Taylor & Francis
Taylor & Francis Group

Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/taut20

GBDTMO: as new option for early-stage breast cancer detection and classification using machine learning

Vibith A. S. & Jobin Christ M C

To cite this article: Vibith A. S. & Jobin Christ M C (2023) GBDTMO: as new option for early-stage breast cancer detection and classification using machine learning, Automatika, 64:4, 858-867, DOI: 10.1080/00051144.2023.2226946

To link to this article: https://doi.org/10.1080/00051144.2023.2226946

Published online: 27 Jun 2023.

Submit your article to this journal

Article views: 572

View related articles

View Crossmark data

AUTOMATIKA
2023, VOL. 64, NO. 4, 858–867
https://doi.org/10.1080/00051144.2023.2226946

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

GBDTMO: as new option for early-stage breast cancer detection and classification using machine learning

Vibith A. S.[a] and Jobin Christ M C[b]

[a]Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Sriperumbudur, Tamilnadu, India;
[b]Department of Biomedical Engineering, Rajalakshmi Engineering College, Thandalam, Tamilnadu, India

ABSTRACT
Breast cancer is the second leading cause of disease death in women, after lung and bronchus cancer. According to measurements, mammography misses breast cancer in 10% to 15% of cases for women aged 50 to 69 years. In the current study, we used the Wisconsin breast cancer dataset to develop a two-stage model for breast cancer diagnosis. The main goal of this study effort is to effectively carry out feature selection and classification tasks. Gradient Boosting Decision Tree-based Mayfly Optimisation (GBDTMO), an innovative and efficient breast cancer diagnostic machine learning system, is provided. In the second stage, we employ a Mayfly search to determine which subset of traits is the best. Two more well-known datasets on breast cancer, the ICCR and the Cancer Corpus, were also compared for classification accuracy. The accuracy of the suggested GBDTMO model was higher than that of the existing GBDT and Practical Federated Gradient Boosting Decision Tree (PFGBDT), which had accuracy values of 93.25% and 94.25%, respectively. Similarly, the recall, F-measure, and ROC area values were 98.52%, 97.52%, and 96.32%, respectively. Furthermore, it demonstrated a lower RMSE of 0.98 than the existing GBDT and PFGBDT.

ARTICLE HISTORY
Received 22 February 2023
Accepted 13 June 2023

KEYWORDS
Breast cancer; machine learning; accuracy; classification; ensemble; Mayfly search

1. Introduction

Breast cancer is a crippling illness that is becoming more common among women, but it can also strike men. Breast cancer can adversely harm one's health and ultimately cause death if it is not detected in time or during the early stages of tumour formation. For effective treatment, the early stage diagnosis of tumour development is essential. Although the disease has some symptoms that can aid in the diagnosis, the majority of the symptoms are universal to many other conditions, making the tumour's progression undetectable. The biologists had just a few traditional techniques for microscopic illness diagnostics, but these techniques could not correctly identify the tumour responsible for the breast cancer. Cancer is now one of the leading causes of morbidity and mortality all over the world. Cancer has resulted in the deaths of approximately 14.5 million people, with the number expected to rise to more than 28 million by 2030. After lung and bronchus cancer, breast cancer is the second leading cause of disease death in women. Furthermore, breast cancer growth accounts for 30% of all new disease cases [1]. From one end of the globe to the other, breast cancer is the most common type of cancer in females, and it is the second most common disease overall, with millions of new cases in 2018. The five-year endurance rate for ladies analysed at earlier stages is more than 90%, and it is around 15% for ladies determined to have the most exceptional stage [2]. Although breast disease can happen in men, it is extremely uncommon.

Breast cancer is examined and classified using a combination of techniques such as imaging, physical examination, and biopsy. Because of the disease's complexity, early detection will improve patient survival. A dependable approach to diagnosis is required. Despite the use of numerous breast cancer datasets, disease prediction has become both interesting and difficult [3]. Machine learning methods greatly assist researchers in the field. However, many classification methods result in a poor diagnosis. Various case patterns are used to forecast the outcome of this case. The gradient boosting decision tree is a sophisticated ensemble model used for classification and regression. However, the method creates a poor prediction model in terms of speed and accuracy [4]. Furthermore, the boosting model constructs a sequential model, but the error rate is high. As a result, the need for a reliable approach remains a significant challenge [5]. With this inspiration, the paper includes the following contributions towards the accurate detection and classification of breast cancer.

CONTACT AS Vibith vibithas.cse@outlook.com, vibithas87@gmail.com Sri Venkateswara College of Engineering, Pennalur Village, Chennai - Bengaluru Highways Sriperumbudur, Tamilnadu- 602 117

1.1. Contribution

- The main objective of this research work is to present a novel and effective breast cancer diagnostic machine learning algorithm, Gradient Boosting Decision Tree-based Mayfly Optimization (GBDTMO), by hybridizing the existing Gradient Boosting Decision Tree algorithm and the Mayfly Optimization algorithm to perform feature selection and classification tasks efficiently.
- Clinical risk prediction is used in a variety of applications. As the number of patient features increases, so does the model's complexity. As a result, the classification accuracy of high-dimensional clinical data is a significant issue. To validate the proposed GBDTMO's accuracy results, it was compared to three well-known datasets: Breast Cancer Wisconsin, ICCR, and Cancer Corpus.
- The performance time, sensitivity, specificity, accuracy, precision, recall, F-Measure, and ROC Area are studied and analysed, allowing healthcare professionals to provide early diagnosis and treatment to patients.
- Furthermore, the experimental results are compared to recently proposed machine learning approaches such as Gradient Boosting Decision Tree Algorithms (GBDT) and Practical Federated Gradient Boosting Decision Trees (PFGBDT) to validate the efficiency of the proposed GBDTMO. The GBDT ensemble model, which is used in various classification and regression problems, is a well-known machine learning algorithm that has demonstrated success in a variety of domains. The proposed GBDTMO performance is validated by using the GBDT algorithm as a benchmark. The proposed GBDTMO outperforms existing algorithms for breast cancer feature selection and classification.

