International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

Multivariate Classification of Drugs using Parametric and Nonparametric Machine Learning Models

N. Priya, G. Shobana

Abstract*: *In pharmaceutical research, traditional drug discovery process is time consuming and expensive, where several compounds are experimentally tested for their biological activities. Series of lab experiments are conducted to analyze newly synthesized drug's pharmaceutical activities and its biological effects on human. With every new drug discovery, the required clinical properties can be determined using machine learning models and this greatly reduces the experimental cost. This paper explores parametric and non-parametric machine learning models to classify administration properties of drugs and its toxicity. The multinomial classification of drugs was based on their physicochemical and ADMET properties. Balanced data samples were drawn from chEMBL and was pre-processed. Features were reduced using Recursive Feature Elimination and the attributes were ranked based on their importance to reduce highly correlated attributes. The performance of parametric and non-parametric machine learning models was analyzed on cheminformatic data that includes physiochemical, biological and pharmaceutical properties of the drug molecules. Selecting the potent drug candidate along with its administration properties greatly reduces wet lab experimental time and cost. Multiclass classification can be determined efficiently using non-parametric machine learning model. Optimal feature engineering, tuning hyperparameters and adopting hybrid algorithms would result in more accurate predictions in future for cheminformatics data.

Keywords*: *Parametric, machine learning, drug discovery, cheminformatics.

I. INTRODUCTION

In preclinical development, the toxicity level of drug plays a very crucial part. If the toxicity of the drug is negligible, then further research for its bioactivity is performed. Several research papers have been published based on toxicity prediction using machine learning models. Very huge data set of drugs are taken with variant features and the prediction (target) is generally bivariate (class zero-Non-toxic, class one- Toxic). Apart from toxicity, we consider other properties of a drug like oral, parenteral, topical, OP (oral and parenteral) and PT (parenteral and topical). Overfitting problems are frequently observed when smaller data samples are used. This issue can be prevented by using a huge volume of diverse chemical drug compounds [1].

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
Dr. N. Priya, Associate Professor, PG Department of Computer Science, SDNB Vaishnav College for Women (Affiliated to University of Madras), Chennai, India. Email: drnpriya2015@gmail.com
G. Shobana*, Assistant Professor, Department of Computer Applications, Madras Christian College (Affiliated to University of Madras), Chennai, India. Email: gmshobana@gmail.com

The data set is acquired from chEMBL open repository. It a database that stores information of drug-like compounds. Calculated properties about several drugs are available in this database. The features taken for study includes several physicochemical and drug-like properties of the drug. The dataset has 300 balanced and comprehensive samples. 21 features have been considered for the study. The features were ranked using RFE (Recursive Feature Elimination) and feature importance was determined. We employ both parametric and non-parametric machine learning models for prediction. Some of the parametric machine learning models are Logistic regression, LDA (Linear Discriminant Analysis), perceptron, simple neural network, Naïve Bayes etc. Examples for non-parametric models are k-nearest neighbor, Decision Tree - CART (Classification and Regression Trees), SVM (Support Vector Machine) etc. The confusion matrix is generated for the both the parametric and non-parametric machine learning models. Model evaluation metrices like Precision, Recall and F1-Score are computed and analyzed.

II. RELATED WORK

Lei et al have evaluated ADMET properties in drug discovery. Using relevant vector machine and consensus modeling, they have predicted oral acute toxicity in rat. In their study, they obtained 7314 diverse chemical data set with rat oral LD_{50} values. Various machine learning algorithms were applied to the data set. When high dimensional data are trained using machine learning models, overtraining and overfitting are likely to occur. Therefore, with increase in the dimension of data, the complexity of the model, increases and this requires tuning several hyper-parameters. In order to reduce computational complexity, (RVM) Relevance vector machine method may be used. RVM employs Bayesian criteria during the learning phase, reducing irrelevant features and generates an appropriate sparse model. RVM model gave better results among other models used. Among 334 descriptors (features) that characterize physicochemical and drug-like properties, 230 descriptors were chosen for QSAR (Quantitative Structure-Activity Relationship) modeling. Dimension reduction was performed using Chi-squared statistics. The regression models were evaluated and validated using adjusted R^2 and ten-fold cross validation R^2 co-efficient. Lei et al concluded that RVM and RF (Random Forest) were best learning models [2]. Mathew.T. E proposed a logistic regression model with recursive feature elimination model for breast cancer diagnosis [3]. Christelle Reynes et al used decision tree model to analyse protein-protein interaction inhibitors. Optimized decision tree was obtained after cross validation [4].

