28
AdditiveRegression	89.49
RandomCommittee	91.49
Randomizable filterd Classifier	90.17
Decision table	90.40
M5P	90.74
Random Tree	89.32

The simplest metric to assess the accuracy of a forecast is Mean Absolute Error. The MAE, or mean of the absolute errors, is precisely what its name implies. The absolute error is the difference among the projected value, the real value, and the real value, presented as a complete and comprehensive number.

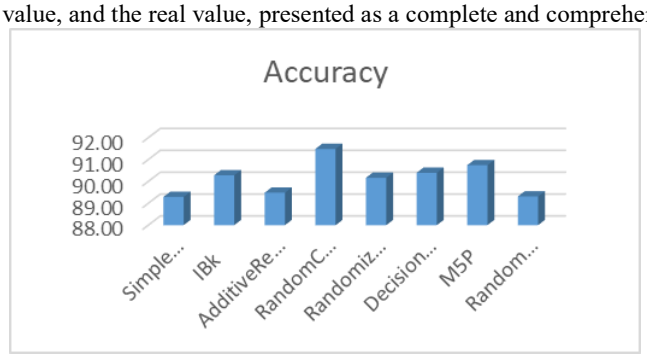

Figure 9 Accuracy

IV. Conclusion and Future Scope

By reducing source characteristics within the training dataset, a machine learning model that is easier and better effective can be created. The effectiveness of a regularly employed machine learning model is examined in this research, and it is discovered that there is room for enhancement.

I. References

[1] Ng,H.W., Shu,M., Luo,H., Ye,H., Ge,W., Perkins,R.,Tong,W., Hong, H. *Estrogenic activity data extraction and in silico prediction show the endocrine disruption potential of bisphenol A replacement compounds.* Chem. Res. Toxicol. 2015, 28, 1784–1795. [CrossRef] [PubMed].

[2] Hong, H., Neamati, N. Winslow, H.E., Christensen, J.L., Orr, A., Pommier, Y., Milne, G. W. A. *Identification of HIV-1 integrase inhibitors based on a four-point pharmacophore.* Antivir. Chem. Chemother. 1998, 9, 461–472. [CrossRef] [PubMed].

[3] Hong, H., Tong, W., Xie, Q., Fang, H., Perkins, R. *An in silico ensemble method for lead discovery: Decision forest*. SAR QSAR Environ. Res. 2005, 16, 339–347. [CrossRef]

[4] Hong, H., Fang, H., Xie, Q., Perkins, R., Sheehan, D.M., Tong, W. *Comparative molecular field analysis (CoMFA) model using a large diverse set of natural, synthetic and environmental chemicals for binding to the androgen receptor.* SAR QSAR Environ. Res. 2003, 14, 373–388. [CrossRef]

[5] Lo, Y.C., Rensi, S.E.; Torng, W., Altman, R.B. Machine learning in chemo informatics and drug discovery. Drug Discov. Today 2018, 23, 1538–1546. [CrossRef] [PubMed]

[6] Lokesh Pawar, Jaspreet Singh, Rohit Bajaj, Gurpreet Singh, Sanjima Rana, *Optimised Ensembled Machine Learning Model for IRIS Plant Classification* published in 6th International Conference on Trends in Electronics and Informatics(ICOEI),2022/4/28 pp 1442-1446

[7] Shubham Kumar Singh, Revant Kumar Thakur, Satish Kumar, Rohit Anand, Deep Learning and Machine Learning based Facial Emotion Detection using CNN, 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom), 530-535, 2022/3/23

[8] MursalDawodi, Tomohisa Wada, Jawid Ahmad Baktash Applicability of ICT, *Data Mining and Machine Learning to Reduce Maternal Mortality and Morbidity: Case Study Afghanistan* , 2020.

[9] Rohit Bajaj, Gaurav Bathla, Abhishek Gupta, Lokesh Pawar, *Optimised Ensemble Model for Wholesale Market Prediction using Machine Learning* published in 3rd International Conference on Electronics and Sustaiable Communication System(ICESE), 2022/8/17, pp 1164-1169.