The rest of the paper's section is organized as follows: Section 2 discusses the various related works and pitfalls in the implementation of machine learning approaches in the prediction and diagnosis of breast cancer disease. Section 3 elaborates on the proposed GBDTMO model for breast cancer diagnosis. Section 4 presents the results and discussion of the proposed GBDTMO model's performance. Finally, Section 5 summarizes the key findings and areas for future development.

2. Related works

Necessity because of the benefits it can bring to clinical patient care management. A significant number of research companies working in the domains of biomedicine and bioinformatics have investigated the application of machine learning strategies to the essential problem of classifying cancer patients into high and low risk categories [6]. Numerous of these techniques, including Decision Trees, Support Vector Machines, and K-Nearest Neighbours, have been extensively used in cancer research to develop prediction models that help decision-makers make more informed and trustworthy choices. As a result, an ML strategy was used to the modelling of cancer development. Different supervised ML methods, as well as a wide range of input characteristics and Data Samples, form the basis of the prediction models covered here.

Cancer progression is a group of conditions in which cells in the body collide to form irregularities known as harmful cancer. Breast cancer is the most commonly studied disease among women in 140 of the world's 184 countries [7]. Mammography fails to detect breast cancer in 10% to 15% of cases for women aged 50 to 69 years, according to measurements [2]. Breast cancer growth may begin in a variety of areas of the bosom, including other organs. Breast disease will spread to different parts of the body if cancer cells enter the lymph framework or circulatory system. Early detection of breast cancer using appropriate methods may result in a decrease in female mortality rates [2]. These cells proliferate haphazardly, spread into surrounding tissues, and crowd out the normal cells. Breast cancer growth investigation used a multi-surface methodology to isolate the example of breast cytology. This study used 11 cytological characteristics of breast disease that differ between harmful and harmless data using a standard Machine Learning approach. Ordinary occurrences are separated and chosen to be stored as idea portrayals. This technique was used to order a harmful growth that included 369 breast disease patients from the Wisconsin dataset, which can be downloaded for free from the UC Irvine Machine Learning repository [8].

Jahangeer [9] focuses on the device used to detect breast cancer, the mammogram, which is a widely used and effective device. It is an imaging technique used to diagnose breast cancer based on mammogram images. The recorded mammogram images are used as input in the study. Further, the fully automatic deep learning technique was used. Specificity, accuracy, recall, sensitivity, Jaccard coefficient, precision, missed classification, and F score was used to calculate the outcome. Finally, comparing mammograms to other cutting-edge approaches will remain a challenge.

Juneja and Rana [10] clarified that if there is an occurrence of breast cancer, the presence can be reduced by distinguishing the growth at an early stage. The endurance rate can be increased if the growth is detected early and does not spread to other organs. Mammography can detect different breast tissues based on region size and criticality boundaries. The Machine Learning Algorithm can be used on these breast tissue provisions to determine the likelihood of growth recurrence. A specific component-based decision tree algorithm is recommended to predict the likelihood of breast cancer growth occurrence. To improve the accuracy of breast cancer growth prediction using breast tissue data, the paper presented a decision tree model. The results demonstrate that the proposed model has improved the accuracy of breast disease forecasting. As a result, determining the correct problem will continue to be difficult.

Murtaza et al. [11] used deep neural networks to study medical imaging multimodalities. For breast cancer, mammograms and histopathologic images were used. The input was taken from 49 academic studies, of which 8 were chosen as a unique academic repository. For searching and selecting studies, a systematic review methodology is used. The limited networks can be properly accessed, and higher data complex patterns must be learned to solve. As a result, critical analysis of breast cancer classification is the higher research finding among scholars in this current researcher's domain. As a result, working on larger networks will continue to be difficult.

Tabrizchi et al. [4] investigated ideal dataset classification using group learning by combining the GBDT and MVO (Multi-Verse Optimizer) into a vigorous classifier. The primary goal of this current study was to improve the precision of breast cancer growth classification. Classification calculations can be used to collect data and information as a directed machine learning strategy. Machine learning strategies can be extremely beneficial to specialists in this field. The GBDT and MVO calculations were thoroughly explained. As a result, the subsequent discussions demonstrated that the proposed group technique outperforms other existing strategies in the field, and the breast cancer classification execution can be improved by utilizing the proposed classifier.

Vijayarajeswari [12] presents an order of mammograms with highlights extracted using the Hough change. The mammograms are pre-processed at the start, which increases the contrast between required articles and undesirable foundation commotion. This strategy was tested on 95 mammogram images collected and arranged using SVM. The results show that the proposed strategy successfully groups the unusual classes of mammograms. To improve precision, bosom malignant growth was discovered using amplification estimation in this work. Better results are obtained by broadening the force class in the assessment amplification. The use of greater force features such as mean, fluctuation, and entropy can affect the results. Therefore obtained a precision range of 94 percent by using the SVM classifier, which is higher when compared to other classifiers such as LDA (Linear discriminant analysis), which has a precision range of only 86 percent. As a result, taking a large number of images will continue to be a challenge.

Hou et al. [13] used the fruit fly optimization algorithm as a search methodology and developed a covering-based component selection strategy known as the twofold binary fruitfly optimization algorithm (BIFFOA). The extensive trials on 25 datasets show that the BIFFOA presentation outperforms a few best-in-class algorithms. The channel-based technique evaluates information highlights based on data or measurable measurements. The grouping precision of the covering-based strategy is generally higher than that of the separated-based strategy, which does not rely on any learning algorithm. Although, in contrast to recent element choice calculations, the work on half-breed calculation, BIFFOA-EPD-Tour has a distinct advantage in managing highlight choice issues. As a result, putting this technique into practice will be a challenge.

