1. Nur Farahaina Idris[1], 2. Mohd Arfian Ismail[1,2]

Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Pahang, Malaysia (1)
Pusat Kecemerlangan Kecerdasan Buatan & Sains Data, Universiti Malaysia Pahang Al-Sultan Abdullah, Lebuh Persiaran Tun Khalil Yaakob, Gambang, Pahang, Malaysia (2)

doi:10.15199/48.2024.01.21

A review of homogenous ensemble methods on the classification of breast cancer data

Abstract. *In the last decades, emerging data mining technology has been introduced to assist humankind in generating relevant decisions. Data mining is a concept established by computer scientists to lead a secure and reliable classification and deduction of data. In the medical field, data mining methods can assist in performing various medical diagnoses, including breast cancer. As evolution happens, ensemble methods are being proposed to achieve better performance in classification. This technique reinforced the use of multiple classifiers in the model. The review of the homogenous ensemble method on breast cancer classification is being carried out to identify the overall performance. The results of the reviewed ensemble techniques, such as Random Forest and XGBoost, show that ensemble methods can outperform the performance of the single classifier method. The reviewed ensemble methods have pros and cons and are useful for solving breast cancer classification problems. The methods are being discussed thoroughly to examine the overall performance in the classification.*

Streszczenie. *W ostatnich dziesięcioleciach wprowadzono nową technologię eksploracji danych, która ma pomóc ludzkości w podejmowaniu odpowiednich decyzji. Eksploracja danych to koncepcja opracowana przez informatyków w celu zapewnienia bezpiecznej i niezawodnej klasyfikacji i dedukcji danych. W medycynie metody eksploracji danych mogą pomóc w przeprowadzaniu różnych diagnoz medycznych, w tym raka piersi. W miarę ewolucji proponuje się metody zespołowe, aby uzyskać lepszą skuteczność klasyfikacji. Technika ta wzmocniła zastosowanie w modelu wielu klasyfikatorów. Przeprowadzany jest przegląd jednorodnej metody zespołowej klasyfikacji raka piersi w celu określenia ogólnej skuteczności. Wyniki recenzowanych technik zespołowych, takich jak Random Forest i XGBoost, pokazują, że metody zespołowe mogą przewyższać skuteczność metody pojedynczego klasyfikatora. Omówione metody zespołowe mają zalety i wady i są przydatne w rozwiązywaniu problemów związanych z klasyfikacją raka piersi. Metody są szczegółowo omawiane w celu sprawdzenia ogólnej wydajności w klasyfikacji.* (***Przegląd jednorodnych metod zespołowych dotyczących klasyfikacji danych dotyczących raka piersi***)

Keywords: ensemble, bagging, boosting, breast cancer
Słowa kluczowe: zespół, pakowanie, wzmacnianie, rak piersi

Introduction

Breast cancer is the most common disease women suffer worldwide and the second largest cause of death among women from cancer [1]. Thus, advanced technology is being applied to improve the speed of breast cancer diagnosis. One of the advanced technologies being utilised in the medical field is data mining. Data mining can be known as acquiring information from data to generate decisions in the subsequent processes using algorithms, database technologies and artificial intelligence [2]. Many notable classification methods, such as decision trees, support vector machines, logistic regression, Naïve Bayes and k-Nearest neighbours, can assist in real-world breast cancer diagnosis [3]. For example, Naïve Bayes acquired 91.18% accuracy in the classification of the WDBC(Diagnostic) data [4]. The existing study shows that data mining methods have high capability and efficiency in diagnosis. Nevertheless, a fully optimised method is needed to acquire the most accurate diagnosis of breast cancer disease. Reliable data mining methods with higher error and outlier tolerance, better generalisation ability, stability and precision are essential to generate optimal classification results that can assist physicians or medical staff in diagnosing breast cancer. Hence, over the years, computer scientists developed many advanced intelligence techniques to acquire the most optimal results in classification. The ensemble method is one of the most common techniques implemented in the breast cancer domain to increase the performance in terms of classification accuracy, stability and fault tolerance of the methods [5], [6].

