Date of publication MAR-31, 2024, date of current version FEB-01, 2024.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.23.1.3439

An Optimal Framework Based on the GentleBoost Algorithm and Bayesian Optimization for the Prediction of Breast Cancer Patients' Survivability

AYMAN AISABRY[1,2], MALEK AIGABRI[1], AMIN MOHAMED AHSAN[2], MOGEEB A.A. MOSLEH[2], F. E. HANASH[2,3], HAMZAH ALI ABDULRAHMAN QASEM[2]

[1] Department of Computer Science, Sana'a University, Sana'a, Yemen
[2] Department of Computer Science, International University of Technology Twintech, Sana'a, Yemen
[3] Emirates International University, Sana'a, Yemen

Corresponding author: Ayman Alsabry (aymanalsabri@su.edu.ye).

ABSTRACT Breast cancer is a primary cause of cancer-associated mortality among women globally, and early detection and personalized treatment are critical for improving patient outcomes. In this study, we propose an optimal framework for predicting breast cancer patient survivability using the GentleBoost algorithm and Bayesian optimization. The proposed framework combines the strengths of the GentleBoost algorithm, which is a powerful machine-learning algorithm for classification, and Bayesian optimization, which is a powerful optimization technique for hyperparameter tuning. We evaluated the proposed framework using the publicly available breast cancer dataset provided by The Surveillance, Epidemiology, and End Results (SEER) program and compared its performance with several popular single algorithms, including support vector machine (SVM), artificial neural network (ANN), and k-nearest neighbors (KNN). The experimental results demonstrate that the proposed framework outperforms these methods in terms of accuracy (mean= 95.16%, best = 95.35, worst = 95.1%, and SD = 0.008). The values of precision, recall, and f1-score of the best experiment were 92.3 %, 98.2 %, and 95.2 %, respectively, with hyperparameters of (number of learners = 246, learning rate = 0.0011, and maximum number of splits = 1240). The proposed framework has the potential to improve breast cancer patient survival predictions and personalized treatment plans, leading to the improved patient outcomes and reduced healthcare costs.

KEYWORDS Data Exploration, GentleBoost algorithm, Hyperparameters Tuning, Machine Learning, SEER breast cancer dataset.

I. INTRODUCTION

BREAST cancer (BC) refers to a cancer that develops within the cells of breast tissue. It is the most prevalent cancer in women worldwide and the second most common cancer worldwide, following lung cancer [1]. According to the World Health Organization (WHO), BC is the most common cancer among women worldwide, with an estimated 2.3 million new cases diagnosed and 685 000 deaths in 2020 [2]. In Yemen, the ranking of cancer types is possibly different from that of nearby Gulf nations. Yemen had the highest BC rate (30.5 per 100,000 population), followed by the colorectum (10.7), stomach (7.1), esophagus (6.4), lung (5.8), liver (5.1), leukemia (4.2), Hodgkin lymphoma (4.0), and ovary (3.4) [3]. The incidence of BC varies with age, with the risk increasing with age. The median age at the time of diagnosis was 62 years. In 2022, approximately 287,850 cases of invasive BC were diagnosed in the U.S. BC can also occur in men, although it is much less common, accounting for less than 1% of all BC cases [4-7].

Several risk factors are associated with breast cancer, including age, sex, family history of breast cancer, genetic mutations, exposure to radiation, and lifestyle factors such as alcohol consumption, physical inactivity, and obesity. Regular screening and early detection are important for improving outcomes, as BC is more treatable when detected at an early stage [8-14].

Early detection of BC allows for a wider range of treatment options, including less-invasive surgeries, radiation therapy, and targeted drug therapies [15]. These treatments can improve the chance of survival. Women with early-stage BC have a higher chance of survival. According to the American Cancer Society, the five-year survival rate of women with BC that has not spread beyond the breast is 99%. Early detection may allow for less aggressive treatment

options, which can reduce the physical and emotional tolls of cancer treatment. Additionally, early detection can reduce the healthcare costs associated with cancer treatment [16-18].

Machine learning (ML) has emerged as a promising tool for improving BC detection and treatment. Numerous studies have investigated the utilization of ML algorithms with the SEER BC dataset and how they can enhance the capacity to diagnose and treat breast cancer [19]. SEER program is a population-based cancer registry in the United States that collects data on cancer incidence, mortality, and survival. The SEER BC dataset is a rich source of information, containing data on patient demographics, tumor characteristics, treatment regimens, and outcomes. ML algorithms can be trained on this dataset to identify patterns and make predictions that can inform clinical decision-making [20].

One of the most significant applications of ML with the SEER BC dataset is the development of predictive models for BC risk assessments. These models use patient information such as age, family history, and genetic markers to estimate the likelihood of developing breast cancer. Several studies have shown that ML algorithms can outperform traditional risk assessment models in terms of accuracy and predictive power [21, 22]. Other studies have explored the use of ML algorithms with SEER BC data to predict patient survival.

In 2019, Lu et al. [23] explored the use of a genetic optimizer to improve the performance of a gradient-boosting machine for BC prognosis. The results showed 28% accuracy improvement over other ML models. Huber et.al. [24] proved that gradient-boosted supervised ML achieves a better performance than linear models.

In 2020, Wang et al. [25] developed an improved random-forest (RF)-based rule extraction method. The method was assessed using three datasets: WDBC, WOBC, and SEER breast cancer. According to the experimental results, the proposed method surpasses various widely used single algorithms, ensemble learning methods, and rule extraction methods in terms of accuracy and interpretability.

In 2022, Kajala and Jaiswal [26] proved that balancing dataset classes using oversampling techniques improved the performance of SVM models, achieving 100% precision and 99.35% AUC. Similarly, Haque et al. [27] applied RF models to the SEER BC dataset, and the results showed an accuracy of 94.64 %.

The structure of this paper is as follows: Section II outlines the methodology, Section III delves into the experimental findings, and Section IV concludes the paper.