Gradient boosting decision trees are widely used in many machine learning tasks. However, to increase the performance, the hyper-parameters must be manually tuned, which is a time-consuming process since it involves repeated training. Li et al. [14] concentrated on Practical Federated Gradient Boosting Decision Trees (PFGBDT), which distributes data samples with the same features among multiple parties under relaxed privacy constraints. When compared to traditional training with each party's local data, the approach significantly improves predictive accuracy. However, improving accuracy remains a challenge.

Khuriwal and Mishra [15] demonstrated how deep learning technology can be used to diagnose breast cancer disease using the UCI dataset. This study demonstrates how they are implementing profound learning innovations on the Wisconsin Breast Cancer Database, which we have found to be extremely useful for determining breast cancer. The research was divided into three sections: first, they gathered datasets and used pre-preparing calculations for scaled and channel information; second, they divided datasets for preparing and testing reasons; and finally, they created some diagrams for perception information. They used convolutional neural networks to determine breast cancer and also ran the same dataset through other AI calculations, such as Neural Network, Support Vector Machine, and Random Forest. As a result, in the study, they used 11 provisions for analysing breast cancer that had after pre-planning. The paper also compared the profound learning calculations to other AI calculations, and the proposed framework outperforms. As a result, implementing this technique into daily life will be difficult.

The proposed novel GBDTMO algorithm is explained in detail in this section. Medical administrations assert that Electronic Health Records have a monetary value for medical interventions. Text notes, results, reports, symptoms, and other clinical data are stored in EHR databases. As a result, the algorithm is needed to validate those clinical data. The emergence of machine learning methods enables us to classify, identify, and compute outcomes to assess risk and support clinical practice. EHR data must now go through several transformations before it can be accessed by humans [16,18–21–22].

3. Proposed GBDTMO model

The primary goal of the proposed model is to present a novel and effective breast cancer diagnostic machine learning algorithm, Gradient Boosting Decision Tree-based Mayfly optimization (GBDTMO), by hybridizing the existing Gradient Boosting Decision Tree algorithm and the Mayfly optimization algorithm to efficiently perform feature selection and classification tasks. The model's performance was investigated in the selection of the critical feature and effectively improved the model's performance. The feature selection is evaluated by the machine learning algorithm by ranking is based on the weighted score. For comparison, the feature set provided by the machine learning model with the highest accuracy was chosen.

The model is divided into phases that classify the benign and malignant categories. Iteratively, multiple weak learners are combined into a single strong learner in GBDTMO. An ensemble machine learning model is used in the algorithm. The gradient boosting decision tree is hybridized with the mayfly optimization algorithm in this study to improve the efficiency of feature selection. As shown in Equation (1), the model is trained for a training set of known a and their corresponding b values.

$$\{(a_1, b_1), \ldots\ldots\ldots (a_n, b_n)\} \tag{1}$$

The objective is to achieve the approximation function $\hat{G}$ (a) that minimizes the loss function $G_L(b, G(a))$, as expressed in Equation (2).

$$\hat{G} = \text{argmin}(G_L(b, G_L(a))) \tag{2}$$

Here b are the real values that are being approximated. The weighted sum function $h_i(a)$, as expressed in Equation (3), is used to classify weak learners for a class of H [4].

$$G(a) = \sum_{i=1}^{N} \gamma_i h_i(a) + Constant \tag{3}$$

As a result $G_n(a)$, as expressed in Equation (4), it reduces the risk and minimizes the loss function.

$$G_n(a) = G_{n-1}(a) + \text{argmin}_{h_n} \in H \left[\sum_{i=1}^{m} G_L(b_i, G_{n-1}(a_i) + h_n(a_i)) \right] \tag{4}$$

As a result, the learner's function $h_n \in H$ is impaired. While selecting the best function of h, an arbitrary loss function G_L occurs, resulting in an infeasible computation. As a result, this computation is infeasible in the proposed novel model and is solved using the mayfly optimization algorithm. Thus, rather than generating a new solution for each stage, the search index is improved, as shown in Equation (5). The Mayfly Optimization Algorithm is an intelligent algorithm that provides evolutionary computation by imitating the foraging behaviours of mayflies. The mayfly can detect various odours and smell food sources up to forty kilometres away. The fly has an acute vision when it comes to food. The food process of discovery is described in the following steps: (1) detects the scent of food and flies to the location; (2) with the assistance of sensitive vision, approaches the food location. (3) The other flies congregate at the identified location and move towards the target [17]. Mayfly's iterative food-searching process is depicted in Figure 1.

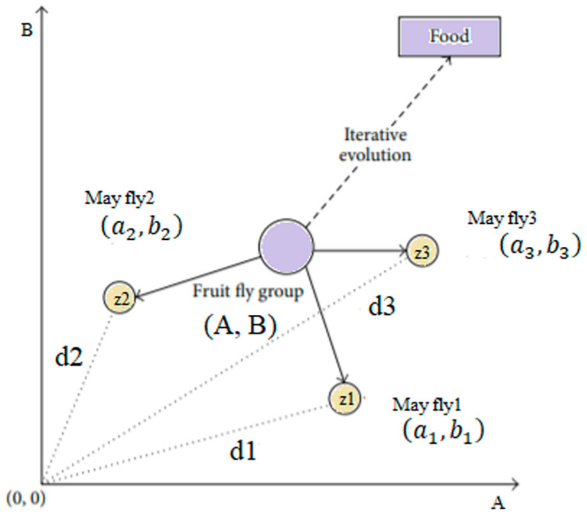

Figure 1. Mayfly search process.

$$\gamma = \gamma_{\max}.\exp\left(\frac{\log\gamma_{\min}}{\gamma_{\max}}\right).\frac{\text{Iter}}{\text{Iter}_{\max}} \tag{5}$$

The mayfly radius γ is expressed for each iteration. What $\gamma_{\max}$ the maximum and $\gamma_{\min}$ minimum radius. Iter is the current iteration, and $\text{Iter}_{\max}$ is the maximum iteration.