Classification

Classification is the data mining process involving performing knowledge extraction tasks [7]. It is defined as categorising undetermined data into specific classes based on attributes [8]. However, the incapability to produce accurate outcomes due to high volumes of data and disparity features becomes a limitation in the classification [9]. Other issues that can arise during classification include feature selection, memory consumption, training time and low test accuracy [10]. Therefore, prominent data mining techniques are being hybridised with the ensemble technique to improve the classification. The data mining techniques would be known as the base classifiers in the ensemble method term. Classification is widely used in the breast cancer domain for diagnosing purposes [11]–[13]. For example, the study by [14] utilised AdaBoost and ANN to detect microcalcifications in breast cancer.

Research Method: Ensemble Method

Ensemble methods are computational learning similar to human behaviour of seeking various opinions before producing crucial decisions [15]. Generally, the ensemble technique applies the multiple classifiers system in which the outcomes of the base classifiers are combined [16]. The previous experimental studies show that the ensemble technique improved the classifier performance as the classifiers' errors are negatively correlated. Furthermore, several other theories explain the successful application of ensemble methods in different domains. For example, Allwein, Schapire and Singer explicated the enhancement in the generalisation ability of ensembles in the framework of large margin classifiers [17], [18] while Breiman explained the improvement due to bias and variance analysis [19].

Ensemble methods can be divided into two, which are homogenous and heterogeneous [20]. Homogenous ensemble methods consist of methods such as Random Forest [21], Bagging [22] and Boosting [23]. The homogenous ensemble methods only consist of a single type of base classifier, as the model would generate the same specific type of base classifiers iteratively. Meanwhile, the heterogenous ensemble, like stacking, can combine different types of base classifiers with different natures (i.e. Combination of the decision tree and KNN) [20], [24]. This

study reviewed homogenous ensembles like Random Forest, Bagging, XGBoost, and AdaBoost, which are notable in the breast cancer domain [16], [25].

BAGGING

Bagging, also known as bootstrap aggregation, was first proposed by Leo Breiman in 1994 [22]. This method aims to reduce the prediction error in the data mining algorithm [26]. Conceptually, bagging implements bootstrapping of samples from the training data, and each base classifier would utilise the bootstrap sample. In bagging, there are two ways two determine the prediction for test data: majority voting of the base classifiers, which class label with the most votes are selected for the test pattern (classification) and averaging all the predictions from the base classifiers (regression) [27]. The resample data *k* may differ from the original sample size *n*, and the resampling process can be done either with replacement or without replacement. [26]. Among the advantages of bagging is reducing the variance of the model without affecting much of the bias [28]. Figure 1 expresses the schematic diagram of the bagging process using the cross-validation technique. According to [29], bagging acquired 94.5169% accuracy when classifying with Wisconsin Diagnostic Breast Cancer (WDBC) data. Then, an existing study by [11] conveys that bagging got 71.5517 % accuracy with Breast Cancer Coimbra data.

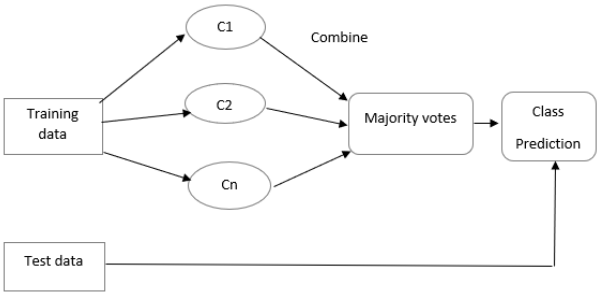

Fig. 1: The schematic diagram of bagging using the cross-validation technique.

RANDOM FOREST

Random forest is an ensemble technique that was also introduced by Leo Breiman with the application of the CART decision tree learning [21]. Random Forest applied the bagging technique that includes both bootstrap and aggregation [30]. It was built using two primary concepts: a subset of samples that undergo bootstrapping and a random subset of features. The random forest consists of the generation of multiple decision trees, implemented using a subset of bootstrapped training samples. The difference between random forest and bagging is that only the best feature (not all features used) was selected among the random subset of features at each split node.

