Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. 1 (2022) 64 - 69 ISSN Media Electronic: 2580-0760

Implementation of Ensemble Method in Schizophrenia Identification Based on Microarray Data

Diya Namira Purba[1], Fhira Nhita[2], Isman Kurniawan[3]
[1,2,3]School of Computing, Telkom University
[3]Research Center of Human Centric Engineering, Telkom University
[1]namirapurba@student.telkomuniversity.ac.id, [2]fhiranhita@telkomuniversity.ac.id, [3]ismankrn@telkomuniversity.ac.id

Abstract

Schizophrenia is a chronic mental illness that leads the patient to hallucinations and delusions with a prevalence of 0.4% worldwide. The importance early detection of Schizophrenia is tracking the pre-syndrome of Schizophrenia during the active phase, and could reduce psychosis symptomatic. However, the method sometimes cannot detect the symptoms accurately. As an alternative, machine learning can be implemented on microarray data for early detection. This study aimed to implement three ensemble methods, i.e., Random Forest (RF), Adaptive Boosting (AdaBoost), and Extreme Gradient Boosting (XGBoost) to identify Schizophrenia. Hyperparameter tuning was performed to improve the performance of the models. Based on the results, we found that the model 6, which is developed by the XGBoost method, performs better than other models with the value of accuracy and F1-score are 0.87 and 0.87, respectively.

Keywords: ensemble method, microarray, schizophrenia, disease detection

1. Introduction

Schizophrenia is a chronic mental illness that leads patients into hallucination and delusion with a prevalence of 0.4% worldwide[1]. The symptoms of these diseases are divided into positive, negative, and cognitive symptoms. Positive symptoms are additional brain activities that are not supposed to be exist, for example, hallucination and delusion[2]. Meanwhile, negative symptoms are the opposite things that are supposed to be exist but are not present, namely apathy and lack emotion [2]. Cognitive symptoms are related to disturbances in memory and difficulty concentrating. The active phase for schizophrenia symptoms is one month and will occur with the mood episode followed by two weeks of delusions or hallucinations[2].

The symptoms of Schizophrenia have a poorer outcome than other psychotic and nonpsychotic diseases[3]. This outcome will lead patients to a danger of suicide and early death of Schizophrenia[4]. Diagnostic criteria for Schizophrenia is asking patients questions to elicit information such as duration of illness and clinical symptoms [5]. Early detection is essential to limit the morbidity of illness [3]. The is to track the pre-syndrome of Schizophrenia during the active phase [3]. This early detection method could reduce the duration of psychosis symptomatic than others without early detection methods [3]. However, the method cannot detect the symptoms accurately [3].

Recently, machine learning has been commonly used on microarray data for early detection of Schizophrenia. In 2019, Karthik and coworkers developed a Deep Neural Network for predicting Bipolar and Schizophrenia [4]. This study achieved 95.65% accuracy on Schizophrenia[4]. In 2016, a study about Schizophrenia microarray gene expression data was proposed by Aristotelis and coworkers[6]. They used a support vector machine, random forests, and an extremely randomized tree classifier. Models' performance was evaluated using accuracy, precision, the area under the curve (AUC), and sensitivity[6]. They found that the best model is obtained from a random forest model with the values of accuracy, precision, sensitivity, and AUC are 0.83, 0.093, 0.89, and 0.98[6].

In 2017, Zhang and coworkers proposed a study about diagnosing Schizophrenia based on gene expression[7]. This study used a combined four microarray dataset from the GEO database and used five methods, i.e., Locally weighted learning (LWL), bayesian network,

Accepted: 12-01-2022 | Received in revised: 27-01-2022 | Published: 01-02-2022

nearest neighbor, naive Bayes, and J48[7]. The evaluation method using a cross-validated score achieved 100% accuracy, making the LWL algorithm the best method[7]. A study about the diagnosis of Schizophrenia based on gene expression in peripheral blood was proposed by Zhu and coworkers in 2020[8]. The data is taken from Guangxi Zhuang Region Brain Hospital and used five machine algorithms: artificial neural network, extreme gradient boosting, support vector machine, decision tree, and random forest[8]. This study shows that the support vector machine is the best method with the area under th e curve, sensitivity, and specificity equal to 0.993, 1.00, and 0.895[8].

