Journal of Scientific & Industrial Research
Vol. 85, February 2026, pp. 128-139
DOI: 10.56042/jsir.v85i2.14711

A Novel Hybrid Stacked Ensemble Model for Breast Cancer Classification

Suraj Arya & Anju*

[1]Department of Computer Sciences and Information Technology, Central University of Haryana, Mahendergarh 123 031 (Haryana)

Received 29 October 2024; revised 24 December 2025; accepted 11 March 2026

The healthcare industry also leverages technology to address various problems. The medical industry continues to adopt artificial intelligence, deep learning, Machine Learning (ML), and big data solutions to automate tasks, improve workflows, and enhance decision-making. Currently, various AI−based solutions are available in the healthcare industry, for example, the analysis of medical images and the identification of patterns in patient data. Thus, these emerging techniques provides many solutions, such as predictive healthcare, and automated drug discovery. The present study proposes the earlier cancer−detection and prediction model to resolve this real-life problem. This study proposed an improved ML model using Synthetic Minority Oversampling Technique (SMOTE) to detect and predict breast cancer at an earlier stage. The four machine learning algorithms achieved the highest test-set accuracy. These algorithms include a novel Hybrid Stacked Ensemble Model (HSEM) and Random Forest (RF), achieving accuracies of 99.12% and 98.59%, respectively; logistic regression, achieving 98.59%; and the support vector classifier, achieving 98.25%. The Area under curve (AUC) for the Breast Cancer (BC) dataset with the HSEM and RF classifier is 99.90%, indicating the model's accuracy. Secondly, cancer treatment exists in expensive types of treatments, and cost has an important role. Therefore, a low-cost solution is required and would be beneficial for the healthcare industry. Thus, this paper developed a novel, low-cost model for cancer prediction for the healthcare industry, enabling people to estimate their cancer risk earlier.

Keywords: Healthcare industry automation, Hybrid stacked ensemble model, Machine learning, ROC−AUC, SMOTE sampling

Introduction

One of the central malignancies threatening women's health is Breast Cancer (BC), which is caused by irregular cells in the breast growing out of control. The US has an alarming expected 313,510 new breast cancer diagnoses in 2024 (310,720 women; 2,790 men), making it a public health concern. Among all these instances in women, the fatality rate is almost 12%, making BC the second most common cause of cancer-related fatalities in women. For early efficient treatment, women over 40 are encouraged to follow screening guidelines, which may include mammography, ultrasound, and Magnetic Resonance Imaging (MRI). The disease is caused by abnormal growth of breast cells within the breast tissue and has a multifactorial origin that includes both genetic and hormonal components.[1] Breast cancer is the collective kind of disease among Indian women, making up 14% of all cancers. In 2018, in India, new cases (1,62,468) and reported deaths (87,090) were recorded from breast cancer. However, in 2020, 685,000 deaths occurred due to BC. The death rate due to BC is increasing day by day.[2] Cancer is a common disease that damages Deoxyribonucleic Acid (DNA) expression by uncontrolled growth of cells. According to the World Health Organization (WHO), in 2022, 670000 women died due to BC, and 2.3 million were suffering. BC is of two types: non-invasive and invasive. Cancers that penetrate healthy breast tissue are classified as invasive. The majority of BC cases that affect women are invasive malignancies. Non-invasive refers to the removal of cancerous cells from their original site. In this type of cancer, normal breast tissues are unaffected and are in good condition, but cancer cells are trapped inside the milk ducts.[3] Other types of BC are Ductal carcinoma in situ (DCIS)[4], Lobular carcinoma in situ (LCIS**)**, Invasive ductal carcinoma (IDC)[4], and Invasive lobular carcinoma (ILC). To predict and detect cancer at an earlier stage. Various Machine Learning (ML) models are used, but classifying cancer remains a critical challenge for researchers. To classify and detect cancer at an early stage, various ML techniques like Extra Tree (ET), Logistic Regression (LR), Random Forest (RF),

*Author for Correspondence
E-mail: anju24sanga@gmail.com

AdaBoost, eXtreme Gradient Boosting (XGBoost), Decision Tree (DT), CatBoost, Gradient Boosting (GB), Support Vector Machine (SVM), Naïve Bayes (NB), K−Nearest Neighbors (KNN), Light Gradient Boosting Machine (LightGBM), and Hybrid Stacked Ensemble Model (HSEM) are used. Using various ML algorithms, several researchers have conducted research to predict cancer. The efficiency and effectiveness of each algorithm have been evaluated in this work. This paper develops a hybrid ML algorithm that accurately detects and classifies BC using the University of Wisconsin Breast Cancer (WBC) dataset. The primary aim of this work is to automate the diagnosis of BC through multiple diagnostic techniques. The second objective of the BC study using ML is to detect and classify BC at early stage. In addition, we will develop new healthcare technologies and conduct ML−based research. To address BC related issues, our research aims to enhance BC screening using ML models.[5] The research reveals several critical gaps in ML for BC, including model performance and data imbalance. These gaps highlight the need for data handling to enhance the predictive accuracy of ML algorithms.[6]

Literature Review

This section discusses an ML−based study for detecting and classifying BC at an early stage. According to our previous studies, the recent research gaps for breast cancer are as follows: The five-year overall survival rate for BC in India reached from 40% to 62%, with 54% of Indian women reporting the disease at a late stage.[7] Batool & Byun's paper did not implement the SMOTE sampling with ML models to predict BC. They proposed a new ML model, ELRL−E, achieving 97.60% accuracy.[8] Eshun *et al.* developed a deep Convolutional Neural Network (CNN) for an imbalanced BreaKHis dataset consisting of 7909 rows. Out of which 2440 belong to the minority class, i.e., benign, and 5429 to the malignant class. In this paper, only deep learning models are tested without the use of sampling techniques. Breast cancer is generally found in women. The root cause of this type of cancer is an abnormal growth of breast cells, which ultimately forms tumors. These tumors may spread throughout the body if not detected initially.[9] Sukmandhani *et al.* examined the classification capabilities of Deep learning (H2O), Neural Network (NN), KNN, DT,SVM, NB and RF models to categorize BC. The RF model had the highest score (92.26%) among the ML models, followed by SVM, NB, KNN, and DT (88.59%, 90.52%, 88.93%, and 90.50%, respectively). H2O and NN, on the other hand, achieved accuracy of 93.14% and 92.97%, respectively.[10] Deep learning had the highest recall (89.62%), and NB had the highest precision (93.89%). To select highly discriminating features for BC classification, Akkur *et al.* proposed a hybrid feature selection method in which binary Harris hawk optimization was used in conjunction with relief to order features. Using a ten-fold cross-validation approach, SVM outperformed other machine learning models, achieving a specificity of 94.56%. After four features were selected from 30 using the addressed feature selection technique, the accuracies of NB, KNN, SVM, and LR were 98.77%, 97.37%, 97.19%, and 96.13%, respectively. SVM, KNN, LR, and NB, on the other hand, achieved an accuracy of 94.73%, 91.56%, 91.92%, and 92.62% respectively, using thirty features.[11] Okundalaye *et al.* predict BC using SVM model achieving 98% on the same dataset.[12] Walsh & Tardy compared various DL techniques for class imbalance in the BC dataset. However, they did not use SMOTE to address class imbalance.[13]

