Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prediction of Micronucleus Assay Outcome using in Vivo Activity Data and Molecular Structure Features

Priyanka Ramesh
Vellore Institute of Technology: VIT University
Shanthi V (shanthi.v@vit.ac.in)
VIT University https://orcid.org/0000-0003-2297-2751

Research Article

Keywords: Machine learning, Fingerprints, Descriptors, Structural alerts, Toxicity prediction

Posted Date: July 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-721125/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Applied Biochemistry and Biotechnology on October 20th, 2021. See the published version at https://doi.org/10.1007/s12010-021-03720-8.

Prediction of micronucleus assay outcome using in vivo activity data and molecular structure features

Priyanka Ramesh and Shanthi Veerappapillai *

Department of Biotechnology, School of Bio Sciences and Technology,

Vellore Institute of Technology, Vellore, Tamil Nadu, India.

*Corresponding author. Email: shanthi.v@vit.ac.in; Phone: +91 94865 36201

Abstract

In vivo micronucleus assay is the widely used genotoxic test to determine the extent of chromosomal aberrations caused by the chemical compounds in human beings, which plays a significant role in the drug discovery paradigm. To reduce the uncertainties of the in vivo experiments and the expenses, we intended to develop novel machine learning-based tools to predict the toxicity of the compounds with high precision. A total of 472 compounds with known toxicity information were retrieved from the PubChem Bioassay database and literature. The fingerprints and descriptors of the compounds were generated using PaDEL and ChemSAR for the analysis. The performance of the models was assessed using three tires of evaluation strategies such as 5-fold, 10-fold, and external validation. The accuracy of the models during external validation lay between 0.57 and 0.86. Note that a combination of fingerprints and random forest showed reliable predictive capability. In essence, structural alerts causing genotoxicity of the compounds were identified using the structural activity relationship model of SARpy tool. This study highlights that the structural alerts such as chlorocyclohexane and trimethylamine are likely to be the leading cause of toxicity in humans, further validated using the Toxtree application. Indeed, the results from our study will assist in scrutinizing the genotoxicity of the compounds with high precision by replacing extensive sacrifice of animal models.

Keywords: Machine learning; Fingerprints; Descriptors; Structural alerts; Toxicity prediction.

1. Introduction

Currently, humans are confronted with various chemicals, including cosmetics, food additives, pesticides, and drugs. These are the significant causes of mutagenicity or irreversible damage to genetic material, causing multiple negative impacts on human health, including cancer [1]. Recently, appropriate toxicity tests are implemented by the regulatory to measure the genotoxicity of compounds causing mutations at both genetic and chromosome levels. Initially, in vitro methods were implemented for examining the toxicity of the compounds. Subsequently, the positive compounds were further validated through in vivo studies [2]. For instance, chromosome aberration assay, micronucleus assay, comet assay, bacterial reverse mutation test, and sister chromatid exchange test were reported in the literature [3]. However, in vivo micronucleus assay successfully determines the genotoxicity of the compounds obtained as positive during in vitro tests.

In general, animal assays are regarded as the fate of chemicals and are precise than in vitro studies. For instance, in vivo assays are unethical, expensive, and time-consuming [4]. Moreover, some chemicals, including cosmetic constituents, are already banned for animal testing [5]. Hence, they are replaced by other strategies, including in silico approaches such as machine learning (ML) and quantitative structure-activity relationship (QSAR) approaches.

Recently, computational approaches capture more attention even by the regulatory authorities due to their resource and time-saving attributes. Initially, QSAR based technique is used for evaluating the toxicity of compounds. Nonetheless, these conventional strategies were not applicable for big data, and the extraction of features from high-volume data has become difficult [6]. Moreover, building features for the QSAR model from diverse resources has become grim due to its incapability of mapping the activity of the compounds with their corresponding targets.

Further, this strategy was time-consuming in properly splitting compounds into train and test sets [7]. In addition, developing a precise QSAR model has become difficult due to the rise in the number of samples in the test set [8]. Hence developing a successful QSAR model with high accuracy has become challenging. For instance, the QSAR model developed by Bossuyt et al using 718 micronucleus assay data from five different resources achieved the highest accuracy of about 0.68 for the test set [4]. Thus, incorporating machine learning techniques to replace this method, as proposed by Lavecchia et al., should overcome such limits in forecasting the toxicity of chemicals [9].

Wu et al., highlighted the building of machine learning models using various descriptors that are different from the traditional chemical descriptors. The author also highlighted the essence of machine learning in toxicity prediction analysis. The article also listed the available online tools built using machine learning and deep learning strategies [10]. Moreover, Chen et al developed a non-linear QSAR model to predict acute toxicity in fathead minnow species. Further, they validated it using 482 chemical compounds using a support vector machine and compared their results with other literature reported earlier [8]. Recently, Fan et al predicted toxicity compounds using different combinations of chemical descriptors and fingerprints. For instance, the results of this study highlighted that the support vector machine achieved the highest accuracy of 0.882 on using PubChem descriptors. In contrast, random forest outperformed with 0.882 accuracy during five-fold cross-validation [11].

However, two glitches were identified from previously reported studies: 1) identification of significant descriptors and fingerprints, 2) lack of cross-validation, and lower accuracy of models in predicting the toxicity of the test compounds. Most importantly, none of the studies identified the structural fragments associated with the toxicity of the compounds. Keeping these

issues in mind, high precision machine learning models were built using compounds retrieved from PubChem Bioassay database for predicting the toxicity [12]. Further, the frequency of structural fragments in micronucleus positive chemicals was also investigated to identify the highly toxic structural alerts. Overall, we believe that this study will provide a valuable tool to predict genetic toxicity and facilitate the designing of less toxic compounds in the mere future.

