International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.224.03

Optimized Ensembled Model to Predict Drug Toxicity using Machine Learning

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*— Over the years, predicting toxicity has proven to be a difficult task. The technique of developing a novel medication or pharmaceutical molecule is both costly and time-consuming. Yearly, a great amount of chemical substances are created around the universe, and majority of those are discarded because to their toxicity; this highlights the necessity of drug toxicity forecasts. Drug toxicity predictions gains unique perspectives thanks to artificial intelligence and machine learning approaches. Machine learning feature selection methods assist in assigning scores to identified attributes based on their preferences and significance. Furthermore cluster assessment aids in increasing the predictions model's performance. In many ways, machine learning is crucial to drug toxicity; if we can forecast drug toxicity, the medical field will benefit greatly. The Optimised Model, our intended framework, is the main emphasis of this study. In this research, we will focus on an optimised ensemble model as machine learning and related methodologies play a vital part in forecasting the same. Since there is room for enhancement, the efficiency of the traditional model was enough, therefore we combined and streamlined the outcomes. By using operational research, we strengthened our model.**

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

I. INTRODUCTION AND BACKGROUND

In current world, our skin is regularly exposed to a variety of chemical substances such as cosmetics, particles, and everyday dangerous and toxic chemicals. However, we don't realize exactly chemicals generate side reactions or, in the worst-case scenario, non-acute or sub-acute poisoning, which can lead to allergic. It has the potential to cause organ damage and even mortality. Because of the drug's toxicity, this is happening. Mostly in Big Data and artificial intelligence generation, machine learning, that is currently routinely used throughout numerous domains including natural language processing, speech recognition, picture recognition, computation chemistry, and bioinformatics to beneficial performance, can assist with toxicity forecasting. Calculating toxicity levels is critical for identifying any damaging consequences generated by substances. Humans, plants, and even animals are affected by these substances. Clinical trials are required for all drugs, as is common information before being approved for sale. Drug studies, regrettably, are indeed accompanied with some degree of danger. In late human clinical studies, roughly two - thirds medications have been discovered to be hazardous or useless, according to reports. Clinical trial insecurity emphasises the necessity of preclinical evaluations, which are essential for preventing harmful medications in undergoing clinical testing. In medication development, toxicity predicting is extremely important. Animal techniques are frequently employed to toxicity assessment, although in vivo animal tests are limited by time, cost, and moral concerns. In light of these considerations, professionals chose analytical modelling over traditional methods for forecasting levels of toxicity. The world largest pharmaceutical organizations began to employ an allopathic approach to treatment and rehabilitation for last decade. This shift resulted in greater progress in disease treatment and prevention, but it also resulted in higher drug expenditures, which became a societal cost. The cost of drug discovery and development has constantly and increased drastically[1], although being extremely variable and specific to possibilities. Endpoint selection and classification, breakthrough exploration, and desired behaviour are all generalised pieces of initial drug design. For the generation and enhancement of lead molecules, a variety of computer-based techniques have now been utilised, include chemical docked[2,3], combinatorial modelling[4,5], selection rainforests[6], the roughly similar biochemical external analysis case[7].

Machine Learning methods are used to educate and analyse the examples in order to construct a pattern. When new data is fed into the machine learning process, it makes predictions based on the model. Under this study, the sentiments will be recognized and individual behaviour may be retrieved using Machine Learning and Deep Learning[8].Machine Training is based around a idea is teaching and training machines by providing them facts and identifying attributes. The majority of the problems with fatalities and morbidity can thus be resolved utilising ICT, as demonstrated in this research. The article suggests using data mining and machine learning algorithms to forecast the likelihood that a pregnant woman may experience maternal difficulties[9,10]. Despite relying on particular code, machines teach, expand, modify, and extend when updated and relevant facts are provided to systems. Computers may remember less if they don't have recordings. The Device examines statistics, finds patterns within it, trains through behaviour obediently, or generates forecasts. Machine learning approaches use trained information to construct a model that predicts or judge without anyone being explicitly instructed. Numerous situations wherever conventional procedures are too time-consuming or challenging to implement can benefit from the usage of machine learning algorithms. Pharmaceutical investigations, phishing identification, human speech analytics, and consumer machine communication are just a few examples. Determining the type of an individual seems to be the very basic requirement for conducting categorization then use a machine learning model. This task comment's information comes again from UCI Machine Learning repository.