The bootstrapped sample data is divided into *m* feature samples to train a single decision tree. Normally, in the random forest, to acquire the *m* value of the feature, the best-split point is computed from randomly selected$\sqrt{}$ m features at each split node of the tree. The randomness in the technique produces better performance as it reduces the correlation of the tree [31]. Like bagging, after *d* trees are computed, the test data class would be determined via majority voting (classification). A previous study by [29] identified that Random Forest obtained a classification accuracy of 94.7295% with WDBC data; meanwhile, the study by Khuriwal and Mishra determined that Random Forest acquired the accuracy, precision and recall of 97.01% with the same WDBC data [32]. The study by [11] shows that random forest got 79.3103% accuracy with Breast Cancer Coimbra data.

ADABOOST

AdaBoost, the short term for Adaptive Boosting, is a supervised learning introduced by Freund and Schapire in 1995 [23]. In AdaBoost, the set of training data would be independently sampled from some unspecified distribution, and each distribution would be equally distributed with 1/*m* initially [25]. The model would implement *t* iterations of weak classifiers, and weak learners would be trained using distribution Dt. The distribution weights also are preserved after every iteration [33]. The weight of each observation would increase in every iteration to enable the weak learners to focus on the observations that cannot classify correctly. The weak learner would identify an appropriate weak hypothesis that affects the classification and be able to acquire predicting function sequences. Each predicting function provides weight, and the weight of a learner with a better predicating effect would be greater [33]. Hence, minimise the errors in every iteration. According to [14], AdaBoost acquired the classification accuracy and sensitivity of 82% and 98%, respectively, outperforming single classifier ANN, which only got 64% of sensitivity and 97% accuracy with MIAS Breast Cancer Mammogram data. Based on the study by [7], AdaBoost M1 obtained the precision and recall of 75% with breast cancer data outperformed other ensemble methods, Random Forest.

XGBOOST

XGBoost, or can be known as Extreme Gradient Boosting, is the ensemble learning for decision tree boosting that can be applied to both classification and regression problems [34]. One of the advantages of applying XGBoost is the scalability because of the algorithmic optimisations. For example, the model runs approximately ten times faster than other well-known techniques on a single machine [35]. Based on [35], XGBoost's other advantage is its ability to handle sparse data. It uses advanced regularisation compared to Gradient Boosting. The value *k* that represents the number of trees in the XGBoost model needs to be used to find the best set of functions by reducing the loss and regularisation objective. Then, *l,* which represents the loss function that holds the difference value between the predicted output yˆi and the actual output yi and *l,* is also used to measure the prediction ability [36]. Next, the measurement of the regularisation term to find the model's complexity is computed, which is used to prevent overfitting and simultaneously control the complexity in the generated model [37]. In the generated decision trees, to reduce the objective function boosting, adding a new function, *f* needed as the model keeps training. Hence, in the t-th iteration, a new function (tree) would be added. The study by [29] also identified that the XGBoost ensemble achieved an accuracy of 95.1691% with Wisconsin Diagnostic Breast Cancer data. The study by [38] identified that XGBoost outperformed other methods, such as Random Forest, KNN and SVM, by achieving a mean AUC of 0.82 for the prediction of metastatic status in breast cancer.

Comparative analysis of the homogenous ensemble

This study implements the comparative analysis to identify the most suitable ensemble method for breast cancer data. Many factors must be considered to identify the most reliable methods, including processing speed, classification accuracy (existing breast cancer study), generalisation ability, and training time. Table 1 presents a thorough comparison between the ensemble methods.

Table 1. Comparison between the ensemble method.

Method	Advantage	Disadvantage
	-Can be applied to any classifier [39]	-Lower performance than Random Forest [11].
Bagging	-Acquired better accuracy than AdaBoost [40]. -Perform better than Random Forest and AdaBoost when there are extreme outliers and imbalanced classes [40]. -Minimise variance of the classifiers [22], [41].	-Longer training time [42].
Random Forest	-Classification performance is better than XGBoost [43]. -Minimise variance of classifiers [21].	-Classifiers can be used only limited to decision trees [44]. -Longer training time [42].
AdaBoost	-Works well with large datasets [5].	-It takes longer to build compared to Bagging [45].
	-Can be applied to any classifiers [5], [25], [46]	-It has a higher tendency to overfit [47].
XGBoost	-Do regularisation that avoids overfitting; better generalisation [34] -Fast training time [35], [49].	-Sensitive to outliers [48]. -Classifiers can be used only limited to decision trees [50]