$$a_{i,j} = \{\beta_j \pm \gamma\ .rand(); if\ j = d\} \tag{6}$$

$$a_{i,j} = \{\beta_j\ otherwise\ j = 1, 2, \ldots .Z\} \tag{7}$$

where $d \in \{1, 2, \ldots .Z\}$ is the variable of the randomly indexed uniformly distributed gradient boosting decision tree; Z is the solution dimension, and rand () returns a random number between [0,1]. $a_{i,j}$ as expressed in Equations (6) and (7) is the updated iteration and optimal solution for the jth dimension. As a result, the minimization problem is solved, and the steepest descent is used. Thus, Equations (8) and (9) are used to update the novel model of GBDTMO.

$$G_n(a) = G_{n-1}(a) - \gamma_n \sum_{i=1}^{m} \emptyset_{G_{n-1}} G_L(b_i, G_{n-1}(a_i)) \tag{8}$$

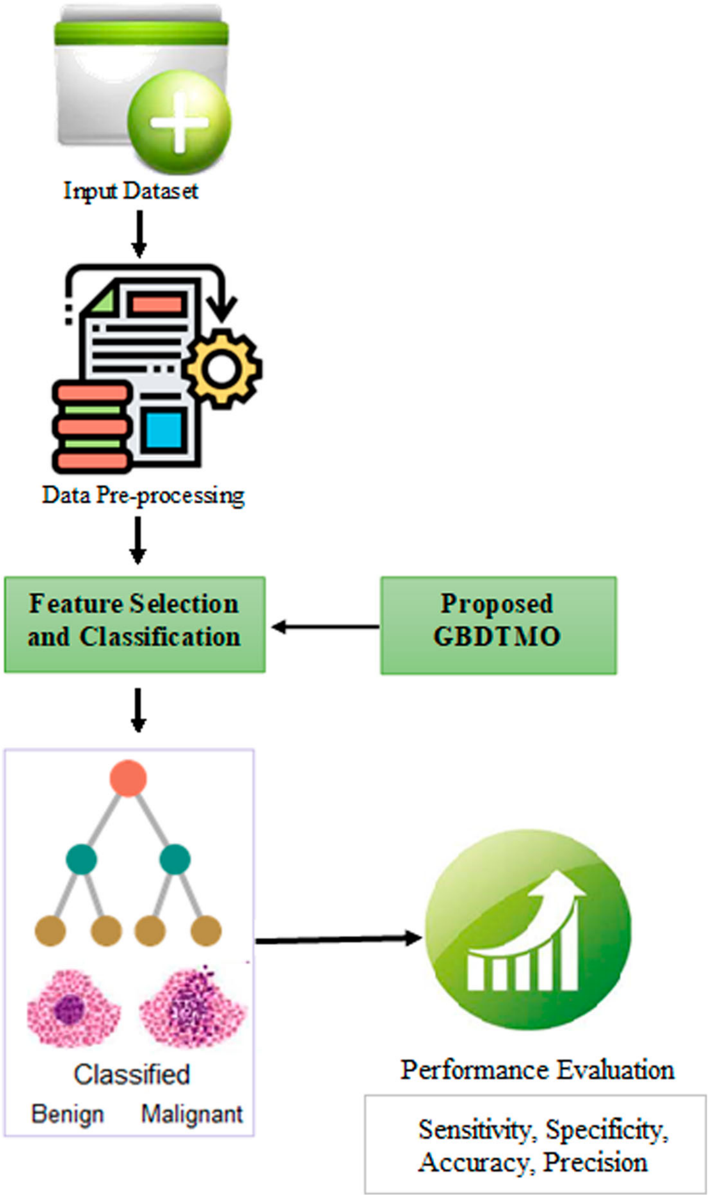

Figure 2. Proposed GBDTMO model framework.

$$\gamma_n = argmin\sum_{i=1}^{m} G_L(b_{i,}G_{n-1}(a_i)) - \gamma \emptyset_{G_{n-1}} G_L(b_{i,}G_{n-1}(a_i)) \tag{9}$$

The derivatives are taken concerning function G_i for $i \in \{1, 2, \ldots .n\}$. In this case, G_L is the closest gradient for h, the candidate function, and H, the finite set. The mayfly optimization is used to compute the coefficient γ.

The derivatives G_{n-1} for the loss function are calculated for each iteration, and the goal is to select the current learner that best fits the γ_n. In subsequent iterations, the step length of this gradient produces the lowest loss, and the loss function gradually decreases.

Figure 2 depicts the proposed GBDTMO framework. The datasets were transformed into a set of CSV-based data that was preferred for the conditions of each patient. The conditions are clump thickness, cell size-uniformity, cell shape-uniformity, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses, and class. Thus, researchers were able to assess and classify the patient files from the EHR into classes coded as Malignant, value 0, and Benign, value 1 as presented in Table 1. Each instance was examined for correctness, and incorrect samples were identified and removed from the classification.

Table 1. Breast cancer patient class.

Confusion Matrix	Predicted class	
Actual class	Malignant	Benign
Malignant	True Positive (TP)	False Positive (FP) Error
Benign	False Negative(FN) Error	True Negative (TN)

Performance time, sensitivity, specificity, accuracy, precision, recall, F-measure, and ROC area are all calculated during the performance analysis. Accuracy is a measure of weighted arithmetic mean, expressed as an inversion of precision in Equation (10)

$$Accuracy = \frac{TP}{(TP + TN + FP + FN)} \tag{10}$$

The ratio of true positives to the sum of false negatives and true positives is referred to as sensitivity. This is also known as the true positive rate (TPR), which is expressed in Equation (11).