Hasan *et al.* employed a grammatical evolution-based approach for diagnosing BC and employed SMOTE to balance the dataset.[21] Rana & Bajwa used chemicals for cancer treatment. This article provides an overview of using phytotherapy and apitherapy as alternatives to treating patients with cancer by using chemotherapeutic agents, while specifically mentioning the following bioactive compounds, i.e., melittin, hesperidin, and caffeic acid phenethyl ester, which induce apoptosis and inhibit the proliferation of tumor cells, thereby providing potential for the treatment of cancer-related ailments.[22] Furthermore, ML can improve the accuracy of BC diagnosis through the development of more accurate diagnostic methods. This study uses the BC Wisconsin (Diagnostic) dataset as a source for prior research, but notes that it has limitations that may affect the ability to generalize results.[23–25] Here, we investigate freely accessible datasets related to BC.

According to the aforementioned literature, no researcher has balanced the unbalanced dataset of breast cancer using the SMOTE sampling technique. Thus, the current research work contributes to existing knowledge by implementing the SMOTE technique with machine learning algorithms to achieve more accurate results and balanced datasets, which are lacking in previous studies. Secondly, the

Breast cancer dataset (University of Wisconsin) was analysed and concluded with an accuracy of 97.60% by Batool & Byun in 2024.[8] The present study improves the model's performance by increasing accuracy to 99.12%, as discussed in the results section. Another unique contribution of the current study is the development of novel HSEMs and the determination of how different features of the datasets are related to one another, which helps detect and predict cancer at an early stage. This study exclusively uses machine learning classifiers with the SMOTE sampling technique to achieve better results. The overview of published cancer literature is shown in Table 1.

Main Contributions

Breast cancer can be prevented if the doctor is aware of this at an earlier stage. It is possible through machine learning models. The present study provides a solution in the form of an ML model that not only detects but also predicts cancer at an earlier stage at a very low cost, with the highest accuracy. The model's performance could be compared with that of other cancer classification methods. Our research contribution includes:

- Developed a three−level novel HSEM using Synthetic Minority Oversampling Technique (SMOTE) for balancing at the data level, SVM−RF−XGBoost for cost−sensitive at the algorithm level, and LR at the decision level for a stacked Meta learner.
- The performance metrics such as precision, ROC−AUC, recall, f1 score, and accuracy of this paper have been compared with the existing literature, and we find that our two models, HSEM and RF, performed best.
- To enhance model performance GridSearchCV and RandomSearchCV techniques are used.

Therefore, to identify cancer, we must employ both supervised and unsupervised ML models. Consequently, SMOTE used in this investigation to extract important characteristics from both supervised and unsupervised models. Finally, our target is to recognize the most effective ML models for detecting cancer vulnerability. With this background, the current study was conducted to employ ML to enhance the BC screening and implement a system for identification and management of suspicious cases.

Novelty

The current study addresses class imbalance and hyperparameter optimization to provide a novel approach for categorizing ML algorithms for BC dataset classification. It also introduced a novel HSEM that enhanced the accuracy of model, as shown in the results, discussion, and findings sections.

Table 1 — Overview of published works on Breast Cancer (BC)

Machine learning classifiers and Datasets[Ref]	Pros	Cons
Breast Self-Supervised Temporal Learning Framework (BSTNet)[14]	High AUC makes it more generalised	Requirement of longitudinal MRI
Averaged perception classifier[15]	The investigation has played a noteworthy role in the threshold for FP and FN prediction.	The threshold's importance in false positive and false pessimistic predictions
DieleNet, CNN, Long Short Term Memory (LSTM), attention mechanism[1]	Use of real-time data	No comparison with ML models
GB. Link to download dataset (https://www.kaggle.com/datasets/gunesevitan/breast-cancer-metabric)[16]	Use of a large dataset	Overfitting due to lower test accuracy than training accuracy.
SVM, RF, ET. The dataset name is (BreakHis)[17]	Classify BC data	Limited optimal features and best accuracy classification
RF, SVM, k−NN. Dataset downloaded from (WBCD)[12]	Use of SHapley Additive ex Planations (SHAP)	Not using DL and computer vision techniques
RF. Dataset link (https://www.cancerimagingarchive.net/collection/duke-breast-cancer-mri/)[18]	Use of an innovative approach (contralateral breast texture features)	An imbalanced dataset was used, and a moderate accuracy was achieved.
DT, SVM, EL, KNN, NB. Dataset downloaded from (MBCD and WBCD) [19]	Use LASSO with ML	Do not use DL models for early prediction on BC
RF. Link to download dataset (https://www.kaggle.com/datasets/fatemehmehrparvar/breast-cancer-prediction/data)[20]	Use ML models	An imbalanced dataset was used, and the RF model achieved a moderate AUC−ROC.
LDA, RC, ET, LGBM. Dataset downloaded from (WBCD)[8]	ELRL−E, a unique model, is used with ML models	Evaluation of small datasets is a limitation

Materials and Methods

The actual source of BC data is the University of Wisconsin. The BC dataset, downloaded from the publicly available (open-access) website Kaggle (https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset/data), is used for analysis in this study. The breast cancer dataset initially contains 569 rows and 32 columns, including a diagnosis feature (target variable).The BC dataset includes M-Malignant (212) and B-Benign (357), indicating the presence of cancer. Thus, this BC dataset is imbalanced due to (M < B). This type of data is dangerous for medical diagnosis. Therefore, to balance the dataset, the SMOTE sampling technique was implemented on the training dataset. Choose y_i is the instance of the minority class of the BC dataset, i.e., M, SMOTE generates a new sample, i.e., y_{new}, as given in Eq. (1).

$$y_{new} = y_{en} + \lambda (y_{am} - y_{en}) \quad \ldots (1)$$

where, $\lambda \in (0,1)$ is the random scalar, y_{en} instance of the minority class, y_{am} k-nearest neighbours of y_{en}. Feature selection is applied after resampling, and three ML models are used as base learners: RF, SVM, and XGBoost. Out-of-fold predictions are generated using k-fold cross-validation for Meta feature generation. The matrix is defined as:

$$X = [\text{Predicted probability of (SVM, RF, and XGBoost)}] \quad \ldots (2)$$

This matrix ensures unbiasedness and prevents overfitting. For Meta, the learner LR model is used, and the model's performance is finally evaluated. Before and after SMOTE, the dataset visualizations are shown in Fig. 1. During the study, classification was the key challenge between Malignant and benign. All features of the BC dataset, except the ID and the target column, are shown in Fig. 2.