Conclusion

This study presents an overview of breast cancer classification using the well-known homogenous ensemble technique. The reviewed methods include Random Forest, Bagging, AdaBoost and XGBoost. The methods' precision, accuracy, stability and sensitivity with outliers are among the supreme components when choosing the best method that can be used in the breast cancer domain. In this presented paper, the ensemble methods are deeply discussed. The discussion covers the methods' advantages and disadvantages, which revolves around the mechanism, time taken, ability to handle large data, outliers sensitivity, accuracy, and overfitting issue. Based on the analysis, it is inscrutable to acknowledge which ensemble method is the best in this particular domain because each discussed method has advantages and disadvantages. The best method would depend on the problem, situation, data and features, as no method can work best and perform excellently in every problem.

Acknowledgement

This study was supported by Fundamental Research Grant (FRGS) with FRGS/1/2022/ICT02/UMP/02/2 (RDU220134) from the Ministry of Higher Education Malaysia.

Authors: *Nur Farahaina Idris, Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Pahang, Malaysia, Email: farahaina@ump.edu.my.*

Mohd Arfian Ismail, Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Pahang, Malaysia. Pusat Kecemerlangan Kecerdasan Buatan & Sains Data, Universiti Malaysia Pahang Al-Sultan Abdullah, Lebuh Persiaran Tun Khalil Yaakob, Gambang, Pahang, Malaysia, Email: arfian@ump.edu.my.

REFERENCES

[1] H. B. Lee and W. Han, "Unique features of young age breast cancer and its management," *J. Breast Cancer*, vol. 17, no. 4, pp. 301–307, 2014, doi: 10.4048/jbc.2014.17.4.301.

[2] H. I. Bülbül and Ö. Ünsal, "Comparison of classification techniques used in machine learning as applied on vocational guidance data," in *Proceedings - 10th International Conference on Machine Learning and Applications, ICMLA 2011*, 2011, vol. 2, pp. 298–301, doi: 10.1109/ICMLA.2011.49.

[3] M. M. Pyingkodi *et al.*, "Performance Study Of Classification Algorithms Using The Breast Cancer Dataset Microarray," vol. 13, no. 2, pp. 1238–1245, 2020.

[4] N. F. Idris and M. A. Ismail, "Breast cancer disease classification using fuzzy-ID3 algorithm with FUZZYDBD method : automatic fuzzy database definition," *Peerj Comput. Sci.*, 2021, doi: 10.7717/peerj-cs.427.

[5] L. Zhao, S. Lee, and S.-P. Jeong, "Decision Tree Application to Classification Problems with Boosting Algorithm," *Electronics*, vol. 10, no. 1903, 2021, doi: https://doi.org/10.3390/electronics10161903.

[6] R. O. Odegua, "An Empirical Study of Ensemble Techniques (Bagging, Boosting and Stacking)," in *Deep Learning IndabaX*, 2020, vol. 12, no. 10, p. 1683.

[7] F. Sardouk, A. D. Duru, and O. Bayat, "Classification of breast cancer using data mining," *Am. Sci. Res. J. Eng. Technol. Sci.*, vol. 51, no. 1, pp. 38–46, 2019.

[8] S. Taneja, B. Suri, S. Gupta, and H. Narwal, "A Fuzzy Logic Based Approach for Data Classification," *Data Eng. Intell. Comput. Adv. Intell. Syst. Comput.*, pp. 605–616, 2018, doi: 10.1007/978-981-10-3223-3.

[9] R. Longadge, S. S. Dongre, and L. Malik, "Class imbalance problem in data mining: review," *Int. J. Comput. Sci. Netw.*, vol. 2, no. 1, pp. 83–87, 2013, doi: 10.1109/SIU.2013.6531574.

[10] K. Kowsari, N. Bari, R. Vichr, and F. A. Goodarzi, "FSL-BM : Fuzzy Supervised Learning with Binary Meta-Feature for Classification," no. April, 2018.

[11] A. Fauzi, R. Supriyadi, and N. Maulidah, "Deteksi Penyakit Kanker Payudara dengan Seleksi Fitur berbasis Principal Component Analysis dan Random Forest," *J. Infortech*, vol. 2, no. 1, 2020.