2. Materials and methods

2.1. Dataset construction

The dataset used in the study were obtained from two different streams: (i) A list containing 3074 compounds and its bioassay results were downloaded from the PubChem Bioassay database (https://www.ncbi.nlm.nih.gov/guide/chemicals-bioassays/) using the query as micronucleus assay as "Micronucleus assay in humans" [12]. The metadata sheet contained PubChem CID number, PubChem SID number, activity outcome, species/cell type, assay type, and results. Using the assay type as a filter, we have retrieved 412 compounds that were tested by micronucleus assay. Notably, these compounds were classified based on the assay results as "Positive," "Negative," and "No conclusion." Around 140 compounds with "No Conclusion" were discarded. Notably, the 272 compounds obtained from the PubChem Bioassay database were used for model development. (ii) Moreover, around 100 compounds analyzed with micronucleus assay were retrieved from the literature published between 2004 and 2020 and utilized for external validation of the generated models. The resultant spatial data file of the 372 compounds was downloaded using the PubChem CID number to extract features. Moreover, the compounds' molecular weight and tanimoto coefficient were calculated using MACCS fingerprints to evaluate the distribution of compounds using rdkit tool of Python [13]. In our study, tanimoto coefficient was calculated by averaging the proportion of standard features shared between two compounds

divided by their union. A threshold of 0.25 was set during the analysis to prove the non-redundancy nature of the dataset.

2.2. Molecular fingerprint and descriptor generation

Molecular fingerprints are the vectors encoding the structural characteristics of the chemical compounds. They are widely used for virtual screening, machine learning, virtual chemical maps, and QSAR modeling [14]. In the current investigation, six different categories of molecular fingerprints were generated using PaDEL Descriptors. The developed fingerprints are MACCS fingerprint (166 bits), Substructure fingerprint (307 bits), PubChem fingerprint (881 bits), Estate fingerprint (79 bits), CDK (1024 bits), and CDK Extended fingerprints (1024 bits). Each bit of a fingerprint represents the chemical property of the compounds.

On the other hand, the ChemSAR tool was implemented in this study to generate 325 features belonging to five different clusters such as Basak descriptors, Constitutional descriptors, Burden Descriptors, MOE descriptors. These feature clusters are the numerical descriptions depicting the structural and chemical properties of the molecules [15]. Altogether, 3806 features were retrieved for each compound and used for model development.

2.3. Dataset preprocessing and Feature selection

Over-fitting is the most prevalent issue in machine learning, particularly in models generated with a high number of variables. In addition, the predictive model built using all the features produces non-readable results and is complicate for visualization. Hence, data preprocessing and feature selection strategies are incorporated in the machine learning pipeline to reduce the detrimental effects of over-fitting and produce significant predictive models [16]. Initially, variance and correlation were calculated for all the 3806 features to preprocess the data. Variance depicts the sensitivity of the model towards the taken data, which reduces error during

model generation. Moreover, correlation establishes the relationship between the features and the target variable, which in turn assists us in selecting the essential features. Subsequently, three different criteria were implemented to clean the dataset [11]:

1. Features with all zero values are removed.
2. Features with variance less than 0.05 are eliminated.
3. Correlation values less than 0.01 (less correlated) and above 0.99 (highly redundant) are removed.

Consequently, the crucial features of the dataset were selected using recursive feature elimination with decision tree estimator kernel in combination with cross-validation strategy (RFECV). This strategy identifies the optimal number of features and ranks them accordingly based on the level of correlation [17]. Finally, the top-ranking fingerprints and descriptors were selected for model development and validation.

2.4. Model development using a machine learning strategy

Six classifier algorithms, namely Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting Machines (GBM) were employed for model development with train-test ratio distribution of about 80% and 20% respectively [18-21]. These models were chosen due to their simple interpretation strategy and are highly based on data-driven predictions rather than static program instructions [22]. Moreover, models were developed using three different approaches: 1) using only the fingerprints of compounds, 2) Using only the descriptors of the compounds, and 3) models were developed using both fingerprints and descriptors.

a) Naïve Bayes

Naïve Bayes algorithm is a fast and simple classifier that works on a well-known Bayes theorem. The Bayes theorem is given by:

$$P(A|B) = \frac{P(A|B) \cdot P(A)}{P(B)} \quad (1)$$

Where A and B are the events,

A|B is the probability of A given by B is true

B|A is the probability of B given by A is true

P(A), P(B) are the independent probabilities of A and B, respectively.

Using the above rule, the probability of fitting to each subset is learned by eliminating the marginal probabilities determined based on the specific conditions [23]. This classification was carried out by using GaussianNB sub-package imported from the sklearn. Naive bayes package [24].

b) Logistic regression

The sigmoid function is the most widely recognized function used in neural networks due to its non - linearity and computational simplicity. Thus, logistic regression (LR) uses the sigmoid function to estimate the probability threshold by mapping the dependent and independent variables. LR works on the given below sigmoid function:

$$p = \frac{1}{1+ e^{-z}} \quad (2)$$

P is the probability threshold, and z is a real number ranging between -∞ and +∞ [22]. Although it is more appropriate for binary classification, it provides satisfactory results even for categorical and continuous variables. Moreover, LR model was implemented in our analysis as the target variable (survival rate) used in this study is a binary variable. In this model generation, the patients' survival is considered the dependent variable, and all 36 features were regarded as the independent variables. This algorithm was carried out by using LogisticRegression sub-package imported from the sklearn.linear_model package [24].

c) Decision tree

Decision Tree (DT) is a predictive model that maps and develops the relationship between nodes and leaf nodes. The object (feature) is represented as a node in the tree, and the leaf node demonstrates the object value (class of the corresponding feature). DT has become advantageous because of its wide usage in data mining, data analysis, and data prediction [25]. Gini is used to measure the distribution of the subsets of the data in the generated tree. The Gini is calculated using the given below formula:

$$Gini = 1 - \sum(P_i)^2 \quad (3)$$

The P represents the probability of each class in the respective feature. Note that the DecisionTreeClassifier sub-package of sklearn,tree package was implemented to perform the decision tree analysis. This sub-package processed the data and produced a decision tree at the end and the other assessing values, including the accuracy and precision of the model [24].

d) Random forest

Random forest (RF) generates considerable numbers of trees from the point of a random seed, which is split during the model's training. This algorithm generates trees randomly to overcome data overfitting, which is ensemble by the RF classifier resulting in one final model with higher accuracy [26]. It is achieved by identifying and classifying the essential features into binary outcomes using the mean decrease accuracy and Gini score. Here, we employed RandomForestClassifier imported from sklearn.ensemble package for the development of random forest model [24].