ML Algorithms Used in Drug Discover

Machine learning methods have considerably improved drug development. The discovery of drugs using a variety of machine learning techniques has greatly enhanced the biopharmaceutical industry. Various Machine learning (ML) techniques were widely applied in the observation of diabetes in earlier decade[11]. Machine learning techniques are employed in the creation of numerous methods for predicting the chemical, biological, and physiological properties of molecules[12–16]. Using typical lifespan research and machine learning approaches centered on bottlenecks, the presented approach presents an innovative technique to determining the seriousness of cardiac illnesses. In the modern day, machine learning, is crucial to the area of medicine, as well as during the managing of new medical procedures, patient information, and clinical history[17]. Machine learning approaches could be advantageous at various phases of the drug creation process. For instance, machine learning strategies has widely utilised to seek out additional medication uses, detect drug relationships, identify drug performance, assure security monitoring, and optimise molecular biocompatibility. The machine learning techniques Random Forest, Naive Bayesian, and Support Vector Machine were often used in pharmaceutical research[18–20].

The methods and approaches used in machine learning are not just a single, consistent part of AI. Both supervised and unsupervised learning fall within the two main categories of machine generation techniques. Supervised learning uses current labelling of trained examples to predict the tags of fresh patterns. Unsupervised learning finds similarities in a group of items that are typically unlabelled. Before to recognising similarities in increased input, this information is constantly translated together into lower resolution via unsupervised learning strategies. Scale decrease is advantageous that's not because unsupervised learning seems to be better productive in a low-dimensional field, however mainly when an identified trend is easier to analyse. During last 20 years, the general consumption of rising testing and decoding, updated “repositories, and machine learning approaches has generated a thriving atmosphere for various areas such as evidence production, gathering, and support need for drug discovery. Analytics improvements has been beneficial in attempting to explain and comprehend the created facts. This effort, which is now routinely employed for all stages of drug discovery, is aided by machine learning strategies and interlinked database via different apps. The potential of newer data analysis to combine with traditional methodologies and earlier assumptions to generate innovative assumptions and theories has shown effective in localization, goal discoveries, bimolecular invention, formulation, and other fields [21–23].

Scientific and inter evidence is multifunctional. In terms of style and sources, the input is frequently fragmented and varied. The challenges of investigation and comprehension of high dimensionality might be unconstrained by applying ML approaches, such as generalised sequential designs with NB. Machine learning algorithms and ideas including as regression, clustering, regularisation, neural networks, decision trees, dimensionality reduction, ensemble techniques, optimization techniques, and parameterization procedures are frequently used in a variety of research domains[24].