Yang Li et al classified HIV-1 Protease Inhibitors using Decision Tree. They applied DT model for 3 different descriptor data set and achieved training set accuracy as 86.00, 89.52 and 80.89 [5]. Huge data are generated from high throughput screening (HTS) and machine learning models are applied to analyse and build effective cheminformatics models for the screening of various diseases [6]. Maibam Mangalleibi Chanu et al have used SVM classifier for the classification of Brain tumour. MRI (Magnetic Resonance Imaging) images were employed for the classification purpose and the classifier achieve 97.12% accuracy [7]. Setu Basak et al used Machine Learning models to predict kidney disease. The dataset was obtained from the UCI repository and 24 features were employed for the investigation. Five different machine learning techniques like Naive Bayes, Decision stump, Random Forest, J48 and IBK were applied to the dataset. IBK achieved 98.25% accuracy, when compared to other learning models [8]. Heena Nankani et al used various algorithms like Linear Regression, K-Nearest Neighbor, Logistic Regression and Support vector Machine for detection of cancer types. Logistic Regression learning model was found to have better prediction accuracy [9].

III. DATA SETS

A chEMBL is a chemical database that stores data about molecules with drug-like properties. It is an open repository and several aspects for drug discovery can be accessed. The European Bioinformatics Institute (EBI) maintains this database. This Institute belongs to the European Molecular Biology Laboratory (EMBL) based at the Wellcome Trust Genome Campus, Hinxton, UK [10]. chEMBL holds information on more than 1.4 million compounds. The database includes information about ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) properties, structures, bioactivity, calculated properties of drug-like small molecules. ADMET is one of the different methods and techniques used for Lead identification process [11]. It holds Structure-activity relationship (SAR) data that has been manually extracted and curated and the data was obtained from medicinal chemistry and pharmacological domains [12]. In this paper, calculated properties of drugs have been investigated. Fifty samples for each type of drug (oral, parenteral, topical, toxic, oral and parenteral (OP) and parenteral and topical (PT) has been added to the dataset. All the fifty samples of each drug type (e.g., oral) are 85% similar. The data set is divided in to training and testing data.70% of the data is taken for training and 30% of the data is taken for testing the models. The data is shuffled and split in to training and testing samples. The data set with 300 samples is split in to 210 for training and 90 for testing. The predicted results of the parametric and non-parametric machine learning models are investigated.

IV. FEATURE DESCRIPTION

Twenty-one features were extracted from the chEMBL database. These were the calculated properties of the drug. They define both physicochemical and drug-like property of the compound. The properties are Mol.Weight, MW. Monoisotopic, ALogP, Number of rotatable bonds, Polar Surface Area, Molecular species, HBA, HDB, Ro5 violations, HBA(Lipinski), HBD(Lipinski), Ro5 violations (Lipinski), ACD Acidic pKa, ACD Logp, ACD LogD pH7, Aromatic Rings, Heavy atoms and QED weighted. The 21 features are acd_logd, acd_logp, mw_freebase, qed_weighted, acd_most_bpka, hba_lipinski, alogp, molecular_species, hbd_lipinski, hba, full_mwt, acd_most_apka, aromatic_rings, rtb, heavy_atoms, num_lipinski_ro5_violations, hbd, psa, ro3_pass, num_ro5_violations, and mw_monoisotopic. qed_weighted and aromatic_rings were found to be the most important features. Fig. 1 and Fig.2 shows the violin plot for the most important features qed_weighted and aromatic_rings respectively. Features are ranked and the value of each attribute is computed using RFE (Recursive Feature Elimination). Some data which had string values were converted to numerical data. The feature molecular_species had data like Acid, Base, Neutral and Zwitterion. The categorical data were converted to numerical values 1, 2, 3 and 4. Data, that had the string of yes or no was converted to 1 and 0 respectively. The Type attribute had six classes 1, 2, 3, 4, 5 and 6.