One of the machine learning methods commonly used in prediction tasks is the ensemble method[9]. Ensemble learning is an effective algorithm that combines all learning algorithms to improve accuracy[9]. This algorithm techniques advantage that can alleviate small sample size problem by average and incorporate from the model to prevent overfitting from training data. Hence, the ensemble method is promising to be used to improve prediction accuracy in Schizophrenia identification[9].

In this study, we aim to predict Schizophrenia by implementing an ensemble method on microarray data. We used three ensemble methods to predict Schizophrenia, i.e., Random Forest, Adaptive Boosting, and Extreme Gradient Boosting.

2. Research Methods

2.1. Dataset

We used a microarray dataset is microarray from Geo Datasets with GSE code is GSE17612 [6]. The dataset is a 54, 675 brain postmortem gene expression derived from anterior prefrontal cortex consists of 51 samples with two classes, i.e., 28 Schizophrenia and 23 normal[6]. Then, the dataset is split into train and test set with the ratio of 70:30.

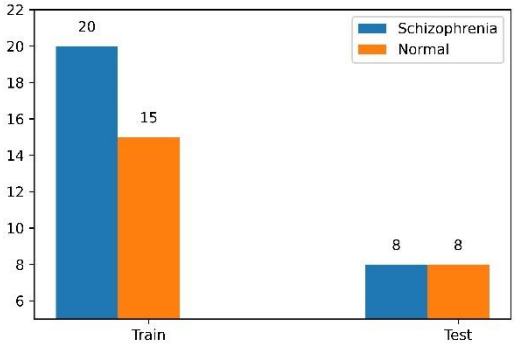


Figure 1. The number of samples in the test and train set

The comparison of samples number for each class in train and test are shown in Figure 1. In the train set, the number of Schizophrenia is equal to 20 and normal samples is equal to 15. The number of samples is not the same it means the difference is not too significant. In the test set, the number of Schizophrenia and normal samples has the same value. So, the samples are balanced. For the distribution of samples, we calculated the two principal components from the dataset by using principal component analysis (PCA) method as shown in Figure 2.

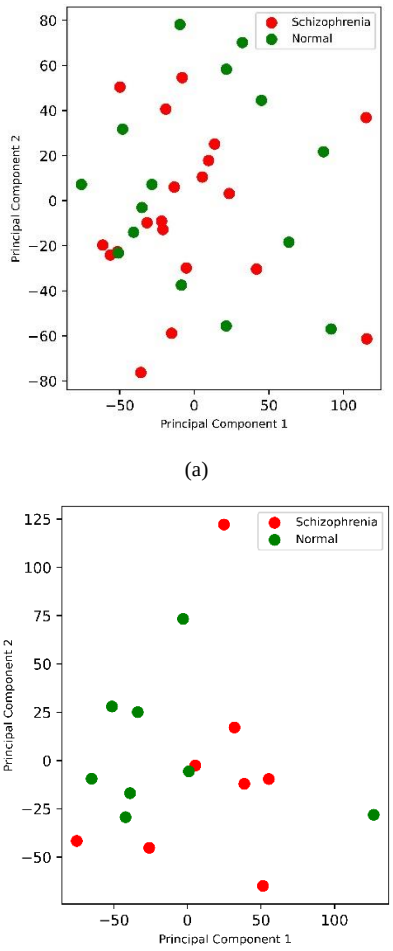


Figure 2. Sample distribution in (a) train Set and (b) test set

From the principal component analysis results as shown in figure 2, In the train set, we found that the distribution of samples is still in the same region. It means the classification process cannot perform well in the train set. As for the test set, we found that the sample distribution is almost separated, and the classification can perform well in the test set.

