ENHANCING ACCURACY FOR CLASSIFYING DRUGS BASED ON PATIENT DETAIL USING NOVEL ADABOOST ENSEMBLE CLASSIFIER OVER RANDOM FOREST CLASSIFIER

S. Monish Kumar[1], Rashmita Khilar[2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: To classify drugs based on patients' health-related data using Novel Adaboost Ensemble Classifier over Random Forest classifier.
Material and Methods: Classification is performed by content-based Novel Adaboost Ensemble Classifier (N=10) over random forest classifier (N=10). The sample size is calculated using GPower with pretest power as 0.9 and alpha 0.05.
Results: Mean accuracy of content-based AdaBoost (98.47%) is high compared to the Random forest classifier (96.45%). The significance value for accuracy and loss is 0.331 ($p>0.05$).
Conclusion: The mean accuracy of drug classifying based on patient detail-based AdaBoost is better than the random forest classifier.

Keywords: Novel Adaboost Ensemble Classifier, Drugs, Random Forest Classifier, Accuracy, Classification, Prediction.

[1]Research Scholar, Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India, Pincode:602105.
[2*]Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India, Pincode: 602105.

1. Introduction

The appropriate selection of chemical characteristics is a necessary preprocessing step for the effective use of computational intelligence approaches in virtual screening for bioactive molecule identification in drug discovery (Maeda et al. 2021). The choice of molecular descriptors has a significant impact on the precision of affinity prediction. In the investigation of Random Forest (RF)-based strategy to automatically choose molecular descriptors of training data for age, sex, BP, cholesterol, drugs, and other data to improve this prediction (Kolárik et al. 2007). Using RF in two separate ways: feature ranking and dimensionality reduction; and classification using the automatically selected feature subset, is the main originality of this work in the realm of drug development (Ubels et al. 2020). The positive findings are obtained in terms of accuracy and reduction of computational resources, it is also concluded that this technology can be utilized to improve drug design and discovery, thus aiding biomedical research significantly (Page, Baysari, and Westbrook 2017). The application of drug classification application form, an index, a summary, five or six technical sections, case report tabulations of patient data, case report forms, drug samples, and labeling, including (Cook, Addicks, and Wu 2008)**.**

In this research work, there have been 278 articles in Science Direct and 135 in scholars. Comparison is done using AdaBoost classifiers such as content-based AdaBoost classifier and random forest classifier. This focuses on existing techniques by developing novel multi-step algorithms that build models of drug response using random forest (Riddick et al. 2011)**.** Random Forest-based approach to improve the selection of molecular descriptors in automatic features selection improves drug discovering methods accuracy (Cano et al. 2017)**.** Furthermore, we have tried and confirmed that our strategy not exclusively could be applied to anticipate the new communications yet in addition could get a good outcome on the new dataset (Shi et al. 2019). The presented models are fast to generate and may serve as easily implemented screening tools for personalized oncology medicine, drug repurposing, and drug discovery (Lind and Anderson 2019). Another strategy to anticipate the medication target corporations precisely. Examine the impact of the four distinct classifiers on the outcomes. The proposed strategy expands the forecast execution more than a few techniques.

Our institution is passionate about high quality evidence based research and has excelled in various domains (Vickram et al. 2022; Bharathiraja et al. 2022; Kale et al. 2022; Sumathy et al. 2022; Thanigaivel et al. 2022; Ram et al. 2022; Jothi et al. 2022; Anupong et al. 2022; Yaashikaa, Keerthana Devi, and Senthil Kumar 2022; Palanisamy et al. 2022).The drug classification has the drawback of having low prediction and accuracy rates. In general, these methods are fast to train but quite slow to create predictions once they are trained. In the above problem, complexity will be decreased once a model is developed (Urista et al. 2020). Many researchers have discovered various prediction models that have low accuracy compared to other models. This work aims to enhance accuracy for classifying drugs based on patient details using the Novel Adaboost Ensemble Classifier, thereby improving accuracy and prediction and reducing time complexity.