The features were ranked using RFE (Recursive Feature Elimination) as:

[6 5 17 1 12 11 4 8 10 13 20 16 2 14 15 3 7 19 21 9 18]

The feature importance generated using logistic regression classifier is as follows: [0.029, 0.016, 0.061, 0.074, 0.077, 0.042, 0.019, 0.070, 0.036, 0.038, 0.054, 0.091, 0.048, 0.047, 0.099, 0.027, 0.059, 0.052, 0, 0.010, 0.041]

A violin plot resembles box plot but gives better information. The density of the data can be analyzed using this plot. When there is multiclass prediction, violin plot gives clear understanding about the data density of each class. Violin plots can be generated using R and Python languages. Python libraries like matplotlib, plotly and seaborn are used to generate a violin plot. The distribution of a specific attribute can be compared against various classes using this plot [13]. The attributes, qed_weighted and aromatic_rings are the two features that are ranked as important with logistic regression model. In Fig.1 and Fig. 2 the distribution of these attributes against various classes are compared.

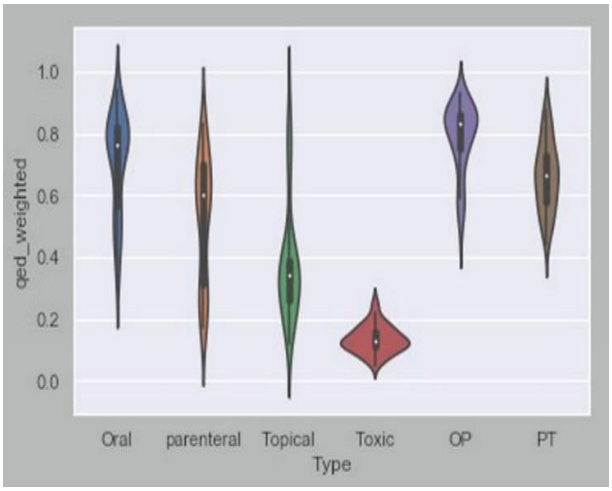

Fig. 1. Violin plot for the feature qed_weighted

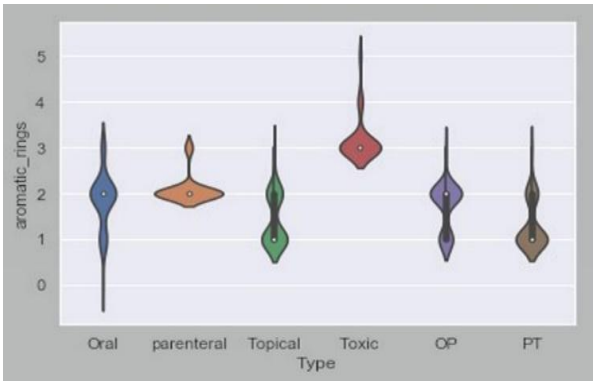

Fig. 2. Violin plot for the feature aromatic_rings

There are several chemical databases, where the required properties of the drugs can be retrieved. Cheminfo is a website that provides information on the structure, property, chemical activity, bioactivity etc. Similarity search and structure search is also available. Molinspiration offers a range of several tools to calculate the molecular properties of the drug molecules. Molinspiration tools are developed using Java. It provides free on-line services for calculating several important molecular properties of the drug candidate.

V. METHODOLOGY

In the proposed methodology, the dataset is obtained from chEMBL as shown in Fig. 3.

Procedure:

Step 1: Obtain 300 samples (50 samples for each class) from chEMBL with 21 features.
Step 2: Perform pre-processing and convert categorical data to numerical data.
Step 3: Investigate for highly correlated attributes and missing data.
Step 4: Apply RFE on the dataset.
Step 5: Reduce features to 19.
Step 6: Apply parametric and non-parametric ML Classifiers.
Step 7: Compare the prediction accuracy.
Step 8: Select the efficient ML Classifier

