Clinical Decision Support Meets Machine Learning: Statistical Evaluation of Optimized Feature Selection

Hanaa Elgohari [(1),(4)] , Mohamed Zakariaa Fouda [(2)] , and Omar M. Elzeki [(3)]

(1) Department of Applied Statistics, Faculty of Commerce, Mansoura University, Mansoura, Egypt
(2) Researcher at Department of Applied Statistics, Faculty of Commerce, Mansoura University, Egypt
(3) Faculty of Computer Science and Engineering, New Mansoura University, Mansoura, Egypt
(4) Faculty of Business, Horus University, New Damietta, Egypt

Abstract

Breast cancer (BC) is one of the most prevalent diseases and ranks as the second leading cause of mortality in both developed and developing nations.. BC is a serious problem and causes a considerable mortality rate worldwide for women. The diagnosis of BC includes differentiation between malignant and benign tumors, accurate differentiation is crucial. As a result, automating this classification process is essential to reduce reliance on a physician's experience and subjective judgment. The primary objective of this study is to apply multiple machine learning techniques to classify tumors as malignant or benign using the Wisconsin Breast Cancer Diagnostic dataset and to determine the most accurate classifier among them.. The performance comparison of classifiers with different feature selection methods is summarized as follows. We conclude that Gradient Boosting and XGBoost provide robust results, but their performance does not surpass that of simpler models like Logistic Regression.

Keywords: Breast Cancer, Feature selection, Univariate selection (K- Best), Filter method, PCA, classification, machine learning

1. Introduction

Recent studies have highlighted a significantly high mortality rate linked to breast carcinoma. World Health Organization (WHO) data demonstrates that breast cancer impacts in excess of 1.5 million women globally on an annual basis. It is one of the oldest known forms of cancer, with its earliest documentation traced back to ancient Egypt around 1600 BC [1] . Treatment approaches for breast cancer generally fall into two categories: local and systemic, which include methods

like chemotherapy and hormone therapy. Physicians often employ both types simultaneously to achieve the most effective outcomes. Early detection plays a vital role in timely diagnosis and treatment, which are key factors in improving long-term survival rates. With early intervention, breast cancer can be effectively managed, increasing the chances of saving lives. The healthcare sector increasingly relies on data mining to extract meaningful insights from extensive medical databases. This process supports informed decision-making and enhances the quality of healthcare services. In the context of breast cancer, data mining proves valuable in prevention, personalized treatment based on symptoms, and correcting errors in hospital records [2] .

Data mining and machine learning (ML) techniques have been widely utilized across various areas of medical and healthcare (MHC) systems, offering valuable outcomes by leveraging advanced technologies to identify contributing factors to breast cancer (BC) and uncover hidden patterns within clinical data. These approaches help address numerous challenges and provide critical support to clinicians and medical researchers. Numerous studies on BC have employed computer-aided diagnostic systems based on imaging techniques such as magnetic resonance imaging (MRI), thermography, mammography, ultrasound, and histopathological biopsy images either individually or in combination. Most MHC applications for BC focus on tumor detection and classification into benign or malignant categories. Typically, these systems follow a five-step pipeline: data acquisition, image preprocessing, segmentation, feature extraction, feature selection and optimization, and classification. However, the specific methods and algorithms used in each step may vary across different implementations [4]. The use of ML algorithms functions optimally for the early detection of signs and symptoms of BC, but developing an MHC system in the medical domain raises problems pertaining to privacy, confidentiality, ethics, and security which make dataset acquisition quite challenging. Not long ago, the Wisconsin Breast Cancer Dataset (WBCD) was developed as a benchmark dataset for breast cancer detection and classification tasks. We provide the dataset online in UCI ML Repository and it contains the features extracted from the images after necessary pre-processing, segmentation and extraction is done over the acquired images. Data Classification is however complicated stage as it relies on the selected and extracted features as well as the selected ML model and its hyperparameters. Features selection and/or optimization methods which were applied to enhance the classification accuracy and prediction results by discarding the redundant and irrelevant features to mitigate the computational complexities. Although feature trans- formation and

dimensionality reduction are used for the same purpose, these tech-niques are mostly used for high dimensional data which contain the features in large amount.

he proposed approach employs a max-voting ensemble, where predictions from multiple models are combined and the final output is selected based on majority agreement [3].