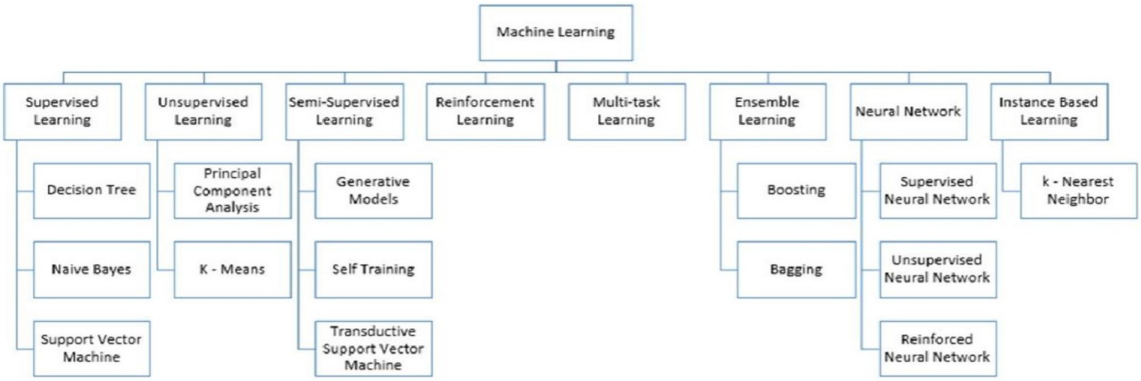

Fig. 1 Machine Learning Techniques

Lazy and Eager Learning

Eager learning occurs if a machines having to learn program constructs a prototype shortly once obtaining classification model. That's named eager as it main step it acts while it obtains the input source is develop the prototype. The professional development facts is finally forgotten. Whenever new raw material arrives, the prototype is used to analyse it. The vast majority of machine learning strategies remain eager to grow. Lazy learning, by the other hand, occurs whenever an algorithm doesn't quite develop a framework directly after obtaining data for training, but instead delays until it is given additional records to analyse. It's named lazy because it waits and when it's quite essential to produce a prototype, if it constructs some at all. This merely saves input points once it receives it. Once the raw data is received, it employs the previously recorded facts to analyse the output. So rather than developing a racially biased product out from tracking approach, the lazy learning process "memorises" the original sample. The eager learning process, but in the other hand, discovers its prototype elements (parameters) during training.

The advantages and disadvantages of enthusiastic and lethargic learning are distinct. Lazy learning, on the other hand, takes minimal duration throughout preparation but much longer when forecast.

Users look again for closest neighbours with in serious fitness group every occasion users really like to build a predictive model. So at start, eager learning creates a framework for the entire input set, i.e., it makes assumptions the original sample. In comparison to lazy learning, which seems to have further alternatives in due to the allocation of the entire records set as well as the structures to produce something of it, it may undergo from repeatability issues. Here just a few instances:

Nearest Neighbour, Case-Based Reasoning, K – Lazy Decision Tree, SVM, and logistic regression are some of the terms used by Eager.

Ensembling

There had apparently numerous techniques of create lazy learning. Ensemble classifications are among of the greatest prominent applications of lazy learning. Ensemble procedures were also unsupervised learning who build a collection of classifications and later use a majority of its estimates to identify fresh data items. Integrating the outcomes of numerous models minimises the probability of choosing a poor performing classifier when a collection of detectors having equivalent instructional performances have varied generalisation outcomes. A classifying ensemble was already demonstrated to commonly outperform an option to keep.

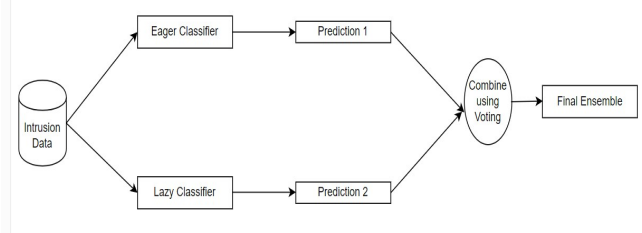

Fig. 2 Ensemble Model

In Machine Learning, the ensemble approach is stated as a heterogeneous framework in when many classifiers and methodologies are deliberately integrated to create a prediction model. Moreover, the ensemble method aids in reducing dispersion in expected data, minimising bias in the predictive model, and accurately classifying and predicting statistics from complicated issues.

Types of Ensemble Methods in Machine Learning

Ensemble procedures aid inside this creation of many concepts, which are eventually combined to give better solutions. Ensemble technologies are divided into the various categories

Sequential Methods have any additional piece of input in the base learner is dependent on preceding input in another way. As a result, the previously mismarked input is tweaked depending on its relevance to optimize the entire structure's effectiveness. Boosting is a case. In Parallel Method the basic trainee is built in concurrent manner, so that is no dataset connection. Each piece of input in the core trainer is developed on its own. Case: Stacking Homogeneous Ensemble approach is a collection of learners of that comparable kind. However, every classifier's sample is unique. Following the consolidation of outcomes of every approach, the composite design will perform better accurately. This kind of ensemble approach is capable of handling a huge amount of samples. The uniform technique uses the equivalent attribute choosing strategy for all data sets. It takes a long time to compute. Prominent approaches such as bagging and boosting, for example, are included in this uniform ensemble[25]. The danger and intensity of the participant's conditions are taken into account while designing a multilayer decision-making system for medical care[26]. For the best results, it is necessary to handle the data on a various layer[27].

The collection of cells that develop improperly or spread into a tumour[28]. With the feature selection procedure, the meta-heuristic dragonfly optimization method was used to improve the chosen features. Support vector machines were then used to classify the results further[29]. Heterogeneous ensemble technique is a collection of various kind of classifier or machine learning algorithms, every of these is based on the similar input. For tiny samples, this strategy functions well. For the similar trained input, the attribute extraction procedure in diversity is distinct. The ensemble technique's total outcome is calculated by combining the findings of each integrated models. Stacking is an example.