The proposed methodology can be divided into five components, as seen in Fig. 3(a). The data acquisition component obtained data from Kaggle. The data preprocessing component modified the dataset for ML using encoding, normalizing, and feature scaling, and then split the dataset into 80% (training) and 20% (testing) subsets. The dataset was classified by using ML algorithms to classify the dataset. The ML implementation was accomplished using the following Python packages: NUMPY, PANDAS, SKLEARN, and MATPLOTLIB as well as using ROC curves, f1−score, recall, accuracy, and precision to evaluate the performance of the ML models. Exploratory Data Analysis (EDA) will be an important part of classifying cancer datasets. During this analysis, we understand the dataset's features, labels, and data types, check for class imbalance, identify missing values, and remove outliers and null values. Data visualization is used to identify patterns using histograms, scatterplots, boxplots, etc. Dummy variables created with the help of Data transformation. In feature selection, the prediction power of machine learning algorithms is increased by eliminating redundant and irrelevant data. The complete EDA process is shown in Fig. 3 (b). To classify the problem, machine learning models: DT, XGBoost, GB, CatBoost, SVM, LR (Logistic Regression), KNN, NB, LightGBM, RF, AdaBoost, ET, and HSEM are implemented with and without the SMOTE technique, which ultimately enhances the model's performance.