Following data preprocessing, we utilized several feature selection techniques and employ various algorithms. The results of the best-performing classification models are highlighted in bold in Table. Then, we identify the most effective feature selection technique. Based on the predictions of the three top-performing classifiers, we introduce voting classifier. The workflow of the proposed methodology is illustrated in Fig. (1)

To examine and compile the MHC systems utilized for BC detection and classification, a number of surveys were carried out. The four primary processes of MHC systems—image pre-processing, picture segmentation, feature extraction, and classification were used to investigate various strategies. Each section concludes with a comparison table that includes the name, definition, reference, and benefits and drawbacks of each strategy. Additionally, the accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of six contemporary categorization methods were examined.

2. 2. Literature Review

Globally, breast cancer remains a leading cause of mortality among women. Timely detection significantly enhances treatment outcomes by enabling early therapeutic intervention. Recent advances have incorporated machine learning algorithms into computer-assisted diagnostic tools to facilitate early breast cancer identification.. The primary goal of a CAD system is to classify patients into different categories of benign and malignant cancer, or to classify tumors into different categories of likelihood of development into life-threatening cases. In supervised learning, outcomes are known for a set of data, and the goal is to construct a classifier that can accurately predict the outcome of a new case. The data are classified into a set of attributes or feature that, describe the data. The first step in developing a machine-learning based breast cancer diagnosis system is data collection. The dataset instances of patient with breast cancer for whome outcomes are already known. The Wisconsin Breast Cancer Database (WBCD) is a common source of breast cancer data. This data is partitioned into a training set to construct, the classifier and a test set to evaluate the performance of the learned classifier. The WBCD contains 699 instances described by 9 attributes a piece, and a single 2 class outcome attribute. Using

these data, varios machine learning algorithms can be employed to construct a classifier. In analyzing these historical data and building a classifier to breast cancer, an important aspect is to employ an effective feature selection technique. The goal is to identify a minimal set of features that achieves high classification accuracy, as this will aid the physician in making an accurate diagnosis based on the simplest set of tests to be performed.

Breast cancer is a significant global health concern, highlighting the need for accurate diagnosis to distinguish between malignant and benign tumors. Automation of this diagnostic process is essential for effective tumor recognition. This paper compares the performance of three machine learning algorithms (Support Vector Machine, K-nearest neighbors, and Decision tree) using the Wisconsin Breast Cancer dataset, aiming to identify the most accurate classifier for breast cancer classification. The study revealed that the quadratic support vector machine achieves the highest accuracy (98.1%) with the lowest false discovery rates, highlighting its effectiveness in diagnosis. [5]

Breast cancer has emerged as a growing concern, with a notable increase in incidence among women. Late-stage diagnosis often necessitates drastic measures such as limb amputation to prevent fatality, emphasizing the critical need for accurate prediction models. This study employed various machine learning classification algorithms on the Wisconsin Breast Cancer dataset (WDBC) to predict diagnosis, assess attribute effectiveness, and optimize model performance. Naïve Bayes emerged as the most effective algorithm based on performance metrics, and comparisons were performed with existing state-of-the-art approaches on the same dataset to further validate the proposed model. [6]

Model W, also called the University of Wisconsin Breast Cancer Epidemiology Simulation Model (UWBCS), is a discrete-event microsimulation model that mimics breast cancer trends in the U.S. It tracks women's lifetimes across four key areas: natural disease progression, detection, treatment, and survival outcomes. The model is calibrated using data from the SEER registry and Wisconsin's cancer reporting system, providing detailed insights into disease stages, clinical results, costs, and quality-of-life impacts related to screening and treatment. Recent enhancements now account for variations in detection, treatment, and survival based on molecular subtypes, allowing researchers to analyze new screening approaches and subtype-specific outcomes. [7]

Breast cancer continues to be a major global health challenge, driving the need for better prediction and early detection methods to enhance patient survival. In this study, machine learning and deep learning techniques—specifically XGBoost and deep neural networks—are explored using the Wisconsin Breast Cancer Diagnostic dataset. The research assesses and contrasts the effectiveness of these models, employing performance metrics like confusion matrices, precision, and accuracy to determine the most reliable approach for improving breast cancer diagnosis.[8]