[10] KS Betts, S Kisely, RAlati *Predicting common maternal postpartum complications: leveraging health administrative data and machine learning.* 20 february, 2019(pp.702-703)

[11] Harinder Singh, Tasneem Bano Rehman, Ch Gangadhar, Rohit Anand, Nidhi Sindhwani, M Babu, Accuracy detection of coronary artery disease using machine learning algorithms, Applied Nanoscience, Springer International Publishing, 2021/8/27

[12] Machine Jaspreet Singh, Shruti Agarwal, Piyush Kumar, Divyans Rana, Rohit Bajaj, Prominent Feature based Chronic Kidney Disease Prediction Model using Machine, published in 3rd International Conference on Electronics and Sustaiable Communication System(ICESE), 2022/8/17, pp 1193-1198.

[13] 20Khamis,M.A., Gomaa,W., Ahmed,W.F. *Machine learning in computational docking.* Artif. Intell. Med. 2015, 63, 135–152. [CrossRef]

[14] Leelananda, S.P., Lindert, S. *Computational methods in drug discovery.* Beilstein J. Org. Chem. 2016, 12, 2694–2718. [CrossRef] [PubMed]

[15] Lokesh Pawr, nuj Kumar Sharma, Dinesh Kumar, Rohit Bajaj, Advanced Ensemble Machine Learning Model for Balanced BioAssays, published in the book Artificial Intelligence and Machine Learning in 2D/3D Medical Image Processing, 2022/12/22, pp 171-178.

[16] Maia, E.H.B., Assis, L.C., de Oliveira, T.A., da Silva, A.M., Taranto, A.G. *Structure-Based Virtual Screening: From Classical to Artificial Intelligence.* Front. Chem. 2020, 8, 343. [CrossRef] [PubMed]

[17] Talambedu, U., Shanmugarajan, D., Goyal, A.K., Kumar, C.S., Middha, S.K. *Recent Updates on Computer-aided Drug Discovery: Time for a Paradigm Shift.* Curr. Top. Med. Chem. 2017, 17, 3296–3307. [CrossRef]

[18] Lokesh Pawar, Pranshul Agrwal, Gurjot Kaur, Rohit Bajaj, Elevate Primary Tumor Detection Using Machine Learning, published in Journal of Cognitive Behaviour and Human Computer Interaction Based on Machine Learning Algorithm, 2021/12/1, pp 301-313.

[19] Bioassays Dinesh Kumar, Anuj Kumar Sharma, Rohit Bajaj, Lokesh Pawar, *Feature Optimized Machine Learning Framework for Unbalanced Bioassays,* , published in Journal of Cognitive Behaviour and Human Computer Interaction Based on Machine Learning Algorithm, 2021/12/1, pp 167-178.

[20] Réda, C., Kaufmann, E., Delahaye-Duriez, *A. Machine learning applications in drug development.* Comput. Struct. Biotechnol. J. 2020, 18, 241–252. [CrossRef]

[21] Pnkaj Rahi, Sanjay P Sood, Rohit Bajaj, Yogesh Kumar, *Air quality monitoring for Smart eHealth system using firefly optomization and support vector machine, published in International Journal of Information Technology, 2021/10*

[22] Lokesh PAwar, Anuj Kumar Sharma, Dinesh Kumar, Rohit Bajaj, Advanced Ensemble Machine Learning Model for Balanced BioAssays, Artificial Intelligence and Machine Learning in 2D/3D Medical Image Processing, pp 171-178, 2021.

[23] Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32. [CrossRef]

[24] Webb, G.I. Naïve Bayes. In Encyclopedia of Machine Learning Sammut, C., Webb, G.I., Eds.; Springer: Boston, MA, USA, 2010. [CrossRef]

[25] Cortes, C., Vapnik, V. Support-vector networks. Mach. Learn. 1995, 20, 273–297. [CrossRef]

[26] Dugger, S.A.; Platt, A., Goldstein, D.B. Drug development in the era of precision medicine. Nat. Rev. Drug Discov. 2018, 17, 183–196. [CrossRef] [PubMed]

[27] Hulsen, T., Jamuar, S.S., Moody, A.R., Karnes, J.H., Varga, O., Hedensted, S., Spreafico, R., Hafler, D.A., McKinney, E.F. *From Big Data to Precision Medicine.* Front. Med. 2019, 6, 34. [CrossRef] [PubMed]

[28] Liu, B., He, H., Luo, H., Zhang, T., Jiang, J. *Artificial intelligence and big data facilitated targeted drug discovery.* Stroke Vasc. Neurol. 2019, 4, 206. [CrossRef]