Feature Selection Ranking

In reality, a dataset can have thousands of different features. However, not all of the characteristics are required to perform the extraction process. Feature selection algorithms are used to determine the relevance of attributes. In the extraction procedure, only important features are processed, rather than everything that data. This will shorten the operating duration and improve the extraction task's productivity. As a result, before performing data mining tasks like classification, clustering, and outlier analysis, attribute selection techniques are used.
Attribute choosing is a two-part procedure, with the first phase being collection production and the second being rating. Subset production is a method that compares the candidate subgroup that has previously been identified. The current candidate subgroup is considered the optimum if it produces greater outcomes in terms of a specific examination. This procedure is repeated until the closure criteria is met. The second method is feature rating that can be used to determine the relevance of features. Numerous rating methods exist, the majority of them are dependent on statistical or information theory.

Table 1: Feature Selection Ranking

Sr.No	Attributes	Correlation Attribute	Principal Components
1	TPSA(Tot)	4	0.5355
2	Saacc	5	0.3906
3	H-050	1	0.038
4	MLOGP	8	0.0243
5	RDCHI	7	-0.0702
6	GATS1p	2	-0.098
7	nN	3	-0.1891
8	C-040	6	-0.3118

As we can observe in above table, we have shown many attributes with their ranking according to their attribute selector.

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.

- **Li Zhang et al. 2018** problem identified in this paper is conventional toxicity testing procedures are moment, labor-intensive, and costly. The research suggests using machine learning techniques to build a structure for computation forecasts as a response. For the optimal forecasting, additional interpretative molecule features and an appropriate feature selecting method can be applied.
- **Anna O. Basile et al.** 2019 Toxicity Issues and Security Concerns is discussed in this paper. Examining recent advancements in initial drug security with a focus on machine learning as well as deep learning techniques. AI is predicted to be used in drug development as well as security in the next.
- **Borrero et al.** This research analyses ignorance regarding absorption, distribution, metabolism, excretion, and toxicity.. Examine the most effective machine learning algorithms for forecasting poison like an ADME-Tox characteristic. For the future scope reliability of predictions can be enhanced.
- **Nishtha Hooda et** al. 2017 problem identified in this paper is to enhance and generate novel treatments, an effective model for predicting drug toxicity must be developed. To construct a stronger approach, an ensemble architecture for categorization of toxicity compounds are used by using unbalanced and high-dimensional complicated drug data. To improve the B2FSE Framework by building it on upper edge of new big data approaches such as Hadoop, Spark, and others.
- **Agnieszka et al.** 2021, In this paper forecast of aquatic toxicity, dataset heterogeneity importantance is discussed. Two similarity-based machine learning approaches are used to forecast the acute aquatic toxicity of a variety of chemical compounds. For future scope different ways could be used to create models having great total efficiency.
- **Sharma et al.** 2019 in this paper use of mathematical expertise for investigate the activities of bioactive chemicals and medications in high-dimensional, unbalanced bio assays are analysed. Scholar created a system for assessing the functionality of biological chemicals and medications that is both quick and reliable. For future scope enhance by adding machine learning models to big data approaches such as Hadoop and others, the ML methods will be improved.
- **Hooda et al.** 2019 in this paper inappropriate business tactics without concern of judicial repercussions, e.g. scam investigation business is discussed. While beginning inspection field work for major companies, scholar designed a judgement methodology for inspectors. It is advised to combine different groups for further scope.
- **Collado et al.** 2020 A difficulty with class balancing is analysed in this paper. To forecast chemical toxicity, an appropriate Feature Selection approach is presented for class imbalance samples. It can also be applicable to different challenges in cheminformatics by extending this with various FS approaches.
- **Asnat et al.** 2020 in this paper marketing management of merchants with regular prices is a difficult challenge to solve is analysed. Scholar developed a new method for forecasting the influence of item cost flexibility on e-commerce merchants. In addition to estimate past pricing flexibility influence for additional items, scholar shared a gather competition data and optimise it.
- **Austin** 2020 in this paper the impact of lacking individuals on reliability forecasting is analysed. The impact of lacking individuals on the forecast efficiency of a Stacking-based ensemble as well as a Voting-based ensemble is investigated. For the future scope it is necessary to investigate the impact of lacking individuals on different types of classifier ensembles, like Grading, Boosting, and Bagging.
- **Han et al.** 2020 in this paper to anticipate the behaviours of a complicated process, a solitary factor grey prediction model is used. During long term power production forecast, an unique multimodal grey prediction model depending on first order linear differentiation equation is suggested. For future it will be necessary to research first order non-homogeneous differential equations with variable coefficients in the ahead.
- **Kurz et al.** 2020 the majority of stacking algorithms are dependent on sequential models, which can cause issues when forecasts are highly connected. Scholar designed a greedy stacking technique for the model that solves this problem while remaining fast and simple to understand. When using a broad range of facts and circumstances, the findings can vary when applied to various individuals and whenever alternative assessment variables are employed.
- **Tuladhar** 2020 In this paper the lack of patent data hampered the Machine learning model is discussed. By combining ANN, SVM, and RV, we are able to create and test simulated distributed learning models. Expand your research to include multiclass classification, regression problems, and image segmentation.
- **Filho et al.** 2019 In this paper uneven distribution of wealth is discussed. In this paper, scholar introduced imbalance learning strategies for boosting statistic model accuracy in algorithmic grading. Other statistical models, like neural networks as well as various ensembling techniques, are being investigated.
- **Son et al.** 2021 Uneven distribution of wealth is anlysed in this paper. Suggested an oversampling method centered on a probabilistic generated artificial network and a borderline class (CGAN). To locate extreme minority class, optimise K values for every piece of data in the procedure.