II. METHODOLOGY

This section discusses the proposed framework, research aims, and objectives that will be addressed. The primary objective of this study is to develop a framework that can accurately predict the survival of BC patients based on clinical and pathological features. The secondary objectives are as follows:

1. To evaluate the performance of the proposed framework in terms of accuracy, recall, specificity, and f1-score.
2. To compare the performance of the proposed framework with other existing models.

To accomplish these objectives, a framework is presented in Figure 1. The framework comprises of eight notable steps, as outlined below:

1. Acquiring the BC patient dataset from the November 2017 update of the SEER Program of the NCI (https://ieee-dataport.org/open-access/seer-breast-cancer-data#).
2. Exploring the dataset.
3. Cleaning the dataset (instances that contain missing values and duplicates will be eliminated).
4. Using the SMOTE technique to balance the target class.
5. Dividing the dataset into two groups (training and testing data)
6. Applying GentleBoost model with Bayesian optimization.
7. Evaluating the proposed model using the matrices of accuracy, precision, recall, and F1-Score.
8. Comparing the performance of the proposed model to the state-of-the-art models' performance.

A. DATA EXPLORATION AND PREPROCESSING:

The SEER BC dataset is a publicly available database containing information on BC patients diagnosed between 2006-2010. Patients with unknown tumor size, examined regional lymph nodes, regional positive lymph nodes, and those with less than one month of survival were excluded. Consequently, 4024 patients were included in the dataset. Table 1 presents a description of SEER BC features.

Table 1. SEER BC Dataset Features.

No.	Feature	Type	No. of instance
F1	AGE	Intervals	4024
F2	RACE	Categorical	
F3	MARITAL STATUS	Categorical	
F4	T STAGE	Categorical	
F5	N STAGE	Categorical	
F6	6TH STAGE	Categorical	
F7	GRADE	Categorical	
F8	A STAGE	Categorical	
F9	TUMOR SIZE	Intervals	
F10	ESTROGEN STATUS	Categorical	
F11	PROGESTERONE STATUS	Categorical	
F12	REGIONAL NODES EXAMINED	Intervals	
F13	REGIONAL NODES POSITIVE	Intervals	
F14	SURVIVAL MONTHS	Intervals	
F15	STATUS	Categorical	

As shown in Figure 2, there are some data points that are significantly different from the rest (outliers), an imbalanced distribution of the target class within the data, and some features with overlapping data. These problems are addressed as follows:

- **Handling Outliers:** Outliers are removed using three standard deviations (3 SD) above and below the mean. The use of the outlier removal method of 3-SD above and below the mean can be an effective way to deal with extreme data points that may skew the analysis or modeling results. This method is based on the assumption that the data follows a normal distribution, as shown in Figure 3.
- **Handling Imbalanced Distribution:** Utilizing the SMOTE technique is beneficial for addressing the imbalance problem in machine learning. By generating synthetic data points for the minority class, this technique can balance the data distribution and enhance the performance of the ML models. The

balanced data obtained after implementing SMOTE are shown in Figure 4.