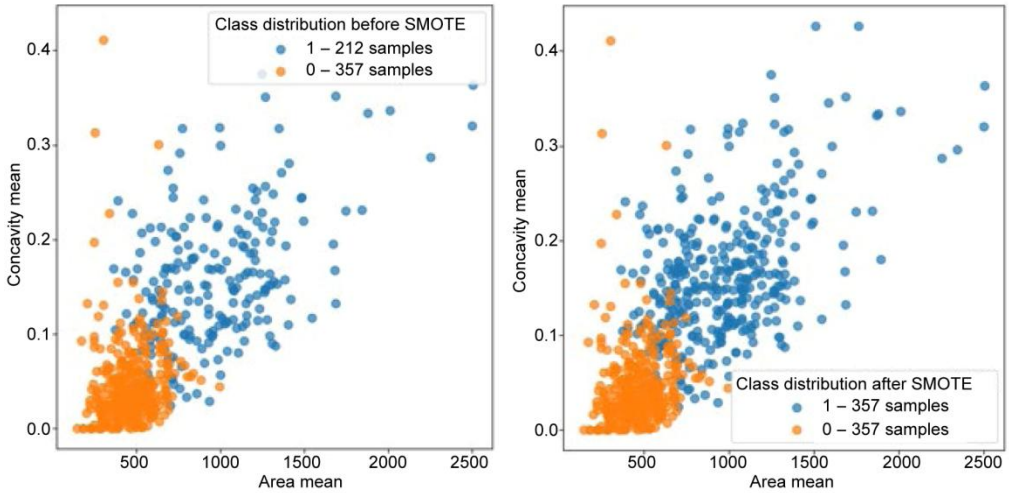

Fig. 1 — Data visualization before and after SMOTE

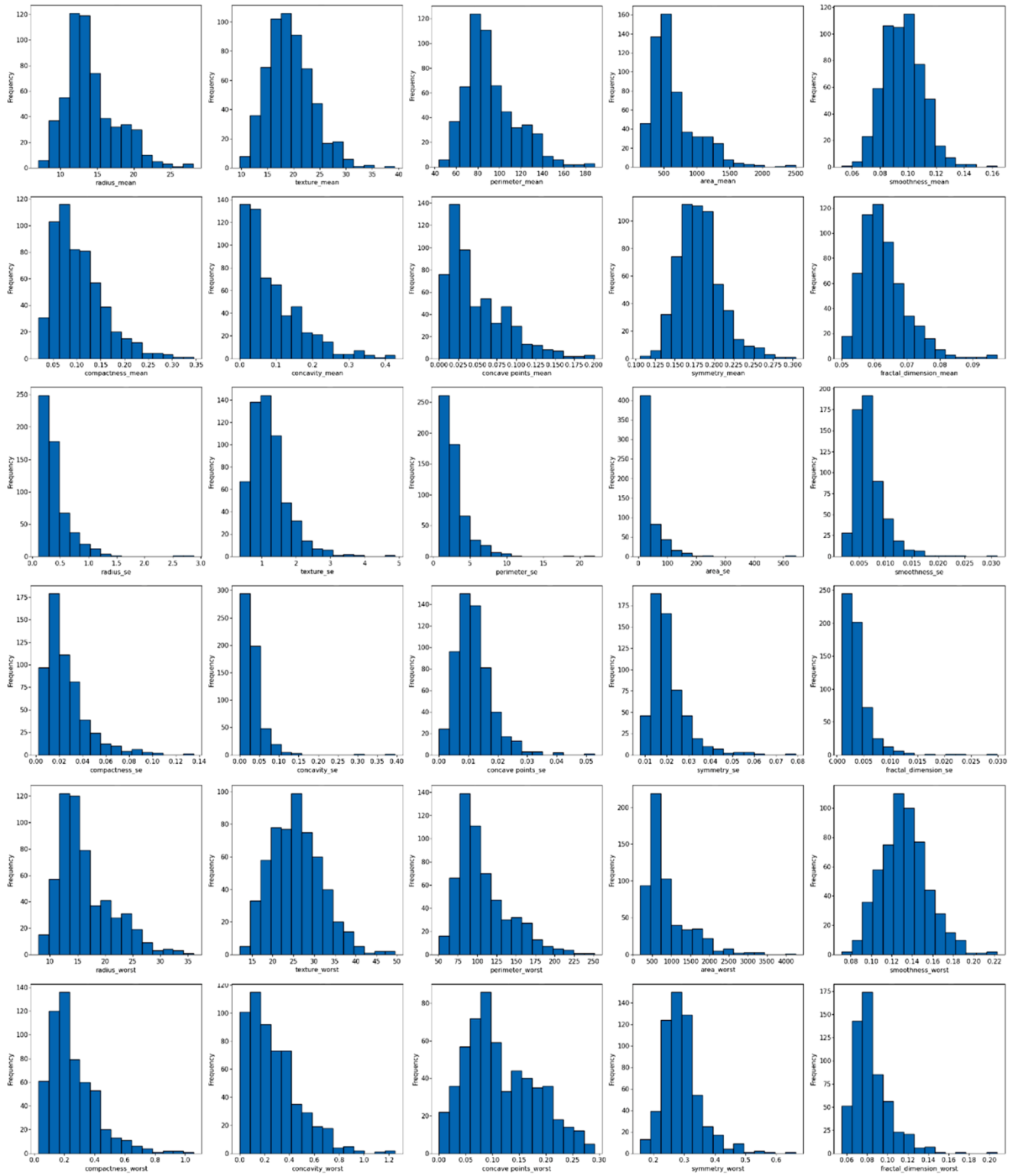

Fig. 2 — Histogram visualization of BC dataset features

Fig. 3 — (a) Proposed methodology and (b) EDA process of BC dataset

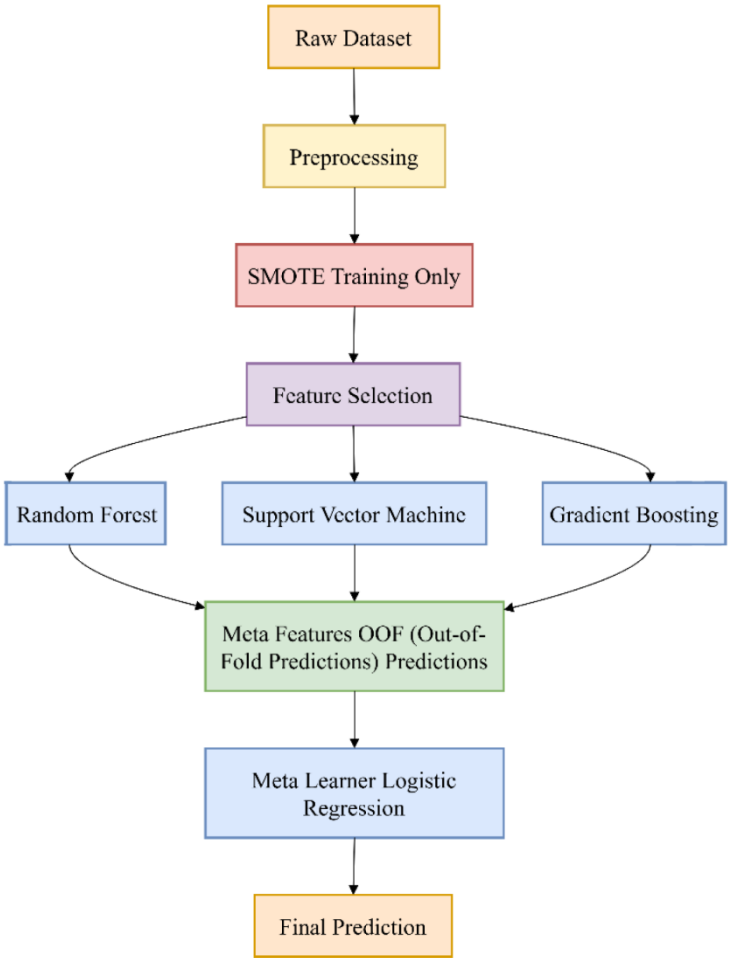

Fig. 4 — Proposed methodology for novel HSEM of BC dataset

Various classification models were used to predict BC accurately. The current study proposed a novel model, HSEM, that used four ML models (SVM−RF−XGBoost and LR). The HSEM used (SMOTE for balancing at the data level, SVM−RF−XGBoost for cost-sensitive at the algorithm level, and LR at the decision level for the stacked Meta learner) to detect BC at an earlier stage. The proposed methodology for HSEM is shown in Fig. 4.

Results & Discussion

The best algorithms for the BC dataset with and without SMOTE were performed on the basis of ROC-AUC curves, confusion matrices, and various metrics that reported the accuracy, recall, precision, f1-score of the algorithms.

The Random Forest algorithm in ML is a very effective tree learning approach that builds a large number of Decision Trees in the training phase by utilizing some of the data in the dataset to attempt to estimate random subsets of features in each partition that go into building each decision tree. When predicting, the Random Forest averages the output by all trees if it is for regression purposes, or selects the largest vote count if it is for a classification type of outcome.[26]

Using logistic regression, which is a supervised ML method, is applicable for classification problems

where you want to predict the likelihood of an instance belonging to a specific class/label. Logistic regression works well for cases that only have two possible outcomes.[27] However, it can also work for classification problems with three or more potential outcomes.[28]

A SVC (Support Vector Classifier) is a type of SVM used for classification tasks. SVM is a supervised ML model, which handle linear as well as nonlinear data.[29]

AdaBoost is a boosting algorithm that uses the stagewise addition approach, which trains several weak learners to produce strong learners by combining multiple classifiers to improve their accuracy.[30] The mathematical formulation of metrics given as:

$$Accuracy\ (a) = \frac{tpvalue\ + tnvalue}{tpvalue\ + tnvalue\ + fpvalue\ + fnvaue} \quad \ldots (3)$$

$$Precsion\ (p) = \frac{tpvalue}{tpvalue\ + fpvalue} \quad \ldots (4)$$

$$Recall\ (r) = \frac{tpvalue}{tpvalue\ + fnvalue} \quad \ldots (5)$$

$$f1 - score = 2 * \frac{p*r}{p+r} \quad \ldots (6)$$

$$True\ Positive\ Rate = \frac{tpvalue}{tpvalue\ + fnvalue} \quad \ldots (7)$$

$$False\ Positive\ Rate = \frac{fpvalue}{fpvalue\ + tnvalue} \quad \ldots (8)$$

Confusion matrix is shown in Table 2.

where, tp value = true positive, tn value = true negative, fp value = false positive, fn value = false negative

Table 2 — Confusion matrix

	predicted positive	predicted negative
actual positive	$tp\ value$	$fn\ value$
actual negative	$fp\ value$	$tn\ value$

RF algorithm provides accurate results with or without SMOTE techniques. The test accuracy, recall, precision, and f1-score for the selected models on BC dataset presented in Table 3. The RF model achieved highest value of accuracy (0.9912) on an imbalanced dataset. Second, the outperformed model is SVM, with accuracy 0.9825 and precision, recall, and f1-scores ((1.0000, 0.9535, and 0.9762 respectively).

Comparison results of the four best-performing models, with classification metrics shown in Fig. 5, without SMOTE. Among all these models, RF performed best with an accuracy 99.12%, recall rate 97.67%, precision value of 100%, and f1−score of 98.82%.