Cancer remains a major worldwide health concern and one of the leading causes of mortality, with breast cancer disproportionately affecting women. Timely detection plays a vital role in enhancing survival outcomes and minimizing healthcare expenses, though existing diagnostic approaches face challenges like high resource demands and limited accessibility. This research investigates the application of Artificial Intelligence particularly artificial neural networks, support vector machines, and Random Forest algorithms to analyze breast cancer imaging data for improved early detection. [9]

Decision Support Systems (DSSs) aid decision makers by providing intelligent solutions. This study implemented and validate clinical decision support systems using Mamdani fuzzy set theory, incorporating clustering and dynamic tables. The outcomes were compared with those of existing literature, confirming the efficacy of the suggested fuzzy systems for classifying the Wisconsin breast cancer dataset, with superior precision observed in most performance metrics compared with previous studies. [10]

The aim of this study was to compare the performance of various machine learning-based optimized feature selection approaches for breast cancer diagnosis. This is achieved by applying different feature selection techniques coupled with various classifiers to be built and compared in terms of classification accuracy. The WBCD is used to determine an effective feature selection method, and the best combination of attributes obtained from this feature selection process is used to construct a classifier for implementation in a breast cancer diagnosis system. The focus will be on comparing the classification accuracy of diagnosis systems and the performance of different classifiers. This comparison of classification results will enable an assessment of the best method for selecting features, which is invaluable for future development of machine learning-based breast cancer diagnosis systems. These techniques, including Filter Methods, Univariate Selection, Feature Importance, and Principal Component Analysis (PCA). Filter

Methods like Correlation and Variance Threshold simplify datasets by eliminating highly correlated or low-variance features. Univariate Selection, exemplified by Select-KBest and Select Percentile, ranks features based on individual significance to ensure that they are, suitable for diverse applications. Feature Importance techniques such as Random Forest and Gradient Boosting provide scores reflecting features' contributions to model performance, while L1 Regularization (Lasso) encourages sparsity in coefficients.In addition, PCA transforms features into a lower-dimensional space, thereby reducing dimensionality while retaining critical information.

3. Materials and Methods

This section presents the experimental framework, detailing the materials and methodologies employed. The discussion is organized into six key components: dataset characteristics, the proposed approach, data preprocessing procedures, feature selection methods, comparative algorithms, and evaluation metrics.

3.1 Data Description

This study utilizes breast cancer diagnostic data obtained from the UCI Machine Learning Repository. The dataset was originally compiled by Dr. William H. Wolberg at the University of Wisconsin Hospitals (Madison, WI, USA) between 1989 and 1991 [11]. The features were derived from digitized images of fine needle aspirate (FNA) samples collected from breast masses. These attributes characterize cell nuclei properties in a three-dimensional space as described in[12].

Category	**Details**
Dataset Components	
ID Number	Unique identifier for each case
Diagnosis	M = Malignant, B = Benign
Cell Nucleus Features	10 core measurements (each with mean, SE, and worst value = 30 total features):
Radius	Mean distance from center to perimeter points
Texture	Standard deviation of gray-scale values
Perimeter	-
Area	-

Smoothness	Local variation in radius lengths
Compactness	(perimeter^2 / area) - 1.0
Concavity	Severity of contour concave portions
Concave Points	Number of contour concave portions
Symmetry	-
Fractal Dimension	"Coastline approximation" - 1
Feature Computation	Each core feature has:
	- Mean value (e.g., Field 3 = Mean Radius)
	- Standard error (e.g., Field 13 = Radius SE)
	- "Worst" value (mean of 3 largest, e.g., Field 23 = Worst Radius)
Data Characteristics	
Precision	All values recorded with 4 significant digits
Missing Values	None
Class Distribution	357 Benign, 212 Malignant

3.2 Data Normalization

A data preprocessing method called normalization is used toconver the feature values in a dataset to accordance with a common scale.

3.3 Feature selection techniques

The rapid growth of computational techniques has facilitated accurate and precise breast cancer diagnosis accurate and precise breast cancer diagnosis by physicians. In general, cancer diagnosis is a two-step process that involves the identification of affected area i.e. tumor and its classification in malignant (having potential to spread) or benign (limited to original position) form. Thus, pattern recognition techniques have provided a consistent and reliable tool for malignancy classification. A pattern recognition system generally comprises two steps: the first step is feature selection, where a subset of the most relevant features is selected from the entire set of available features. The second step is classifier induction, which uses the selected features to construct a model for prediction or decision making. In the past, many data mining and machine learning researchers have employed a variety of feature selection and classification techniques to diagnose breast cancer. The most widely used dataset for these experiments was

the Wisconsin Breast Cancer Diagnostics dataset (WBCD) from the UCI repository. The proposed study selects relevant features using the following feature selection techniques.