2. Materials and Methods

This study setting was done in the Data Analytics Lab, Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. The sample size for this project is 20 (Group 1=10, Group 2=10). In the classifying drugs system, to modify the problem of low accuracy rate content-based Novel Adaboost Ensemble Classifier and Random forest classifier is used. The mean accuracy of the Novel Adaboost Ensemble Classifier is 98.47% and the mean accuracy of the Random Forest classifier is 96.45%. Dataset for this article is collected from (https://www.kaggle.com/ibrahimbahbah /drug200) website with 6 attributes and 201 rows. The Independent variables are blood pressure, age, sex and cholesterol. Dependent variables are precision and accuracy.

Adaboost Ensemble Classifier

A Novel Adaboost Ensemble classifier could be a meta-estimator that begins by fitting a classifier on the first dataset and so fits extra copies of the classifier on a similar dataset however wherever the weights of incorrectly classified instances area unit adjusted such ensuant classifiers focus additional on troublesome cases. AdaBoost is straightforward to implement. It iteratively corrects the mistakes of the weak classifier and improves accuracy by combining weak learners. AdaBoost is not liable for overfitting. The Pseudocode for the AdaBoost classifier is described in Table 1.

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \qquad (1)$$

Random Forest

Breiman proposed the Random Forest (RF) method in 2001 as a new machine learning method (Breiman 2017). It creates a decision tree-based Bagging integration and adds random attribute selection to the decision tree training process. RF has excellent classification abilities as a classifier. The RF approach has been widely used in recent years to solve a variety of issues, including classification, prediction, determining the importance of variables, dimensionality reduction, and abnormal point detection. When selecting an attribute, a typical decision tree chooses the best option from the set of options available to the current node. In RF, a subset of attributes is randomly selected from the node's attribute set for each node in the base decision tree, and then an optimal attribute is determined for the division from this subset. The Pseudocode for a Random Forest classifier is described in Table 2.

$$RFfi_i = \frac{\Sigma_{\ j} normfi_{ij}}{\Sigma_{\ j\epsilon all\ features, k\epsilon all\ trees} normfi_{jk}} \qquad (2)$$

Statistical Analysis

The minimum requirement to run the software used here is intel core I3 dual core cpu@3.2 GHz, 12GB RAM, 64 bit OS, 1TB Hard disk Space Personal Computer and Software specification includes Windows 8, 10, 11, Python 3.8, and MS-Office. Statistical Package for the Social Sciences Version 26 software tool was used for statistical analysis. An independent sample T-test was conducted for accuracy. Standard deviation, standard mean errors were also calculated using the SPSS Software tool. The significance values of proposed and existing algorithms contain group statistical values of proposed and existing algorithms.

3. Results

In statistical tools, the total sample size used is 20. This data is used for the analysis of the Novel Adaboost Ensemble Classifier and Random Forest Classifier. Statistical data analysis is done for both the prescribed algorithms namely Novel Adaboost Ensemble Classifier and Random Forest Classifier. The group and accuracy values are being calculated for given AdaBoost systems. These 20 data samples used for each algorithm along with their loss are also used to calculate statistical values that can be used for comparison. Table 3, shows that group, accuracy, and loss values for two algorithms content-based Novel Adaboost Ensemble Classifier and Random Forest Classifier are denoted. The Group statistics table shows the number of samples that are collected. Mean and the standard deviation is obtained and accuracies are calculated and entered.

Table 4, shows group statistics values along with mean, standard deviation and standard error mean for the two algorithms are also specified. Independent sample T-test is applied for data set fixing confidence interval as 95%. Table 5, shows independent t sample tests for algorithms. The comparative accuracy analysis, mean of loss between the two algorithms are specified. Figure 1 shows a comparison of the mean accuracy and mean loss between content Novel Adaboost Ensemble Classifier and Random Forest Classifier.

4. Discussions

The accuracy of random forest classifiers is 96.45% whereas content-based Novel Adaboost Ensemble Classifiers have higher accuracy of 98.47% with p = 0.331 because, a large number of datasets with fewer parameters. which shows that content-based Novel Adaboost Ensemble Classifiers are better than random forest classifiers. Mean values for content-based Novel Adaboost Ensemble Classifiers are 98.47 respectively. Similarly for random forest classifier mean values are 96.45 respectively.