2.2. Feature Selection

We reduced the number of features by eliminating a feature with a value of standard deviation equal to 0.5 and a variance threshold equal to 0.5. Then, the feature selection is followed by calculating statistical parameters that represent the correlation of individual features to the target. Here, we consider two statistical

DOI: https://doi.org/10.29207/resti.v6i1.3788

parameters, i.e., mutual information and analysis of variance (ANOVA). Mutual information (MI) is a feature selection that measures the relationship between two variables [10]. Suppose a class *c* and probabilities define as *p(c)* and *p(t)* [10]. Then, mutual information, *I (t,c)* is defined as follows:

$$I(t,c) = \log \frac{I\ (t,c)}{p(t) \times p(c)} = \log \frac{p(t\ \wedge c)}{p(t) \times p(c)} \quad (1)$$

Analysis of variance (ANOVA) is a filter feature selection method that can be used in multi-class features by measuring their variance[11]. The formula is defined as follows:

$$F = \frac{\sum_{C=1}^{C} \left(\bar{f}_c - \bar{f} \right)^2 / (c-1)}{\sum_{C=1}^{C} \sum_{c=1}^{N_c} \left(f_{ci} - \bar{f}_c \right)^2 / N - c} \quad (2)$$

2.3. Prediction Model

We utilized six models by combining the combination of three ensemble methods i.e., Random Forest, AdaBoost, and XGBoost and wo statistical parameters i.e., Mutual Information (MI) and Analysis of variance (ANOVA) as shown in Table 1

Table 1. The Model and Feature Selection Criteria

Model No	Criteria	Method
1	Mutual Information	Random Forest
2	Anova	Random Forest
3	Mutual Information	AdaBoost
4	Anova	AdaBoost
5	Mutual Information	XGBoost
6	Anova	XGBoost

Random forest is a tree-based ensemble technique that depends on random variables and was introduced by Leo Breiman [8]. This algorithm can be used for categorical variables referred to in classification and continuous response referred to as regression. The samples data and randomly construct decision trees to avoid overfit from train data [12]. Random Forest's additional features include measuring variable importance, missing value imputation, outlier detection, etc [13].

XGboost is a classification algorithm based on an ensemble of classification and regression trees optimized by gradient boosting [14]. This algorithm has achieved considerably in classification because of its performance on label-imbalanced data[15]. XGboost can handle large-scale machine learning tasks can be proved by performance superiority [16]

AdaBoost is a learn a set of classifiers to produce the final classifier. The weak classifiers are obtained with the use of training data, with the weight depending on the accuracy [16]. This algorithm's main idea is to use a weighted version from train data instead of random sampling [16]. It can adaptively adjust the weight from the classifier group and give a better result because of the diversity of each group [12]. The difference between Random Forest with XGBoost and AdaBoost is the tree in Random Forest work sequentially [13]. However, the difference between AdaBoost and XGBoost is in parallel processing. The parallel processing in XGBoost is to increase performance. Then, we try to improve the performance of the models by conducting a hyperparameter tuning procedure. The parameters involved in the hyperparameter tuning for each model are presented in Table 2

Table 2. Hyperparameter Tuning Ranges

Method	Parameters	Ranges
Random Forest	N estimators	[200, 300, 400, 500]
	Min_samples leaf	[2, 3, 4, 5]
	Min_samples split	[4, 6, 8, 10]
	Criterion	[‘gini, ‘entropy]
AdaBoost		[0.1, 1.0]
	N estimators	[150, 200, 250, 500]
	Algorithm	[‘SAMME’, ‘SAMME.R]
XGBoost	Learning rate	[0.6, 0.7, 0.8, 0.9]
	Max depth	[0.7, 0.8]

2.4. Model Validation

We evaluated the models by calculating several validation parameters to measure the performance of each method. Those parameters are Accuracy (Q), Precision (PR), Recall (RC), and Receiver operating characteristic curve (ROC) and formulated in Equation (3) – (6)

$$Q = \frac{TP\ + TN}{TP + FP + TN + FN} \quad (3)$$

$$RC = \frac{TP}{TP + FN} \quad (4)$$

$$PR\ = \frac{TP}{TP + FP} \quad (5)$$

$$F1\ = \frac{2\ \times\ (\ PR\ \times\ RC)}{PR + RC} \quad (6)$$

ROC graph is constructed by the value of true positive rate and false-positive rate with the sample proportion. This model is also evaluated by using the area under the curve (AUC)[12].