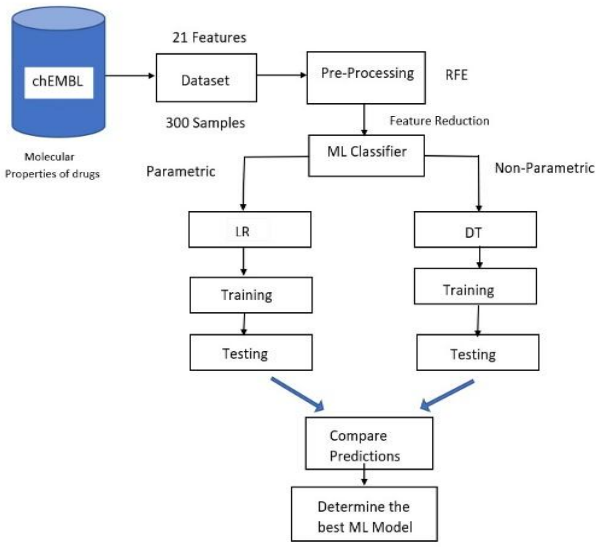

Figure 3. The Proposed Approach for Prediction

Twenty-one features that describe the drugs molecular properties were drawn for 300 samples from chEMBL. Multivariate classification includes six classes. In the pre-processing stage, the data was investigated for the presence of categorical data and was converted to numerical data. Feature selection technique was applied and the features were ranked for its importance. The most redundant or collinear attributes were reduced using the Recursive Feature Elimination (RFE). The features were reduced from 21 to 19. Next to the pre-processing stage is the application of the Machine Learning Classifier (ML Classifier) to the dataset. Parametric and Non-parametric Machine Learning Model were applied to the dataset.

A. *Parametric Machine Learning Model*

Multiclass classification using Logistic regression is also known as multinomial logistic regression or softmax regression. In binary classification, the model is computed with one weight vector w and the target is 1 [14].

Let function input s be the weighted sum of the features. Let a be a data sample that has n features $a_1, a_2, a_3,\ldots, a_n$. Let a be a feature vector, where $a = (a_1, a_2, \ldots, a_n)$. Let w be a vector that represents weights or coefficients

$S = w_1a_1+w_2a_2+\ldots\ldots+w_na_n = w^ta$ (1)

Whenever, there is a bias or an intercept w_0 in the model, the linear relationship becomes Equation (2).

$S=w_0+w_1a_1+w_2a_2+\ldots\ldots\ldots+w_na_n = w^ta$ (2)

The output is the probability of the target being 1 or positive and the logistic classifier or a probabilistic classifier is given as Equation (3).

$\hat{b} = P\left(b = \frac{1}{a}\right) = \frac{1}{1+\exp(-w^ta)}$ (3) The

cost function is given by Equation (4).

$c(w) = \frac{1}{k}\sum_{j=1}^{k} -\left[b^{(j)} \log\left(\hat{b}\left(a^{(j)}\right)\right) + \left(1-b^{(j)}\right) \log\left(1 - \right.\right.$ baj (4)

The step Δw_i for the i weight vectors is given by Equation (5).

$$\Delta w_i = \frac{1}{k}\sum_{j=1}^{k}\left(-1\{b^{(j)} = i\} + \hat{b}_m\left(a^{(j)}\right)\right)a^{(j)} \quad (5)$$

Hence in multiclass classification, when there are m classes, the model is represented by m weight vectors $w_1, w_2,\ldots..w_m$. The probability of the m class which is the target is represented as Equation (6).

$$\hat{b}_m = P(b = m / a) = \frac{exp(w_m^t a)}{\sum_{i=1}^{m} exp\,(w_i^t a)} \quad (6)$$

The probabilities $\hat{b}_m$ (1 to m) is normalized by $\sum_{i=1}^{m} exp\,(w_i^t a)$

The cost function is given as Equation (7).

$$c(w) = \frac{1}{k}\sum_{j=1}^{k} -\left[\sum_{i=1}^{m} 1\{b^{(j)} = i\} \log\left(\hat{b}_m\left(a^{(j)}\right)\right)\right] \quad (7)$$

Where the value of the function $1\{b^{(j)} = i\}$ is one, when $b^{(j)} = i$ is true.

The value is otherwise zero. In each iteration all the m vectors get updated and after completion of the iterations, the learned weight vectors w_1, w_2,…..,w_m are used for the classification of new samples a'.

$$b' = arg\ \underset{m}{max}\hat{b}_m = arg\ \underset{m}{max}P(b = m/a') \quad (8)$$

The relationship between the dependent variables is determined by regression machine learning models [14]. In 1958, David Cox introduced Logistic regression. Linear and polynomial regression are other two regression types [3]. Various evaluation metrices for the logistic regression model like precision, recall and F1-Score have been generated as shown in Table 1. Logistic regression model had a prediction accuracy of 0.867, as shown in Table II when the train and test partition was 70-30. Two of the least important collinear features were eliminated by RFE (Recursive Feature Elimination) and the logistic regression model was applied. The prediction accuracy obtained was 0.867. Further reduction in features, resulted in less prediction accuracy.