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{11}$$

Specificity is defined as the proportion of correctly identified negatives, also known as the false positive rate, as expressed in Equation (12).

$$\text{sensitivity} = \frac{TN}{TN + FP} \tag{12}$$

Precision is defined as the fraction of instances in the field of information retrieval and the fraction of features retrieved, as expressed in Equation (13)

$$precision = \frac{TP}{TP + FP} \tag{13}$$

The proportion of recovered related instances is referred to as a recall. As a result, both accuracy and recall are dependent based on an understanding of significance and measurement. The formula in Equation (14) is used to estimate it.

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

The F-measure is defined as the harmonious mean of accuracy and recall, as expressed in Equation (15).

$$F - measure = \frac{(2 * precision * Recall)}{precision + Recall} \tag{15}$$

The accuracy of the data was calculated utilizing the RMSE analysis. RMSE measures the observed value differences with the predicted value. The deviations of the RMSE are the residuals. Suppose the calculation of the data samples used for estimation leads to errors. RMSE is a significant accuracy measure utilized in many best models for a particular dataset as expressed in Equations (16) and (17). Mean Square Error assesses the quality of the Predictor. For n prediction vector, y observed vector and $\hat{y}$ the predicted value. The MSE calculation expressed in Equation (16)

$$MSE = \frac{\sum_{t-1}^{N} (y_P - \hat{y}_P)}{N} \tag{16}$$

Where $\frac{1}{N}\sum_{i=1}^{N}$ is the error calculation for the observed and predicted vector. Taking the square root for Equation (16), we calculate the RMSE, expressed in Equation (17)

$$RMSE = \sqrt{\frac{\sum_{t-1}^{N} (y_P - \hat{y}_P)}{N}} \tag{17}$$

Section 4 discusses the evaluation results of the proposed GBDTMO.

4. Results and discussion

The model uses a machine-learning algorithm to identify useful features and classifications for breast cancer diagnosis using the GBDTMO, demonstrating its performance using the Wisconsin breast cancer dataset. The dataset contains 699 instances, and the attributes chosen for evaluation include tumour size, Inv-nodes, node-caps, deg-malign, and irradiated. Clump thickness, cell size and shape uniformity, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli. The proposed model's classification accuracy is also compared to two other popular datasets, ICCR (International Collaboration on Cancer Reporting) and Cancer Corpus. The International Society of Breast Pathology and Singapore General Hospital collaborated to create the ICCR dataset, which allows for the reporting of surgically removed ipsilateral lymph nodes from breast tumours.

4.1. Experimental setup

The experiment was carried out on a PC with a 3.6 GHz CPU, 12 GB of RAM, and the Windows 10 operating system. All codes are implemented in MATLAB 2018a to extract breast cancer features and assist clinical experts in making accurate decisions.

The aim is to assess the proposed GBDTMO model for categorizing patient profiles into two groups: malignant and benign. This algorithm's goal, in this case, is to perform feature selection and classification on the given dataset and predict the patient category. Based on the attributes provided, the algorithm iterates for each patient. The proposed GBDTMO model's performance is evaluated and presented in Table 2. The performance time, sensitivity, specificity, accuracy, precision, recall, F-Measure, and ROC Area are studied and analysed, allowing healthcare professionals to provide early diagnosis and treatment to patients. Furthermore, the experimental results are compared to recently proposed machine learning approaches such as Gradient Boosting Decision Tree Algorithms (GBDT) and Practical Federated Gradient Boosting Decision Trees (PFGBDT) to validate the efficiency of the proposed GBDTMO. The GBDT ensemble model, which is used in various classification and regression problems, is a well-known machine learning algorithm that has

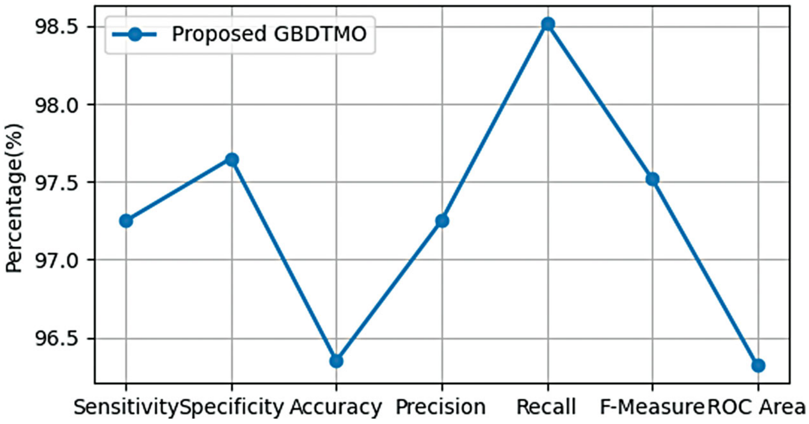

Figure 3. Performance evaluation – proposed GBDTMO.

Table 2. Evaluation results: proposed GBDTMO's performance in comparison to the existing model.

Performance Metrics	GBDT	PFGBDT	Proposed GBDTMO
Performance Time(secs)	18.2	16.2	14.25
Sensitivity	92.32	93.21	97.25
Specificity	91.25	92.32	97.65
Accuracy	93.25	94.25	96.35
Precision	89.32	90.25	97.25
Recall	92.32	93.25	98.52
F-Measure	91.25	94.32	97.52
ROC Area	93.6	94.25	96.32

Table 3. Error measure: proposed GBDTMO in comparison to the existing model.

Parameters	GBDT	PFGBDT	Proposed GBDTMO
RMSE	1.25	1.02	0.98

demonstrated success in a variety of domains. The proposed GBDTMO performance is validated by using the GBDT algorithm as a benchmark.