III. PURPOSED METHODOLOGY

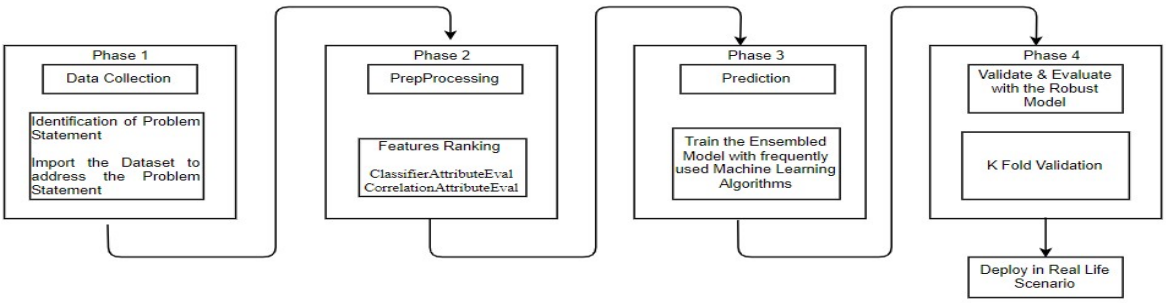

Fig. 3 Proposed Methodology

This part concentrates on developing a framework that predicts pharmaceutical toxicity using machine learning techniques. The complete process consists of four essential parts. Selecting the data we must work with is the first step, which is followed by determining the main problem. We do features assessment on the dataset using the Classifier Attribute Eval and Correlation Attribute Eval in regard to determining the qualities that stand out most in our data set in order to highlight the qualities which were very important in the next stage.

In the following step, we use an ensemble framework that has been trained applying well machine learning methods to produce recommendations. In mixing numerous models rather than relying just on one, ensemble approaches seek to increase the reliability of findings in systems. The precision of the results is significantly increased by the combined models. As a result, ensemble techniques in machine learning have become more popular.

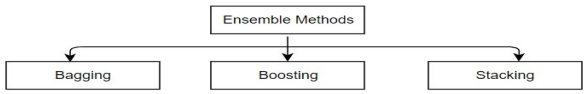

Fig. 4 Categories of Ensemble Methods

Ensemble techniques fall into two primary groups: sequential ensemble methods and parallel ensemble approaches. Sequential ensemble techniques, such adaptive boosting, are used to create basic learners. Similar to random forest, parallel ensemble techniques construct base learners in a parallel manner. Parallel approaches use parallel generations of base learners to encourage independence among them. The freedom of base learners significantly reduces the averages-related error.