3.3.1 Filter methods

Filter methods rank features based on a predefined scoring criterion to perform variable selection. These approaches are widely adopted due to their computational efficiency and proven effectiveness in real world scenarios.The process involves:1) Scoring features using an appropriate metric (e.g., correlation, mutual information).2) Sorting features by their relevance scores. 3) Applying a threshold to discard low-ranking variables. Since these methods operate independently of the classifier (executed as a preprocessing step), they are categorized as filter methods. A key advantage is their ability to efficiently eliminate irrelevant features, reducing dimensionality before model training.[13]

3.3.2 Univariate selection (K- Best)

We have two methods:

- Select K-Best Selects the top k features based on statistical tests like chi-squared test, ANOVA, or mutual information.
- Select Percentile The top feature are selected based on a specified percentile of the highest scores.

3.3.3 Feature importance

We have two methods:

- Tree-based methods (e.g., Random Forest, Gradient Boosting) These models provide feature importance scores based on the contribution of each feature to the model's performance.
- L1 Regularization (Lasso) This condition encourages sparsity by penalizing the absolute magnitude of the coefficients, there by allowing some coefficients to become zero and effectively selecting features.

3.3.4 Principal component analysis

The PCA technique performs dimensionality reduction through orthogonal transformation, producing a set of linearly uncorrelated variables that account for the greatest variance in the data, which are subsequently employed as features.

3.4 Models used in the study

3.4.1 Naïve bayes

Are a group of straightforward probabilistic models that utilize Bayes' theorem under the key assumption that all features are conditionally independent. Despite this simplifying (and often unrealistic) assumption, they remain highly efficient, scaling linearly with the number of features in the dataset. The mathematical foundation of Naïve Bayes relies on Bayes' theorem: [14]

$$P(C/X) = \frac{P(X/C).P(C)}{P(X)}$$

Where:

P(C|X) = posterior probability of class C given feature vector X.

P(X|C) = likelihood of feature vector X in class C.

P (C) = prior probability of class C.

P(X) = prior probability of feature vector X.

The "naïve" assumption states that features contribute independently to the class probability, allowing the likelihood P(X|C) to be computed as the product of individual feature probabilities:

$$P(X/C) = \prod_{i=1}^{n} P(x_i/C)$$

Where x_i is the value of feature in X. This simplification enables fast training and prediction, making Naïve Bayes effective for high-dimensional data despite its strong independence assumption.[15]

3.4.2 Logistic regression

Logistic regression is a binary classification method that models the probability of an outcome using a linear combination of input features. As a specialized generalized linear model, it employs the logistic function as its link function logit $\left(\frac{1}{1+e^{-t}}\right)$ or logistic function in the equation

$$LR = \frac{1}{1 + e^{-(Y=\beta_0+\beta_1 X_1+\cdots\beta_n X_n)}}$$

3.4.3 KNN

The k-nearest neighbors (k-NN) method, first proposed by [16], is a nonparametric approach that classifies a query point by finding its k most similar instances in the training data and determining the output label through majority voting [17]. The training set $T = \{(y_i, c_i)\}_{i=1}^{N}$ consists of N instances from M classes, where (y_i)represents a training instance in a D-dimensional space, and (c_i) is the corresponding class label for y_i , with $C = \{c_1, c_2, \dots\dots, c_M\}$. An instance $(x, \tilde{c}) \notin T$ has an unknown class label $\tilde{c}$.

3.4.4 SVC

Support Vector Machines (SVMs) are powerful ML algorithms, especially effective for classification. The ideal decision boundary is determined by maximizing the margin the shortest distance between the hyperplane and the closest training samples (support vectors) of each class. The margin *M* is defined as:

$$M = \frac{2}{\|w\|}$$

Where ‖w‖ is the Euclidean norm of the weight vector. To maximize the margin, we need to minimize ‖w‖ subject to the constraints imposed by the data points.