Table 2 compares the proposed GBDTMO model to the existing models in terms of performance. When compared to the existing GBDT and PFGBDT accuracy of 93.25% and 94.25%, respectively, the proposed GBDTMO model had a higher accuracy of 96.35%. In terms of sensitivity, specificity, and precision, the proposed GBDTMO model scored 97.25%, 97.65%, and 97.25%, respectively. GBDT and PFGBDT have values of 92.32%, 91.25%, 89.32%, and 93.21%, 92.32%, 90.25%, respectively. Similarly, when compared to the other two algorithms, the proposed GBDTMO model performed better in terms of recall, F-measure, and ROC area, with values of 98.52%, 97.52%, and 96.32%, respectively. As a result, the proposed GBDTMO outperforms existing breast cancer feature selection and classification algorithms.

Figure 3 depicts the proposed GBDTMO model's performance in terms of sensitivity, specificity, accuracy, precision, recall, F-Measure, and ROC Area. The proposed model produced a sensitivity of 97.25%, specificity of 97.65%, the accuracy of 96.35%, precision of 97.25%, recall of 98.52%, F-Measure of 97.52%, and ROC Area of 96.32%.

Figure 4 compares the proposed GBDTMO model to the existing two models, Practical Federated Gradient Boosting Decision Trees (PFGBDT) [14] and Gradient Boosting Decision Tree Algorithm (GBDT) (Anghel *et al.* 2018). The proposed GBDTMO model has a sensitivity of 97.25%, while the GBDT and PFGBDT have 92.32% and 93.21%, respectively. The proposed GBDTMO model's specificity, accuracy, precision, recall, F-Measure, and ROC Area are 97.65%, 96.35%, 97.25%, 98.52%, 97.52%, and 96.32%. The other two existing models, on the other hand, produced lower results. As a result, the proposed model's efficiency is demonstrated.

Figure 5 compares the proposed GBDTMO model's performance time evaluation to the existing two models, PFGBDT and GBDT. In terms of computation time, the lower the value, the better the performance. The existing GBDT and PFGBDT computation times are 18.2 and 16.2 s, respectively. In contrast, the proposed GBDTMO outperforms with a shorter computation time of 14.5 s, demonstrating the model's efficiency.

Table 3 compares the proposed GBDTMO model's RMSE (Root Mean Square Error) to the existing two models, Practical Federated Gradient Boosting Decision Trees (PFGBDT) and Gradient Boosting Decision Tree Algorithm (GBDT). The existing GBDT and PFGBDT have RMSE values of 1.25 and 1.02, respectively. The proposed GBDTMO, on the other hand, outperforms with a lower RMSE of 0.98. The lower the RMSE, the more accurate the prediction.

Figure 6 compares the RMSE of the proposed GBDTMO model to the RMSE of the existing two models, PFGBDT and GBDT. In RMSE, errors are squared before being averaged, giving large errors a high weight. The RMSE can be anywhere between 0 and infinity. Lower values, on the other hand, are desirable and preferable. Since RMSE is a good metric for assessing the prediction model's accuracy. The RMSE values of the existing GBDT and PFGBDT are 1.25 and 1.02, respectively. In contrast, the proposed GBDTMO outperforms with a lower RMSE of 0.98.

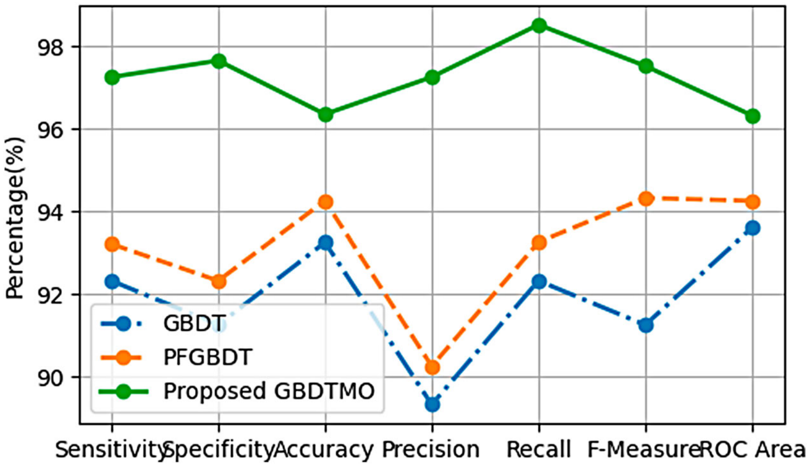

Figure 4. Performance evaluation – proposed GBDTMO in comparison to the existing model.

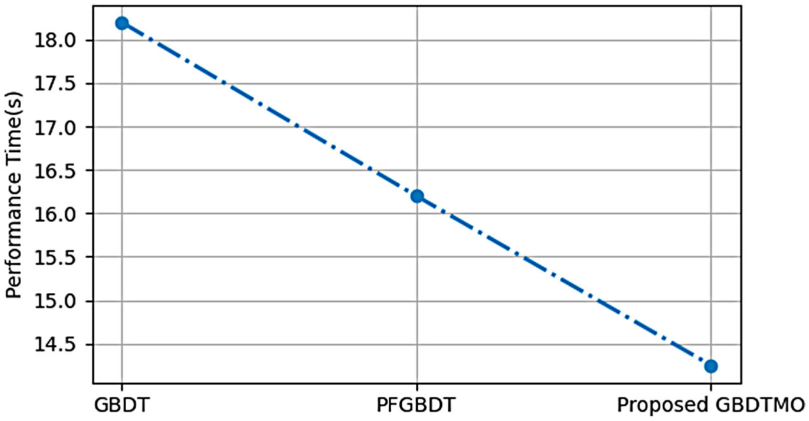

Figure 5. Performance time evaluation – proposed GBDTMO in comparison to the existing model.