Table I. Precision, recall and f1-score of logistic regression

Type	Precision	Recall	F1-Score
Class-1	0.75	0.92	0.83
Class-2	0.88	1	0.93
Class-3	1	0.8	0.89
Class-4	1	1	1
Class-5	0.82	0.7	0.76
Class-6	0.8	0.86	0.83

Table II. Accuracy of the logistic regression model

Data Set	Accuracy
Training data set	0.929
Test data set	0.867

B. *Non-Parametric Machine Learning Model*

A decision-tree is a greedy algorithm. It has a root node with no incoming edges. Every leaf node in decision tree is assigned a class label. The root node and other internal nodes that are called non-terminal nodes contain test-conditions. CART-DT (Classification and regression trees) uses Gini index. The Gini index measures the impurity of D. D is a data partition or set of training tuples.

It is defined as

$$\text{Gini(D)} = 1 - \sum_{i=1}^{m} p_i^2$$

p_i is the probability that a tuple in D belongs to the class C_i. p_i is estimated by $|C_{i,D}| / |D|$. Over m classes, the sum is computed [15].

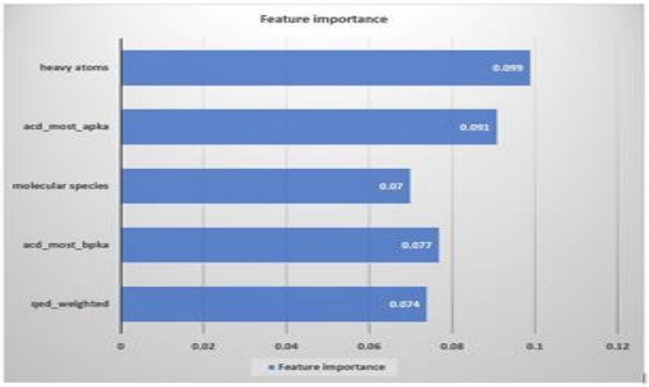

Fig. 4. Feature importance computed by decision tree classifier

Evaluation metrices for the decision tree model like precision, recall and F1-Score have been generated as in Table III. The Decision model had a prediction accuracy of 0.900, as shown in Table IV when the train and test partition was 70-30.

Table III. Precision, Recall and F1-Score for decision tree classifier

Type	Precision	Recall	F1-Score
Class-1	0.92	0.85	0.88
Class-2	0.88	1	0.93
Class-3	1	0.87	0.93
Class-4	1	1	1
Class-5	0.84	0.8	0.82
Class-6	0.81	0.93	0.87

Table IV. Accuracy of he decision tree model

Data Set	Accuracy
Training data set	1
Test data set	0.9

RMSE (Root Mean Squared Error) value of dataset with 21 features is 1.13853 and RMSE value of dataset with 19 features is 1.13351. RMSE is one of the analysis parameters that can be used in performance comparison between different learning models. Figure 4. Shows the important features generated by the decision tree classifier. acd_most_apka and heavy_atoms are considered to be the most important features, while decision tree learning model is applied to the dataset. Precision is the ability of the learning model, not to predict a negative result as positive.

$$\text{Precision} = T_p / (T_p + F_p)$$

True positives are indicated by T_p and False positives are indicated by F_p. Recall is the ability of the learning model, to predict positively.

$$\text{Recall} = T_p / (T_p + F_n)$$

Here, True positives are indicated by T_p and False negatives are indicated by F_n. The best and worst values are 1 and 0 respectively. F1-Score is also called as F-measure or F-Score. It is the weighted average of the precision and recall. The best and worst scores are 1 and 0 respectively. The contribution of precision and recall in computing F1-Score are relatively equal [16].

$$\text{F1-Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

Classification Rules were derived from the decision tree as in Fig. 5. The method that classifies records using a collection of IF-THEN rules is called a rule-based Classifier [17].

R1: IF full_mwt <= 684.79 AND
acd_most_apka <= 9.02 AND
hbd>1.5 AND
aromatic rings > 1.5 THEN
CLASS =1(Oral)
R2: IF full_mwt <= 684.79 AND
acd_most_apka > 9.02
AND

acd_most_bpka > 7.675 AND
rtb > 2.5 AND
hba > 3.5 AND
qed_weighted <=0.835 THEN
CLASS = 2 (Parenteral)

R3: IF full_mwt> 684.79 THEN
CLASS = 3 (Topical)

R4: IF full_mwt <= 309.935 AND
acd_most_apka <= 7.295 AND
psa <= 155.7 AND
heavyatoms > 16.5 THEN
CLASS = 4 (Toxic)