The similar research increases prediction for recommendation systems to find drugs based on patient details with their data. With a hybrid database, the chances for correct prediction is also greatly increased (Liu et al. 2015). This model has a slow processing rate with better accuracy (Ng and Linn 2017). The slow processing rate is due to the usage of a large database but in the case of a smaller database, both the processing and accuracy are faster and better. The opposite problem's complexity will be reduced once a model is built (Sumner et al. 2004). Despite the fact that many researchers have discovered various prediction models, many of them are unable to accurately predict better drugs for patients (Eitrich et al. 2007). Many applications can be developed to predict accurately for sensitivity from various platforms.

The limitations of a significant amount of computing power as well as resources because it constructs several trees and combines their outcomes. It also takes a long time to train because it uses several decision trees to select the class. It also lacks interpretability due to the ensemble of decision trees and fails to evaluate the significance of each variable. The most common reasons for patient non-compliance to medications are intentional and include: high drug costs, fear of

adverse events, being prescribed multiple medications, and experiencing either instant relief or medication ineffectiveness leading to self-discontinuation of medications.

5. Conclusion

From this study of college recommendation systems, the mean accuracy of random forest classifier is 96.45% whereas content-based Novel Adaboost Ensemble Classifier has a higher mean accuracy of 98.47%. Hence it is inferred that content-based Novel Adaboost Ensemble Classifiers appear to be better in accuracy when compared to random forest classifiers.

Declarations

Conflict of Interest

No conflict of interest in this manuscript.

Authors' Contribution

Author MK was involved in data collection, data analysis, and manuscript writing. Author RK was involved in conceptualization, data validation, and critical reviews of the manuscript.

Acknowledgment

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete this study.

1. Best Enlist, Chennai.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

6. References

Anupong, Wongchai, Lin Yi-Chia, Mukta Jagdish, Ravi Kumar, P. D. Selvam, R. Saravanakumar, and Dharmesh Dhabliya. 2022. "Hybrid Distributed Energy Sources Providing Climate Security to the Agriculture Environment and Enhancing the Yield." Sustainable Energy Technologies and Assessments. https://doi.org/10.1016/j.seta.2022.102142.

Bharathiraja, B., J. Jayamuthunagai, R. Sreejith, J. Iyyappan, and R. Praveenkumar. 2022. "Techno Economic Analysis of Malic Acid Production Using Crude Glycerol Derived from Waste Cooking Oil." Bioresource Technology 351 (May): 126956.

Breiman, Leo. 2017. Classification and Regression Trees. Routledge.

Cano, Gaspar, Jose Garcia-Rodriguez, Alberto Garcia-Garcia, Horacio Perez-Sanchez, Jón Atli Benediktsson, Anil Thapa, and Alastair Barr. 2017. "Automatic Selection of Molecular Descriptors Using Random Forest: Application to Drug Discovery." Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2016.12.008.

Cook, Jack, William Addicks, and Yunhui Henry Wu. 2008. "Application of the Biopharmaceutical Classification System in Clinical Drug Development--an Industrial View." The AAPS Journal 10 (2): 306–10.

Eitrich, T., A. Kless, C. Druska, W. Meyer, and J. Grotendorst. 2007. "Classification of Highly Unbalanced CYP450 Data of Drugs Using Cost Sensitive Machine Learning Techniques." Journal of Chemical Information and Modeling 47 (1): 92–103.

Jothi, K. Jeeva, K. Jeeva Jothi, S. Balachandran, K. Mohanraj, N. Prakash, A. Subhasri, P. Santhana Gopala Krishnan, and K. Palanivelu. 2022. "Fabrications of Hybrid Polyurethane-Pd Doped ZrO2 Smart Carriers for Self-Healing High Corrosion Protective Coatings." Environmental Research. https://doi.org/10.1016/j.envres.2022.113095.