Algorithm 1. Different types of Voting algorithm

- Hand voting classifier voting.

```
Vothard=Voting_Classifier(calculators=calculate, votingtype='soft')
Votsoft_fit(A_train, B_train)
B_pred=votsoft_predict(A_test)
```

- Soft voting classifier voting

```
Votsoft=voting_Classifier(calculators=calculate, votingtype='soft')
Votsoft_fit(A_train, B_train)
B_pred=votsoft_predict(A_test)
```

Algorithm 2. Prediction algorithm

- READ Dataset => {D}
- {D}train => {D} [0:x]
- {D}test=>{D}[x+1:n]
- SET P1,P2

- Define Prediction({D}, type={EAGER}{LAZY}) as Predictions

Return Classifier. {type}-predict({D};

- Define VotingClassifier(calculators = calculate, type = {hard}{soft},{P1},{P2})

VotingClassifier,fit({P1},{P2})
return VotingClassifier.{type}.predict({D});

- Ensemble=VotingClassifier(calculate,type,Predictions);
- End procedure;

In the next section, we compare the efficiency of the recommended Robust Model to that of the conventional, k-fold cross-validation-reliant Machine Learning approach. The efficiency of every prototype is assessed utilizing randomly chosen samples varying from 0 to 10 folds in order to understand how reliable our technique is. We arrived at our estimations regarding the worst case, best case, and average viable situation while running the 10 folds. The outcome is enhanced and rendered more reliable whenever the average is established across all conditions. A model's robustness can be used to gauge its efficiency when it is evaluated on a brand-new, separate database which is comparable to the initial one.

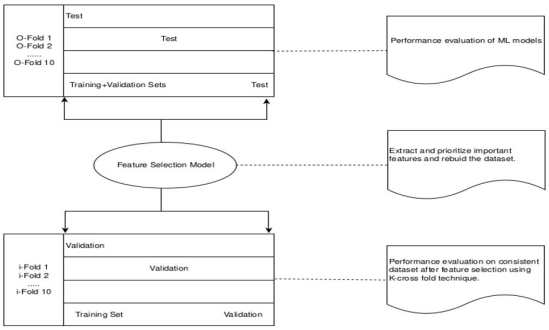

Fig. 5 K-Fold Cross Validation

The database is split into k-folds, however each folding also has two additional sections: once for understanding or building a technique and another for assessing it. A quantitative method for evaluating and contrasting learning methods is cross-validation. Every fold is divided into two segments to do this. The training and testing collections should be crossed several occasions in a standard cross-validation technique. This makes ensuring that every bit of data has the chance to be evaluated. The model's mean ability rating is frequently used as an informative metric while examining the findings of a k-fold cross-validation run. It is suggested to evaluate the skill scores' cyclical nature using a metric like the standard variation or standard variance.

IV. Result and Discussion

This part focuses on a comparison of our suggested approach with commonly utilised machine learning models.

Table 2: Correlation coefficient

Regeration Model	**Correlation coefficient**
Linear regression	0.6865
Multi-layer perceptron	0.5516
SMOreg	0.6757
Kstar	0.6781
Bagging	0.6974
Decision stump	0.375
Random Forest	0.7439
Optemised Ensembled	0.7331

A mathematical indicator of the degree of the association among the comparative changes of two elements is the correlated coefficient. We can see from the given table that the correlation coefficient values for the various classifiers vary.

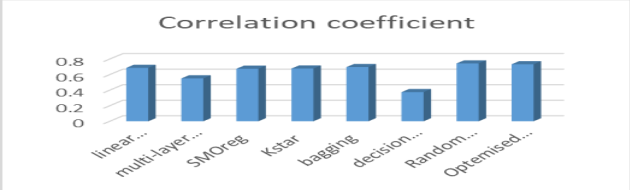

Fig 6: Correlation coefficient

In this figure we can compare different Correlation coefficient values.

Table 3: Mean absolute error

Regeration Model	**Mean absolute error**
linear regression	9.19
multi-layer perceptron	11.53
SMOreg	9.19
Kstar	8.60
bagging	8.93
decision stump	11.83
Random Forest	8.11
Optemised Ensembled	7.96

The **Mean Absolute Error** (MAE) The average of all absolute mistakes is known as the Mean Absolute Error. Despite taking into account the source of the mistakes, the Mean Absolute Error calculates the average size of errors in a set of forecasts. It's the measurement standard that has an impact on correctness.

