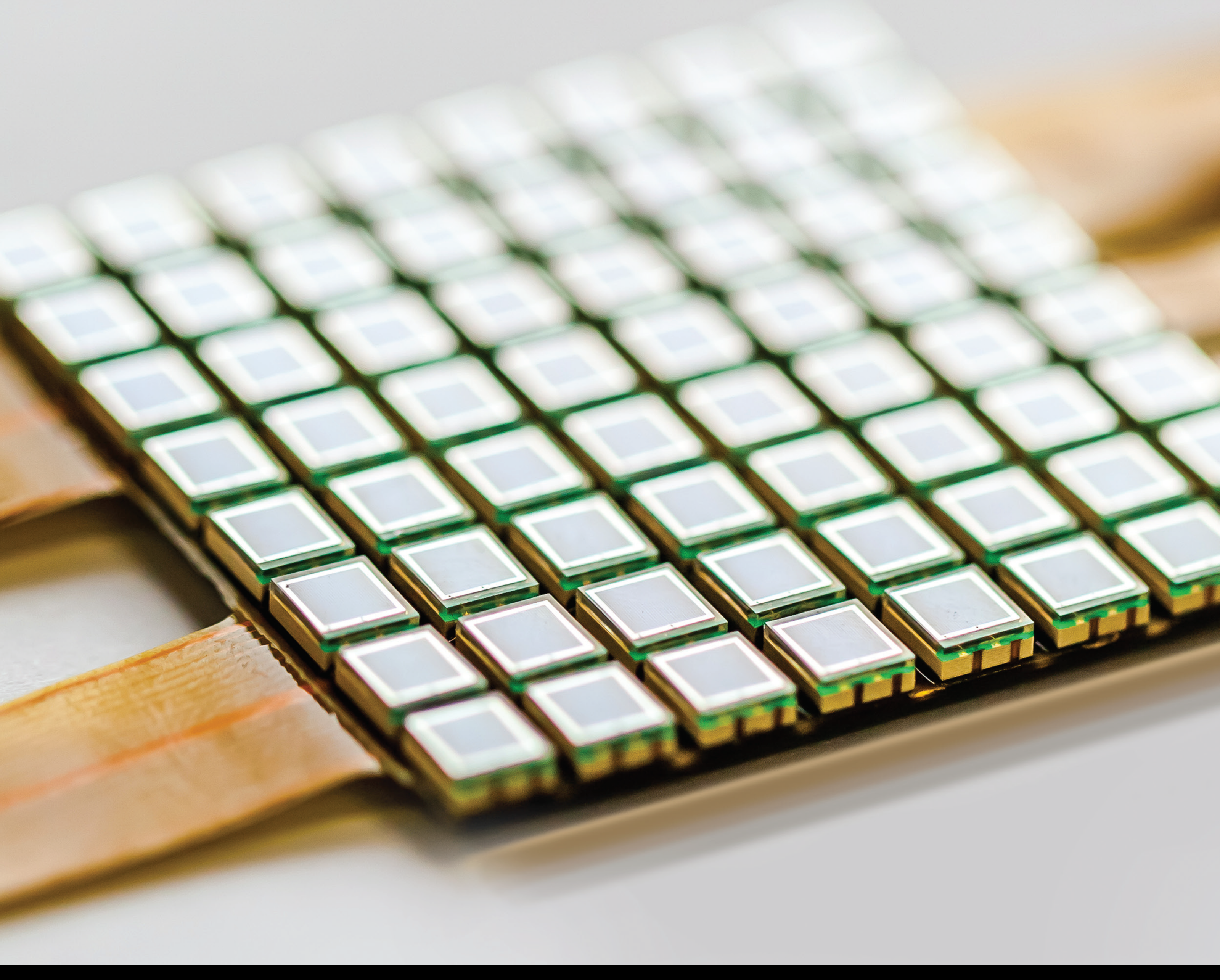
Journal of Sensors

Advanced Sensors and Sensing Technologies for Telehealth Monitoring Systems

Lead Guest Editor: Danilo Pelusi
Guest Editors: Masood Ur-Rehman and Isabel de la Torre



Advanced Sensors and Sensing Technologies for Telehealth Monitoring Systems

Journal of Sensors

Advanced Sensors and Sensing Technologies for Telehealth Monitoring Systems

Lead Guest Editor: Danilo Pelusi
Guest Editors: Masood Ur-Rehman and Isabel de la Torre