e) Support vector machine

Support vector machine (SVM) is the most widely used algorithm for solving classification and regression problems. SVM minimizes over-fitting by using a structural risk minimization

strategy during model generation. This strategy reduces the generalization error instead of the mean square error of a model [26]. SVM classifies the input data using both linear and non-linear hyperplanes. Noises in the model are finally reduced using the slack variables [27]. A linear hyperplane with statistical characteristics is created in this study to distinguish two different cohorts using the svm sub-package from sklearn package [24]. The equation given below describes the linear hyperplane that divides the input space.

$$W^T X_i + b = 0 \quad (4)$$

Where the separation hyperplane is defined by W and the bias is defined by b.

f) Gradient Boosting Machines

Gradient boosting machine, an ensemble and non-parametric approach, associate predictions of feeble learners to yield robust outcomes than the single learner. Unlike other parametric machine learning models, GBM implements additive expansion for model developments providing more freedom for the researchers. This algorithm trains the model in a gradual, additive, and sequential manner. Initially, it trains a decision tree using the user-defined dataset and later on which new tree is fit into the modified dataset sequentially. Since GBM works based on the functional-based optimization strategy, the performance of GBM is comparatively better than other bagging-based techniques [28]. GBM is implemented by importing the sub-package GradientBoostingClassifier from the sklearn ensemble of sklearn [24].

2.5. Performance evaluation of the models

The criteria such as accuracy, precision, recall, F1 score, and average precision-recall score (AP score) were calculated using the given equations to evaluate the predicting ability of the models used.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (5)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (6)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (7)$$

$$\text{F1 score} = 2 * \frac{\text{Precision*Recall}}{\text{Precision+Recall}} \quad (8)$$

$$\text{AP Score} = \sum_n (R_n - R_{n-1}) P_n \quad (9)$$

Where TP and TN denote true positive and true negative. Similarly, FP and FN denote false positives and negatives, respectively [28]. Moreover, R_n and P_n represent the recall and precision of the n^{th} threshold.

Additionally, Receiver operating curve (ROC) scores were estimated for all the developed machine learning models to visualize the connection between the sensitivity and specificity of the models. It is plotted between FP rate in the x-axis and TP rate in the y-axis. The ROC AUC score and the metrics for the plot were measured using roc_auc_score and metrics sub-packages of sklearn.metrices package. These measurements were used to plot the ROC curve using a matplot library of python packages [24].

2.6. Analysis of Structural alerts

The relationship of the presence of toxic compounds causing adverse effects on human organs and their sub-structures are defined by the structural alerts. In general, SA can be identified by using three different strategies, namely, (i) graphical method, (ii) fragment-based, and (iii) SMILES-based tools. Among them, the SMILES based prediction using SARPy was found to be more efficient in determining the SA of the structure with high accuracy [29]. Hence, in the current investigation, SARPy was implemented to identify the SA with increased toxicity. The privileged

substructures were analyzed using the frequency of the SA among the taken dataset. The frequency of the substructures is given by

$$f_m = \frac{N_{\text{fragment class}} X\, N_{\text{total}}}{N_{\text{fragment total}}\, X\, N_{\text{class}}} \quad (10)$$

Where $N_{\text{fragment class}}$ is the number of compounds with micronucleus positive assay result and the N_{class} denotes the total number of micronucleus assay positive compounds. On the other hand, N_{total} represents the total number of compounds in the dataset. $N_{\text{fragment total}}$ indicates the number of compounds containing the fragments in their structure [11]. The following settings were fixed to identify the toxic substructures [4]:

1) SA comprising of minimum 2 atoms

2) SA comprising of maximum 18 atoms

3) SA occurring in a minimum of 10 training set compounds

2.7. Cross-validation of the models

Cross-validation is a mathematical approach used to evaluate the performance of the machine learning models. Foremost, cross validation, a standard tool in ML pipeline, is commonly implemented to provide statistically unbiased results among the six machine learning models [30]. In the current analysis, the dataset (n=272) is randomly partitioned into training and test set in the ratio of 8:2, respectively, for each fold of validation. Then, the average of each scoring matrix, including accuracy, precision, recall, and F1 score, was calculated to compare and select the best models. Finally, the sub-packages k-fold and cross_val_score are imported from sklearn.model_selection package to carry out the cross validation process [21].

3. Results and discussion

3.1. Dataset analysis and feature selection

272 non-duplicated chemical compounds with toxicity data were collected from the PubChem database for model development and validation. Around 3481 bits of molecular fingerprints and 325 features of descriptor clusters were generated for each compound using PaDEL and ChemSAR tools, respectively. The molecular weight of the compounds varied from 18.9 to 1250.7. To elucidate the diversity of the dataset, tanimoto coefficients between all the compounds were calculated using MACCS fingerprints. The average tanimoto coefficient of 0.191 diminutive than the threshold value depicts the diversity of compounds retrieved for our analysis [13].

3.2. Feature selection of descriptors and fingerprints

Feature selection is the automated process of selecting highly relevant features to improve the predictive performance of the models. Hence, a low variance feature filtering and high correlation feature filtering strategy were established in our investigation to eliminate the highly redundant features. For instance, 143 descriptors and 1067 fingerprints having low variance value of less than 0.05 were removed. This yielded a total of 2591features, including 182 descriptors and 2409 fingerprints. Subsequently, feature selection based on correlation resulted in the selection of 139 descriptors and 116 fingerprints which were further considered for recursive feature elimination with 5-fold cross-validation strategy. This process resulted in the selection of 42 crucial descriptors and 6 fingerprints for machine learning model generation. Initially, model generation was accomplished by implementing the selected 6 fingerprints and 42 descriptors distinctly. Towards the end, the fingerprints and descriptors were combined to generate the model. For instance, fingerprints such as FP18, FP29, FP70, FP193, FP236 and FP426 were selected for model generation. The distribution of selected descriptors over different categories is consolidated in supplementary table 1. The 20 constitutional features describe the

physicochemical and structural information of the compounds. The 10 Basak descriptors depicted the polarity nature of the molecules. Besides, 9 MOE descriptors reflected the contribution of surface area and partial charges by the compounds. The other 3 topological descriptors bcutm4, bcutm15 and bcutv15 were related to atomic Van der Waal's forces and the atomic masses of the compounds [31].