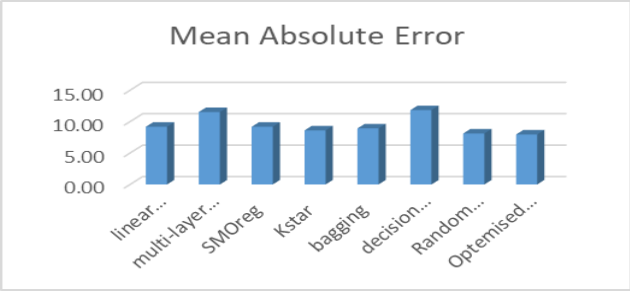

Fig 7 : Mean Absolute Error

In the above figure we can evaluate different Mean Absolate Error, it directly effect accuricy of model.

Table 4: Root mean squared error

Regeration Model	**Root Mean Squared Error**
linear regression	1.2099
Multi-layer perceptron	1.4513
SMOreg	1.2369
Kstar	1.2635
Bagging	1.1933
Decision Stump	1.5479
Random Forest	1.1124
Optemised Ensembled	1.1333

Root mean squared error (RMSE) The square root of the mean of the square of every one of the errors is known as the root mean squared error. RMSE is frequently employed and is regarded as a superior all-purpose error metric for numerical forecasts. As we'll see, each classifier has a different RMSE value.

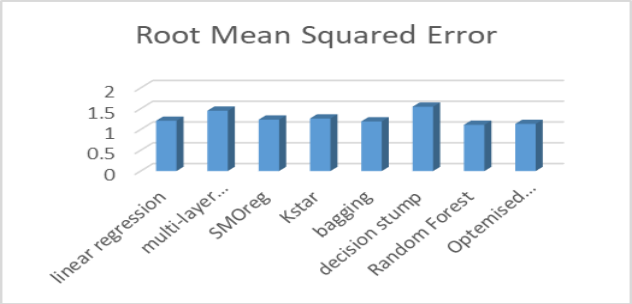

Fig 8 : Root mean squared error

As we can observe in the above figure different classifiers have different RMSE values.

Table 5: Accuracy

Regeration Model	**Accuracy**
linear regression	90.81
multi-layer perceptron	88.47
SMOreg	90.81
Kstar	90.40
bagging	91.07
decision stump	88.17
Random Forest	91.05
Optemised Ensembled	92.04

Mean Absolute Error, the easiest way to gauge prediction reliability. Its mean of the absolute errors, or MAE, is just what its name implies. The distinction between both the the predicted value and the true value and the actual value, expressed as an utter and total number, is the absolute error. As seen in the accompanying table, our accuracy has grown to 92.04.

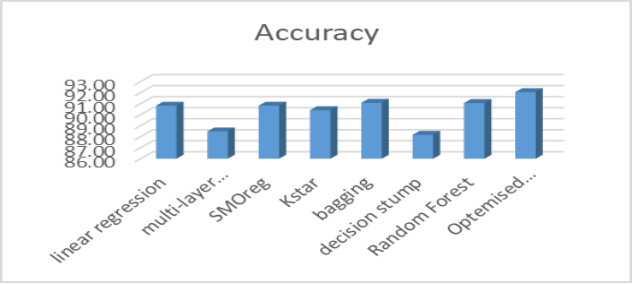

Figure 9 Accuricy

V. Research Methodology

Building a solid simulation to forecast the model's reliability is the main goal of this part. There are four key steps to the entire operation. Collecting data and applying any necessary pre-processing to that dataset constitute the first phase. We execute characteristic choice just on information in the acquired dataset's second phase in attempt to choose the traits that are particularly noticeable in our given dataset. The third part of the project involves testing and training the system using the standard machine learning technique. In the fourth phase, we contrast the suggested Robust Model's performance against those of the traditional Machine Learning technique, namely is predicated on k-fold cross-validation.

Strengthen the Model

In this stage we are giving strength in our model by WSAW score. As we can observe WSAW is high and balanced error is comparatively less. Here we are considering more than one parameters and calculated the WSAW Score that is compared with Balanced Error.