Chief Editor

Associate Editors

Academic Editors

Ehsan Namaziandost, Iran
Heinz C. Neitzert, Italy
Sing Kiong Nguang, New Zealand
Calogero M. Oddo, Italy
Tinghui Ouyang, Japan
SANDEEP KUMAR PALANISWAMY, India
Alberto J. Palma, Spain
Davide Palumbo, Italy
Abinash Panda, India
Roberto Paolesse, Italy
Akhilesh Pathak, Thailand
Giovanni Pau, Italy
Giorgio Pennazza, Italy
Michele Penza, Italy
Sivakumar Poruran, India
Stelios Potirakis, Greece
Biswajeet Pradhan, Malaysia
Giuseppe Quero, Italy
Linesh Raja, India
Maheswar Rajagopal, India
Valerie Renaudin, France
Armando Ricciardi, Italy
Christos Riziotis, Greece
Ruthber Rodriguez Serrezuela, Colombia
Maria Luz Rodriguez-Mendez, Spain
Jerome Rossignol, France
Maheswaran S, India
Ylias Sabri, Australia
Sourabh Sahu, India
José P. Santos, Spain
Sina Sareh, United Kingdom
Isabel Sayago, Spain
Andreas Schütze, Germany
Praveen K. Sekhar, USA
Sandra Sendra, Spain
Sandeep Sharma , India
Sunil Kumar Singh Singh, India
Yadvendra Singh, USA
Afaque Manzoor Soomro, Pakistan
Vincenzo Spagnolo, Italy
Kathiravan Srinivasan, India
Sachin K. Srivastava, India
Stefano Stassi, Italy
Danfeng Sun, China
Ashok Sundramoorthy, India
Salvatore Surdo, Italy
Roshan Thotagamuge, Sri Lanka
Guiyun Tian, United Kingdom
Sri Ramulu Torati, USA
Abdellah Touhafi, Belgium
Hoang Vinh Tran, Vietnam
Aitor Urrutia, Spain
Hana Vaisocherova - Lisalova, Czech Republic
Everardo Vargas-Rodriguez, Mexico
Xavier Vilanova, Spain
Stanislav Vítek, Czech Republic
Luca Vollero, Italy
Tomasz Wandowski, Poland
Bohui Wang, China
Qihao Weng, USA
Penghai Wu, China
Qiang Wu, United Kingdom
Yuedong Xie, China
Chen Yang, China
Jiachen Yang, China
Nitesh Yelve, India
Aijun Yin, China
Chouki Zerrouki, France

Contents

Hindawi
Journal of Sensors
Volume 2022, Article ID 4649510, 8 pages
https://doi.org/10.1155/2022/4649510

Research Article

Performance Analysis of XGBoost Ensemble Methods for Survivability with the Classification of Breast Cancer

T. R. Mahesh,[1] **V. Vinoth Kumar**,[2] **V. Muthukumaran**,[1] **H. K. Shashikala**,[1] **B. Swapna**,[3] **and Suresh Guluwadi**[4]

[1]*Department of Computer Science and Engineering, Faculty of Engineering and Technology, JAIN (Deemed-to-be University), Bangalore, India*
[2]*Department of Mathematics, College of Engineering and Technology, SRM Institute of Science and Technology, Tamilnadu, India*
[3]*Department of ECE, Dr MGR Educational and Research Institute, Chennai, India*
[4]*Adama Science and Technology University, Adama, Ethiopia*

Correspondence should be addressed to Suresh Guluwadi; suresh.guluwadi@astu.edu.et

Received 23 June 2022; Accepted 1 September 2022; Published 13 September 2022

Academic Editor: Masood Ur Rehman

Breast cancer (BC) disease is the most common and rapidly spreading disease across the globe. This disease can be prevented if identified early, and this eventually reduces the death rate. Machine learning (ML) is the most frequently utilized technology in research. Cancer patients can benefit from early detection and diagnosis. Using machine learning approaches, this research proposes an improved way of detecting breast cancer. To deal with the problem of imbalanced data in the class and noise, the Synthetic Minority Oversampling Technique (SMOTE) has been used. There are two steps in the suggested task. In the first phase, SMOTE is utilized to decrease the influence of imbalance data issues, and subsequently, in the next phase, data is classified using the Naive Bayes classifier, decision trees classifier, Random Forest, and their ensembles. According to the experimental analysis, the XGBoost-Random Forest ensemble classifier outperforms with 98.20% accuracy in the early detection of breast cancer.

1. Introduction

One of the most as well as dangerous diseases existing on the planet is breast cancer (BC). There are two types of BC: invasive and noninvasive. The first type invasive cancer is malignant, and it spreads to other organs. The second type noninvasive cancer is precancerous and does not spread beyond the native organ. In the end, it progresses to invasive BC. The glands along with the milk ducts that convey the milk are the parts of the body where breast cancer can be found. Breast cancer frequently spreads to other organs, turning them aggressive. BC can be categorized into four categories: the first type of cancer is a prestage breast cancer called carcinoma in situ. The second type of BC is the most common, accounting for 70-80 percent of all diagnoses. Inflammatory BC is another type of BC that develops fast and strongly. Inflammatory BC cells enter the skin and lymph veins of the breast. One more type is metastatic BC that spreads to other regions of the body.

Disease diagnosis is a difficult and time-consuming task in medicine. A great amount of medical diagnostic data can be found in many diagnostic institutions, hospitals, research organizations, and websites. To automate and speed up disease diagnosis, however, categorizing them is not absolutely necessary. As per the American Cancer Society [1], BC affects more women than any other malignancy. According to estimates, 252,710 women in the USA were identified with invasive BC in 2017 and 63,410 women were diagnosed with in situ BC.

So avoiding BC is quite difficult; however, if identified early, proper diagnosis and treatment can be provided to cure the disease. This also reduces the treatment expenses as well. However, since symptoms of cancer might be uncommon at times, prior detection can be challenging.

Mammograms and also self-breast examinations are essential for detecting and identifying any prior anomalies before the malignancy progresses [2].