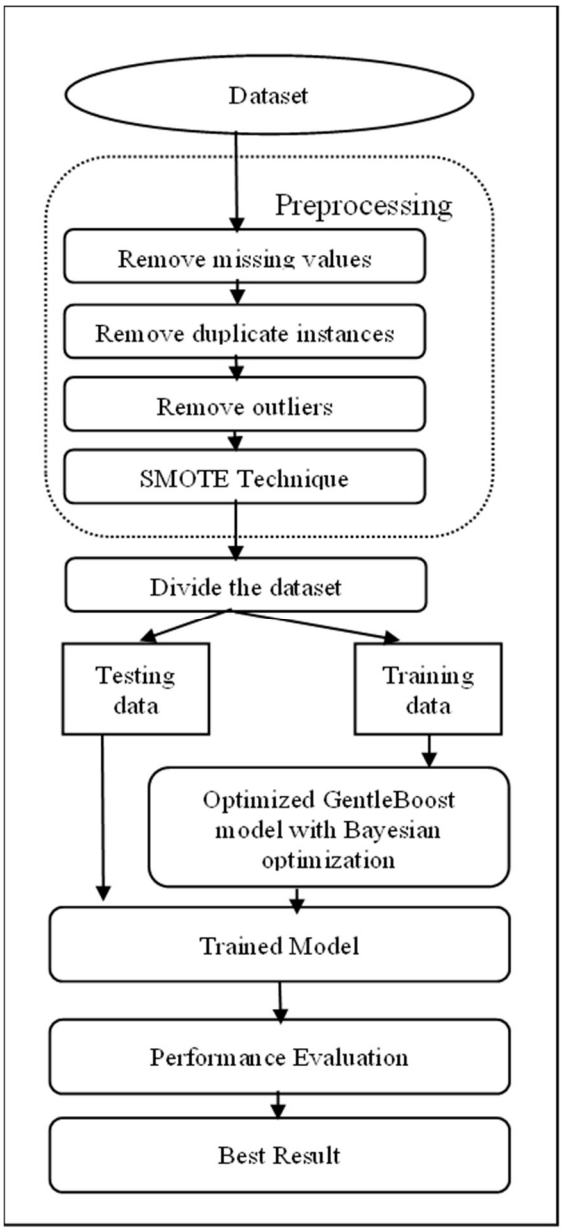

Figure 1. The proposed framework block diagram

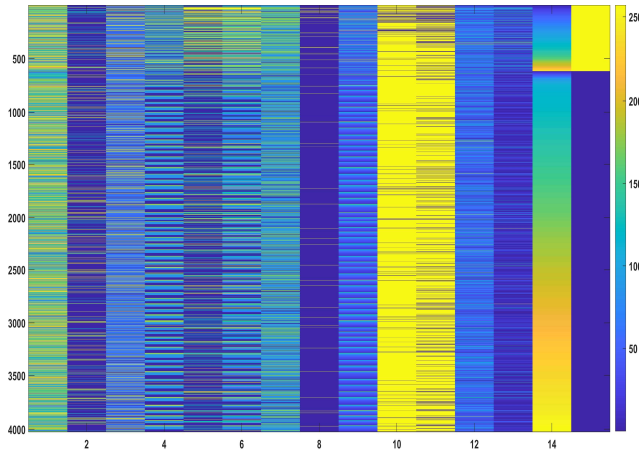

Figure 2. Colormap Visualization Technique

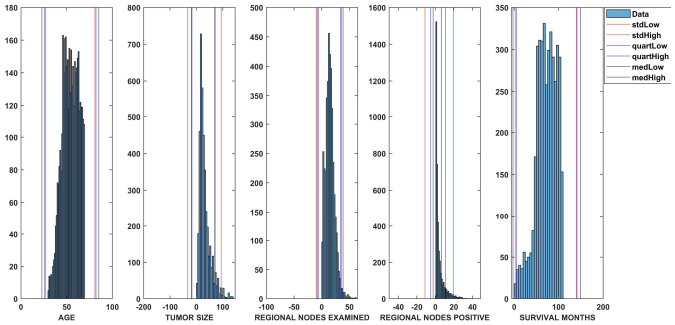

Figure 3. Outliers Detection

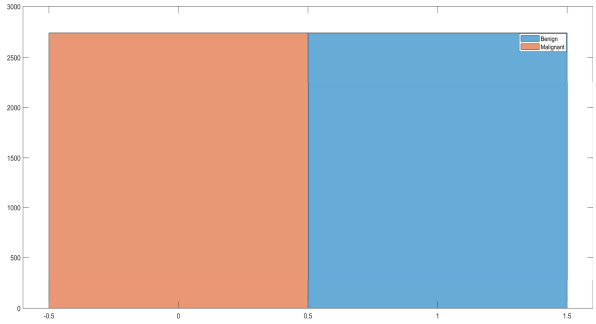

Figure 4. Target Class Distribution

B. BAYESIAN HYPERPARAMETER OPTIMIZATION:

Bayesian hyperparameter optimization is a powerful approach for tuning the hyperparameters of ML models. It is a probabilistic method that uses Bayesian optimization to search for the optimal hyperparameters of a model. The goal of Bayesian hyperparameter optimization is to find the set of hyperparameters that maximizes the expected improvement (EI) of the objective function. EI is defined as the difference between the expected value of the objective function at the current best set of hyperparameters and the expected value of the objective function at the candidate set of hyperparameters. The candidate set of hyperparameters was selected based on the probabilistic model of the objective function.

The equations for Bayesian hyperparameter optimization can be written as follows:

1. Define the prior distribution over the objective function:

$$f \sim GP(m, k), \quad (1)$$

where m is the mean function and k is the covariance function.

2. Evaluate the objective function at the current best set of hyperparameters:

$$y\_best = max(y\_1, y\_2, ..., y\_n). \quad (2)$$

3. Compute the expected improvement:

$$EI(x) = E[max(y - y\_best, 0)], \quad (3)$$

where y is the value of the objective function at x.

4. Update the probabilistic model based on the observed data:

$$f \mid D \sim GP(m\_post, k\_post), \quad (4)$$

where D is the set of observed data and m_post and k_post are the posterior mean and covariance functions, respectively.

5. Select the next set of hyperparameters to evaluate:

$$x\_next = argmax(EI(x)). \quad (5)$$

B. GENTLEBOOST:

The GentleBoost algorithm is an ML algorithm used to transform weak classifiers into strong classifiers. It is a variant of the AdaBoost algorithm, designed to be more robust to noisy data and outliers. The GentleBoost algorithm works by iteratively adding weak classifiers to the ensemble, with each new classifier trained on a weighted version of the training data. The weights are adjusted after each iteration to give more importance to the misclassified examples, which helps improve the performance of the ensemble. The algorithm is called "gentle" because it places less emphasis on misclassified examples than AdaBoost, which can be more prone to overfitting. The equations for the GentleBoost algorithm are as follows:

1. Initializing the weights for the training examples:

$$w_i = 1/N, \text{ for } i = 1, 2, ..., N, \quad (6)$$

where N is the number of training examples.

2. For t = 1, 2, ..., T, do:

$$a_t = \text{argmin\_a } sum_i\ W_i \exp(-y_i\ a_t\ h_t((x_i))), \quad (7)$$

where h_t is the weak classifier being trained, y_i is the label of the i-th training example, and a is a scalar parameter that controls the contribution of the weak classifier to the ensemble.

Updating the weights:

$$W_i = W_i \exp(-y_i\ a_t\ h_t((x_i))) \quad (8)$$

Normalizing the weights:

$$W_i = W_i / sum_j\ W_j \quad (9)$$

3. Returning the final classifier:

$$H(x) = sign\ (sum_t\ a_t\ h_t(x))) \quad (10)$$

C. EVALUATION METRICS:

The proposed model is evaluated using accuracy, recall, F-measure, and precision, which are calculated through the following:

$$Accuracy = \frac{TP+TN}{TP+TN+FN+F} \quad (11)$$

$$Precision = \frac{TP}{TP+FP} \quad (12)$$

$$Rcall = \frac{TP}{TP+} \quad (13)$$

$$F1-Score = \frac{2\times Precision\times Recall}{Precision+Rec}, \quad (14)$$

where:

TP = True Positive

FP = False Positive

TN = True Negative

FN = False Negative.

III. RESULTS AND DISCUSSION

Two distinct approaches are employed to assess the performance of the proposed framework. In the first approach, the original dataset without preprocessing was used to train the proposed optimized GentleBoost model. The resulting performance of this model was then compared with that of three other cutting-edge models.

In the second approach, outliers were eliminated and the target class distribution was balanced using the SMOTE technique. Subsequently, the ML models were trained.

The outcomes of all the approaches under different circumstances are given below.