3.4.5 Random forest

The proposed model is a supervised machine learning classification ensemble method based on multiple decision trees. This ensemble model builds several decision trees that, are combined to enhance performance. The bootsatrap aggregating technique is primarily used, bootstrap aggregating technique is primarily used, for tree learning. Given a dataset, $X = \{x_1, x_2, \dots\dots, x_n\}$ represents the input vectors, and $Y\{y_1, y_2, \dots\dots, y_n\}$ is the response variable, with bagging repeated from b=1 to B. The final prediction for a new, unseen sample $\acute{x}$ is obtained by averaging the predictions of all individual decision trees.[18]

$$j = \frac{1}{B}\sum_{b=1}^{B} fb(\acute{x})$$

The variability in predictions across these trees, as defined by the following equation, is quantified using standard deviation to assess prediction uncertainty.

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(fb(\acute{x}) - \acute{f})^2}{B-1}}$$

3.4.6 XG Boost

This method builds decision trees sequential. Prior to each independent variable being introduced into a decision tree to predict outcomes, it is assigned a weight. Variables that were incorrectly predicted by the first tree received increased weighting before being fed into the subsequent tree. By combining these different classifiers, a robust and accurate model is obtained. Weak base learners, which have a high bias and slightly better predictive power than random guessing, are combined using the boosting approach to create strong learners, thereby reducing both bias and variance. This technique is enabled by parallel and distributed computing. [19]

3.4.7 Ada Boost

AdaBoost is a widely-used machine learning technique effective for selecting relevant features in the context of breast cancer diagnosis. It works by creating a highly accurate classifier by combining multiple weak classifiers. The ada boost algorithm is a machine learning approach where multiple weak models are combined to form a stronger, more accurate model, which has been widely applied in various fields. In the context of feature selection for breast cancer diagnosis, ada boost has demonstrated promising results in improving classification performance. One of the key advantages of ada boost is its ability to handle high-dimensional data and noisy features, which are common in breast cancer datasets. It achieves this by repeatedly training weak classifiers on various data subsets and increasing the emphasis on incorrectly predicted samples in every iteration. This iterative process allows ada boost to focus on the most informative features for breast cancer diagnosis [20]. The final classifier is a weighted sum of the weak leaners:[21]

$$H(x) = sign\left(\sum_{m=1}^{M} \propto_m . h_m(x)\right)$$

Where:

$H(x)$: The final strong classifier after combining all weak learners

$h_m(x)$: $The\ weak\ learner\ in\ iteration\ m$

$\propto_m$: The weight of the weak learner in the final classifier.

3.4.8 Neural network

This is an overseen machine learning classification algorithm encourged by the workings of the human brain. The proposed model comprises three main components: an input layer (x_i)hidden layers $X = [x_1, x_2, \dots\dots, x_3]$ with n features and an output layer (y_i).

$$y_i = \sigma\left(\sum_{j=1}^{h} \omega_{ij}.a_j + b_i\right)$$

Where, ω_{kj} are the weights connecting the hidden neuron j to the output neuron i, and $b_i is\ the\ bias\ term\ for\ the\ output\ neuron$

Any artificial neural network (ANN) must have at least one hidden layer [22]. The effectiveness of a neural network depends on its accociatet weights. The activation function is crucial in neural networks for producing the final result by incorporating a bias value [23].

3.4.9 Gradient Boosting

The gradient boosting algorithm is widely utilized across various domains because of its effectiveness in managing large, high-dimensional datasets while maintaining strong predictive performance. In medical diagnostics, particularly for breast cancer analysis, it has proven valuable. Researchers have explored multiple machine learning techniques such as genetic algorithms and particle swarm optimization to identify the most discriminative features for breast cancer detection. Given the importance of accurate diagnosis, evaluating the performance of these optimized feature selection methods is crucial for enhancing both precision and computational efficiency in breast cancer screening**.** [24]. The model is given by [25]:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \nu.h_m(x)$$

$F_0(x)$: The initial model, typically a constant prediction.

$h_m(x)$: The weak learner fitted to the residuals.

ν: The learning rate, which controls the step size in the gradient descent process.