BC outcomes are classified using a number of methods. This disease can be classified and predicted using a variety of approaches. The XGBoost ensemble method developed in this paper can be used to classify breast tumors. For the proposed XGBoost ensemble model, we used Naive Bayes (NB) and decision tree (DT) as base learners. In addition, the accomplishments of the suggested models are being evaluated using the Kaggle Wisconsin BC dataset and the UCI ML Repository. The aim of the research work is to increase prediction accuracy by detecting and categorizing malignant and benign individuals.

2. Related Work

This section contains information about related research that has already been completed. The model that was recommended in this work [3] uses a hybrid method employing machine learning. It used feature selection approach called MRMR with four different classifiers to figure out the optimal results for this method. SVM, Naive Bays, End Meta, and Function Tree were the four classifiers utilized by the author, and they were all compared. It was discovered that SVM was an effective classifier. RFE and SVM are together combined in the SVM classifier technique [4]. RNNs are a type of neural network (NN) [5–7] that has a large number of layers in the sequential dimension and has been widely used in the modeling of time sequence. RNNs, unlike regular NNs, may analyze data objects where actually, the activation at each step is dependent on the previous step. CNN relies on “discrete convolution” since it makes use of spatial data [8] among picture pixels. As a result, it is assumed that the image is grayscale.

In this work [9], one more hybrid model based on ML was proposed. The authors claimed through experimental results that SVM was a good classifier with higher accuracy than others. They compared SVM with KNN and ANN as well DT algorithms. It was applied to the blood and picture datasets. As a result, the authors [10] suggested a machine learning model but with a different classifier. Extreme Learning Machine, SVM, KNN, and ANN were the classifiers employed by the author. To get better results, the classifier has to be tweaked a little. Extreme Learning Machine, on the other hand, produced better results.

In this study [11], different ML techniques were compared. WEKA was used to perform the comparison, and the dataset used was the Wisconsin BC dataset. According to their findings, SVM produced improved performance matrices. Deep learning (DL) methods were originated after ML to overcome the difficulty of ML. The paradigm of a DL-based CNN was proposed in this work [12]. CNN employed a variety of models, and after comparison, they concluded that Inception V3 provided better accuracy than others.

In healthcare, machine learning and its associated approaches have been identified as crucial in improving patient outcomes and wellbeing. An accuracy of 96.4 percent was found using logistic regression [13]. SVM and KNN were employed to classify breast cancer in this study [14], and the accuracy was 96.85 percent. RF [15] was used, and the accuracy was 92.2 percent. To figure out the optimal classifier in the BC dataset [16], researchers compared the performance of NB, SVM-RBF kernel, DT, basic CART classifiers, and RBF neural networks. AdaBoost was used, and it performed 97.5 percent better than Random Forest. Ensemble methods were used in this study [17] to achieve 96.25 percent accuracy, compared to 96.2 percent accuracy in earlier studies [18] using the back propagation strategy. The results showed 96.84% accuracy using the Wisconsin dataset for BC. As classification algorithms, they used SVM, KNN, RF, NB, and ANN.

On the acquired BC dataset, we used XGBoost ensemble learning employing NB and DT algorithms as base learners, and a significant boost in accuracy and recall was found. Machine learning models use a variety of approaches to improve the performance of classic models by integrating multiple models. By generating numerous models, the goal is to introduce ensemble learning and comprehend basic techniques. Compared to the individual classifiers, the ensemble learning method provides prominent accurate results. Our methodology employs the ensemble method, which employs in predicting good accuracy findings that correct issues or any restrictions according to the research study.

3. Methodology

Medical treatment as well as the accuracy of the diagnosis has a significant impact on the likelihood of survival and cancer recurrence. In this experiment, arbitrary extracted data was employed, with a 70 : 30 split between training as well as testing data. Training sets were used to provide the training to the model, and its effectiveness was assessed using test data. The dataset has 143 instances and contains 10 variables or attributes whose values will indicate whether or not a person is likely to get breast cancer. The output variable, also known as the target variable, is a binary variable that can be either malignant or benign. The dataset taken from Kaggle consists of these independent variables, sample code number, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses, and one dependent or output variable. However, the first feature sample code number is not considered for processing as it does not have any significance. Figure 1 represents the different stages of the procedure.

The authors [19] describe A-SMOTE, an advanced strategy to deal with the data imbalance problem. The steps are described below.

Step 1. A-SMOTE method is utilized to generate a synthetic object by using the following equation:

$$N = 2 * (r - z) + z, \tag{1}$$

where r is majority class samples, z minority class sample number, and N newly generated synthetic instance number.

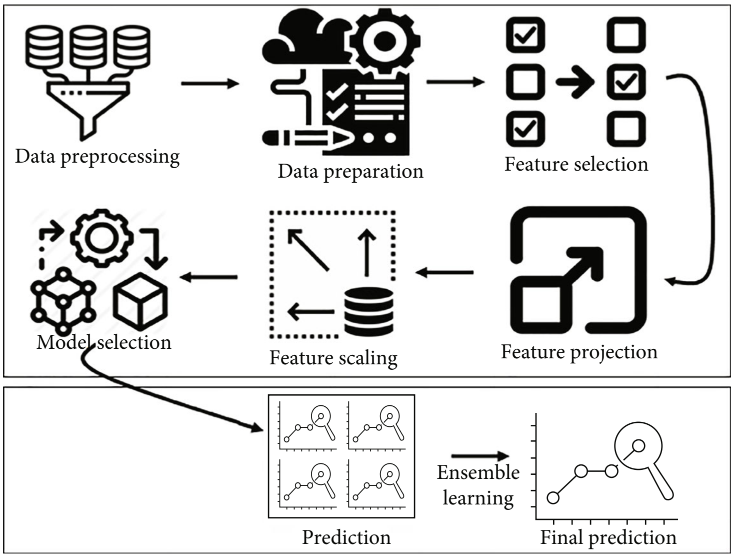


Figure 1: Process of the proposed methodology.