A. THE FIRST APPROACH (DOING MODELS TRAINING WITH ORIGINALDATASET):

In this approach, every model is trained using the original dataset. Using 5-fold cross-validation, the study validated the performance of each model. The performance outcomes for all models are discussed below:

- ***SVM Classifier with Bayesian Optimization:***

Table 2 shows that the accuracy of the optimized SVM classifier varied from 90.2% to 91%. The mean accuracy was 90.6%, with the highest score being 91% and the minimum score falling to 90.2%. A standard deviation of 0.23 was noted.

Table 2. First approach SVM classifier performance analysis

Experiments	Accuracy (%)	Mean (%)	SD
#1	91	90.6	0.23
#2	90.7		
#3	90.5		
#4	90.2		
#5	90.7		
#6	90.5		
#7	90.7		
#8	90.7		
#9	90.3		
#10	90.8		

The confusion matrix for the top-performing SVM is displayed in Figure 5, and Table 3 presents the ideal hyperparameters for the SVM classifier. This classifier achieved a recall of 89.3%, precision of 47.2%, and F1-score of 61.8%.

Table 3. Evaluation metrics and SVM optimal hyperparameter of the first approach

Hyperparameter	Value	Evaluation metrics		
		Recall	Precision	F1-score
Multiclass method	One-VS-One	89.3%	47.2%	61.8%
BOX constraint level	965.3563			
Kernel function	Gaussian			
Kernel scale	115.5548			
Standardize data	False			

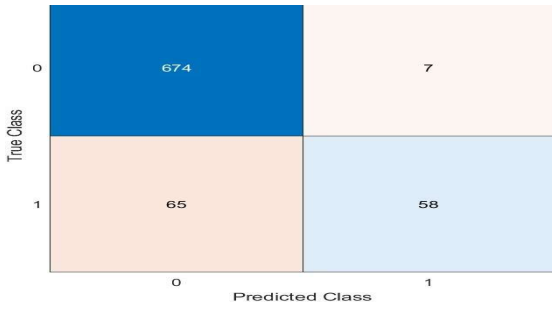

Figure 5. Confusion matrix of SVM classifier

- ***ANN classifier with Bayesian Optimization:***

As shown in Table 4, the accuracy of the optimized ANN classifier ranges between 90.2% and 91%, with an average

accuracy of 90.8%. The maximum accuracy achieved is 91%, whereas the lowest is 90.2%. A standard deviation of 0.3 is observed.

Table 4. First approach ANN classifier performance analysis

Experiments	Accuracy (%)	Mean (%)	SD
#1	90.9	90.8	0.3
#2	90.2		
#3	91		
#4	91		
#5	91		
#6	90.4		
#7	91		
#8	90.5		
#9	90.5		
#10	91		

Figure 6 displays the confusion matrix of the ANN that performs the best, and the optimal hyperparameters for the ANN classifier are presented in Table 5. The ANN classifier attained a recall of 90.5%, precision of 46.3%, and F1-score of 61.3%.

Table 5. Evaluation metrics and ANN Classifier optimal hyperparameter of the first approach

Hyperparameter	Value	Evaluation metrics		
		Recall	Precision	F1-score
Number of fully connected layers	1	90.5%	46.3%	61.3%
Activation	None			
Standardize data	Yes			
Regularization strength (Lambda)	1.711e-08			
First layer size	7			

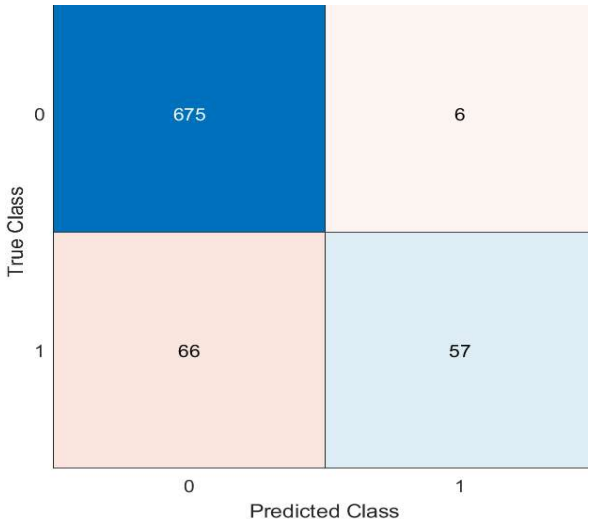

Figure 6. Confusion matrix of ANN Classifier

- ***KNN classifier with Bayesian Optimization:***

As shown in Table 6, the accuracy of the optimized KNN classifier fluctuates between 89.3% and 91%. The average accuracy is 90.48%, with the peak score reaching 91% and the lowest score dipping up to 89.3%. A standard deviation of 0.44 is observed.

The best-performing KNN confusion matrix is shown in Figure 7, and the optimal hyperparameters for the KNN classifier are listed in Table 7. The classifier achieved a recall of 82.3%, a precision of 52.8%, and an F1-score of 64.3%.

Table 6. First approach KNN performance analysis

Experiments	Accuracy (%)	Mean (%)	SD
#1	90.5	90.48	0.44
#2	90.5		
#3	90.7		
#4	89.3		
#5	90.5		
#6	90.3		
#7	90.7		
#8	90.8		
#9	91		
#10	90.5		

Table 7. Evaluation metrics and KNN Classifier optimal hyperparameter of the first approach

Hyperparameter	Value	Evaluation metrics		
		Recall	Precision	F1-score
Number of neighbors	15	82.3%	52.8%	64.3%
Distance metrics	Chebyshev			
Distance weight	Equal			
Standardize data	False			

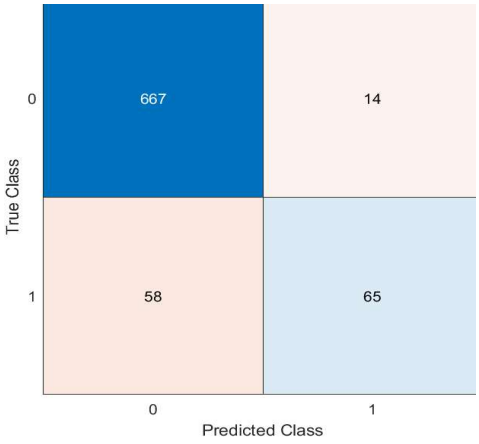

Figure 7. Confusion matrix of KNN Classifier

- ***GentleBoost with Bayesian Optimization:***

As shown in Table 8, the accuracy of the optimized GentleBoost classifier varies from 90.7% to 91.7%. The mean accuracy is 91.1%, with the highest score reaching 91.7% and the lowest score dropping to 90.7%. A standard deviation of 0.35 is noted.