3.3. Performance evaluation of models

3.3.1. Fingerprint based models

The significant fingerprints identified using RFECV method were implemented for model generation in the initial stage of our analysis. In the current investigation, six different models were built and evaluated using 6 significant fingerprints identified during recursive feature elimination process. The models are estimated using different metrices and the results are tabulated in Table 1. The accuracy of the models ranged from 0.66 to 0.85. Moreover, the precision values of all the models were found to be in the range between 0.73 and 0.86.

On the other hand, recall ranged from 0.68 to 1.00, and F1-score ranged from 0.75 to 0.91. Based on the accuracy, the random forest model (0.86) outperformed in predicting the toxicity of the chemicals than the other models investigated in our analysis. By considering other performance metrices, F1-score served as a benchmark evaluating metric due to its most delicate blend between precision and recall [32]. The result from the fingerprint-based ML model portrays that random forest exhibited good predicting capability with accuracy of 0.91 to predict the genotoxicity of the compounds (Table 1).

3.3.2. Descriptors based models

The machine learning models were generated in the second stage using the significant descriptors identified using REFCV strategy. The accuracies of all the models were found to be

LR (0.77), RF (0.82), DT (0.67), SVM (0.76), NB (0.75) and GBM (0.76) respectively. The precision and recall value of the models varied between 0.77 to 0.87 and 0.71 to 0.94, respectively. Moreover, the F1-score lied from 0.78 to 0.9. It is to note that random forest achieved good performance metrices than other models developed earlier using the descriptors. For instance, the precision of 0.85, recall of 0.94, and F1-score of 0.9 were observed in the descriptor-based RF model. On comparing the results of descriptor-based models, random forest exhibited exceptional predicting capability than other models.

3.3.3. Fingerprints and descriptor-based models

To date, no literature was reported on predicting the genotoxicity of the compounds using the combined features. Hence, in our analysis, we have merged the significant fingerprints and descriptors using the feature selection process in all possible to evaluate the efficiency of the model. Although, the accuracy was comparatively lower than fingerprint and descriptor-based models. The other metrices, such as precision, recall, and F1-score was equivalent to previously built models. Even though random forest excelled with high accuracy of 0.8 in this analysis, the performance of the combined feature models was not satisfactory.

In addition, the ROC of the models was evaluated as it is highly correlated with the accuracies of the classifiers [33]. It is evident from Fig. 1 that the area under the curve value (AUC) of the random forest model built was found to be 0.82, 0.74, and 0.76 for fingerprints, descriptors, and combination of features, respectively. Using fingerprints to develop random forest models has increased the accuracy by 0.08 times than the descriptor-based RF model and 0.06 times than the combination model. Ultimately, the fingerprint-based random forest model is the most suitable strategy for predicting the genotoxicity of the compounds.

3.3.4. Performance evaluation of cross-validation

5-fold and 10-fold cross-validation techniques were implemented with the same dataset to validate the generated machine learning models and identify the best ML model to predict the toxicity of the compounds. The results of six models with cross-validation were shown in Fig. 2 and Fig. 3. Other performance metrics such as precision and recall are represented in supplementary Fig. 1. It is to be noted that no significant variation in accuracy was observed for all the developed models throughout the 5-fold cross-validation process. On the other hand, the accuracy of models generated using fingerprints hiked above 0.88 during 10-fold cross-validation. Notably, random forest achieved the highest accuracy of about 0.97 during validation. Similarly, all the models built using descriptors achieved accuracy greater than 0.9 except decision tree (0.81) and naïve bayes (0.81) algorithms. Remarkably, random forest achieved higher accuracy of 0.96. Similar pattern of variation in accuracy was noted for the models developed using both descriptors and the fingerprints with the highest accuracy of 0.95 achieved by random forest classifier. On analyzing the F1-score of the cross-validation process (Fig. 3), a random forest model developed using fingerprints, and a combination of descriptors and fingerprints achieved the highest score of about 0.98, which depicts the ability of the model to predict the data with high precision [34].

3.3.5. External validation of the models

The accuracy of the ML models was validated using an external dataset of 100 compounds retrieved from the literature. For instance, the literature information in the last twenty years were searched to build the validation dataset. The results of external validation were shown in Fig. 2 and Fig. 3. It is worth mentioning that the generated model performance for the external dataset correlates well with the training set data. For instance, random forest model is performed in prediction in terms of accuracy, precision, recall, and f1-score. Note that

accuracy and F1-score of 0.86 and 0.95 were achieved by the random forest model developed by fingerprints. Notably, the prediction performance of the fingerprint-based model was superior to the model generated using other methods. These observations consistent well with the model generated using the training set. Therefore, we postulate that random forest combined with fingerprints to be the right choice for replacing the animal-based toxicity prediction.

3.4. Structural fragment analysis

To examine the presence of privileged substructures in micronucleus positive compounds, substructure frequency analysis was also performed using SARpy software. This algorithm will retrieve the relationship between the toxic compounds and also identifies the crucial substructure responsible for their toxicity. The comma separated value (CSV) file containing the smiles and toxicity status of the compounds were provided as input for identifying the structural alerts. The existence of substructure in a minimum of 10 micronucleus positive chemical entities is regarded as the toxic compound. Accordingly, nine substructures such as formic acid, polyethylene, chlorocyclohexane, cyclohexane, pyrimidinone, trimethylamine, methoxy-cyclohexane, methylamine and phosphorous monoxide were identified with high frequency in micronucleus positive compounds than the negative chemicals. The substructure and its frequency are presented in table 2. From this study, we hypothesize that a new molecule containing one or more structural alerts mentioned earlier likely to have a high probability of being favorable to micronucleus assay [11]. Although the SARpy tool identifies various substructures, understanding the toxicity of compounds at DNA level is very important. Therefore, the substructures were further validated using Toxtree software, a decision tree-based strategy [35].

In our investigation, among the 9 identified substructures, the toxicity of 2 structural alerts i.e., Chlorocyclohexane and trimethylamine were determined to cause toxic effects at the DNA level. Chlorocyclohexane is a highly flammable substance that causes skin corrosion, severe eye damage, specific target organ toxicity, and respiratory tract irritation. In addition to the aforementioned toxic effects, trimethylamine causes acute oral toxicity in humans [36, 37]. These evidences highlight the existence of chlorocyclohexane and trimethylamine moieties have a higher probability of being positive to in vivo micronucleus assay in our dataset.