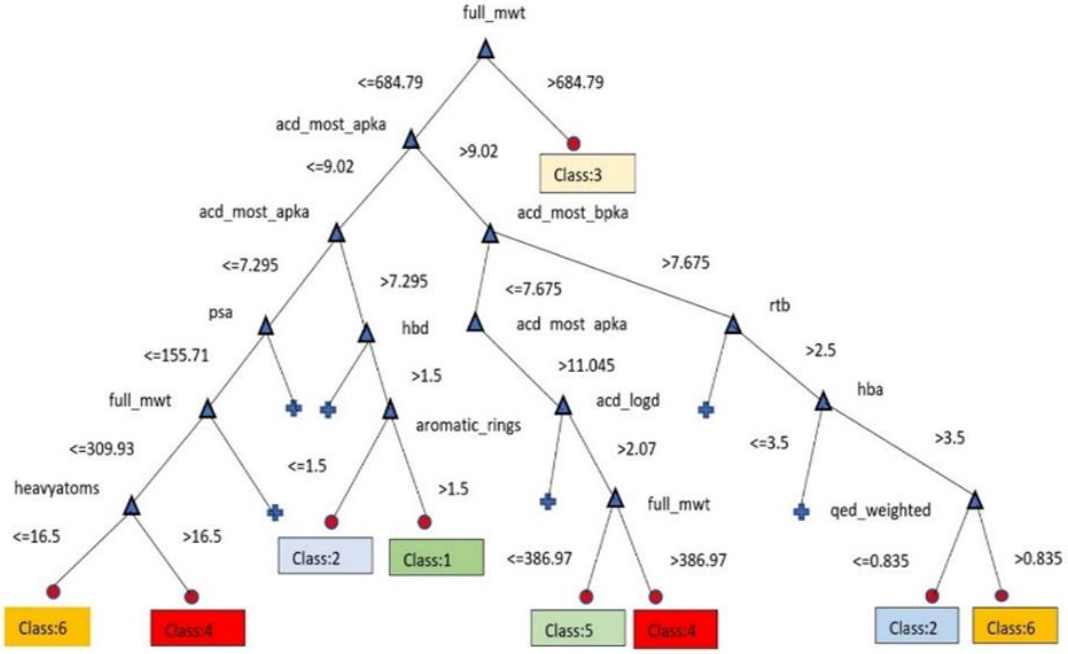

Fig. 5. Decision tree generated for the multiclass prediction (+ represents other branches of the tree)

R5: IF full_mwt <= 386.97 AND
acd_most_apka > 9.02 AND
acd_most_bpka <= 7.675 AND
acd_logd > 2.5 THEN
CLASS = 5 – OP (Oral and Parenteral)

R6: IF full_mwt <= 309.935 AND
acd_most_apka <= 7.295 AND
psa <= 155.71 AND
heavyatoms <= 16.5 THEN
CLASS = 6 – PT (Parenteral and Topical)

The confusion matrices generated by Parametric and Non-Parametric models:

Logistic Regression

	Oral	Parenteral	Topical	Toxic	OP	PT
Oral	12	0	0	0	1	0
Parenteral	0	14	0	0	0	0
Topical	1	1	12	0	0	1
Toxic	0	0	0	14	0	0
OP	3	1	0	0	14	2
PT	0	0	0	0	2	12

Decision Tree

	Oral	Parenteral	Topical	Toxic	OP	PT
Oral	11	1	0	0	1	0
Parenteral	0	14	0	0	0	0
Topical	0	1	13	0	1	0
Toxic	0	0	0	14	0	0
OP	1	0	0	0	16	3
PT	0	0	0	0	1	13

Each matrix shows the result predicted by the models. The diagonal elements of the matrices represent the accurate predictions of the parametric and non-parametric models. The matrix has 90 predicted values that is 30% of the dataset. Toxicity of the drugs has been predicted accurately in both the models. Several classification metrics are available in sklearn.metrics, which helps to learn about the performance of the machine learning models. Some of them are auc, average_precision_score,brier_score_loss,cohen_kappa_score, dcg_score, fbeta_score, hamming_loss, hinge_loss, Matthews_corrcoef, ndcg_score, precision_recall_curve, precision_recall_fscore_support, roc_auc_score, roc_curve, zero_one_loss etc available with the Anaconda package. MAE (Mean Absolute Error), MSE (Mean Squared Error), MSLE (Mean Squared Log Error) and MAE (Median Absolute Error) are some of the error computation functions that can be imported from sklearn.metrics for evaluation [16]. Current drug discovery process is slow and less efficient [18]. KEGG and protein Data Bank are large databases that have abundant cheminformatic information about drugs and the data from these repositories are integrated and classification algorithms are applied [19].