Kale, Vaibhav Namdev, J. Rajesh, T. Maiyalagan, Chang Woo Lee, and R. M. Gnanamuthu. 2022. "Fabrication of Ni–Mg–Ag Alloy Electrodeposited Material on the Aluminium Surface Using Anodizing Technique and Their Enhanced Corrosion Resistance for Engineering Application." Materials Chemistry and Physics. https://doi.org/10.1016/j.matchemphys.2022.125900.

Kolárik, Corinna, Martin Hofmann-Apitius, Marc Zimmermann, and Juliane Fluck. 2007. "Identification of New Drug Classification Terms in Textual Resources." Bioinformatics 23 (13): i264–72.

Lind, Alex P., and Peter C. Anderson. 2019. "Predicting Drug Activity against Cancer Cells by Random Forest Models Based on Minimal Genomic Information and Chemical Properties." PloS One 14 (7): e0219774.

Liu, Zhongyang, Feifei Guo, Jiangyong Gu, Yong Wang, Yang Li, Dan Wang, Liang Lu, Dong Li, and Fuchu He. 2015. "Similarity-Based Prediction for Anatomical Therapeutic Chemical Classification of Drugs by Integrating Multiple Data Sources." Bioinformatics 31 (11): 1788–95.

Maeda, Kazuya, Akihiro Hisaka, Kiyomi Ito, Yoshiyuki Ohno, Akihiro Ishiguro, Reiko Sato, and Naomi Nagai. 2021. "Classification of Drugs for Evaluating Drug Interaction in Drug Development and Clinical

Management." Drug Metabolism and Pharmacokinetics. https://doi.org/10.1016/j.dmpk.2021.100414.

Page, N., M. T. Baysari, and J. I. Westbrook. 2017. "A Systematic Review of the Effectiveness of Interruptive Medication Prescribing Alerts in Hospital CPOE Systems to Change Prescriber Behavior and Improve Patient Safety." International Journal of Medical Informatics. https://doi.org/10.1016/j.ijmedinf.2017.05.011.

Palanisamy, Rajkumar, Diwakar Karuppiah, Subadevi Rengapillai, Mozaffar Abdollahifar, Gnanamuthu Ramasamy, Fu-Ming Wang, Wei-Ren Liu, Kumar Ponnuchamy, Joongpyo Shim, and Sivakumar Marimuthu. 2022. "A Reign of Bio-Mass Derived Carbon with the Synergy of Energy Storage and Biomedical Applications." Journal of Energy Storage. https://doi.org/10.1016/j.est.2022.104422.

Ram, G. Dinesh, G. Dinesh Ram, S. Praveen Kumar, T. Yuvaraj, Thanikanti Sudhakar Babu, and Karthik Balasubramanian. 2022. "Simulation and Investigation of MEMS Bilayer Solar Energy Harvester for Smart Wireless Sensor Applications." Sustainable Energy Technologies and Assessments. https://doi.org/10.1016/j.seta.2022.102102.

Riddick, Gregory, Hua Song, Susie Ahn, Jennifer Walling, Diego Borges-Rivera, Wei Zhang, and Howard A. Fine. 2011. "Predicting in Vitro Drug Sensitivity Using Random Forests." Bioinformatics. https://doi.org/10.1093/bioinformatics/btq628.

Shi, Han, Simin Liu, Junqi Chen, Xuan Li, Qin Ma, and Bin Yu. 2019. "Predicting Drug-Target Interactions Using Lasso with Random Forest Based on Evolutionary Information and Chemical Structure." Genomics 111 (6): 1839–52.

Sumathy, B., Anand Kumar, D. Sungeetha, Arshad Hashmi, Ankur Saxena, Piyush Kumar Shukla, and Stephen Jeswinde Nuagah. 2022. "Machine Learning Technique to Detect and Classify Mental Illness on Social Media Using Lexicon-Based Recommender System." Computational Intelligence and Neuroscience 2022 (February): 5906797.

Sumner, B. E. H., L. A. Cruise, D. A. Slattery, D. R. Hill, M. Shahid, and B. Henry. 2004. "Testing the Validity of c-Fos Expression Profiling to Aid the Therapeutic Classification of Psychoactive Drugs." Psychopharmacology 171 (3): 306–21.