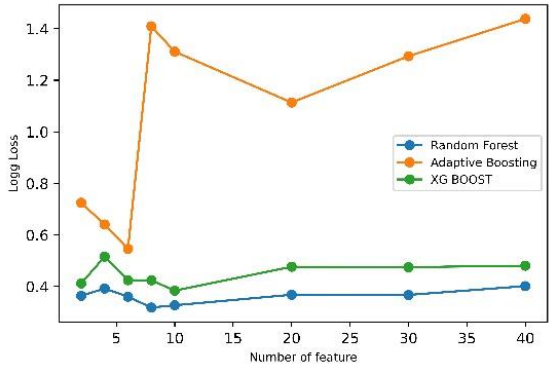
3. Result and Discussions

3.1 Feature Selection

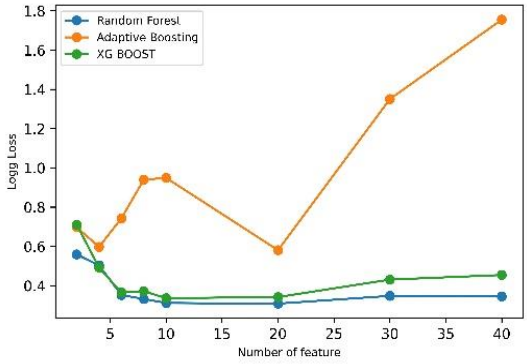
The impact of feature number on log loss for each statistic is shown in Figure 3. Meanwhile, the summary of the feature selection process is presented in Table 3.

DOI: https://doi.org/10.29207/resti.v6i1.3788

As for Mutual Information (MI), the log loss of random forest and AdaBoost is increases when the number of features is less than five, while XGBoost log loss is increased when the number of features is greater than five. For analysis of variance (ANOVA), the log loss of random forest and AdaBoost has an upwards when the number of features is less than five and downwards when the feature number is greater than five. At the same time, XGBoost has an upwards when the number of features is greater than 40 and downwards when number of features is less than five.



(a)



(b)

Figure 3. The contribution of feature number to log loss by using a statistical parameter (a) MI and (b) ANOVA

From the feature selection process, we obtain the average of the standard deviation of each model, and the result is that all models have the same average score except for model 4.

Table 3. Summary of Feature Selection Criteria

Model	Number of feature	Log Loss	Avg.score± Stde
1	2	0.40	±0.03
2	2	0.55	±0.08
3	40	1.43	±0.33
4	40	1.75	±0.38
5	2	0.48	±0.04
6	4	0.71	±0.11

Random forest and AdaBoost has same optimal number of features from mutual information and analysis of variance. But, XGBoost has different number of features from the two parameters. The highest of log loss value is obtained by AdaBoost method with the value of average score is ±0.38 as shown in Table 3.

3.2 Model Development

Hyperparameter tuning is used to obtain the best parameters for all models. In the RF models, we found that the criterion for model 1 and model 2 is gini, the minimum samples split, and n_estimators are all different. In AdaBoost models have same produced the same result for all model parameters. XGBoost also produces the same parameters for all models. The result of hyperparameter tuning is presented in Table 4.

Table 3. Summary of Hyperparameter Tuning Result

Method	Parameters	Best Values
Random Forest	N estimators	[300]
	Min samples leaf	[8., 2]
	Min samples split	[6, 4]
	Criterion	[‘gini’]
AdaBoost	Learning rate	[1.0]
	N estimators	[500]
	Algorithm	[‘SAMME’]
XGBoost	Learning rate	[0.6]
	Max depth	[0.7]

We present a comparison of the F1 score between non-tuned and tuned models in Figure 4. The result shows there is no difference between the tuned and non-tuned models except for model 4. The improvement is occurred in the model 4 because model 4 has the greatest value of Average score

3.3 Model Validation

We consider the F1- score as an overall measurement to determine the best model. The model validation results are presented in Table 5. In the training set, we found that the best F1-score is from models 3, 4, and 5, it's because models 3, 4, and 5 have a high log loss score are 1.78, 1.41, and 0.71.

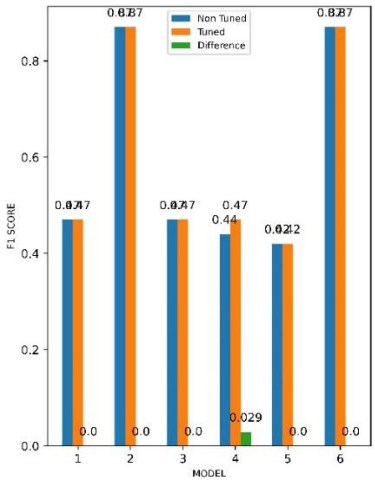


Figure 4. Comparison of F1-Score of Tuned and Non-Tuned Models

DOI: https://doi.org/10.29207/resti.v6i1.3788

In the test set, we found that the best F1-score is models 2 and 6. The performance can be affected by the low number feature. Since the model 3 and 6 number of features is less than 40. Model 2 is the worst one since it has a larger number of features. The summary of validation is present in Table 5.