Table 1: TTMB, accuracy, and F1 score comparison of single classifiers.

Performance metrics	NB	AltDT	RF	RedEPT
TTMB (sec)	12.56	60.38	2.26	12.25
Accuracy (%)	88.50	95.60	94.50	89.23
F1 score	0.3	0.85	0.84	0.83

The synthetic objects obtained by SMOTE can be accepted or rejected and depend upon 2 criteria. For instance, consider a set of synthetic instances which are new as $\hat{x} = \{\hat{x}_1, \hat{x}_2, \hat{x}_3, \cdots . \hat{x}_N\}$ and $\hat{x}_i^{(j)}$ ➔jth feature value of $\hat{x}_i$, $j \in [1, M]$.

Let $S_m = \{S_{m1}, S_{m2}, \cdots, S_{mz}\}$ be minority sample collection and $S_\alpha = \{S_{\alpha 1}, S_{\alpha 1}, S_{\alpha 1}, \overset{,,,}{} S_{\alpha r},\}$ majority sample collection. Distance is calculated between $\hat{x}_i$ and S_{mk}, i.e., $D_{\text{minority}}(\hat{x}_i, S_{mk})$ and between $\hat{x}_i$ and $S_{\alpha l}$, i.e., $D_{\text{majority}}(\hat{x}_i, S_{\alpha l}$). Using equations (2) and (3), the distance is computed as shown.

$$DD_{\text{minority}}(\hat{x}_i, S_{mk}) = \sum_{j=1}^{M} \sqrt{\left(\hat{x}_i^{(j)} - \hat{S}_{mk}^{(j)}\right)^2}, \quad k \in [1, z], \tag{2}$$

$$D\,D_{\text{majority}}(\hat{x}_i, S_{al}) = \sum_{j=1}^{M} \sqrt{\left(\hat{x}_i^{(j)} - \hat{S}_{al}^{(j)}\right)^2}, \quad l \in [1, r]. \tag{3}$$

As per equations (2) and (3), we compute arrays A_{minority} and A_{majority} using equations (4) and (5).

$$A_{\text{minority}} = \left(DD_{\text{minority}}(\hat{x}_i, S_{m1}), \cdots . DD_{\text{minority}}(\hat{x}_i, S_{mz})\right), \tag{4}$$

$$A_{\text{majority}} = \left(DD_{\text{majority}}(\hat{x}_i, S_{\alpha 1}), \cdots . DD_{\text{majority}}(\hat{x}_i, S_{\alpha r})\right). \tag{5}$$

Then, minimum value will be chosen among A_{minority}, min (A_{minority}) and the minimum value out of A_{majority}, min (A_{majority}). If min (A_{minority}) is a lesser than min (A_{majority}), the new samples will be accepted else, rejected.

$$\min\left(A_{\text{majority}}\right) < \min\left(A_{\text{majority}}\right)(\text{accepted}), \tag{6}$$

$$\min\left(A_{\text{majority}}\right) \geq \min\left(A_{\text{majority}}\right)(\text{rejected}). \tag{7}$$

Step 2. The steps to remove the noise are listed below.

For example, if $\hat{S} = \{\hat{S}_1, \hat{S}_2, \hat{S}_3, \cdots . \hat{S}_n\}$ is a latest synthetic minority acquired by Step 1, then, we will compute the distance among $\hat{S}_i$ with every original minority S_m, $i_{\text{Rap}}(\hat{S}_i, \hat{S}_m)$ defined using equation (6).

$$\text{Min}_{\text{Rap}}\left(\hat{S}_i, \hat{S}_m\right) = \sum_{k=1}^{z} \sum_{j=1}^{M} \sqrt{\left(\hat{S}_i^{(j)} - S_{mk}^{(j)}\right)^2}, \tag{8}$$

where $\text{Min}_{\text{Rap}}(\hat{S}_i, \hat{S}_m)$ is sample rapprochement that includes all minority and L is obtained as follows using the following equation:

$$L = \sum_{i=1}^{n} \left(\text{Min}_{\text{Rap}}\left(\hat{S}_i, S_m\right)\right). \tag{9}$$

Step 3. Calculate the distance among $\hat{S}_i$ and every original majority S_a, $\text{Maj}_{\text{Rap}}(\hat{S}_i, S_a)$, described using equation (9).

$$\text{Maj}_{\text{Rap}}\left(\hat{S}_i, S_a\right) = \sum_{i=1}^{r} \sum_{j=1}^{M} \sqrt{\left(\hat{S}_i^{(j)} - S_{al}^{(j)}\right)^2}, \tag{10}$$