3.5 Performance measures

The classification evaluation metrics utilized include accuracy, precision, recall, specificity, and the F-measure. These metrics are derived from the elements of the confusion matrix, which provides information about predicted versus actual values. The performance metrics are expressed as follows:

Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$

Precision $\frac{TP}{TP+FP}$

Recall $\frac{TP}{TP+FN}$

F-Score $\frac{2*precision*Recall}{Precision+Recall}$

Specificity $\frac{TN}{FP+TN}$

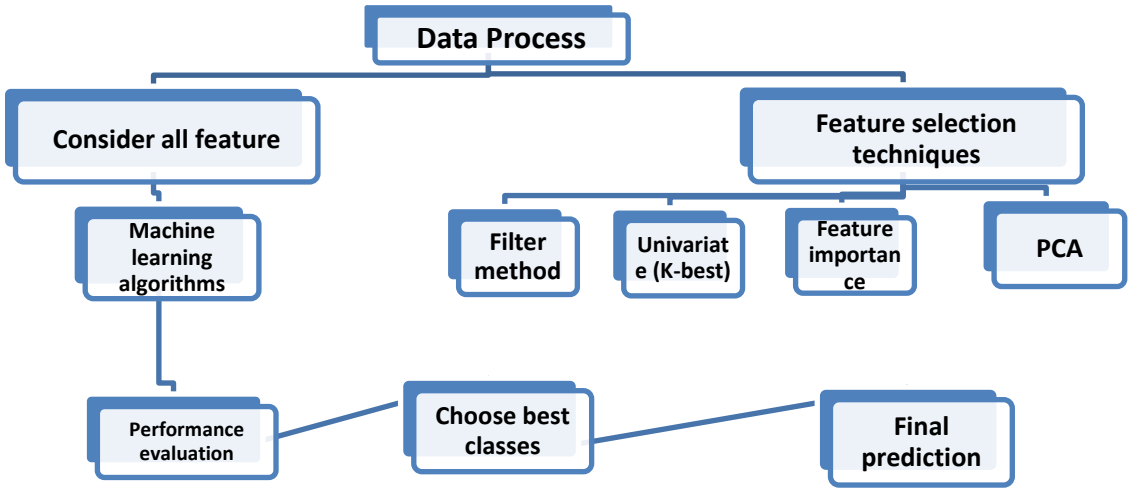

Fig.1 Experiment proposed methodology

Table (1) Performance comparison of classifiers with different feature subsets.

	Models	Accuracy	Precision	Recall	F-Score
Classifier performance with filter methods	**Naive Bayes**	94.74	93.75	88.24	90.91
	Logistic Regression	98.25	100.00	94.12	96.97
	KNN	94.74	88.89	94.12	91.43
	SVC	98.25	100.00	94.12	96.97
	Random Forest	96.49	94.12	94.12	94.12
	XGBoost	94.74	93.75	88.24	90.91
	AdaBoost	96.49	94.12	94.12	94.12
	Neural Network	98.25	100.00	94.12	96.97
	Gradient Boosting	94.74	93.75	88.24	90.91
Classifier performance with Univariate Selection (Select K-Best)	**Naive Bayes**	98.25	100.00	94.12	96.97
	Logistic Regression	**100.00**	**100.00**	**100.00**	**100.00**
	KNN	94.74	88.89	94.12	91.43
	SVC	98.25	100.00	94.12	96.97
	Random Forest	**100.00**	**100.00**	**100.00**	**100.00**
	XGBoost	98.25	94.44	100.00	97.14
	AdaBoost	96.49	100.00	88.24	93.75
	Neural Network	96.49	89.47	100.00	94.44
	Gradient Boosting	98.25	94.44	100.00	97.14
Classifier performance with feature importance	**Naive Bayes**	98.25	100.00	94.12	96.97
	Logistic Regression	**100.00**	**100.00**	**100.00**	**100.00**
	KNN	94.74	88.89	94.12	91.43
	SVC	98.25	100.00	94.12	96.97
	Random Forest	96.49	94.12	94.12	94.12
	XGBoost	96.49	94.12	94.12	94.12
	AdaBoost	92.98	93.33	82.35	87.50
	Neural Network	**100.00**	**100.00**	**100.00**	**100.00**
	Gradient Boosting	96.49	94.12	94.12	94.12
Classifier performance with PCA	**Naive Bayes**	94.74	93.75	88.24	90.91
	Logistic Regression	98.25	100.00	94.12	96.97
	KNN	96.49	94.12	94.12	94.12
	SVC	98.25	100.00	94.12	96.97
	Random Forest	98.25	100.00	94.12	96.97
	XGBoost	96.49	100.00	88.24	93.75
	AdaBoost	96.49	94.12	94.12	94.12
	Neural Network	98.25	100.00	94.12	96.97
	Gradient Boosting	98.25	100.00	94.12	96.97