Table 5:

Regeration Model	WSAW Score	Balanced Error
Linear regression	45.75	4.95
Multi-layer perceptron	44.51	6.21
SMOreg	45.74	4.95
Kstar	45.54	4.63
Bagging	45.88	4.81
Decision stump	44.27	6.37
Random Forest	45.90	4.37
Optemised Ensembled	46.39	4.29

In the above table we can observe that we have given strength by WSAW score and Kstar vs RandomForest have highest WSAW score and have lowest balanced error.

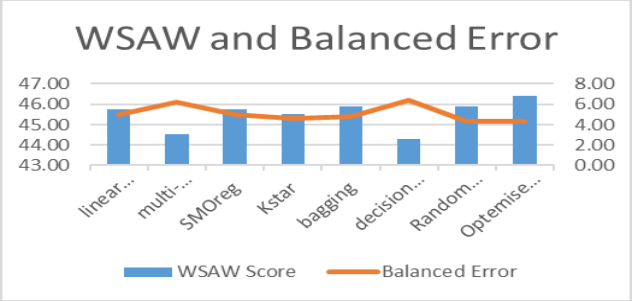

Fig 10 WSAW Score and Balanced Error

In this figure we used two parameters WSAW score and balanced error that can be easily observed.

VI. Conclusion and Future Scope

A simpler and more performant machine learning model can be developed by removing input features from the training dataset In this paper performance of frequently used machine learning model is analyzed and we found that there is scope of improvement also. An optimized ensemble machine learning model is proposed having the higest accuracy 92.04.to strengthen the proposed model. We compared feature selection strategies that decreased cost and increased efficiency.

Additional areas, such as the WSAW Score, an experimental investigation approach that aids in strengthening our model, can also be included in the Concentration of Performance evaluation.

I. References

[1] Morgan, S.; Grootendorst, P., Lexchin, J., Cunningham, C., Greyson, D., *The cost of drug development: A systematic review.* Health Policy 2011, 100, 4–17. [CrossRef] [PubMed].

[2] Ng, H.W., Zhang, W., Shu, M.; Luo, H., Ge, W., Perkins, R., Tong, W., Hong, H. *Competitive molecular docking approach for predicting estrogen recaptor subtype α agonists and antagonists.* BMC Bioinform. 2014, 15, S4. [CrossRef] [PubMed].

[3] 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].

[4] 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].

[5] 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

[6] 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]

[7] 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.

[8] 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

[9] 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.

[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] 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.

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

[14] 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]

[15] 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.

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

[17] 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.

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

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

[21] 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]

[22] 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]

[23] 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*

[24] R Kamalraj, S Neelakandan, M Ranjith Kumar, V Chandra Shekhar Rao, Rohit Anand, Harinder Singh, *Interpretable filter based convolutional neural network (IF-CNN) for glucose prediction and classification using PD-SS algorithm,* Measurement, Vol 183, Pages 109804, Publisher Elsevier, 2021/10/1.

[25] Dinesh Kumar, Anuj Kumar Sharma, Rohit Bajaj, Lokesh Pawar, 2021. Feature Optimized Machine Learning Framework for Unbalanced Bioassays. *Cognitive Behavior and Human Computer Interaction Based on Machine Learning Algorithm, pp. 167-178,* John Wiley & Sons Inc.

[26] Deepika Sharma, Gagangeet Singh Aujli, Rohit Bajaj, *Deep neuro-fuzzy approaches for risk and severity prediction using recommendation systems in connected health care* , Transactions on Emerging Telecommunications Technologies, Vol Vol2,, Issue 7,pp e4159, 2021.

[27] Lokesh Pawar, Rohit Bajaj, Jaspreet Singh, Vipinpal Yadav, *Smart city IoT: Smart architectural solution for networking, congestion and heterogeneity* International Conference on Intelligent Computing and Control System(CCS), pp 124-129, 2019.

[28] Lokesh Pawar, Pranshul Agrwal, Gurjot Kaur, Rohit Bajaj, *Elevate Primary Tumor Detection Using Machine Learning*, Cognitive Behaviour and Human Computer Interaction Based on Machine Learning Algorithm, pp 301-313, 2021.

[29] Pankaj Rahi, Sanjay P Sood, Rohit Bajaj, Yougesh Kumar, *Air Quality monitoring for Smart eHealth system usig firefly optimization and support vector machine*, International journal of Information Technology, Vol 13, Issue 5, pp 1847-1859, 2021.