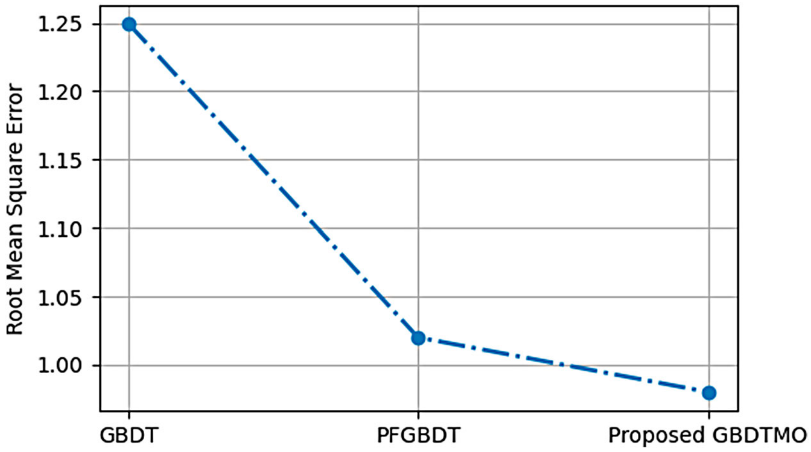

Figure 6. RMSE error measure – proposed GBDTMO in comparison to the existing model.

Clinical risk prediction is used in a variety of applications. As the number of patient features increases, so does the model's complexity. As a result, the classification accuracy of high-dimensional clinical data is a significant issue. To validate the proposed GBDTMO's accuracy results, it was compared to three well-known datasets: Breast Cancer Wisconsin, ICCR, and Cancer Corpus presented in Table 4.

The dataset comparison for classification accuracy is shown in Table 4. In this study, three well-known breast cancer datasets were evaluated: Wisconsin, ICCR, and Cancer Corpus. In the ICCR dataset, the proposed GBDTMO algorithm achieves a classification accuracy of 98.58%. Wisconsin and Cancer Corpus, on the

Table 4. Classification accuracy: dataset-based comparison.

Dataset	Accuracy
Wisconsin	96.35
ICCR	98.85
Cancer Corpus	97.25

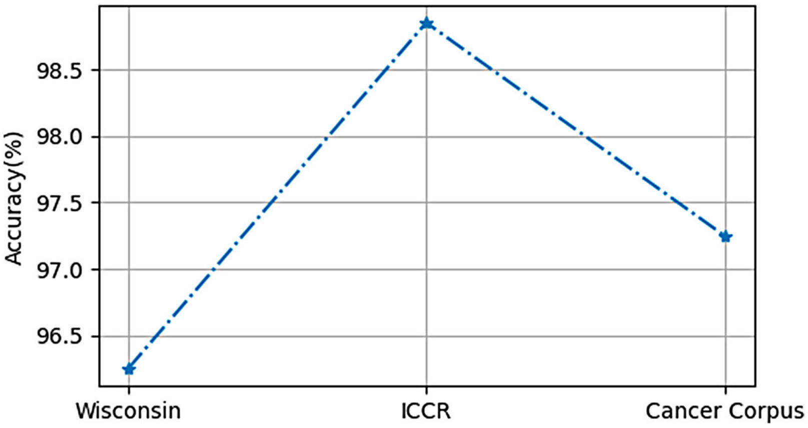

Figure 7. Classification accuracy: dataset-based comparison.

Algorithm: GBDTMO

```
1. Initialize $G_o(a) = argmin \sum_{i=1}^{m} G_L(b_i\gamma)$
2. For $n = 1$ to $N$;
3. Define a switch probability $p \varepsilon [0,1]$
4.         For $i = 1,2, \ldots M$ compute
5. $\gamma_{in} = \left[\frac{\partial G_L(b_i, G(a_i))}{\partial G(a_i)}\right] G = G_{n-1}$
6.         End For
7. Fitting a gradient boosting decision tree each iteration is calculated as
8.         $\gamma = \gamma_{\max}.\exp\left(\frac{\log\gamma_{\min}}{\gamma_{\max}}\right)\cdot\frac{\mathrm{Iter}}{\mathrm{Iter}_{\max}}$
9.         For the derivates $G_i$ fori $\in \{1, 2, \ldots .n\}$
10.          Initalize Population(A,B)
11.                $A_i = A_0 + rand$
12.                $B_i = B_0 + rand$
13.                Compute Distance ($d_i$)
14.                Calculate the derivates ($G_i$)
15.                $\gamma_n = argmin \sum_{i=1}^{m} G_L(b_{i,}G_{n-1}(a_i)) - \gamma \emptyset_{G_{n-1}} G_L(b_{i,}G_{n-1}(a_i))$
16.                Best solution ($z_i$) $= G_L$
17.                Repeat 10 and 16 until condition satisfied
18.          End For
19. Update
20. $G_n(a) = G_{n-1}(a) - \gamma_n \sum_{i=1}^{m} \emptyset_{G_{n-1}} G_L(b_{i,}G_{n-1}(a_i))$
         $\hat{G}(a) = G_n(a)$
21. Output
22. End For
```

other hand, have accuracy rates of 96.35% and 97.25%, respectively.

Figure 7 shows a comparison of the proposed GBDTMO model's classification accuracy with three well-known datasets: Wisconsin, ICCR, and Cancer Corpus. Higher accuracy indicates that the model's prediction performance has improved. In the case of the proposed GBDTMO model, it shows superior performance in distinguishing between malignant and benign tumours in all three datasets. However, the ICCR dataset has a higher accuracy of 98.58%, whereas Wisconsin and Cancer Corpus have an accuracy of 96.35% and 97.25%, respectively. As a result, the implemented GBDTMO model extracts useful features and distinguishes between malignant and benign tumours. As a result, the algorithm is effective in making accurate predictions and assisting the healthcare field in assessing the possibility of breast cancer diseases.