4. Conclusion

In the current investigation, supervised classification models were developed using a dataset of 272 compounds to predict the chemical compounds' genotoxicity. Different types of fingerprints and descriptors were generated for model development and validation. Three tires of evaluation, such as 5-fold and 10-fold cross-validation and an external dataset of 100 compounds, were utilized to evaluate the robustness of the developed models. The results from our study highlight that the random forest algorithm with fingerprint features would be the right choice to replace the in vivo micronucleus assay. Importantly, privileged structural alerts such as chlorocyclohexane and trimethylamine were identified using SARpy and Toxtree algorithms. Based on our findings, we hypothesize that compounds containing these substructures have a high possibility of genotoxicity. Indeed, these findings can assist researchers in identifying genotoxic compounds without the assistance of animal models in the near future.

Declarations

Funding: Not applicable

Conflict of Interests

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Availability of data

https://www.ncbi.nlm.nih.gov/guide/chrmicals-bioassays/

Code availability: Not applicable

Author Contribution

PR performed data collection, machine learning model generation and validation. SV conceived this study and is responsible for overall design, result interpretation, manuscript preparation and communication.

Acknowledgment

The authors thank management of Vellore Institute of Technology for providing the necessary facility to carry out this research work.

References

1. Basu, A. K. and Nohmi T. (2018) Int. J. Mol. Sci. 19(4), 970.
2. Eastmond, D. A., Hartwig, A., Anderson, D., Anwar, W. A., Cimino, M. C., Dobrev, I., ... and Vickers, C. (2009) Mutagenesis. 24(4), 341-349.
3. Kang, S. H., Kwon, J. Y., Lee, J. K., and Seo, Y. R. (2013) J.Cancer. Prev. 18(4), 277.
4. Van Bossuyt, M., Raitano, G., Honma, M., Van Hoeck, E., Vanhaecke, T., Rogiers, V., ... and Benfenati, E. (2020) Toxicol. Lett. 329, 80-84.
5. Kamath, P., Raitano, G., Fernández, A., Rallo, R., and Benfenati, E. (2015) SAR QSAR Environ. Res. 26(12), 1017-1031.
6. Khalifa, N., Kumar Konda, L. S., and Kristam, R. (2020) Future Med. Chem. 12(20), 1829-1843.
7. Tsou, L. K., Yeh, S. H., Ueng, S. H., Chang, C. P., Song, J. S., Wu, M. H., ... and Ke, Y. Y. (2020) Sci. Rep. 10(1), 1-11.
8. Chen, X., Dang, L., Yang, H., Huang, X., and Yu, X. (2020) RSC Adv. 10(59), 36174-36180.
9. Lavecchia, A. (2015) Drug Discovery Today. 20(3), 318-331.
10. Wu, Y., and Wang, G. (2018) Int. J. Mol. Sci. 19(8), 2358.
11. Fan, D., Yang, H., Li, F., Sun, L., Di, P., Li, W., ... and Liu, G. (2018) Toxicol. Res. 7(2), 211-220.
12. Wang, Y., Bryant, S. H., Cheng, T., Wang, J., Gindulyte, A., Shoemaker, B. A., ... and Zhang, J. (2017) Nucleic Acids Res. 45(D1), D955-D963.
13. Bero, S. A., Muda, A. K., Choo, Y. H., Muda, N. A., and Pratama, S. F. (2018) arXiv preprint. arXiv:1806.05237.

14. Capecchi, A., Probst, D., and Reymond, J. L. (2020) J. Cheminf. 12(1), 1-15.
15. Mauri, A., Consonni, V. and Todeschini, R. (2016) In Handbook of Computational Chemistry; Leszczynski J; Springer: Dordrecht, pp 1-29.
16. Wu, Y., Liu, J., Han, C., Liu, X., Chong, Y., Wang, Z., ... and Li, S. (2020) Front. Oncol. 10, 743.
17. Jiang, C., Zhao, P., Li, W., Tang, Y., and Liu, G. (2020) Toxicol. Res. 9(3), 164-172.
18. Bhattacharjee, P., Dey, V., and Mandal, U. K. (2020) Saf. Sci. 132, 104967.
19. Zhou, S., Wang, S., Wu, Q., Azim, R., and Li, W. (2020) Comput. Biol. Chem. 85, 107200.
20. Wang, J., Deng, F., Zeng, F., Shanahan, A. J., Li, W. V., and Zhang, L. (2020) Am. J. Cancer Res. 10(5), 1344.
21. Suresh K. (2020) Artif. Intell. Data. Eng. 1133:329-347.
22. Khairunnahar, L., Hasib, M. A., Rezanur, R. H. B., Islam, M. R., and Hosain, M. K. (2019) Inform. Med. Unlocked. 16, 100189.
23. Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O., and Poorolajal, J. (2019) Clin. Epidemiology Glob. Health. 7(3), 293-299.
24. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... and Duchesnay, E. J. (2011) Mach. Learn. Res. 12, 2825-2830.
25. Lynch Chip M, Behnaz A, Fuqua Joshua D, de Carlo Alexandra R, Bartholomai James A, and Balgemann Rayeanne N. (2017) Int. J. Med. Inform. 108:1-8.
26. Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., and Lopez, A. (2020) Neurocomputing. 408, 189-215.
27. Dwivedi, A. K. (2018) Neural. Comput. Appl. 29(12), 1545-1554.

28. Chen, Y., Jia, Z., Mercola, D., and Xie, X. (2013) Comput. Math. Methods. Med.
29. Yang, H., Li, J., Wu, Z., Li, W., Liu, G., and Tang, Y. (2017) Chem. Res. Toxicol. 30(6), 1355-1364.
30. Wang, G., Sun, Y., Chen, Y., Gao, Q., Peng, D., Lin, H., ... and Zhuo, S. (2020) J. Biophotonics. 13(9), e202000050.
31. Dong, J., Yao, Z. J., Zhu, M. F., Wang, N. N., Lu, B., Chen, A. F., ... and Cao, D. S. (2017) J. Cheminformatics. 9(1), 1-13.
32. Alabi, R. O., Elmusrati, M., Sawazaki-Calone, I., Kowalski, L. P., Haglund, C., Coletta, R. D., ... and Leivo, I. (2020) Int. J. Med. Inform. 136, 104068.
33. Barlow, H., Mao, S., and Khushi, M. (2019) Data. 4(3), 129.
34. Ma'ruf, F. A., and Wisesty, U. N. (2019) In Journal of Physics: Conference Series, March, IOP Publishing, 1192(1), 012011.
35. Han, Y., Zhang, J., Hu, C. Q., Zhang, X., Ma, B., and Zhang, P. (2019) Front. Pharmacol. 10, 434.
36. Ufnal, M. (2020) J. Nutr. 150(2), 419.
37. Xu, J., Qiao, X., Cui, M.F., Tang, J. H., and Zhang, J. P. (2005) Chin. J. Process Eng. 5(6), 643.