The Receiver Operating Characteristic − Area Under Curve (ROC – AUC) is a graph for assessing the efficiency of a binary classification algorithm. It allows us to analyze the ability of the model to discriminate between positive outcomes (e.g., individuals with a particular medical condition) and negative outcomes (e.g., individuals without that condition) across several thresholds. The performance of the ML models also demonstrated through the ROC curve shown in Fig. 6. DT has the lowest ROC value 0.9140 whereas the RF classifier has the highest ROC−AUC value 0.9990 (without SMOTE).After balancing the dataset, the obtained accuracy of models are shows in Table 4. Logistic regression achieved the second−highest accuracy (98.59%), but HSEM performed best overall, with 99.12% accuracy. The original contribution of this research is the implementation of ML models using the SMOTE technique to enhance model accuracy, enabling earlier classification of cancer.

Grid Search CV and Random Search CV hyperparameter optimization techniques of ML are

Table 3 — Classification report of the BC dataset before SMOTE

Classifier	BC Dataset			
	Test accuracy	precision	recall	f1−score
CatBoost	0.9561	0.9524	0.9302	0.9412
AdaBoost	0.9474	0.9302	0.9302	0.9302
LightGBM	0.9561	0.9524	0.9302	0.9412
RF	0.9912	1.0000	0.9767	0.9882
XGB	0.9649	0.9756	0.9302	0.9524
DT	0.9386	0.9286	0.9070	0.9176
GB	0.9561	0.9524	0.9302	0.9412
ET	0.9737	0.9762	0.9535	0.9647
SVM	0.9825	1.0000	0.9535	0.9762
NB	0.9649	0.9756	0.9302	0.9524
KNN	0.9649	0.9535	0.9535	0.9535
LR	0.9737	0.9762	0.9535	0.9647

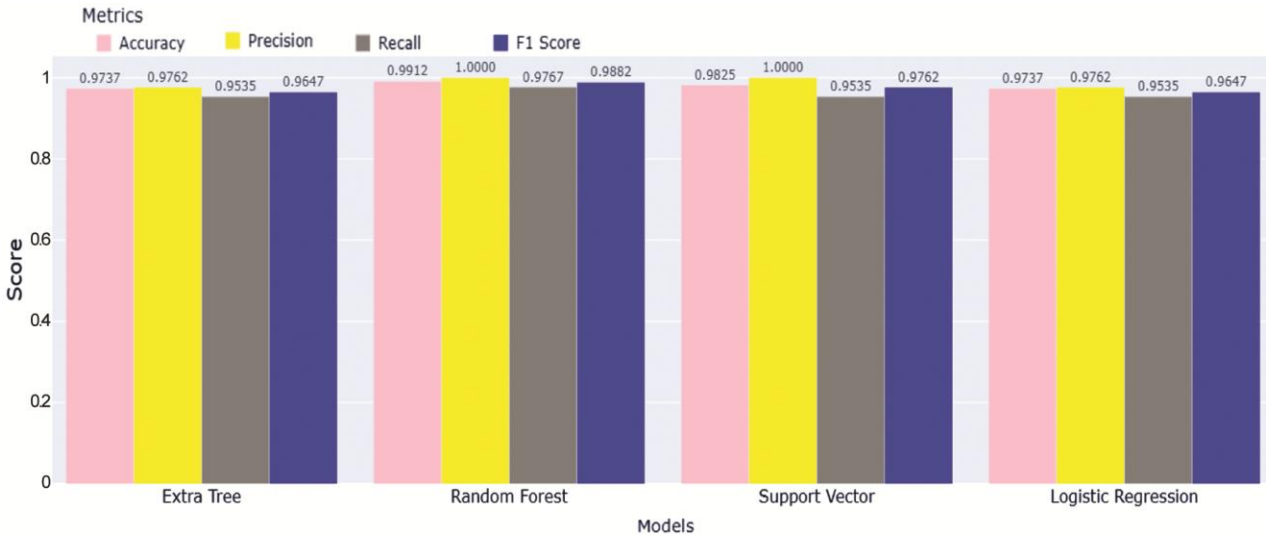

Fig. 5 — Comparison results of the best four models

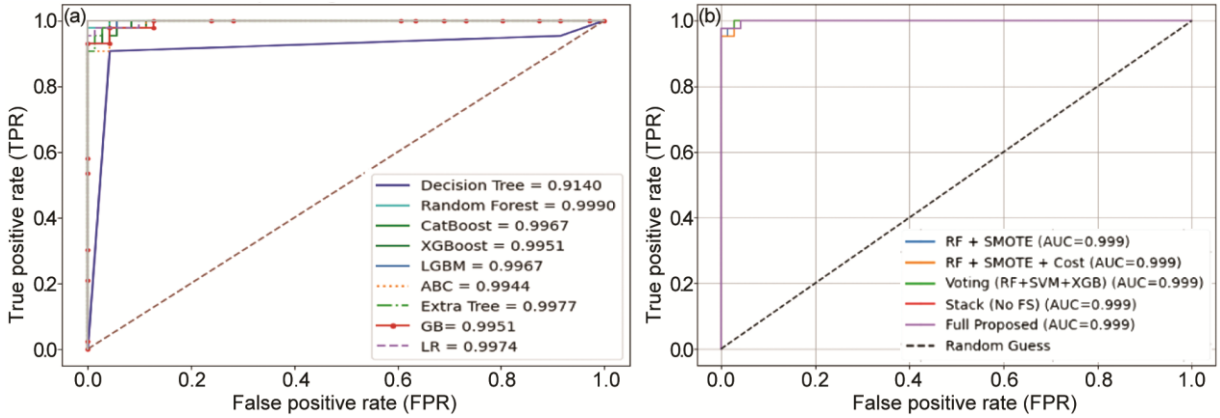

Fig. 6 — ROC-AUC curve of BC dataset: (a) without SMOTE and (b) HSEM with SMOTE

Table 4 — ML algorithms after SMOTE

ML models	Test accuracy after SMOTE
CatBoost	97.88%
AdaBoost	97.88%
LightGBM	97.88%
RF	97.88%
XGB	97.18%
DT	95.07%
GB	96.47%
ET	97.88%
SVM	97.88%
NB	97.18%
KNN	96.47%
LR	98.59%
HSEM	99.12%

applied to the dataset that improve ML models performance and identify best scores and parameters for each algorithm. From Table 5, we find the four best−performing ML models (LR, SVM, XGB, and CatBoost), and the least-performing algorithms are NB, DT, RF, and GB. Using Table 6, we find that LR also performed best with RandomSearchCV, with parameters 'solver': 'liblinear', 'penalty': 'l2', 'C': 0.1, whereas NB is the least performing model.