Table 8. First approach GentleBoost performance analysis

Experiments	Accuracy (%)	Mean (%)	SD
#1	91.3	91.1	0.35
#2	91.2		
#3	90.7		
#4	91.7		
#5	91.4		
#6	90.7		
#7	90.7		
#8	90.7		
#9	91.4		
#10	91.2		

The optimal hyperparameters for the GentleBoost classifier, resulted in an accuracy of 91.7%, recall of 88.8%, precision of 52%, and F1-score of 65.6%, are presented in Table 9. The best-performing confusion matrix for GentleBoost is shown in Figure 8.

Table 9. Evaluation metrics and GentleBoost optimal hyperparameter of the first approach

Hyperparameter	Value	Evaluation metrics		
		Recall	Precision	F1-score
Number of learners	10	88.8%	52%	65.6%
Learning rate	0.0274			
Maximum number of splits	4			

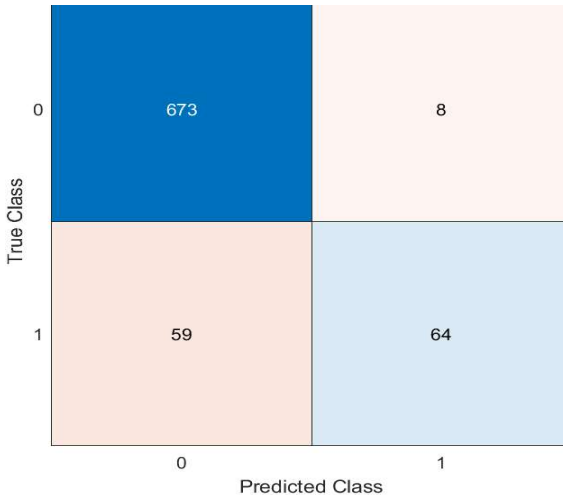

Figure 8. Confusion matrix of GentleBoost Classifier

B. SECOND APPROACH (DOING OUTLIERS DETECTION AND DOING SMOTE TECHNIQUE):

SMOTE technology and outlier removal are two important techniques used in data preprocessing for ML models. These techniques help improve the model performance by balancing imbalanced datasets and eliminating extreme values that may skew our results.

A discussion of the results showed by the models is given below.

- ***SVM Classifier with Bayesian Optimization:***

As indicated in Table 10, the accuracy of the optimized SVM classifier ranges from 87.7% to 95.6%. The accuracy rate of the model is 90.8%, with the highest accuracy at 95.6% and the worst accuracy at 87.7%. The standard deviation is 2.3.

Table 10. Second approach SVM performance analysis

Experiments	Accuracy (%)	Mean (%)	SD
#1	89.6	90.8	2.3
#2	90.1		
#3	88.3		
#4	89.5		
#5	92.3		
#6	92.5		
#7	92.6		
#8	89.5		
#9	95.6		
#10	87.7		

Figure 9 shows the confusion matrix of the best performance of SVM, and Table 11 lists the optimal hyperparameters of the SVM classifier, which has the best accuracy of 95.6%, recall of 94%, precision of 97.4%, and F1-score of 95.7%.

Table 11. Evaluation metrics and SVM optimal hyperparameter of the second approach

Hyperparameter	Value	Evaluation metrics		
		Recall	Precision	F1-score
Multiclass method	One-VS-All	94%	97.4%	95.7
BOX constraint level	2.4105			
Kernel function	Gaussian			
Kernel scale	5.2326			
Standardize data	False			

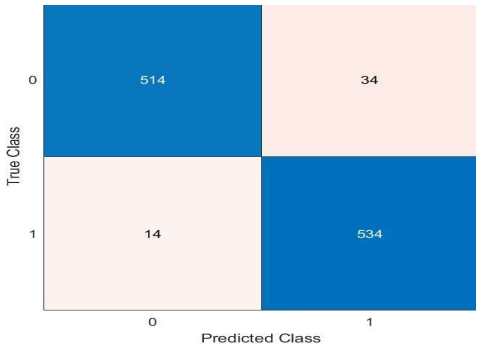

Figure 9. Confusion matrix of SVM classifier

- ***ANN classifier with Bayesian Optimization:***

Table 12 shows that the accuracy of the optimized ANN classifier varies from 91.6% to 94%. The accuracy rate of the model is 92.7%, with the highest accuracy at 94% and the worst accuracy at 91.6%. The standard deviation is 0.79.

Table 12. Second approach ANN performance analysis

Experiments	Accuracy (%)	Mean (%)	SD
#1	92.5	92.7	0.79
#2	91.9		
#3	92		
#4	92.3		
#5	92.5		
#6	93.8		
#7	93.2		
#8	91.6		
#9	93.5		
#10	94		

The optimal hyperparameters for the ANN classifier, which has the best accuracy of 94%, recall of 92.9%, precision of 95.1%, and F1-score of 94%, are listed in Table 13, and Figure 10 shows the confusion matrix of the best performance of the ANN.