Legends

List of Tables

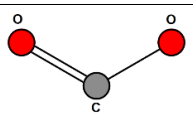
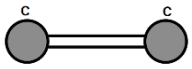
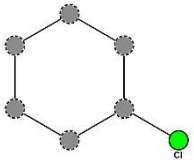
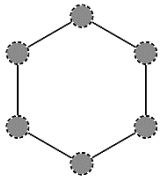
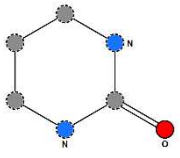
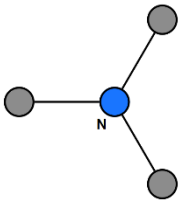
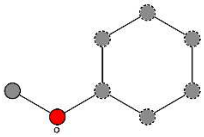
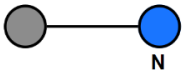
List of Figures

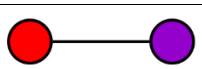
Table1: Performance metrices of machine learning models developed using different molecular representation of compounds

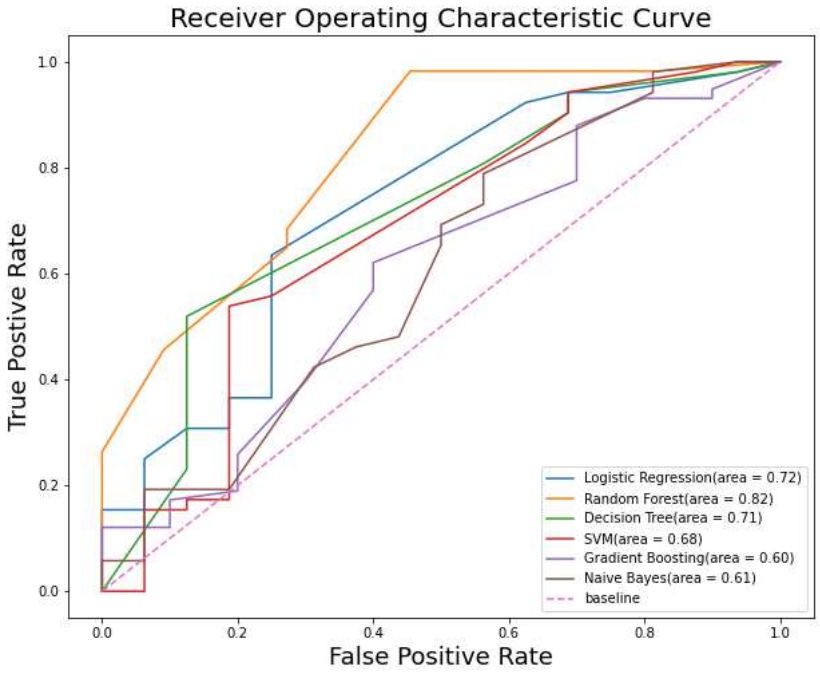
Molecule representation	Machine learning Model	Accuracy	Precision	Recall	F1-score
Fingerprints	Logistic Regression	0.8	0.85	0.93	0.89
	Random Forest	**0.85**	**0.86**	**0.98**	**0.91**
	Decision Tree	0.8	0.81	0.95	0.88
	Support Vector Machine	0.75	0.75	1.00	0.85
	Naïve Bayes	0.66	0.83	0.68	0.75
	Gradient Boosting	0.7	0.73	0.94	0.82
Descriptors	Logistic Regression	0.77	0.82	0.92	0.87
	Random Forest	**0.82**	**0.85**	**0.94**	**0.9**
	Decision Tree	0.67	0.87	0.71	0.78
	Support Vector Machine	0.76	0.77	0.98	0.86
	Naïve Bayes	0.75	0.79	0.9	0.84
	Gradient Boosting	0.76	0.79	0.92	0.85
Combination of fingerprints and descriptors	Logistic Regression	0.79	0.84	0.92	0.88
	Random Forest	**0.8**	**0.85**	**0.93**	**0.89**
	Decision Tree	0.7	0.87	0.75	0.81
	Support Vector Machine	0.76	0.79	0.94	0.86

	Naïve Bayes	0.73	0.82	0.85	0.84
	Gradient Boosting	0.75	0.84	0.85	0.84

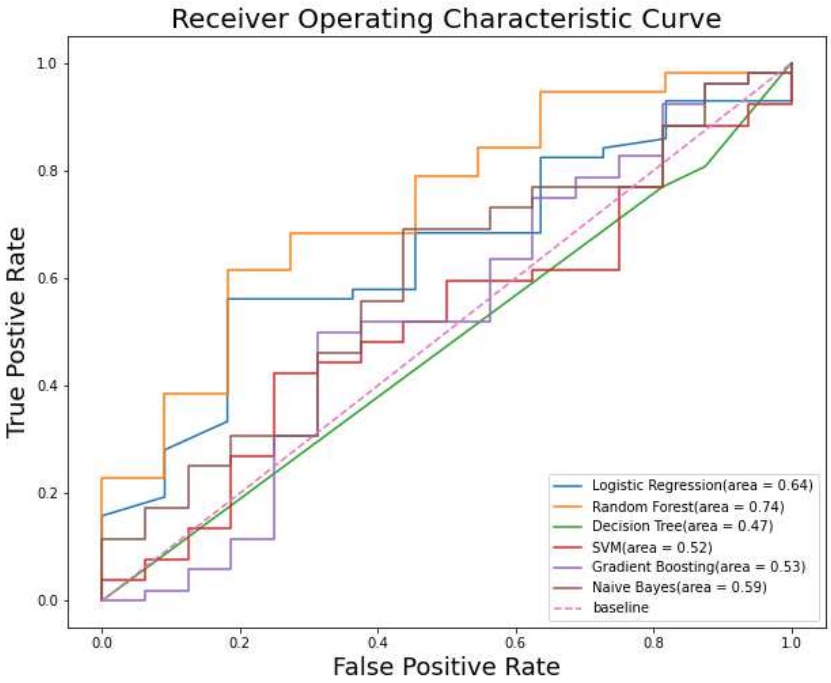
Table 2: Significant structural alerts and its frequency in substructures