After hyperparameter tuning, we found that the accuracy of all models improved as shown in Tables (5 and 6). The statistical test Chi−Square (χ^2) used to show the relationship between each feature and target variable. The feature whose p−value < (0.05) is more significant for classification. The top and least significant value of features shown in Table 7. The feature “radius_wrost” shows strong dependency due to high value of Chi−Square (389.98) and low value p−value. Similarly, “smoothness_se” not strongly associated with target feature.

The investigation for BC disease prediction using the same BC dataset compared with the newly developed models used in Table 8. It was found that in 2024, Nafea *et al.* [30] proposed an XGB model with

Table 5 — ML models after GridSearchCV

Algorithms	Best Score after GridSearchCV	Best Parameters
CatBoost	0.9758	{'depth'=6, 'iterations'=200, l2 'leaf_reg'=3, 'learning_rate'= 0.2}
AdaBoost	0.9736	{'learning_rate'=0.9, 'n_estimators'=105}
RF	0.9670	{'max_depth'=12, 'max_features'=0.2, 'max_samples'=1.0, 'n_estimators'=20}
XGB	0.9758	{'max_depth'=3, 'n_estimators'=100, 'subsample'=-0.8, 'learning_rate'=0.1, 'colsample_bytree' =1.0}
DT	0.9538	{'criterion'=entropy, 'max_depth'=5, 'min_samples_leaf'=2, 'min_samples_split'=2}
GB	0.9692	{'learning_rate'=0.5, 'max_depth'=3, 'n_estimators'=105}
ET	0.9604	{'criterion'=entropy, 'max_depth'=10, 'min_samples_leaf'=6, 'min_samples_split'=6, 'n_estimators'=105}
SVM	0.9758	{'C'=10, 'gamma'=scale, 'kernel'=rbf}
NB	0.9296	{'var_smoothing'=1e-09}
KNN	0.9626	{'metric'=euclidean, 'n_neighbors'=5, 'weights'=uniform}
LR	0.9780	{'C'= 0.1, 'penalty'=l2, 'solver'=liblinear}

Table 6 — ML models after RandomSearchCV

Algorithms	Best Score after Random Search CV	Accuracy	Best Parameters
CatBoost	0.9714	0.9736	{'random_strength'=3, 'learning_rate'=0.1, l2 'leaf_reg'=5, 'iterations'=100, 'depth'=4, 'border_count'=64}
AdaBoost	0.9714	0.9736	{'n_estimators'=100, 'learning_rate'=1.0, 'algorithm'=SAMME}
LightGBM	0.9714	0.9561	{'subsample'=1.0, 'n_estimators'=200, 'max_depth'=9, 'learning_rate'=0.1, 'colsample_bytree'=1.0, 'reg_lambda'=0, 'num_leaves'=31, 'reg_alpha'=0}
RF	0.9626	0.9649	{'max_depth'=20, 'n_estimators'=100, 'min_samples_split'=2, 'min_samples_leaf'=2, 'max_features'=sqrt}
XGB	0.9736	0.9736	{'subsample'=0.8, 'reg_alpha'=0, 'n_estimators'=150, 'max depth'=11, 'gamma'=0, 'colsample bytree'=0.8, 'min_child_weight'=5, 'reg_lambda'=0.1, 'learning_rate'=0.1}
DT	0.9340	0.9649	{'min samples split'=2, 'min samples leaf'=1, 'max_features'=log2, 'max_depth'=15, 'splitter'=best, 'criterion'=gini}
GB	0.9714	0.9649	{'learning_rate'=0.5, 'n_estimators'=50, 'min_samples_split'=2, 'min_samples_leaf'=4, 'max_depth'=5, 'subsample'=0.8}
ET	0.9648	0.9736	{'max_depth'=50, 'n_estimators'=150, 'min_samples_leaf'=2, 'max_features'=sqrt, 'min_samples_split'=2 }
SVM	0.9758	0.9736	{'kernel'=rbf, 'gamma'=scale, 'degree'=4, 'C'=10}
NB	0.9340	0.9561	{'var_smoothing'=0.12328467394420659}
KNN	0.9626	0.9473	{'weights'=distance, 'n_neighbors'=5, 'metric'=euclidean}
LR	0.9780	0.9912	{'solver'=liblinear, 'penalty'=l2, 'C'=0.1}

Table 7 — Summary of Chi−Square test of BC dataset

Index	Feature	χ^2 value	p−value	Significant ($p < 0.05$)
21	radius_worst	389.988621	3.263040e-84	Yes
24	area_worst	388.909811	5.588362e-84	Yes
23	perimeter_worst	385.898118	2.509492e-83	Yes
28	concave points_worst	374.088770	9.062320e-81	Yes
8	concave points_mean	358.423588	2.236927e-77	Yes
10	fractal_dimension_mean	12.236962	6.613980e-03	Yes
12	texture_se	8.399456	3.843876e-02	Yes
19	symmetry_se	5.455808	1.413058e-01	No
15	smoothness_se	4.619883	2.018431e-01	No