Table 13. Evaluation metrics and ANN optimal hyperparameter of the second approach

Hyperparameter	Value	Evaluation metrics		
		Recall	Precision	F1-score
Number of fully connected layers	3	92.9%	95.1%	94%
Activation	Relu			
Standardize data	Yes			
Regularization strength (Lambda)	0.00018859			
First layer size	290			
Second layer size	109			
Third layer size	34			

- ***KNN classifier with Bayesian Optimization:***

As demonstrated in Table 14, the accuracy of the optimized KNN classifier fluctuates between 93.3% and 94.8%. The average accuracy is 94.1%, with the peak score reaching 94.8% and the lowest score dipping to 93.3%. A standard deviation of 0.56 is observed.

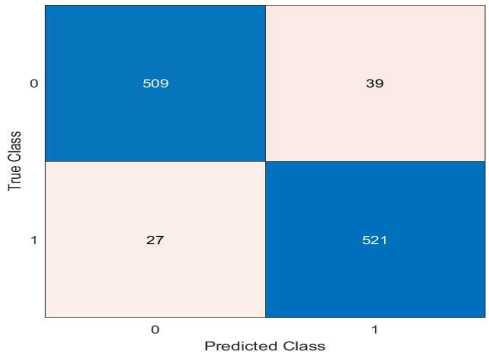

Figure 10. Confusion matrix of ANN Classifier

Table 14. Second approach KNN classifier performance analysis

Experiments	Accuracy (%)	Mean (%)	SD
#1	94.8	94.1	0.56
#2	94.3		
#3	93.3		
#4	94.3		
#5	93.3		
#6	94.3		
#7	94.4		
#8	93.3		
#9	93.9		
#10	94.8		

The ideal hyperparameters for the KNN classifier yielding an accuracy of 94.8%, recall of 93.4%, precision of 96.2%, and F1-score of 94.8% are listed in Table 15. Additionally, Figure 11 illustrates the confusion matrix for the KNN top-performing results.

Table 15. Evaluation metrics and KNN Classifier optimal hyperparameter of the second approach

Hyperparameter	Value	Evaluation metrics		
		Recall	Precision	F1-score
Number of neighbors	2	93.4%	96.2%	94.8%
Distance metrics	City block			
Distance weight	Inverse			
Standardize data	True			

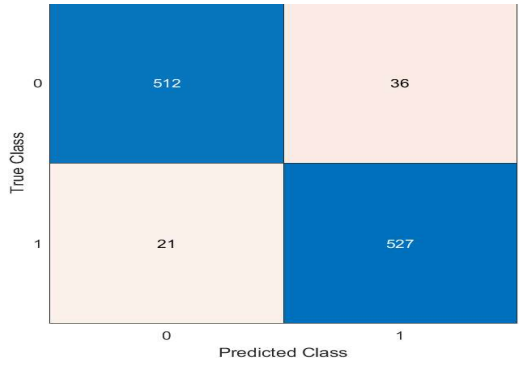

Figure 11. Confusion matrix of KNN Classifier

- ***GentleBoost with Bayesian Optimization:***

As evidenced in Table 16, the performance of the optimized GentleBoost classifier exhibits a range of 95.1% to 95.3% in terms of accuracy. The average accuracy marks at 95.2%, with the peak at 95.3% and the nadir at 95.1%. A standard deviation of 0.008 is documented.

Table 16. Second approach GentleBoost classifier performance analysis

Experiments	Accuracy (%)	Mean (%)	SD
#1	95.1	95.16	0.008
#2	95.1		
#3	95.3		
#4	95.1		
#5	95.2		
#6	95.3		
#7	95.1		
#8	95.1		
#9	95.1		
#10	95.2		

The optimal hyperparameters for the GentleBoost classifier, resulted in an accuracy of 95.3%, recall of 98.2%, precision of 92.3%, and F1-score of 95.2%, are presented in Table 17. The best-performing confusion matrix for GentleBoost is shown in Figure 12.

Table 17. Evaluation metrics and GentleBoost Classifier optimal hyperparameter of the second approach

Hyperparameter	Value	Evaluation metrics		
		Recall	Precision	F1-score
Number of learners	246	98.2%	92.3%	95.2%
Learning rate	0.0011			
Maximum number of splits	1240			

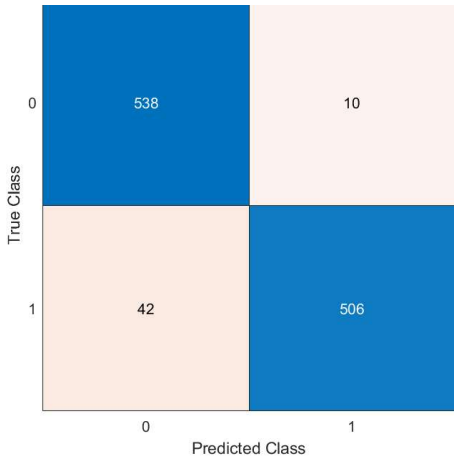

Figure 12. Confusion matrix of GentleBoost Classifier

C. PERFORMANCE COMPARISION:

As demonstrated in Tables 18 and 19, the performance evaluation of the proposed model and three other state-of-the-art models is conducted based on their test accuracy rates. The optimized GentleBoost model outperforms the others with an average test accuracy of approximately 95.16% (mean), reaching 95.3% at its best, 95.1% at its lowest, a recall of 98.2%, and a precision of 92.35. f1-score is 95.2%, and the standard deviation is 0.008.

Table 18. Performance analysis of the second approach in terms of test accuracy

Models	Mean (%)	Best (%)	Worst (%)	SD
SVM classifier with Bayesian Optimization	90.8	95.6	87.7	2.3
ANN classifier with Bayesian Optimization	92.7	94	91.6	0.79
KNN classifier with Bayesian Optimization	94.1	94.8	93.3	0.56
Proposed	**95.16**	**95.3**	**95.1**	**0.008**

Table 19. Performance comparison of the second approach

ML Models	Precision (%)	Recall (%)	F1-Score (%)
SVM classifier with Bayesian Optimization	97.4	94	95.7
ANN classifier with Bayesian Optimization	95.1	92.9	94
KNN classifier with Bayesian Optimization	96.2	93.4	94.8
Proposed	92.3	98.2	95.2

Furthermore, Table 20 shows the superiority of the proposed model compared with previous studies, with an accuracy rate of 95.2%.