S. No	Names	SMARTS	Frequency	Structure
1	Formic acid	C(=O)O	1.27	
2	Polyethylene	C=C	2.65	
3	Chlorocyclohexane	C1CCC(CC1)Cl	3	
4	Cyclohexane	C1CCCCC1	1.11	
5	Pyrimidinone	C1CN(C(=O)NC1)	2.31	
6	Trimethylamine	CN(C)C	2.25	
7	Methoxy-cyclohexane	COC1CCC(CC1)	4.49	
8	Methylamine	NC	1.93	

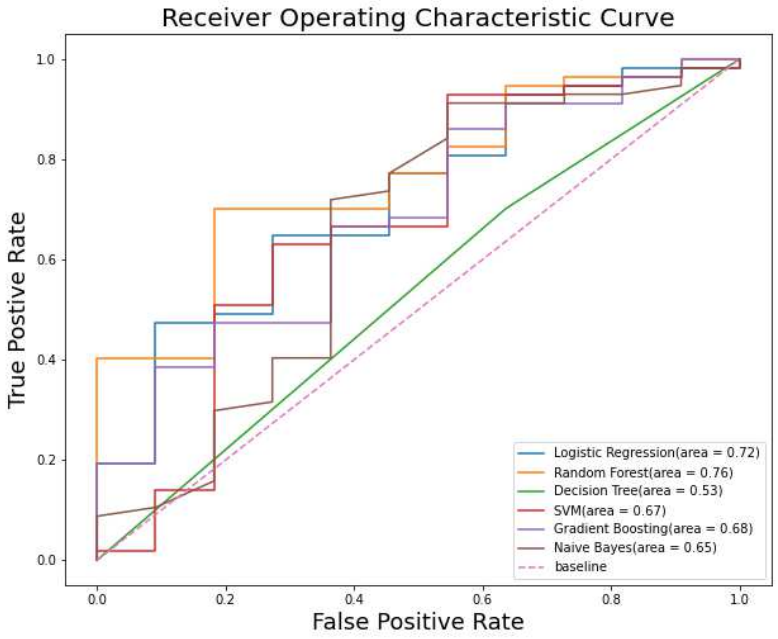
9	Phosphorous Monoxide	P(O)	4.63	



(a) ROC Curve analysis of models generated using fingerprints of the compounds

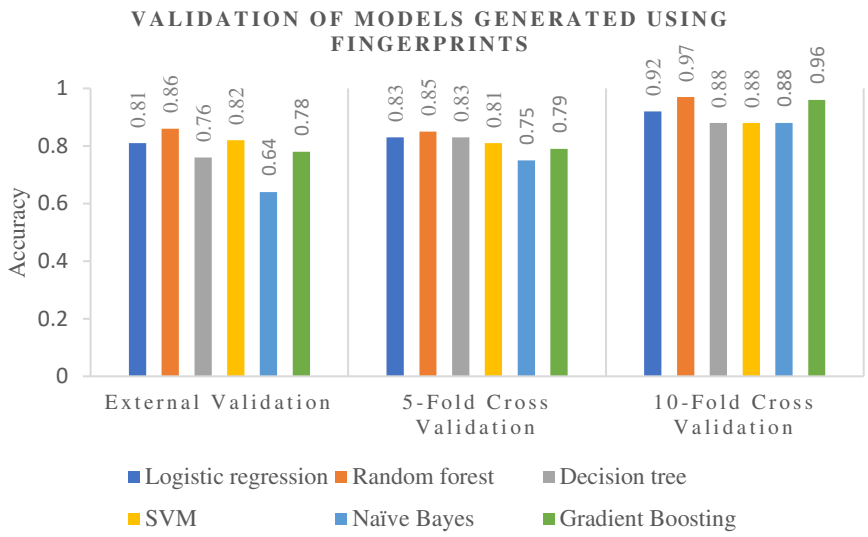


(b) ROC Curve analysis of models generated using descriptors of the compounds

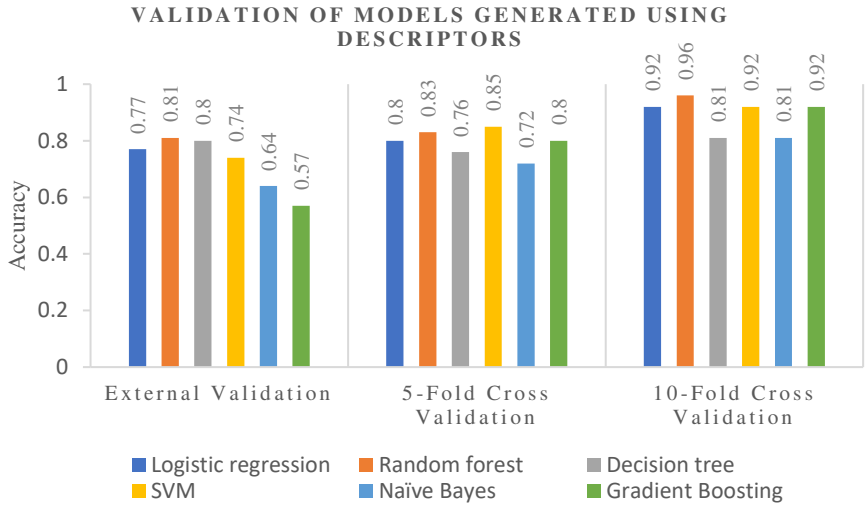


(c) ROC Curve analysis of models generated using both fingerprints and descriptors of the compounds