Table 4. Summary of Validation Result

Model	TP	FP	TN	FN	Q	PR	RC	F1	AUC_ROC
Train									
1	13	2	18	2	0.85	0.90	0.90	0.90	0.96
2	13	2	18	2	0.85	0.90	0.90	0.90	0.98
3	**15**	**0**	**20**	**0**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
4	**15**	**0**	**20**	**0**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
5	**15**	**0**	**20**	**0**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
6	14	1	20	0	0.97	0.95	0.99	0.94	0.96
Test									
1	3	5	4	4	0.43	0.44	0.50	0.47	0.56
2	**7**	**1**	**7**	**1**	**0.87**	**0.87**	**0.87**	**0.80**	**0.84**
3	3	5	4	4	0.43	0.40	0.50	0.47	0.46
4	2	4	4	6	0.37	0.40	0.50	0.40	0.31
5	5	3	3	5	0.50	0.50	0.37	0.42	0.43
6	**7**	**1**	**7**	1	**0.87**	**0.87**	**0.87**	**0.80**	**0.86**

The numbers with bold print represent the best values amongst all models.

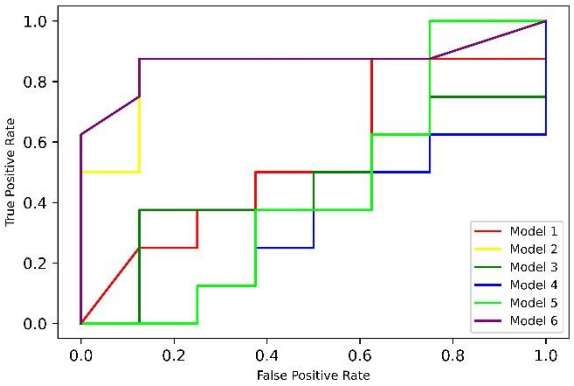


Figure 4. ROC Curves

To determine the best model between Model 2 and Model 6, we consider the value of AUC. We present that the AUC was calculated by using the ROC in Figure 5. The ROC value of model 6 (0.86) is better than model 2 (0.84). It can be confirmed that model 6 prediction ability is better than model 2 since XGBoost advantages that is can perform well in label-imbalanced data.

We also compared our results with the results of ref [6]. As shown in Table 6. We presented only the best two models for each study. We found that the random forest method that we proposed had an increase in performance in the value of accuracy but had a decrease in precision and recall value. In the second method, the method that we propose, namely the XGBoost method, has good performance compared to the value of the AdaBoost method from the study we compared with the better accuracy, precision, and recall values, are 0.87, 0.87, and 0.87

Table 6. Comparison of The Proposed Method with Previous Studies

Author	Method	Q	PR	RC
Chatziiannou, 2016 [6]	Random Forest	0.83	0.93	0.89
	AdaBoost	078	0.80	0.77
Our Proposed Method	Random Forest	0.87	0.87	0.87
	XGBoost	0.87	0.87	0.87

4. Conclusions

We have developed three ensemble methods i.e., random forest, AdaBoost, and XGBoost to identify Schizophrenia. The number of features was reduced with standard deviation and variance threshold. Then, the feature selection is followed by calculating two statistical parameters are MI and ANOVA. We found that the performance improvement of the models by hyperparameter tuning procedure is not too significant. According to the validation results, we found that the best model is model 6 which is developed by XGBoost, with the value of F1-score and AUC are 0.86 and 0.87, respectively.

Reference

[1] M. Dong *et al.*, "Quality of Life in Schizophrenia: A Meta-Analysis of Comparative Studies," *Psychiatr Q*, vol. 90, no. 3, pp. 519–532, Sep. 2019, doi: 10.1007/s11126-019-09633-4.

[2] N. Ueda, K. Maruo, and T. Sumiyoshi, "Positive symptoms and time perception in schizophrenia: A meta-analysis," *Schizophrenia Research: Cognition*, vol. 13, pp. 3–6, Sep. 2018, doi: 10.1016/j.scog.2018.07.002.