Table 20. Evaluation of performance in comparison to similar works of literature

Author	Year	Method	Accuracy (%)
[28]	2018	Twelve Different SVMs, based on the proposed Weighted Area Under the Receiver Operating Characteristic Curve Ensemble (WAUCE)	76.42
[23]	2019	Gradient Boosting with Genetic Algorithm	75.03
[25]	2020	Improved Random Forest (RF)-based rule extraction (IRFRE)	80.45
[29]	2020	J48	93.02
[27]	2022	RF	94.64%
Proposed	2023	GentleBoost with Bayesian Optimization	Test accuracy rate = 95.2

IV. CONCLUSION

According to the above mentioned, we can conclude that the proposed framework based on the GentleBoost algorithm and Bayesian optimization has the potential to improve the accuracy and efficiency of predicting the survivability of BC patients. The technique of removing outliers is used to improve the performance of the ML models. Additionally, the SMOTE technique is applied to balance the target class.

Table 17 exhibited the superiority of the proposed method in terms of accuracy rate compared with a broad spectrum of related studies. With the preprocessed dataset, the proposed optimized GentleBoost algorithm obtained an accuracy rate of 95.2 %, whereas the prediction accuracy rate of the optimized GentleBoost algorithm using the original dataset was 91.1%.

References

[1] A. Iqbal and M. Sharif, "BTS-ST: Swin transformer network for segmentation and classification of multimodality breast cancer images," *Knowledge-Based Systems,* vol. 267, p. 110393, 2023. https://doi.org/10.1016/j.knosys.2023.110393.

[2] World Health Organization. Breast cancer: Scope of the problem. [Online]. Available at: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[3] H. O. Al-Shamsi, I. H. Abu-Gheida, F. Iqbal, and A. Al-Awadhi, *Cancer in the Arab world*: Springer Nature, 2022. https://doi.org/10.1007/978-981-16-7945-2.

[4] "American Cancer Society. Breast Cancer Facts & Figures 2021-2022," [Online]. Available at: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/2022-2024-breast-cancer-fact-figures-acs.pdf.

[5] S. Fox, V. Speirs, and A. M. Shaaban, "Male breast cancer: an update," *Virchows Archiv,* vol. 480, pp. 85-93, 2022. https://doi.org/10.1007/s00428-021-03190-7.

[6] E. Garreffa and D. Arora, "Breast cancer in the elderly, in men and during pregnancy," *Surgery (Oxford),* vol. 40, pp. 139-146, 2022. https://doi.org/10.1016/j.mpsur.2021.11.018.

[7] R. Singh, L. Cao, A. L. Sarode, M. Kharouta, R. Shenk, and M. E. Miller, "Trends in surgery and survival for T1-T2 male breast cancer: a study from the National Cancer Database," *The American Journal of Surgery,* vol. 225, pp. 75-83, 2023. https://doi.org/10.1016/j.amjsurg.2022.09.043.

[8] A. N. Hurson, T. U. Ahearn, R. Keeman, M. Abubakar, A. Y. Jung, P. M. Kapoor, *et al.*, "Systematic literature review of risk factor associations with breast cancer subtypes in women of African, Asian, Hispanic, and European descents," *Cancer Research,* vol. 82, pp. 3670-3670, 2022. https://doi.org/10.1158/1538-7445.AM2022-3670.

[9] B. Smolarz, A. Z. Nowak, and H. Romanowicz, "Breast Cancer—Epidemiology, Classification, Pathogenesis and Treatment (Review of Literature)," *Cancers,* vol. 14, p. 2569, 2022. https://doi.org/10.3390/cancers14102569.

[10] M. O. Abbas and M. Baig, "Knowledge and Practice Concerning Breast Cancer Risk Factors and Screening among Females in UAE," *Asian Pacific Journal of Cancer Prevention,* vol. 24, pp. 479-487, 2023. https://doi.org/10.31557/APJCP.2023.24.2.479.

[11] A. Alsabry, M. Algabri, and A. M. Ahsan, "Breast Cancer-Risk Factors and Prediction Using Machine-Learning Algorithms and Data Source: A Review of Literature," *JAST,* vol. 1, 2023. https://doi.org/10.59628/jast.v1i2.361.

[12] A. Alsabry and M. Algabri, "Iterative tuning of tree-ensemble-based models' parameters using Bayesian optimization for breast cancer prediction," *Informatics and Automatization,* vol. 23, pp. 129-168, 2024. https://doi.org/10.15622/ia.23.1.5.

[13] A. Alsabry, M. Algabri, A. M. Ahsan, M. A. Mosleh, A. A. Ahmed, and H. A. Qasem, "Enhancing Prediction Models' Performance for Breast Cancer using SMOTE Technique," in *2023 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, 2023, pp. 1-8. https://doi.org/10.1109/eSmarTA59349.2023.10293726.

[14] A. Alsabry, M. Algabri, A. M. Ahsan, M. A. Mosleh, A. A. Ahmed, and H. A. Qasem, "Breast Cancer Prediction Framework Based on Iterative Optimization with Bayesian Hyperparameter Tuning," in *2023 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, 2023, pp. 01-08. https://doi.org/10.1109/eSmarTA59349.2023.10293277.

[15] C. E. Holmes and H. B. Muss, "Diagnosis and treatment of breast cancer in the elderly," *CA: a cancer journal for clinicians,* vol. 53, pp. 227-244, 2003. https://doi.org/10.3322/canjclin.53.4.227.

[16] F. Cardoso, S. Kyriakides, S. Ohno, F. Penault-Llorca, P. Poortmans, I. Rubio, *et al.*, "Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up," *Annals of oncology,* vol. 30, pp. 1194-1220, 2019. https://doi.org/10.1093/annonc/mdz173.