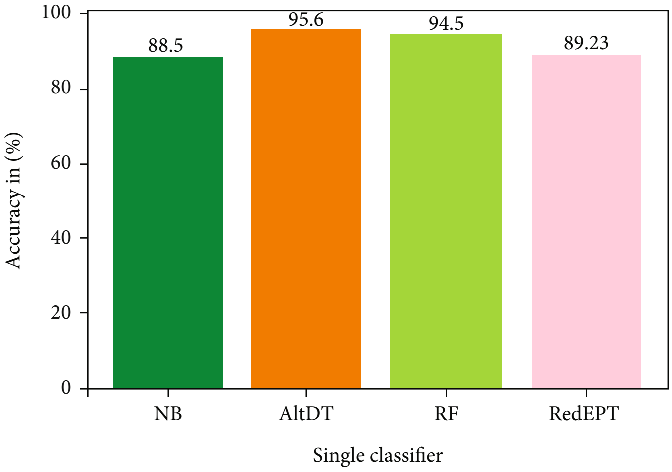


Figure 2: Accuracy prediction for single classifiers.

Table 2: Different error rates' comparison of single classifiers.

Performance metrics	NB	AltDT	RF	RedEPT
MAE	0.60	0.38	0.25	0.25
RMSE	0.75	0.38	0.36	0.40
RAE	78.12	67.71	67.27	79.12
RRSE	83.31	95.33	82.92	97. 89

where $\mathrm{Maj}_{\mathrm{Rap}}(\widehat{S}_i, S_a)$ is sample rapprochement that includes all majority and H is obtained using the following equation:

$$H = \sum_{i=1}^{n} \left(\mathrm{Maj}_{\mathrm{Rap}} \left(\widehat{S}_i, S_a \right) \right). \tag{11}$$

3.1. Decision Tree (DT) Classifier. A basic diagram for categorizing samples is a DT. In DT, the data is constantly divided based on a parameter [20]. The DTs are a group of supervised classification algorithms that are well-known. They perform well on classification tasks, the decisional process are easy to understand, and the algorithm for creating (training) them is quick and simple. It is one of the most well-known modeling strategies because it was one of the first elite regression analysis methods individuals learned when learning predictive modeling.

3.2. Alternating Decision Tree (AltDT). An AltDT consists of a sequence of decision nodes. An AltDT categorizes an instance by summarizing all prediction nodes traversed and pursuing all paths for which all decision nodes are true [21]. Both the root and leaves of AltDT are always prediction nodes. An AltDT classifies an instance by traversing all paths where all decision nodes are true and adding any prediction nodes traversed.

3.3. Reduced Error Pruning Tree (RedEPT). RedEPT is a quick DT learning algorithm that constructs a DT based on the information obtained or by minimizing variance. This algorithm's basic pruning method is REP with back overfitting [22]. It politely arranges numerical attribute values once, and in fractional instances, it handles missing values with an embedded function by C4.5 which is an extension of Quinlan's earlier ID3 algorithm based on wrapper feature selection. Training, validation, and test sets are used until additional trimming is damaged, which is an effective strategy if a substantial amount of data is available.

3.4. Random Forest (RT) Classifier. Random Forest is a ML technique that is part of the supervised ML model. The RF classifier is made up of numerous DTs representing various subjects. It takes the average of each tree's subset to improve predictive accuracy. RF, rather than depending on a single decision tree, uses the majority prediction of voting from every tree and then predicts the result [23]. Every node in the decision tree answers a query about the situation.

For a candidate (nominal) split attribute X_i denoted possible levels as $L_i \cdots \cdots, L_j$. Gini Index for this feature is computed using the following equation:

$$G(X_i) = \sum_{j=1}^{J} \Pr(X_i = L_j) (1 - \Pr(X_i = L_j)) = 1 - \sum_{j=1}^{J} \Pr(X_i = L_j)^2. \tag{12}$$

3.5. Naïve Bayes (NB) Classifier. The Bayes' theorem is a straightforward formula for estimating conditional probabilities. The formula is given as follows:

$$P(S|R) = \frac{P(R|S) * P(S)}{P(R)}, \tag{13}$$

where R, S are events, $P(S|R)$ probability of Y given X is true, $P(R|S)$ probability of X given Y is true, $P(R)$ probability of X , and $P(S)$ probability of Y.

3.6. XGBoost. XGBoost is a high-scalability DT ensemble based on gradient boosting. XGBoost, like gradient boosting, minimizes a loss function to produce an additive expansion of the objective function. Because XGBoost only uses DTs as base classifiers, the complexity of the trees is controlled

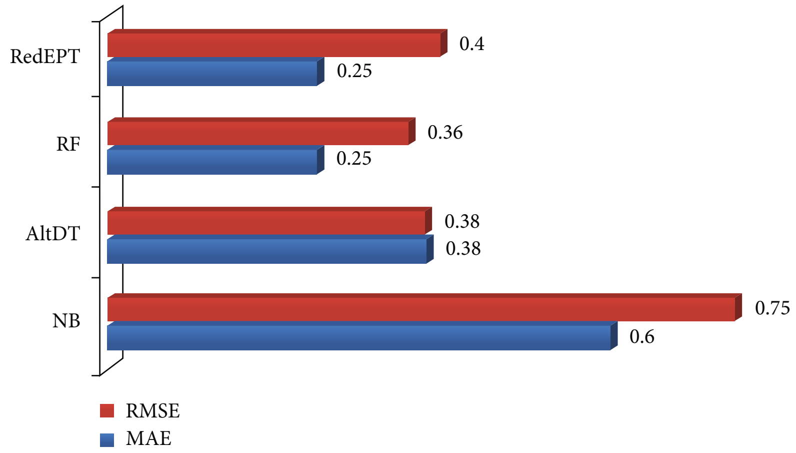


Figure 3: Individual classifiers—error rates.