5. Conclusion

ML-based decision support systems have proven to be very efficient and effective in the diagnosis of breast cancer. Using ensemble learning, this research article presents a novel and effective breast cancer diagnostic machine learning algorithm, Gradient Boosting Decision Tree-based Mayfly Optimization (GBDTMO), by combining the existing GBDT and the Mayfly Optimization algorithm to efficiently perform feature selection and classification tasks. The Wisconsin breast cancer dataset was used in this research paper to demonstrate the proposed model's high reliability and effectiveness. The following parameters are studied and analysed: performance time, sensitivity, specificity, accuracy, precision, recall, F-Measure, and ROC Area. The proposed GBDTMO model had a higher accuracy of 96.35% when compared to the existing GBDT and PFGBDT accuracy of 93.25% and 94.25%, respectively. Similarly, the proposed GBDTMO model outperformed the other two algorithms in terms of recall, F-measure, and ROC area, with values of 98.52%, 97.52%, and 96.32%, respectively. Furthermore, the proposed GBDTMO outperforms with a 14.5-second computation time, demonstrating the model's efficiency. In addition, it showed a lower RMSE of 0.98 than the existing GBDT and PFGBDT values of 1.25 and 1.02, respectively. Thus, the experimental results demonstrated that the proposed GBDTMO performs significantly better than other existing methods in this field for breast cancer classification. Other algorithms can be hybridized in future work to adjust the parameters that suit other biological diseases. It can be extended for breast cancer classification using medical images in the case of tumour diagnosis. Further, the use of GBDTMO with more classifiers like SVM, Artificial Neural Networks (ANN) can be demonstrated to verify the extensibility of the proposed model.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Alom MZ, Yakopcic C, Nasrin MS, et al. Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network. J Digit Imaging. 2019;32(4):605–617. doi:10.1007/s10278-019-00182-7

[2] Rahman MM, Ghasemi Y, Suley E, et al. Machine learning based computer aided diagnosis of breast cancer utilizing anthropometric and clinical features. IRBM. 2021;42(4):215–226. doi:10.1016/j.irbm.2020.05.005

[3] Rakshit P, Zaballa O, Pérez A, et al. A machine learning approach to predict healthcare cost of breast cancer patients. Sci Rep. 2021;11(1):1–13. doi:10.1038/s41598-021-91580-x

[4] Tabrizchi H, Tabrizchi M, Tabrizchi H. Breast cancer diagnosis using a multi-verse optimizer-based gradient boosting decision tree. SN Appl Sci. 2020;2(4):1–19. doi:10.1007/s42452-020-2575-9

[5] Parvathavarthini S, Karthikeyani Visalakshi N, Shanthi S. Breast cancer detection using crow search optimization based intuitionistic fuzzy clustering with neighborhood attraction. Asian Pac J Cancer Prev: APJCP. 2019;20(1):157. doi:10.31557/APJCP.2019.20.1.157

[6] Ravuri V, Subbarao MV, Sudheer Kumar T, et al. Multi-cancer early detection and classification using machine learning based approaches. In: 2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT). IEEE; 2023. p. 1–7.

[7] Toğaçar M, Özkurt KB, Ergen B, et al. Breastnet: a novel convolutional neural network model through histopathological images for the diagnosis of breast cancer. Physica A. 2020;545:123592. doi:10.1016/j.physa.2019.123592

[8] Al FY, Permanasari AE, Setiawan, N. A. A comparative analysis of tree-based machine learning Algorithms for breast cancer detection. In 2019 12th IEEE International Conference on Information & Communication Technology and System (ICTS); 2019. p. 55–59.

[9] Jahangeer GSB, Rajkumar TD. Early detection of breast cancer using hybrid of series network and VGG-16. Multimed Tools Appl. 2021;80(5):7853–7886. doi:10.1007/s11042-020-09914-2

[10] Juneja K, Rana C. An improved weighted decision tree approach for breast cancer prediction. Int J Inf Technol. 2020;12(3):797–804.

[11] Murtaza G, Shuib L, Abdul Wahab AW. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. Artif Intell Rev. 2020;53(3):1655–1720. doi:10.1007/s10462-019-09716-5

[12] Vijayarajeswari R, Parthasarathy P, Vivekanandan S, et al. Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform. Measurement (Mahwah NJ). 2019;146:800–805.

[13] Hou Y, Li J, Yu H, et al. BIFFOA: a novel binary improved fruit fly algorithm for feature selection. IEEE Access. 2019;7:81177–81194. doi:10.1109/ACCESS.2019.2917502

[14] Li Q, Wen Z, He B. Practical federated gradient boosting decision trees. In Proc AAAI Conf Artif Intell. 2020;34(4):4642–4649. doi:10.1609/aaai.v34i04.5895

[15] Khuriwal N, Mishra N. Breast cancer diagnosis using deep learning algorithm. In 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), IEEE; 2018. p. 98–103.

[16] Gupta P, Garg S. Breast cancer prediction using varying parameters of machine learning models. Procedia Comput Sci. 2020;171:593–601. doi:10.1016/j.procs.2020.04.064

[17] Zhang X, Xu Y, Yu C, et al. Gaussian mutational chaotic fruit fly-built optimization and feature selection. Expert Syst Appl. 2020;141:112976. doi:10.1016/j.eswa.2019.112976

[18] Balkenhol MC, Tellez D, Vreuls W, et al. Deep learning assisted mitotic counting for breast cancer. Lab Invest. 2019;99(11):1596–1606. doi:10.1038/s41374-019-0275-0

[19] Ghiasi MM, Zendehboudi S. Application of decision tree-based ensemble learning in the classification of breast cancer. Comput Biol Med. 2021;128:104089. doi:10.1016/j.compbiomed.2020.104089

[20] Hussain L, Aziz W, Saeed S, et al. Automated breast cancer detection using machine learning techniques by extracting different feature extracting strategies. In 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE); 2018. p. 327–331.

[21] Mehmood M, Ayub E, Ahmad F, et al. Machine learning enabled early detection of breast cancer by structural analysis of mammograms. Comput Mater Continua. 2021;67(1):641–657. doi:10.32604/cmc.2021.013774

[22] Punitha S, Amuthan A, Joseph KS. Benign and malignant breast cancer segmentation using optimized region growing technique. Future Comput Inform J. 2018;3(2):348–358. doi:10.1016/j.fcij.2018.10.005