Thanigaivel, Sundaram, Sundaram Vickram, Nibedita Dey, Govindarajan Gulothungan, Ramasamy Subbaiya, Muthusamy Govarthanan, Natchimuthu Karmegam, and Woong Kim. 2022. "The Urge of Algal Biomass-Based Fuels for Environmental Sustainability against a Steady Tide of Biofuel Conflict Analysis: Is Third-Generation Algal Biorefinery a Boon?" Fuel. https://doi.org/10.1016/j.fuel.2022.123494.

Ubels, Joske, Tilman Schaefers, Cornelis Punt, Henk-Jan Guchelaar, and Jeroen de Ridder. 2020. "RAINFOREST: A Random Forest Approach to Predict Treatment Benefit in Data from (failed) Clinical Drug Trials." Bioinformatics 36 (Suppl_2): i601–9.

Urista, Diana V., Diego B. Carrué, Iago Otero, Sonia Arrasate, Viviana F. Quevedo-Tumailli, Marcos Gestal, Humbert González-Díaz, and Cristian R. Munteanu. 2020. "Prediction of Antimalarial Drug-Decorated Nanoparticle Delivery Systems with Random Forest Models." Biology 9 (8). https://doi.org/10.3390/biology9080198.

Vickram, Sundaram, Karunakaran Rohini, Krishnan Anbarasu, Nibedita Dey, Palanivelu Jeyanthi, Sundaram Thanigaivel, Praveen Kumar Issac, and Jesu Arockiaraj. 2022. "Semenogelin, a Coagulum Macromolecule Monitoring Factor Involved in the First Step of Fertilization: A Prospective Review." International Journal of Biological Macromolecules 209 (Pt A): 951–62.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Algal Biofuels: Technological Perspective on Cultivation, Fuel Extraction and Engineering Genetic Pathway for Enhancing Productivity." Fuel. https://doi.org/10.1016/j.fuel.2022.123814.

TABLES AND FIGURES

Table 1. Pseudocode for Adaboost Ensemble Classifier

//I: Input dataset records
1. Import the required packages.
2. Convert the Data Sets into numerical values after the extraction feature.
3. Assign the data to X_train, Y_train, X_test, and Y_test variables.

4. Using the train_test_split() function, pass the training and testing variables.
5. Give test_size and the random_state as parameters for splitting the data using GBEC training.
6. Calculate the accuracy of the model.
OUTPUT: Accuracy

Table 2. Pseudocode for Random forest classifier

INPUT: dataset records
1. Read and test data for enhancement for Classifying Drugs images
2. Extract Classifying Drugs attributes for enhancement
3. Extract attributes to enhance Classifying Drugs data
4. Input Classifying Drugs Based On Patient Detail
5. Apply Random forest classifier
6. Learn user preferences
7. Return accuracy
OUTPUT: Enhanced Accuracy in Classifying Drugs Based On Patient Detail

Table 3. Group, Accuracy and Loss value for classifying drugs based on patient detail

SI.NO	**Name**	**Type**	**Width**	**Decimal**	**Columns**	**Measure**	**Role**
1	Group	Numeric	8	0	31	Nominal	Input
2	Accuracy	Numeric	8	4	31	Scale	Input
3	Loss	Numeric	8	2	31	Scale	Input

Table 4**.** Group Statistical analysis for Novel Adaboost Ensemble Classifier and Random forest classifier Mean, Standard deviation and Standard error mean are determined

	Group	**N**	**Mean**	**Std Deviation**	**Std Error Mean**
Accuracy	**Novel Adaboost Ensemble Classifier**	10	98.1490	1.06584	0.33705
	Random forest Classifier	10	94.5760	1.32623	0.41939
Loss	**Novel Adaboost Ensemble Classifier**	10	1.8510	1.06584	0.33705
	Random forest Classifier	10	5.4240	1.32623	0.41939

Table 5. Independent sample T-test is performed on two groups for significance and standard error determination. P-value is greater than 0.05 (0.331) and it's considered to be statistically insignificant with 95% confidence interval