Table 3: XGBoost classifiers.

Performance metrics	XGBoost-NB	XGBoost-AltDT	XGBoost-RF	XGBoost-RedEPT
TTMB (sec)	18.32	30.01	10.34	60.25
Accuracy (%)	81.55	96.50	98.20	82.25
F1 score	0.81	0.95	0.98	0.83

using a variant of the loss function as shown in equations (14) and (15).

$$L_{\mathrm{xgb}} = \sum_{i=1}^{N} L(y_i, F(x_i)) + \sum_{m=1}^{M} \Omega(h_m), \tag{14}$$

$$\Omega(h) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2, \tag{15}$$

where T is the number of leaves on the tree and ω denotes the leaf output scores. This loss function can be incorporated into decision trees' split criterion, resulting in a prepruning approach. Trees with higher γ values are simpler. γ determines how much loss reduction gain is required to split an internal node. Shrinkage is an additional regularization parameter in XGBoost that reduces the additive expansion step size. Finally, other tactics such as tree depth can be used to limit the complexity of the trees. The models are trained faster and need less storage space as a side effect of reducing tree complexity.

4. Performance Evaluation

Several ML methods, such as Naive Bayes, AltDT, RedEPT, and RF, are used as independent classifiers on the dataset. The implementation was done using Python language. Their performance is compared using numerous metrics, which are detailed in the next section.

Different performance metrics have been used to evaluate the suggested model. "Precision" is the percentage of accurately classified events among those which have been classified as correctly positive [24]. Precision indicates what proportion of the total positive anticipated is genuinely positive. The precision is computed using the following equation:

$$\mathrm{Precision} = \frac{\mathrm{TPs}}{\mathrm{TPs} + \mathrm{FPs}}. \tag{16}$$

Recall indicates what proportion of the total positives is projected to be positive. The proportion of TPs to the sum of TPs and FNs is known as recall [25]. True positive rate and true recall are the same things. Out of all feasible positive predictions, recall is one metric that quantifies how many right predictions that are positive were made. The recall is computed using the following equation:

$$\mathrm{Recall} = \frac{\mathrm{TPs}}{\mathrm{TPs} + \mathrm{FNs}}. \tag{17}$$

In order for a good classifier to be one, both accuracy and recall must be one, which means the number of FPs and FNs must be zero. So, a statistic is needed that takes both precision and recall into account. F1 is calculated using the following equation:

$$\text{F1 score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \tag{18}$$

The accuracy is computed using the following equation [26]:

$$\mathrm{Accuracy} = \frac{\mathrm{TNs} + \mathrm{TPs}}{\mathrm{TNs} + \mathrm{TPs} + \mathrm{FPs} + \mathrm{FNs}}. \tag{19}$$

Table 1 depicts that RF is the optimal model since it consumes only 2.26 seconds for model building (TTMB—Time for Model Building); however, AltDT has consumed 60.38 seconds for model building.

AltDT provides the best accuracy of 95.6%. RF provides 94.5% accuracy, RedEPT provides 89.23%, and prediction of NB classifier is the least with 88.50% accuracy. The accuracy prediction of various classifiers is shown in Figure 2.

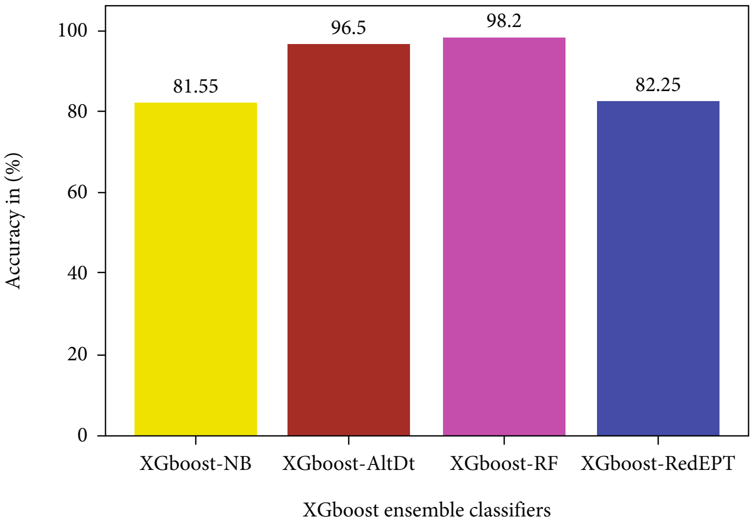


Figure 4: Accuracy of XGBoost ensemble classifiers.

Table 4: Different error rates' comparison of XGBoost classifiers.

Performance metrics	XGBoost-NB	XGBoost-AltDT	XGBoost-RF	XGBoost-RedEPT
MAE	0.44	0.17	0.12	0.22
RMSE	0.56	0.34	0.27	0.35
RAE	67.79	57.78	35.87	45.19
RRSE	92.62	96.23	65.47	91.03

To calculate the error rates in the predicted value, let P^N represent a set of data that has the form (t_1, r_1), (t_2, r_2),…, (t_p, r_p), where t_i represents n-dimensional tuples of test with respective values of r_i for a given response r, as well as represents count of tuples in P^N.