VI. RESULTS AND DISCUSSION

300 samples were obtained from the chEMBL repository. Each class had 50 samples. 21 features defined the molecular properties of the drugs. RFE (Recursive Feature Elimination) was used and the attributes were ranked. The features were reduced to 19 based on their feature importance.

The most important feature is heavy_atoms followed by acd_most_apka and acd_most_bpka as shown in Table V. Root Mean Squared Error (RMSE) of the learning model is shown in Table VI.

Table V. Feature importance

Features	Importance
heavy_atoms	0.099
acd_most_apka	0.091
acd_most_bpka	0.077
qed_weighted	0.074
molecular species	0.070

Table VI. Root Mean Squared Error

RMSE with 21 features	RMSE with 19 features
1.13853	1.13351

Oral, Parenteral, Topical, Toxic, OP and PT are represented as Class 1, Class 2, Class 3, Class 4, Class 5 and Class 6 respectively. Oral represent drugs that can be taken Orally. Parenteral are drugs that can be given as injections. Topical are medicines that can be applied externally. Toxic are toxic drugs. OP represent drugs that can be taken both as Oral and Parenteral and PT has properties of both Parenteral and Topical.

Table VII. Classification by Parametric model

CLASS	Accurate Classification	Misclassifications
Oral	12	OP-1
Parenteral	14	0
Topical	12	Oral-1, Parenteral-1, PT-1
Toxic	14	0
OP	14	Oral-3, Parenteral-1, PT-2
PT	12	OP-2

Table VII shows the multivariate classification of drugs that were predicted accurately by the Logistic Regression model which is a parametric model. Anaconda package was employed and the machine learning models were applied using Jupyter notebook. Among 90 samples of test data set, Parenteral and Toxic were the two classes that were predicted accurately without misclassification. One Oral drug has been misclassified as OP. Three Topical drugs have been misclassified as Oral, Parenteral and PT. Three OP drugs were misclassified as Oral, one as Parenteral and two as PT. Similarly, PT drugs have been misclassified as two OP.

Table VIII. Classification by Non-Parametric model

CLASS	Accurate Classification	Misclassifications
Oral	11	Parenteral-1, OP-1
Parenteral	14	0
Topical	13	Parenteral-1, OP-1
Toxic	14	0
OP	16	Oral-1, PT-3
PT	13	OP-1

Table VIII shows the multivariate classification of drugs that were predicted accurately by the Decision Tree model which is a Non-parametric model. Parenteral and Toxic were the two classes that were predicted accurately without misclassification. One Oral drug has been misclassified as Parenteral and one as OP. Two Topical drugs have been misclassified as Parenteral and OP. Four OP drugs were misclassified as one Oral and three as PT. Similarly, one PT drug was misclassified as OP. Table I and Table II, shows the results of Precision, Recall and F1-Score of Logistic regression and Decision Tree learning models.

Table IX. Prediction accuracy of the models

Dataset	Samples	Parametric model	Non -Parametric model
Training	210	0.927	1.000
Testing	90	0.867	0.900

70% of the dataset with 300 samples were used for training and 30% of the dataset were considered for testing. Decision Tree, which is a Non-Parametric model performed better than the Logistic Regression, which is a parametric model. In the proposed classification, Non-parametric machine learning model performed more accurately than Parametric model and had a prediction accuracy of 0.900. Further, the error rate was reduced when the features were reduced from 21 to 19 as shown in Table VI.

VII. CONCLUSION

In this paper, parametric and non-parametric machine learning models to classify the drug applicability have been analyzed. The chemical drug compounds were acquired from chEMBL database. Data samples drawn from the database defined the physicochemical and molecular properties. Feature reduction was performed on the dataset and the reduced feature set was given as input to the machine learning models. With comprehensive and balanced data samples, non-parametric machine learning model (Decision Tree) resulted in better prediction accuracy of 90%. For multivariate classifications, the increase in data samples would produce best results. Increasing the data samples which are balanced and efficient feature engineering techniques combined with hybrid models would result in accurate classification of multivariate classification of drugs. In future, this would greatly reduce the time consumed and the huge cost incurred, where only the relevant and potent drugs can be identified in the process of drug discovery.