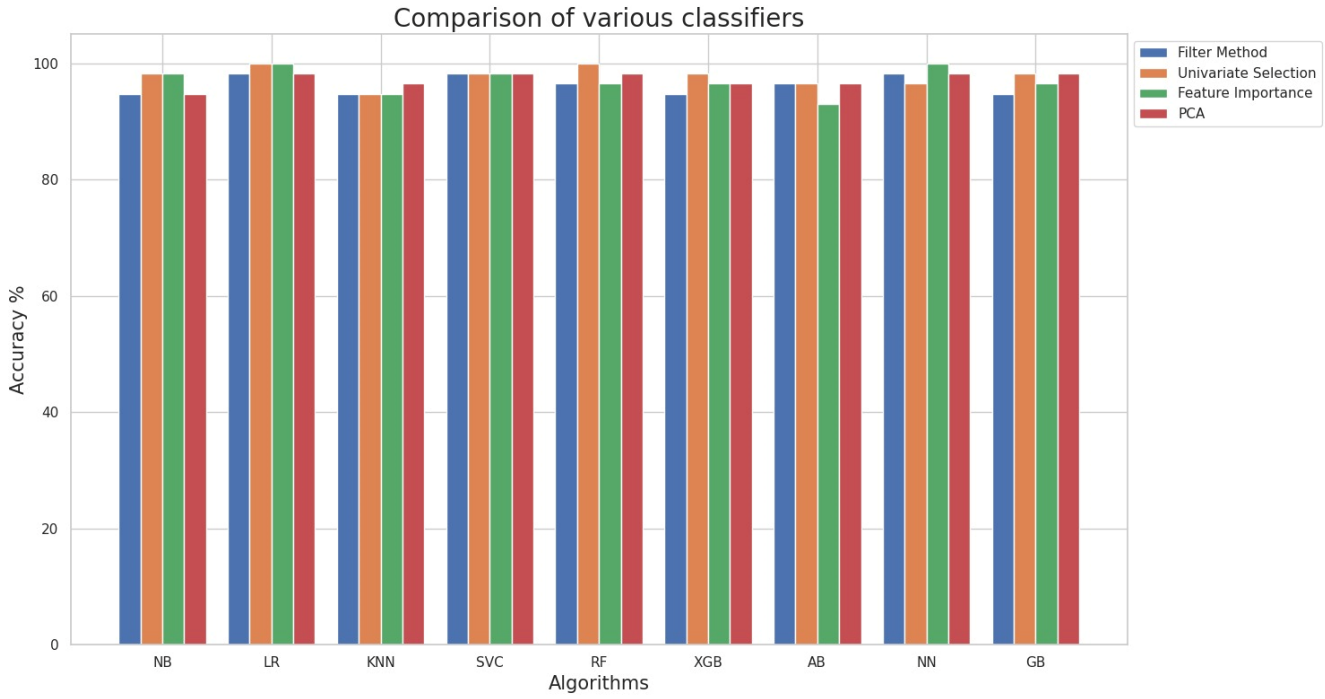

Fig.2 Overall performance comparison of various classifiers

4. Conclusion

In conclusion, the choice of classifier and feature selection method has a substantial impact on the performance of breast cancer diagnosis models. Logistic Regression and Neural Networks, when combined with effective feature selection methods like Univariate Selection and Feature Importance, the two classifiers consistently achieve the highest accuracy 100%. SVC and Random Forest also show strong performance. Ensemble methods like Gradient Boosting and XGBoost provide robust results, but their performance does not surpass that of the simpler models like Logistic Regression.

Acknowledgements

Acknowledgment the authors are highly thankful to the editor and reviewers for their kind suggestions and critical comments on improving the quality of thepaper.