[12] F. S. Nugraha, M. J. Shidiq, and S. Rahayu, "Analisis Algoritma Klasifikasi Neural Network Untuk Diagnosis Penyakit Kanker Payudara," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 149–156, 2019, doi: 10.33480/pilar.v15i2.601.

[13] A. S. P. Angayarkanni and N. B. Kamal, "MRI mammogram image classification using ID3 algorithm," *IET Conf. Publ.*, vol. 2012, no. 600 CP, pp. 1–5, 2012, doi: 10.1049/cp.2012.0464.

[14] G. Saad, A. Khadour, and Q. Kanafani, "ANN and Adaboost application for automatic detection of microcalcifications in breast cancer," *Egypt. J. Radiol. Nucl. Med.*, vol. 47, no. 4, pp. 1803–1814, 2016, doi: 10.1016/j.ejrnm.2016.08.020.

[15] R. Matteo and V. Giorgio, *Ensemble methods: a review*. 2001.

[16] D. Lavanya and K. Usha Rani, "ENSEMBLE DECISION TREE CLASSIFIER FOR BREAST CANCER DATA," *Int. J. Inf. Technol. Converg. Serv.*, vol. 2, no. 1, pp. 12–24, 2012, doi: 10.5121/ijitcs.2012.2103.

[17] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing Multiclass to Binary," *J. Mach. Learn. Res.*, vol. 1, pp. 113–141, 2000, doi: 10.1002/9783527677320.ch16.

[18] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *Ann. Stat.*, vol. 26, no. 5, pp. 1651–1686, 1998, doi: 10.1214/aos/1024691352.

[19] L. Breiman, "Bias, Variance, and Arcing Classifiers," 1996.

[20] X. Zhu, J. Hu, T. Xiao, S. Huang, Y. Wen, and D. Shang, "An interpretable stacking ensemble learning framework based on multi-dimensional data for real-time prediction of drug concentration: The example of olanzapine," *Front. Pharmacol.*, vol. 13, no. September, pp. 1–20, 2022, doi: 10.3389/fphar.2022.975855.

[21] L. Breiman, *RANDOM FORESTS*. 2001, pp. 1–33.

[22] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1023/A:1018054314350.

[23] R. E. Schapire and Y. Freund, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Comput. Syst. Sci.*, vol. 55, pp. 119–139, 1997, doi: 10.1145/2818346.2823306.