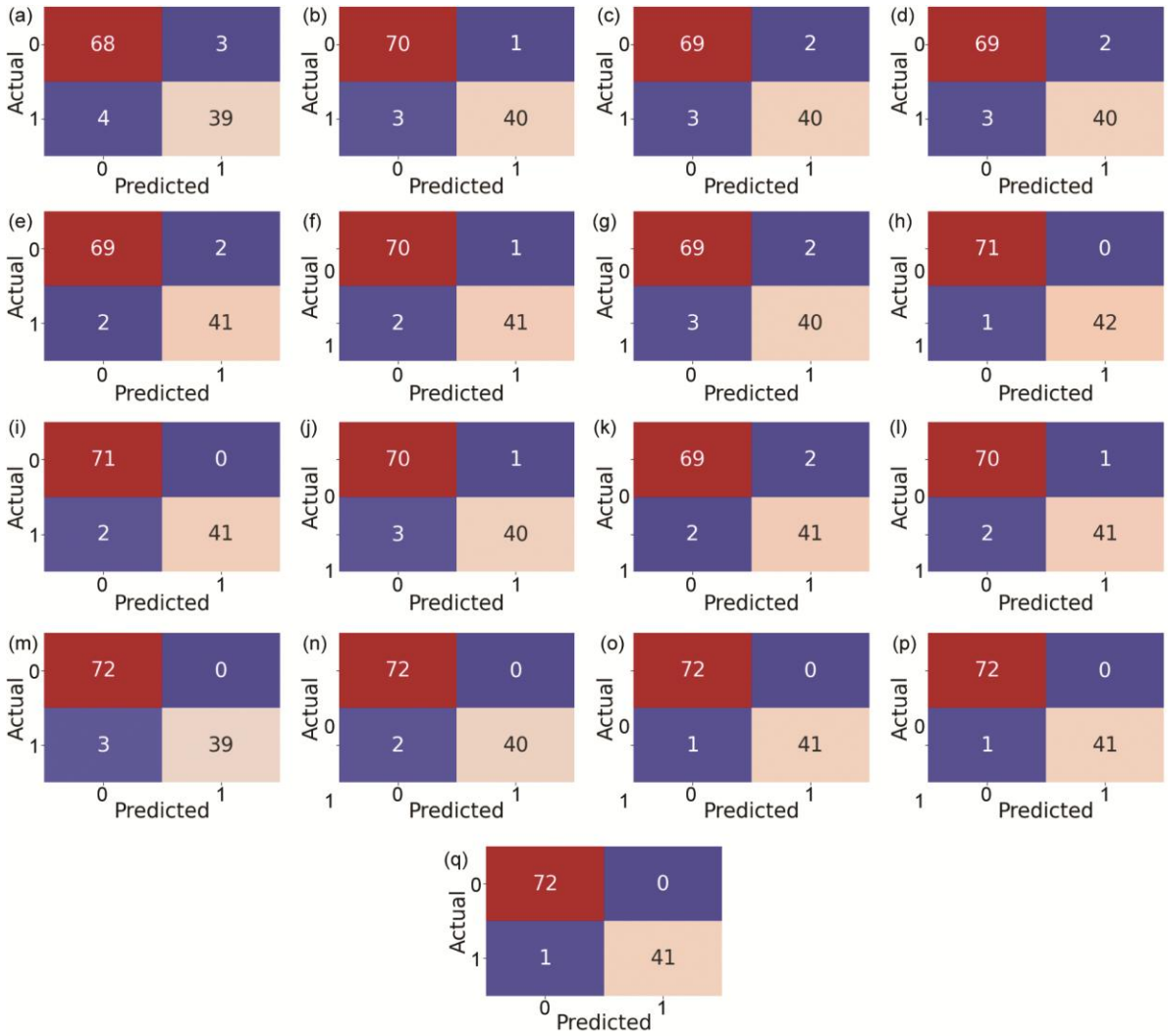

Fig. 7 — Proposed novel HSEM (Full Proposed) as well as all ML models' confusion matrix: (a) Decision Tree, (b) XGBoost, (c) CatBoost, (d) Gradient Boosting, (e) AdaBoost, (f) Extra Tree, (g) LightGBM, (h) Random Forest, (i) SVM, (j) Naive Bayes, (k) KNN, (l) Logistic Regression, (m) RF + MOTE, (n) RF + SMOTE + Cost, (o) Voting, (p) Stacking, (q) Proposed Model

Table 8 — Proposed model comparison with the other ML models

Model[Ref]	Accuracy
XGB[31]	98.24%
ELRL−E[8]	97.60%
XGB[32]	97.40%
Stacking and Logistic Regression[33]	97.37%
H2O[10]	93.14%
SVM[12]	98%
Proposed HSEM	99.12%

an accuracy of 98.24%, which is approximately 1% lower than the proposed HSEM (99.12%), and Batool & Bynn achieved an accuracy of up to 97.60% using the same BC dataset.[8] Still, among the articles listed in the literature, the models developed using the SMOTE sampling strategy show excellent potential to predict breast cancer at an earlier stage, achieving accuracies up to 99.12% (with SMOTE). Thus, the current research work has the highest cancer identification rates among published works in the literature. Thus, the novelty of existing research work is not only in terms of the model's accuracy improvement but also contributes to balancing the unbalanced dataset through the SMOTE technique and proposes a novel HSEM.

By considering the confusion matrix displayed in Fig. 7, the best performer models are Full proposed (HSEM), Stacking, Voting, and RF because all these models have the highest true positive and actual negative values, i.e., 71 + 42 = 113, 72 + 41 = 113, and least performer is DT (i.e., 68 + 39 = 107) have lowest true positive and actual negative value.

Conclusions

The current study presents a cost-effective solution for early cancer detection and prediction. A key contribution of this study is the application of SMOTE for dataset analysis, which enhances the model's performance in BC analysis. The developed ML model achieved an impressive 99.12% accuracy (without SMOTE RF), outperforming other traditional ML methods. In comparison, LR and SVM classifiers achieved an accuracy of 98.59% and 98.25%. This trio of models shows the utility of ML in improving cancer detection outcomes. But the proposed HSEM achieved 99.12% accuracy and an AUC of 0.999. The developed model's accuracy surpasses that of other advanced cancer classification techniques, establishing a noteworthy improvement in the field. Looking ahead, the DL algorithms could further enhance the analysis of similar problems, paving the way for even more robust predictive models in cancer diagnostics. Future works include integrated explainable AI in the proposed HSEM model.

References

1 Abdelmotagally H, Allam J P, Iratni R, Atef M & Hussein M, DieleNet: Dielectric properties based breast cancer classification using deep learning, *Biomed Signal Process Control*, **113** (2026), doi: 10.1016/j.bspc.2025.108981.

2 Zhu W, Lou Q, Vang Y S & Xie X, Deep multi-instance networks with sparse label assignment for whole mammogram classification, *Int Conf Med Image Comput Comput Assist Interv (MICCAI)*, **10435** (2017) 603–611, doi: 10.1007/978-3-319-66179-7_69.

3 World Health Organization, Breast cancer fact sheet, WHO, (2026), Available at: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

4 Shia W C, Hsu F R, Dai S T, Guo S L, Chen & D R, Semantic segmentation of the malignant breast imaging reporting and data system lexicon on breast ultrasound images by using DeepLab v3+, *Sensors*, **22(14)** (2022) 1–14, doi: 10.3390/s22145352.

5 Karunasena G I, Amaratunga D & Haigh R, Capacity building for sustainable post disaster waste management: Construction & demolition waste, *CIB W89 Int Conf Build Educ Res (BEAR)* 2014.

6 Almahmoud Z J, *Forecasting Cyber Threats & Pertinent Alleviation Technologies*, Doctoral dissertation (Birkbeck, University of London) 2024.

7 Sangwan R K, Huda R K, Panigrahi A, Toteja G S, Sharma A K, Thakor M & Kumar P, Strengthening breast cancer screening program through health education of women and capacity building of primary healthcare providers, *Front Public Health*, **11** (2023) 1–10, doi: 10.3389/fpubh.2023.1276853.

8 Batool A & Byun Y C, Towards improving breast cancer classification using an adaptive voting ensemble learning algorithm, *IEEE Access*, **12** (2024) 12869−12882, doi: 10.1109/ACCESS.2024.3356602.