The mean absolute error (MAE) is computed using the following equation:

$$\mathrm{MAE} = \sum_{i=1}^{p} \left| r_i - {r_i}^T \right|. \tag{20}$$

The errors are squared before being averaged in RMSE. This basically means that RMSE gives larger mistakes a higher weight. This suggests that RMSE is far more beneficial when substantial errors exist and have a significant impact on the model's performance. This characteristic is important in many mathematical calculations since it avoids taking the absolute value of the error. In this metric as well, the lower the value, the better the model's performance. RMSE is calculated using the following equation:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{r=1}^{p} \left(r_i - {r^T}_i \right)^2}{p}}. \tag{21}$$

The relative absolute error (RAE) is used for the evaluation of a prediction model's performance. RAE is computed using the following equation:

$$\mathrm{RAE} = \frac{\sum_{r=1}^{p} \left(r_i - {r^T}_i \right)^2}{\sum_{r=1}^{p} \left(r_i - \bar{r}_i \right)^2}. \tag{22}$$

The RRSE is one of the measures to compute to know how good the ML model fits the data. The model does not match the data well if there is a substantial discrepancy between the values. It is calculated using the following equation:

$$\mathrm{RRSE} = \sqrt{\frac{\sum_{r=1}^{p} \left(r_i - {r^T}_i \right)^2}{\sum_{r=1}^{p} \left(r_i - \bar{r}_i \right)^2}}. \tag{23}$$

Table 2 depicts the different error rates' comparison of single classifiers. The error rates of RF are lesser compared to those of other classifiers.

Figure 3 shows the rates of errors of different individual classifiers. MAE and RMSE rates of NB, AltDT, RF, and RedEPT are, respectively, 0.60, 0.38, 0.25, and 0.25 and 0.75, 0.38, 0.36, and 0.40.

Table 3 depicts that XGBoost-RF can be recommended, since it takes as few as 10.34 seconds for model building. However, XGBoost-RedEPT is the least recommended model since it takes 60.25 seconds for model building.

XGBoost-RF provides the best accuracy of 98.20%. XGBoost-AltDT provides 96.50% accuracy, XGBoost-RedEPT provides 82.25%, and the prediction of XGBoost-NB ensemble classifier is the least with 81.55% accuracy. The accuracy prediction of different classifiers is shown in Figure 4.

Table 4 depicts the different error rates' comparison of XGBoost classifiers. The error rates of XGBoost-RF are lesser compared to those of all other ensemble classifiers.

Figure 5 shows the error rates of different ensemble classifiers. XGBoost-RF ensemble classifier is the best one as it provides 0.12 error rates for MAE and 0.27 for RMSE,

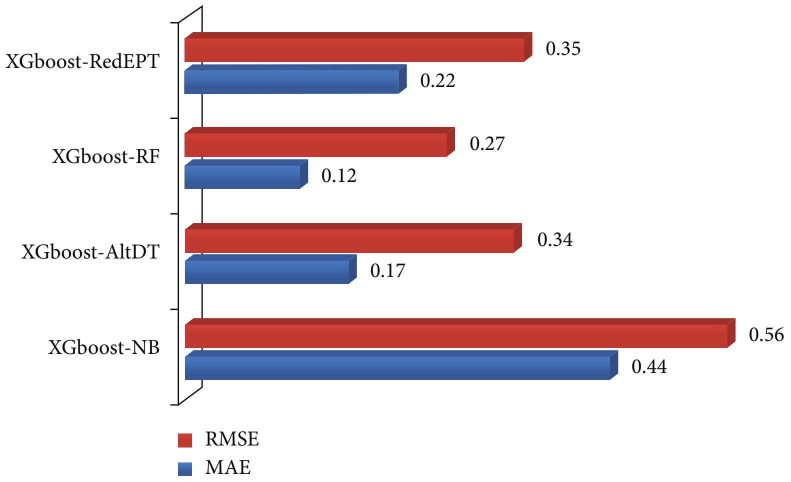


Figure 5: Different error rates of XGBoost ensemble classifiers.

respectively. XGBoost-NB has the highest error rate of 0.44 and 0.56 for MAE as well as RMSE, respectively.

5. Conclusion

This paper proposes the XGBoost ensemble technique for breast cancer prediction based on known feature patterns. It can be compared to traditional data mining methods in terms of disease diagnosis. During the feature extraction process, ensemble classification techniques replace the traditional techniques of retrieving useful information. SMOTE technique has been employed to deal with the problem of data imbalance. According to the experimental results, the time taken to create the model for XGBoost ensemble classifier is only 10.34 seconds for XGBoost-RF, which is the best, and XGBoost-RedEPT takes the worst time of 60.25 seconds. The results show that the XGBoost-RF classifier shows an error rate of 0.12 for MAE and a 0.27 for RMSE. The results show that XGBoost-RF outperforms other ensemble classifiers, with 98.20% accuracy.

Data Availability

The Irvine ML Repository data used to support the findings of this study are available at https://archive.ics.uci.edu/ml/datasets.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] Breast cancer statistics, “Approved by the Cancer.Net Editorial Board,” 2018, http://www.cancer.net/cancer-types/breast-cancer/statistics.