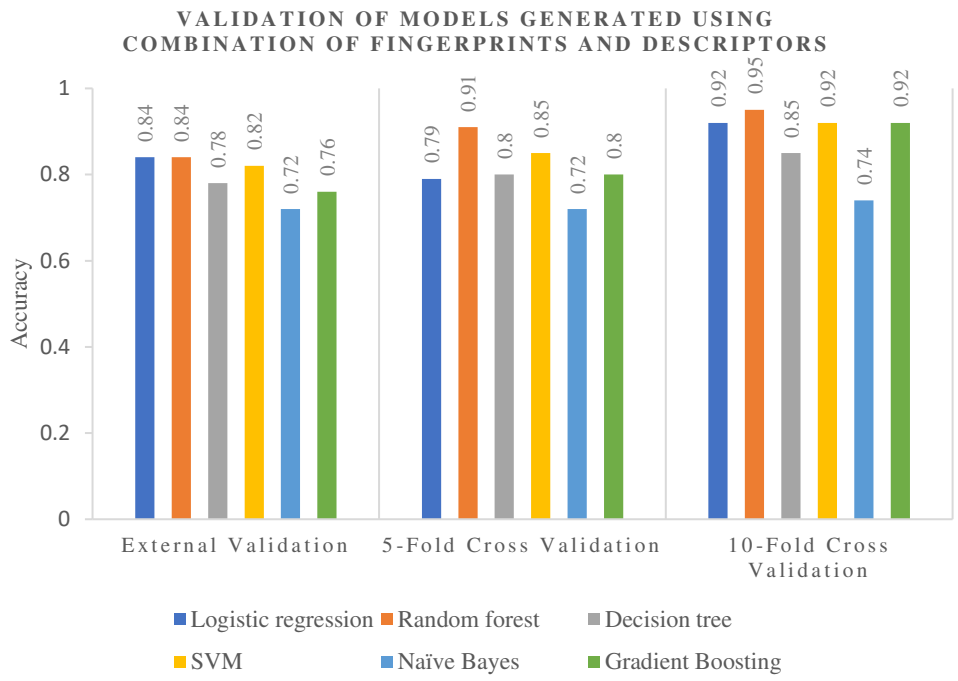
Fig. 1: ROC evaluation of generated models for the prediction of toxicity of the compounds



(a) Validation of models generated using fingerprints of the compounds



(b) Validation of models generated using descriptors of the compounds



(c) Validation of models generated using combination of fingerprints and descriptors of the compounds

Fig. 2: Prediction results of machine learning algorithms during validation process

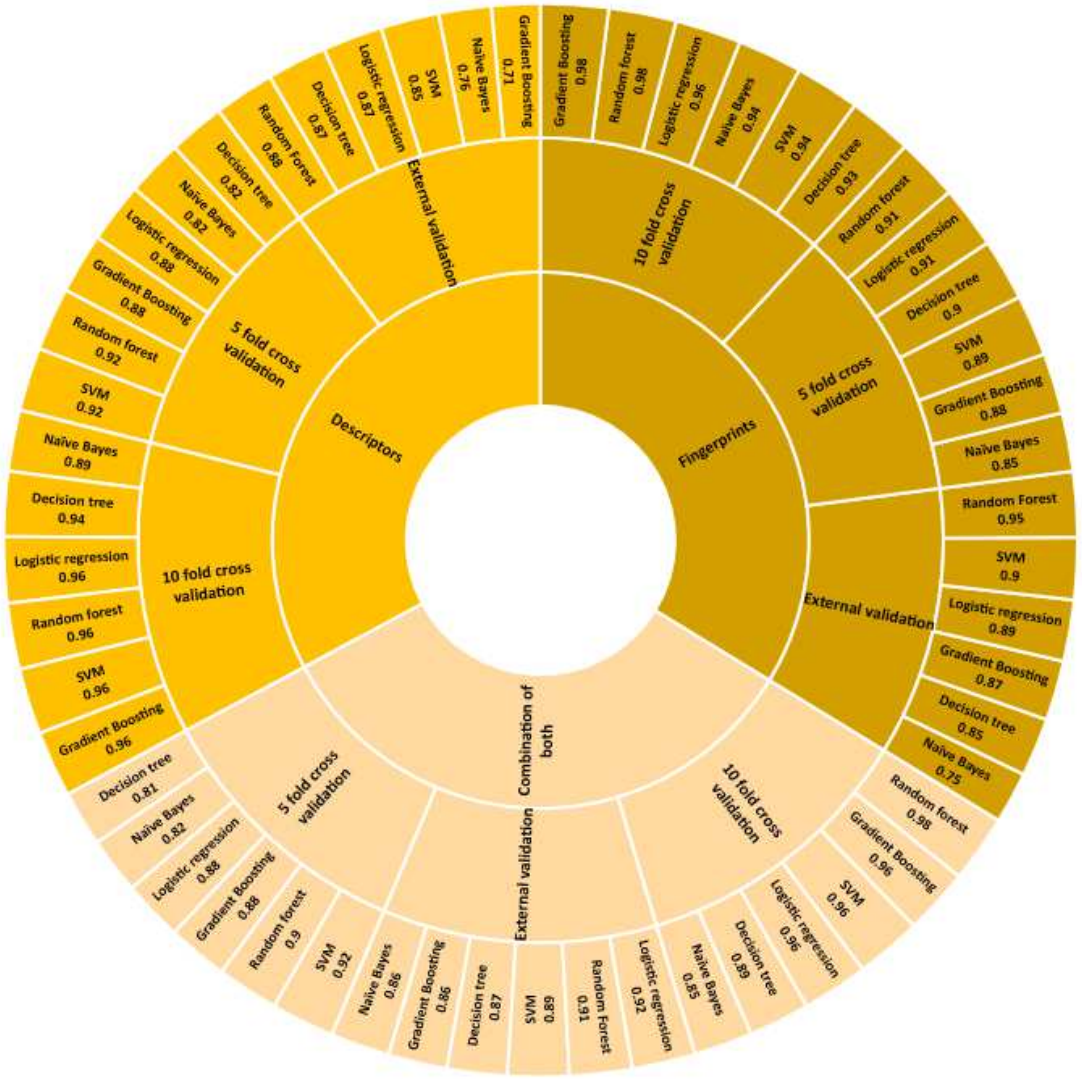


Fig. 3: Doughnut plot comparing the performance of all the models based on F1-score.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFileABAB.docx