		Levene's Test for Equality of Variance		**T-Test for Equality of Means**						
		F	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean Difference**	**Std. Error Difference**	**95% Confidence Interval of the Difference**	
									Lower	**Upper**
Accuracy	**Equal variances assumed**	0.997	0.331	4.782	18	0.000	2.57300	0.53804	1.44261	3.70339
	Equal variances not assumed			4.782	17.204	0.000	2.57300	0.53804	1.43885	3.70715

Error	**Equal variances assumed**	0.997	0.331	-4.782	18	0.000	-2.57300	0.53804	-3.70339	-1.44261
	Equal variances not assumed			-4.782	17.204	0.000	-2.57300	0.53804	-3.70715	-1.43885

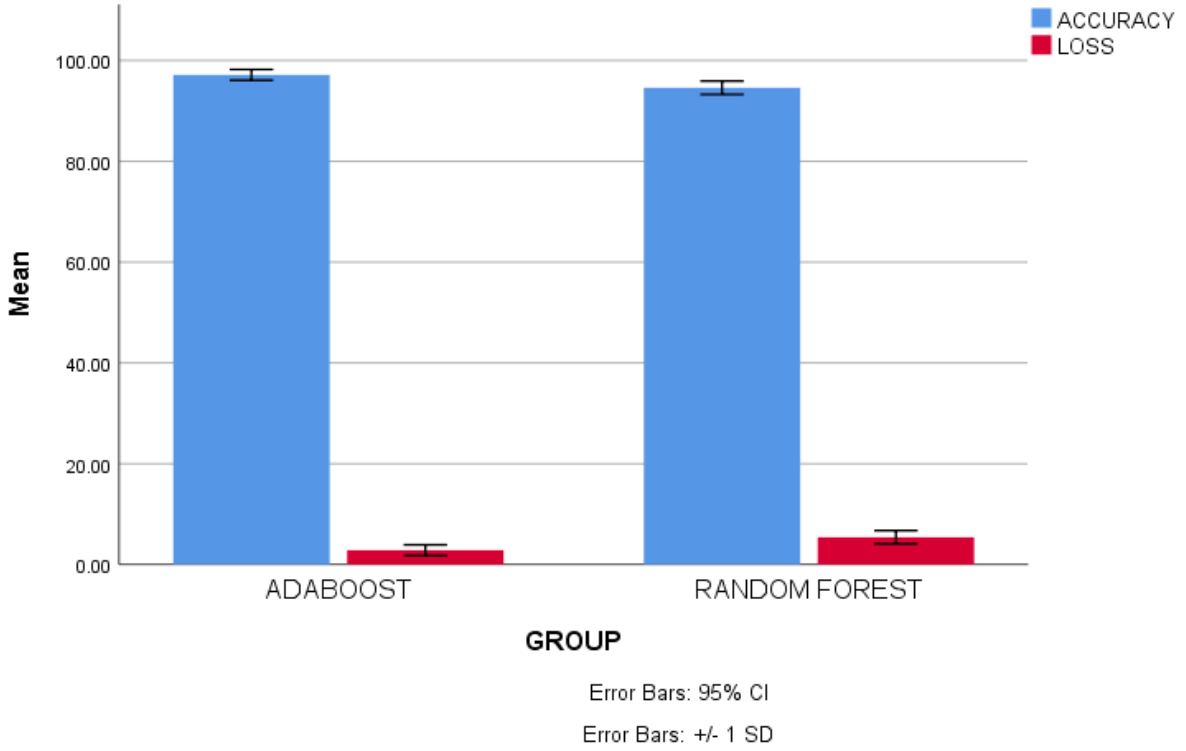

Fig. 1. Comparison of Novel Adaboost Ensemble Classifier and Random forest classifier in terms of mean accuracy. The mean accuracy of the Novel Adaboost Ensemble Classifier is better than the random forest classifier. The standard deviation of the Novel Adaboost Ensemble Classifier is slightly better than the random forest classifier. X-Axis: Novel Adaboost Ensemble Classifier vs Random forest classifier. Y-Axis: Mean accuracy of detection ± 1 SD.