[3] J. A. Lieberman, S. A. Small, and R. R. Girgis, "Early Detection and Preventive Intervention in Schizophrenia: From Fantasy to Reality," *AJP*, vol. 176, no. 10, pp. 794–810, Oct. 2019, doi: 10.1176/appi.ajp.2019.19080865.

[4] S. Karthik and M. Sudha, "Predicting bipolar disorder and schizophrenia based on non-overlapping genetic phenotypes using deep neural network," *Evol. Intel.*, vol. 14, no. 2, pp. 619–634, Jun. 2021, doi: 10.1007/s12065-019-00346-y.

[5] J. Parnas and M. Zandersen, "Self and schizophrenia: current status and diagnostic implications," *World Psychiatry*, vol. 17, no. 2, pp. 220–221, Jun. 2018, doi: 10.1002/wps.20528.

[6] A. Chatziioannou, "Studying Microarray Gene Expression Data of Schizophrenic Patients for Derivation of a Diagnostic Signature through the Aid of Machine Learning," *BBIJ*, vol. 4, no. 5, Sep. 2016, doi: 10.15406/bbij.2016.04.00106.

[7] H. Zhang, Z. Xie, Y. Yang, Y. Zhao, B. Zhang, and J. Fang, "The Correlation-Base-Selection Algorithm for Diagnostic Schizophrenia Based on Blood-Based Gene Expression Signatures," *BioMed Research International*, vol. 2017, pp. 1–7, 2017, doi: 10.1155/2017/7860506.

[8] L. Zhu *et al.*, "The machine learning algorithm for the diagnosis of schizophrenia on the basis of gene expression in peripheral blood," *Neuroscience Letters*, vol. 745, p. 135596, Feb. 2021, doi: 10.1016/j.neulet.2020.135596.

[9] A. Motwani, G. Bajaj, and S. Mohane, "Predictive Modelling for Credit Risk Detection using Ensemble Method," *ijcse*, vol. 6, no. 6, pp. 863–867, Jun. 2018, doi: 10.26438/ijcse/v6i6.863867.

[10] A. Nagpal and V. Singh, "A Feature Selection Algorithm Based on Qualitative Mutual Information for Cancer Microarray Data," *Procedia Computer Science*, vol. 132, pp. 244–252, 2018, doi: 10.1016/j.procs.2018.05.195.

[11] H. Nasiri and S. A. Alavi, "A Novel Framework Based on Deep Learning and ANOVA Feature Selection Method for Diagnosis of COVID-19 Cases from Chest X-Ray Images,"

DOI: https://doi.org/10.29207/resti.v6i1.3788

Computational Intelligence and Neuroscience, vol. 2022, pp. 1–11, Jan. 2022, doi: 10.1155/2022/4694567.

[12] I. Kurniawan, M. Rosalinda, and N. Ikhsan, "Implementation of ensemble methods on QSAR Study of NS3 inhibitor activity as anti-dengue agent," *SAR and QSAR in Environmental Research*, vol. 31, no. 6, pp. 477–492, Jun. 2020, doi: 10.1080/1062936X.2020.1773534.

[13] Z. Xu and Z. Wang, "A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier," in *2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI)*, Guilin, China, Jun. 2019, pp. 278–283. doi: 10.1109/ICACI.2019.8778622.

[14] G. N. Dimitrakopoulos, A. G. Vrahatis, V. Plagianakos, and K. Sgarbas, "Pathway analysis using XGBoost classification in Biomedical Data," in *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, Patras Greece, Jul. 2018, pp. 1–6. doi: 10.1145/3200947.3201029.

[15] C. Wang, C. Deng, and S. Wang, "Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost," *Pattern Recognition Letters*, vol. 136, pp. 190–197, Aug. 2020, doi: 10.1016/j.patrec.2020.05.035.

[16] P. Yang, Y. Hwa Yang, B. B. Zhou, and A. Y. Zomaya, "A Review of Ensemble Methods in Bioinformatics," *CBIO*, vol. 5, no. 4, pp. 296–308, Dec. 2010, doi: 10.2174/157489310794072508.

DOI: https://doi.org/10.29207/resti.v6i1.3788