[24] S. Chatterjee and Y.-C. Byun, "EEG-Based Emotion

Classification Using Stacking Ensemble Approach," *Sensors*, vol. 22, no. 21, p. 8550, 2022, doi: 10.3390/s22218550.
[25] M. Sharifmoghadam and H. Jazayeriy, "Breast Cancer Classification Using AdaBoost- Extreme Learning Machine," in *5th Iranian Conference on Signal Processing and Intelligent Systems, ICSPIS 2019*, 2019, no. December, pp. 1–5, doi: 10.1109/ICSPIS48872.2019.9066088.
[26] A. Buja and W. Stuetzle, "Observations on bagging," *Stat. Sin.*, vol. 16, no. 2, pp. 323–351, 2006.
[27] N. Arsov, M. Pavlovski, and L. Kocarev, "Stacking and stability," *Arxiv*, 2019, [Online]. Available: http://arxiv.org/abs/1901.09134.
[28] M. L. Petersen, A. M. Molinaro, S. E. Sinisi, and M. J. van der Laan, "Cross-validated bagged learning," *J. Multivar. Anal.*, vol. 98, no. 9, pp. 1693–1704, 2007, doi: 10.1016/j.jmva.2007.07.004.
[29] V. Chaurasia and S. Pal, "Applications of Machine Learning Techniques to Predict Diagnostic Breast Cancer," *SN Comput. Sci.*, vol. 1, no. 5, 2020, doi: 10.1007/s42979-020-00296-8.
[30] R. Surendiran, M. Thangamani, C. Narmatha, and M. Iswarya, "Effective Autism Spectrum Disorder Prediction to Improve the Clinical Traits using Machine Learning Techniques," *Int. J. Eng. Trends Technol.*, vol. 70, no. 4, pp. 343–359, 2022, doi: 10.14445/22315381/IJETT-V70I4P230.
[31] S. Janitza and R. Hornung, "On the overestimation of random forest ' s out-of-bag error," *PLoS One*, vol. 13, no. 8, pp. 1–31, 2018, [Online]. Available: https://doi.org/10.1371/journal.pone.0201904 August.
[32] N. Khuriwal and N. Mishra, "Breast Cancer Diagnosis Using Deep Learning Algorithm," in *Proceedings - IEEE 2018 International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2018*, 2018, pp. 98–103, doi: 10.1109/ICACCCN.2018.8748777.
[33] C. Tu, H. Liu, and B. Xu, "AdaBoost typical Algorithm and its application research," *MATEC Web Conf.*, vol. 139, no. 00222, 2017, doi: 10.1051/matecconf/201713900222.
[34] B. Pan, "Application of XGBoost algorithm in hourly PM2.5 concentration prediction," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 113, no. 1, 2018, doi: 10.1088/1755-1315/113/1/012127.
[35] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: doi/10.1145/2939672.2939785.
[36] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *Int. J. Distrib. Sens. Networks*, vol. 18, no. 6, 2022, doi: 10.1177/15501329221106935.
[37] A. Ibrahem Ahmed Osman, A. Najah Ahmed, M. F. Chow, Y. Feng Huang, and A. El-Shafie, "Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia," *Ain Shams Eng. J.*, vol. 12, no. 2, pp. 1545–1556, 2021, doi: 10.1016/j.asej.2020.11.011.
[38] Q. Li, H. Yang, P. Wang, X. Liu, K. Lv, and M. Ye, "XGBoost-based and tumor-immune characterised gene signature for the prediction of metastatic status in breast cancer," *J. Transl. Med.*, vol. 20, no. 1, pp. 1–12, 2022, doi: 10.1186/s12967-022-03369-9.
[39] M. C. Tu, D. Shin, and D. Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms," in *Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*, 2009, pp. 183–187, doi: 10.1109/DASC.2009.40.
[40] G. Tuysuzoglu and D. Birant, "Enhanced bagging (eBagging): A novel approach for ensemble learning," *Int. Arab J. Inf. Technol.*, vol. 17, no. 4, pp. 515–528, 2020, doi: 10.34028/iajit/17/4/10.
[41] N. F. Idris and M. A. Ismail, "The study of cross-validated bagging fuzzy-ID3 algorithm for breast cancer classification," *J. Intell. Fuzzy Syst.*, vol. 43, no. 3, pp. 2567–2577, 2022, [Online]. Available: 10.3233/JIFS-212842.
[42] M. Al Diabat and N. Al-Shanableh, "Ensemble Learning Model for Screening Autism in Children," *Int. J. Comput. Sci. Inf. Technol.*, vol. 11, no. 02, pp. 45–62, 2019, doi: 10.5121/ijcsit.2019.11205.
[43] S. Kabiraj *et al.*, "Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm," *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, pp. 1–4, 2020, doi: 10.1109/ICCCNT49239.2020.9225451.
[44] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," no. September, 2012.
[45] F. Thabtah and D. Peebles, "A new machine learning model based on induction of rules for autism detection," *Health Informatics J.*, vol. 26, no. 1, pp. 264–286, 2020, doi: 10.1177/1460458218824711.
[46] M. M. Baig, M. M. Awais, and E. S. M. El-Alfy, "AdaBoost-based artificial neural network learning," *Neurocomputing*, vol. 248, pp. 120–126, 2017, doi: 10.1016/j.neucom.2017.02.077.
[47] A. Vezhnevets and O. Barinova, "Avoiding boosting overfitting by removing confusing samples," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4701 LNAI, pp. 430–441, 2007, doi: 10.1007/978-3-540-74958-5_40.
[48] Y. Zhao and M. K. Hryniewicki, "XGBOD: Improving Supervised Outlier Detection with Unsupervised Representation Learning," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2018-July, 2018, doi: 10.1109/IJCNN.2018.8489605.
[49] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A Comparative Analysis of XGBoost," no. November 2019, 2019, doi: 10.1007/s10462-020-09896-5.
[50] X. Y. Liew, N. Hameed, and J. Clos, "An investigation of XGBoost-based algorithm for breast cancer classification," *Mach. Learn. with Appl.*, vol. 6, no. August, p. 100154, 2021, doi: 10.1016/j.mlwa.2021.100154.