[17] "National Cancer Institute: Breast Cancer Treatment," [Online]. Available at: https://www.cancer.gov/types/breast/hp/breast-treatment-pdq.

[18] "American Cancer Society. (2022). Breast Cancer Early Detection," [Online]. Available at: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/american-cancer-society-recommendations-for-the-early-detection-of-breast-cancer.html.

[19] J. A. Wernberg, J. Yap, C. Murekeyisoni, T. Mashtare, G. E. Wilding, and S. A. Kulkarni, "Multiple primary tumors in men with breast cancer diagnoses—a SEER database review," *Journal of surgical oncology,* vol. 99, pp. 16-19, 2009. https://doi.org/10.1002/jso.21153.

[20] "E. The Surveillance, and End Results (SEER) program. SEER breast cancer data," [Online]. Available at: https://ieee-dataport.org/open-access/seer-breast-cancer-data.

[21] P. D. Pharoah, A. C. Antoniou, D. F. Easton, and B. A. Ponder, "Polygenes, risk prediction, and targeted prevention of breast cancer," *New England Journal of Medicine,* vol. 358, pp. 2796-2803, 2008. https://doi.org/10.1056/NEJMsa0708739.

[22] S. Bacha and O. Taouali, "A novel machine learning approach for breast

cancer diagnosis," *Measurement,* vol. 187, p. 110233, 2022. https://doi.org/10.1016/j.measurement.2021.110233.
[23] H. Lu, H. Wang, and S. W. Yoon, "A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis," *Expert Systems with Applications,* vol. 116, pp. 340-350, 2019. https://doi.org/10.1016/j.eswa.2018.08.040.
[24] M. Huber, C. Kurz, and R. Leidl, "Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning," *BMC medical informatics and decision making,* vol. 19, pp. 1-13, 2019. https://doi.org/10.1186/s12911-018-0731-6.
[25] S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin, "An improved random forest-based rule extraction method for breast cancer diagnosis," *Applied Soft Computing,* vol. 86, p. 105941, 2020. https://doi.org/10.1016/j.asoc.2019.105941.
[26] A. Kajala and S. Jaiswal, "Breast Cancer Survival Prediction from Imbalanced Dataset with Machine Learning Algorithms," *Mathematical Statistician and Engineering Applications,* vol. 71, pp. 167-172, 2022. [Online]. available at: https://www.philstat.org/index.php/MSEA/article/view/125.
[27] M. N. Haque, T. Tazin, M. M. Khan, S. Faisal, S. M. Ibraheem, H. Algethami, *et al.*, "Predicting characteristics associated with breast cancer survival using multiple machine learning approaches," *Computational and Mathematical Methods in Medicine,* vol. 2022, 2022. https://doi.org/10.1155/2022/1249692.
[28] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *European Journal of Operational Research,* vol. 267, pp. 687-699, 2018. https://doi.org/10.1016/j.ejor.2017.12.001.
[29] G. Y. Özkan and S. Y. Gündüz, "Comparision of Classification Algorithims for Survival of Breast Cancer Patients," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2020, pp. 1-4. https://doi.org/10.1109/ASYU50717.2020.9259846.

***AYMAN ALSABRY** received his BSc degree in Computer Science from the faculty of science at Taiz University, Yemen. He completed his master's degree in information technology from Taiz University in 2020. He is currently a PhD student at Sana'a University. He is currently working as Academic Affairs Director at the International University of Technology Twintech, Sana'a, Yemen. His research interests include computer science, artificial intelligence, and machine learning.*

***MALEK ALGABRI** received his BSc and MSc degrees in Computer Science and Technology from Wuhan University of Technology, Wuhan, China. He completed his PhD at the same university in Wuhan, China, in December 2013. He is currently working as an Associate Professor of Computer Science and Technology in the Faculty of Computing and Information Technology at Sana'a University, Sana'a, Yemen. His research interests include computer science, MANETs, MPLS, artificial intelligence, and machine learning.*

***AMIN MOHAMED AHSAN** received his B.S. degree in Computer Science and Information Systems from the University of Technology, Baghdad, Iraq, in 1999 and his M.S. and Ph.D. degrees in Computer Science and Information Systems from Universiti Teknologi Malaysia, Malaysia, in 2010 and 2015, respectively. He is now an assistant professor and Head of Department of Computer Science in the faculty of Computer Science and Information Technology at the International University of Technology Twintech, Sana'a, Yemen. His research interests include computer vision and image processing, pattern recognition, artificial intelligence, machine learning, and deep learning.*

***ASSOC. PROF. DR. MOGEEB A.A. MOSLEH** is a Vice President of IUTT fir academic affairs and senior lecturer at Software Engineering Dep. Faculty of Engineering and Information Technology- Taiz University. He obtained his Ph.D. and MSc in Artificial Intelligent area at FCSIT – University of Malaya.*

***FOAD HANASH** is currently the CEO of International University of Technology Twintech, Secretary General of the Emirates International University, and senior lecturer of Experimental Physics, Faculty of Science, Saada University. He received his B.Sc. degree in physics from Amran University in 2008. received an M.Sc. degree in experimental physics from the Faculty of Science at Mansoura University, Egypt, in 2012 and a PhD in experimental physics from the same university. His research interests are experimental physics, optics, interference, lasers, nanotechnology, and machine learning.*

***HAMZAH ALI ABDULRAHMAN QASEM** received his Bachelor of Technology in Information Technology from Uttar Pradesh Technical University, India in 2006 and his Master of Technology in Information Science and Engineering from Visvesvaraya Technological University, India in 2010. He has obtained his Ph.D. from Department of Computer Engineering, Aligarh Muslim University, India. He is now an assistant professor and the faculty dean of Computer Science and Information Technology at the International University of Technology Twintech, Sana'a, Yemen. His current research interests include Data Science, artificial intelligence, machine learning, deep learning, and localization in wireless sensor networks*