9 Eshun R B, Islam A K & Bikdash M, A deep convolutional neural network for the classification of imbalanced breast cancer dataset, *Healthc Anal*, **5** (2024) 1–10, doi: 10.1016/j.health.2024.100330.

10 Sukmandhani A A, Lukas L, Heryadi Y, Suparta W & Wibowo A, Classification algorithm analysis for breast cancer, *E3S Web Conf*, **388** (2023) 1–7, doi: 10.1051/e3sconf/202338802012.

11 Akkur E, Türk F & Eroğul O, Breast cancer classification using a novel hybrid feature selection approach, *Neural Netw World*, **33** (2023) 67–83, doi: 10.14311/NNW.2023.33.005.

12 Okundalaye O O, Özdemir N & Awonusika R O, Early breast cancer prediction using optimized machine learning and tumor-immune modeling, *J Comput Appl Math*, **473** (2025), 116875, doi: 10.1016/j.cam.2025.116875.

13 Walsh R & Tardy M, A comparison of techniques for class imbalance in deep learning classification of breast cancer, *Diagnostics*, **13(1)** (2022) 1–19, doi: 10.3390/diagnostics13010067.

14 Huang X, Xu Z, Zhao Y, Wang Y, Liu Y, Hu W, Zhao K, Yao L, He J, Yu Y & Deng T, Longitudinal MRI-based deep learning model for predicting pathological complete response in breast cancer: a multicenter, retrospective cohort study, *NPJ Precis Oncol*, **57** (2026) 1–12, doi: 10.1038/s41698-025-01256-2.

15 De Barros V A M, Paiva H M & Hayashi V T, Using PBL and agile to teach artificial intelligence to undergraduate computing students, *IEEE Access*, **11** (2023) 77737−77749, doi: 10.1109/ACCESS.2023.3298294.

16 Asfaw B B & Tegaw E M, Explainable machine learning to compare the overall survival status between patients receiving mastectomy and breast conserving surgeries, *Sci Rep*, **15(1)** (2025) 1–14, doi: 10.1038/s41598-025-91064-2.

17 Atban F, Ekinci E & Garip Z, Traditional machine learning algorithms for breast cancer image classification with optimized deep features, *Biomed Signal Process and Control*, **81** (2023) 1–12, doi: 10.1016/j.bspc.2022.104534.

18 Nuzla A N, Nabeel A K M, Nirmal W A S, Hewavithana P B & Jayatilake M L, Machine learning-based classification of histological subtypes of invasive breast cancer using MRI contralateral breast texture features, *Sci Rep*, **15(1)** (2025) 1–13, doi: 10.1038/s41598-025-23498-7.

19 Akkur E, Türk F & Erogul O, Breast cancer diagnosis using feature selection approaches and Bayesian optimization, *Comput Syst Sci Eng*, **45(2)** (2023) 1017–1031, doi: 10.32604/csse.2023.033003.

20 Adiga U, Vasishta S, Augustine A J, Farzia K, Venkataravikanth E & Ravi L, Transforming breast cancer prediction: advanced machine learning models for accurate prediction and personalized care, *Int J Stats Med Res*, (2025) 569−577, doi: 10.6000/1929-6029.2025.14.54.

21 Hasan Y, de Lima A, Namjoo E, de Bulnes D F, Albarracín J F & Ryan C, Improving Breast Cancer Diagnosis Using Grammatical Evolution-Based Feature Selection, *SN Comput Sci*, **6(4)** (2025), 1–15, doi: 10.1007/s42979-025-03840-6.

22 Rana A & Bajwa H K, Therapeutics of bioactive compounds from medicinal plants and honeybee products against cancer: phytotherapy and apitherapy as alternatives to chemotherapy,

J Sci Ind Res, **82(08)** (2023) 805−817, doi: 10.56042/jsir.v82i08.1895.

23 Soni G, Sharma A & Singh C, Classification of breast cancer and its diagnosis using machine learning, *5th Int Conf Intell Technol (CONIT) IEEE*, (2025) 1−7, doi: 10.1109/CONIT65521.2025.11166985.

24 Amrane M, Oukid S, Gagaoua I & Ensari T, Breast cancer classification using machine learning, *Electr Electron Comput Sci Biomed Eng Meet (EBBT), IEEE*, (2018) 1−4, doi: 10.1109/EBBT.2018.8391453.

25 Islam M M, Haque M R, Iqbal H, Hasan M M, Hasan M & Kabir M N, Breast cancer prediction: a comparative study using machine learning techniques, *SN Comput Sci*, **1(5)** (2020), 1–14, doi: 10.1007/s42979-020-00305-w.

26 Apat S K, Mishra J, Raju K S & Padhy N, An artificial intelligence-based crop recommendation system using machine learning, *J Sci Ind Res*, **82(05)** (2023) 558−567, doi: 10.56042/jsir.v82i05.1092.

27 da Silva A A, Lovisolo L & Ferreira T N, Machine learning ship classifiers for signals from passive sonars, *Appl Sci*, **15(13)** (2025), 1–33, doi: 10.3390/app15136952.

28 Suraj Arya, A, & Ramli N A, Predicting the stress level of students using supervised machine learning and artificial neural network (ANN), *Indian J Eng*, **21** (2024) 1–24, doi: 10.54905/disssi.v21i55.e9ije1684.

29 Cheng Q, Yang Z, Guodong Y, Ya L, Le L, Xiuying W, Juzhen W, Mingxi W & Qianglin L, AI-Optimised aeration control in SBR systems: an inverse SVM framework toward carbon-neutral wastewater treatment, *Environ Technol*, **47(1)** (2026) 37−51, doi: 10.1080/09593330.2025.2562373.

30 Krishna H V & Samatha B, Intruder detection system employing modified Adaboost for Industrial Internet of Things (IIoT), *Inf Secur J Glob Perspect*, **35(1) (2026)** 189−202, doi: 10.1080/19393555.2025.2484351.

31 Nafea A A, Manar A L, Alheeti K M A, Alsumaidaie M S I & Ani M M, A hybrid method of 1D-CNN and machine learning algorithms for breast cancer detection, *Baghdad Sci J*, **21(10)** (2024) 3333−3343, doi: 10.21123/bsj.2024.9443.

32 Chen H, Wang N, Du X, Mei K, Zhou Y & Cai G, Classification prediction of breast cancer based on machine learning, *Comput Intell Neurosci*, **2023** (2023) 1–9, doi: 10.1155/2023/6530719.

33 Laghmati S, Hamida S, Hicham K, Cherradi B & Tmiri A, An improved breast cancer disease prediction system using ML and PCA, Multimed Tools Appl, **83(11)** (2024) 33785−33821, doi: 10.1007/s11042-023-16874-w.