[2] K. M. Karthick Raghunath, V. Vinoth Kumar, V. Muthukumaran, K. K. Singh, T. R. Mahesh, and A. Singh, “Detection and classification of cyber attacks using XGBoost regression and inception V4,” *Journal of Web Engineering, River Publishers*, vol. 21, no. 4, 2022.

[3] M. Rathi and V. Pareek, “Hybrid approach to predict breast cancer using machine learning techniques,” *International Journal of Computer Science Engineering*, vol. 5, no. 3, pp. 125–136, 2016.

[4] T. R. Mahesh, V. Dhilip Kumar, V. Vinoth Kumar et al., “AdaBoost ensemble methods using K-fold cross validation for survivability with the early detection of heart disease,” *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 9005278, 11 pages, 2022.

[5] Q. Huang, Y. Chen, L. Liu, D. Tao, and X. Li, “On combining biclustering mining and AdaBoost for breast tumor classification,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 4, pp. 728–738, 2020.

[6] S. U. Khan, N. Islam, Z. Jan, I. U. Din, and J. J. P. C. Rodrigues, “A novel deep learning based framework for the detection and classification of breast cancer using transfer learning,” *Pattern recognition Letters*, vol. 125, pp. 1–6, 2019.

[7] Q. Pan, Y. Zhang, D. Chen, and G. Xu, “Character-based convolutional grid neural network for breast cancer classification,” in *2017 International Conference on Green Informatics (ICGI)*, p. 31, Fuzhou, China, 2017.

[8] T. R. Mahesh, V. Dhilip Kumar, V. Vinoth Kumar, J. Asghar, B. M. Bazezew, and R. N. V. Vivek, “Blended ensemble learning prediction model for strengthening diagnosis and treatment of chronic diabetes disease,” *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 4451792, 9 pages, 2022.

[9] G. Sindhu Madhuri, T. R. Mahesh, and V. Vivek, “7 A novel approach for automatic brain tumor detection using machine learning algorithms,” in *Big Data Management in Sensing: Applications in AI and IoT*, pp. 87–102, River Publishers, 2021.

[10] M. F. Aslam, C. Yunus, K. Sabanci, and D. Akif, “Breast cancer diagnosis by different machine learning methods using blood analysis data,” *International Journal of Intelligent System and Applications in Engineering*, vol. 6, no. 4, pp. 289–293, 2018.

[11] I. H. Witten and E. Frank, “Data mining practical machine learning tools and techniques,” Morgan Kaufmann, The United States of America, 2nd edition, 2005.

[12] K. Shwetha, M. Spoorthi, S. S. Sindhu, and D. Chaithra, "Breast cancer detection using deep learning technique," *International Journal of Engineering Research & Technology*, vol. 6, no. 13, pp. 1–4, 2018.

[13] H. K. Shashikala, T. R. Mahesh, V. Vivek, M. G. Sindhu, C. Saravanan, and T. Z. Baig, "Early detection of spondylosis using point-based image processing techniques," in *2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, pp. 655–659, Bangalore, India, 2021.

[14] B. Akbugdev, "Classification of breast cancer data using machine learning algorithms," in *2019 Medical Technologies Congress (TIPTEKNO)*, pp. 1–4, Izmir, Turkey, 2019.

[15] K. K. Jha, R. Jha, A. K. Jha, M. A. M. Hassan, S. K. Yadav, and T. Mahesh, "A brief comparison on machine learning algorithms based on various applications: a comprehensive survey," in *2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pp. 1–5, Bangalore, India, 2021.

[16] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113–127, 2005.

[17] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016.

[18] A. S. Hussein, T. Li, C. W. Yohannese, and K. Bashir, "A-SMOTE: a new pre-processing approach for highly imbalanced datasets by improving SMOTE," *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, pp. 1412–1422, 2019.

[19] M. R. Sarveshvar, A. Gogoi, A. K. Chaubey, S. Rohit, and T. R. Mahesh, "Performance of different machine learning techniques for the prediction of heart diseases," in *2021 International Conference on Forensics, Analytics, Big Data, Security (FABS)*, pp. 1–4, Bengaluru, India, 2021.

[20] M. Tahmooresi, A. Afshar, B. Bashari Rad, K. B. Nowshath, and M. A. Bamiah, "Early detection of breast cancer using machine learning techniques," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 10, no. 3-2, pp. 21–27, 2018.

[21] E. A. Bayrak, P. Kirci, and T. Ensari, "Comparison of machine learning methods for breast cancer diagnosis," in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pp. 1–3, Istanbul, Turkey, 2019.

[22] V. A. Telsang and K. Hegde, "Breast cancer prediction analysis using machine learning algorithms," in *2020 International Conference on Communication, Computing and Industry 4.0 (C2I4)*, Bangalore, India, 2020.

[23] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, New York, NY, USA, 2016.

[24] Y. Chang and X. Chen, "Estimation of chronic illness severity based on machine learning methods," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 1999284, 13 pages, 2021.

[25] V. D. Soni, "Chronic disease detection model using machine learning techniques," *International Journal of Scientific & Technology Research*, vol. 9, no. 9, pp. 262–266, 2020.

[26] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: a review," *Egyptian Informatics Journal*, vol. 19, no. 3, pp. 179–189, 2018.