REFERENCES

1. Mylola Galushka, Fiona Browne, Maurice Mulvenna, Raymond Bond, Gaye Lightbody.Toxicity Prediction Using Pre-trained Autoencoder. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE 2018.
2. Tailong Lei, Youyong Li, Yunlong Song, Dan Li, Huiyong Sun and Tingjun Hou, “ADMET evaluation in drug discovery: 15. Accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling”, Journal of Cheminformatics, 8:6. 2016.
3. Mathew, T.E. A Logistic Regression with Recursive Feature Elimination Model for Breast Cancer Diagnosis. International Journal on Emerging Technologies. 10(3): pages 55–63, 2019.
4. Christelle Reynes et al. Designing Focused Chemical Libraries Enriched in Protein-Protein Interaction Inhibitors using Machine-Learning Methods. PLoS Computational Biology, 6(3), 2010.
5. Yang Li, Yujia Tian, Zijian Qin, and Aixia Yan. Classification of HIV-1 Protease Inhibitors by Machine Learning methods. ACS Omega. 3(11) pages 15837-15849, 2018.

6. Syed Asif Hassan, Syed and Hamza Osman, Ahmed. An Improved Machine Learning Approach to Enhance the Predictive Accuracy for Screening Potential Active USP1/UAF1 Inhibitors. (IJACSA) International Journal of Advanced Computer Science and Applications. 8(4), 2017.
7. Maibam Mangalleibi Chanu and Khelchandra Thongam, "Brain Tumor Segmentation and Classification using Hybrid clustering technique and SVM Classifier", International Journal of Innovative Technology and Exploring Engineering (IJITEE) , Volume 9, Issue-15, Nov 2019.
8. Setu Basak, Md. Mahbub Alam, Aniruddha Rakshit, Ahmed Al Marouf and Anup Majumder, "Predicting and staging chronic kidney Disease of Diabetes (Type-2) Patient using Machine Learning Algorithms", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume 8, Issue-12, Oct 2019.
9. Heena Nankani, Shruti Gupta, Shubjam Singh and S.S.Subashka Ramesh, "Detection Analysis of Various Types of canfer by Logistic Regression using Machine Learning", International Journal of Engineering and Advance Technology (IJEAT), Volume 9, Issue-1, Oct 2019.
10. https://en.wikipedia.org/wiki/ChEMBL.
11. B. Firdaus Begam and Dr.J. Satheesh Kumar, A Study on Cheminformatics and its Applications on Modern Drug Discovery, International Conference on Modeling Optimization and Computing-(ICMOC). Procedia Engineering 38, pages1264-1275, 2012.
12. https://www.ebi.ac.uk/training/online/course/chembl-quick-tour/what-chembl.
13. https://en.wikipedia.org/wiki/violin_plot.
14. Yuxi Hayden Liu. Python Machine Learning by Example. Packt Publishing Ltd, Pages 153-176, 2017.
15. Pang-ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data mining, Pearson India Education Pvt Ltd, 2016.
16. https://www.scikit-learn.org.
17. Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, 2016.
18. Shadi Momtahen, Furat Al-Obaidy and Farah Mohammadi. Machine Learning with Digital Microfluidics for Drug Discovery and Development. In 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE), pages 1-6. IEEE, 2019.
19. Sai Nivedita Chandrasekaran and Jun Huan. Weighted Multiview Learning for predicting Drug-Disease Associations. In 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 699-702. IEEE, 2016.

AUTHORS PROFILE

Dr. N. Priya is a highly skilled academician having more than 15 years of teaching experience. She was awarded Ph. D from Bharathiar University Coimbatore. She completed both her UG and PG degrees with distinction. She has qualified the NET exam. She has been serving as the head of UG Department of Computer Application from 2004 to 2019 in SDNB Vaishnav College for Women, Chennai. In addition, she has also held the post of Additional Controller of Examinations from 2011 to 2014. She is a member of IQAC cell from 2014. She is the resource person for framing the syllabus and designing the Course materials. She has also been a member of Academic Audit and review panel in other institutions. She has chaired several sessions in International Conferences. She has published and presented more than 20 research articles and completed many NPTEL courses. She is a recognized research supervisor under the University of Madras. Her research areas include datamining, image processing, Neural Networks, Network programming and Fuzzy Logic. Currently she is working as Associate Professor in the PG Department of Computer Science, SDNB Vaishnav College for Women.

G. Shobana is currently working as Assistant Professor in Department of Computer Applications, Madras Christian College, Chennai. She completed her M.E (CSE) in 2006 from Sathyabama Deemed University, Chennai. She has more than 15 years of teaching experience and is SET qualified. Her research areas include Machine Learning and Bioinformatics.