References

1. Consoli S, Recupero DR, Petkovic M (2019) Data science for healthcare (methiodides and applicationss . Springer International Publishing, Berlin

2. Mohan S, Thirumalai C, Srivastava G (2019) Effective heart disease prediction using hybrid machine learning techniques. IEEE Access 7:81542–81554

3. Mishra, S , Tripathi AR (2020). Literature review on business prototypes for digital platforms. J Innov Entrepreneurship 9(1):1–19.

4. Kaur, H., & Arora, A. (2020). *A review of machine learning techniques for diagnosis of breast cancer using imaging data*. Computers in Biology and Medicine, 123, 103886.

5. Obaid OI, Mohammed MA, Ghani MKA, Mostafa A, Taha F (2018). Evaluating the Performance of Machine Learning Techniques in the Classification of Wisconsin Breast Cancer. International Journal of Engineering & Technology

6. Ahmed MT, Imtiaz MN, Karmakar A (2020). Analysis of the Wisconsin Breast Cancer original dataset using data mining and machine learning algorithms for breast cancer prediction, Journal of Science Technology and Environment Informatics.

7. Oğuzhan Alagöz; Meltem Ergün; Mücahit Çevik; Brian L. Sprague; Dennis G. Fryback; Ronald E. Gangnon; John M. Hampton; Natasha K. Stout; Amy Trentham-Dietz, (2018). The University of Wisconsin Breast Cancer Epidemiology Simulation Model: An Update, Medical Decision Making, Vol.38, NO. 1.

8. Bakar WAWA, Zuhairi MA, Man M, Jusoh JA, Josdi NLN (2022). Deep learning algorithm vs XGBoost using Wisconsin breast cancer diagnosis, Third International Conference on Computer Science and Communication Technology.

9. Rovshenov A, Peker A. (2022),Performance Comparison of Different Machine Learning Techniques for Early Prediction of Breast Cancer using Wisconsin Breast Cancer Dataset, 3rd International Informatics and Software Engineering Conference.

10. Hernandez-Julio YF, Diaz-Pertuz LA, Prieto-Guevara MJ, Barrios-Barrios MA and Nieto-Bernal W. (2023) Intelligent Fuzzy System to Predict the Wisconsin Breast Cancer Dataset. *Int. J. Environ. Res. Public Health* **2023**, *20*(6), 5103.

11. Bennett KP and Mangasarian OL (1992),Robust Linear Programming Discrimination of Two Linearly Inseparable Sets. Optimization Methods and Software.

12. G Chandrashekar, F Sahin (2014). A survey on feature selection methods, Computers & electrical engineering.

13. De Souza GFM, Melani AHDA, Michalski MADC. , (2022), Reliability Analysis and Asset Management of Engineering Systems

14. Dharyll Prince M. Abellana, Demelo M. Lao. (2023), A new univariate feature selection algorithm based on the best–worst multi-attribute decision-making method, Decision Analytics Journal, Vol.7

15. Cover. T , Hart. P, Nearest neighbor pattern classification, IEEE Trans. Inform. Theory 13 (1) (1967) 21–27.

16. Pan. Z, Wang. Y, Pan. Y, A new locally adaptive k-nearest neighbor algorithm based on discrimination class, Knowl.-Based Syst. (2020) 106-185

17. Sharma1.A, Kumar.PM (2022). Performance analysis of machine learning-based optimized feature selection approaches for breast cancer diagnosis, Int. J. Inf. Technol, 14(4):1949–1960

18. Chen, T., & Gastrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm signede international conference on knowledge discovery and data Mining.

19. Sharma. A, Mishra. PK. (2022) International Journal of Information Technology, Springer. Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis. academia.edu

20. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. Chapter 10 covers boosting and AdaBoost extensively.

21. Shailaja K, Seetharamulu B, Jabbar MA (2018) Machine learning in healthcare: a review. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, pp.910-914

22. Ketu S, Mishra PK (2020) A hybrid deep learning model for COVID-19 prediction and current status of clinical trials worldwide. Comput Mater Contin 66(2).

23. Omoteina.TO , Oceola. DO , Dada. EG, (2023). Healthcare Analytics, Elsevier. A light gradient-boosting machine algorithm with tree-structured parzen estimator for breast cancer diagnosis. sciencedirect.com

24. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine.Annals of Statistics, 29(5), 1189-1232.

25. Sharma, A., Mishra PK (2020). State of-the-art performance metrics and future directions for data science algorithms. J Sci Res 64(2):221–238