OPEN ACCESS
INTERNATIONAL JOURNAL OF IMMINENT SCIENCE & TECHNOLOGY
ISSN.3006-3116-Lead Journal Publisher

INTERNATIONAL JOURNAL OF IMMINENT SCIENCE & TECHNOLOGY
Journal homepage: www.ijisnt.com

ISSN.3006-3116

IJIST 2023 Vol-2 Issue-1

Comparative Performance Analysis of Ensemble Models for Breast Cancer Classification

[1]Nazia Nuzhat, [2]Faisal Islam, [3]Abu Sayed Sikder, [4]Narayan Ranjan Chakraborty
[1]nazia.nuzhat24@gmail.com, [2]Faisalislam1610@gmail.com, [3] PM21496@student.uniten.edu.my,
[4]narayan@daffodilvarsity.edu.bd
[1]Daffodil International University
[2]National University
[3] Universiti Tenaga Nasional (UNITEN)
[4]Daffodil International University

Abstract

Breast cancer remains a critical health issue worldwide, with early and accurate diagnosis playing a pivotal role in improving patient outcomes. This study presents a machine learning-based approach using ensemble methods to enhance breast cancer classification, focusing on distinguishing malignant from benign cases. Utilizing the Wisconsin Breast Cancer (WBC) dataset, six algorithms—K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Random Forest, Gradient Boosting, and XGBoost—were combined using Hard Voting to create an optimized ensemble model. The ensemble model achieved a high classification accuracy of 97.6%, with notable improvements in precision (98.2%), recall (96.4%), and F1 score (97.3%), outperforming individual models. These results underscore the effectiveness of ensemble techniques in enhancing prediction reliability and suggest their potential for aiding in early breast cancer detection. Key findings highlight that ensemble models significantly improve performance by integrating complementary strengths of different algorithms, offering a robust tool for clinical decision-making. Future research could extend these findings by incorporating larger, more diverse datasets and exploring deep learning integrations for further accuracy gains.

Keywords: *Breast Cancer Classification, Machine Learning (ML), Ensemble Learning, Data Mining in Healthcare, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, Random Forest.*

1. Introduction

Breast cancer remains a significant health issue worldwide, particularly among women. In 2022, approximately 2.3 million new cases of breast cancer were reported globally, according to the World Cancer Research Fund (WCRF) [1]. Despite advances in treatment, many deaths still occur due to a lack of awareness and delayed detection. The World Health Organization (WHO) estimates that breast cancer accounted for 670,000 deaths worldwide in 2022 [2]. While commonly associated with women, breast cancer also affects men; in the U.S. alone, an estimated 42,250 women and 530 men are expected to die from the disease in 2024 [3].

Early detection and accurate diagnosis are essential for effective treatment, as timely intervention can significantly improve patient outcomes. With the rapid progress in medical science, breast cancer is treatable if detected early, but delayed diagnosis can lead to life-threatening complications [4]. Several diagnostic techniques are available, including imaging, biopsy, and molecular analysis, which help determine whether a cell is cancerous, guiding further treatment decisions.

Data mining discovers important insights from large datasets, utilizing techniques like machine learning(ML) and neural networks to aid in cancer diagnosis and prediction [5]. For many years, machine learning has become increasingly important in the early detection and diagnosis of various cancers, including breast, lung, prostate, brain cancers and so on [6]. A variety of ML approaches such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, Decision Trees, Logistic Regression, Random Forest, AdaBoost, XGBoost, and Neural Networks, have been widely employed in breast cancer classification [7-15]. These models leverage patient data, such as genetic profiles and imaging results, to effectively classify and predict the presence of breast cancer [16].

Moreover, researchers have increasingly turned to ensemble techniques to further boost the performance of these methods [17, 18]. Early research, such as [19], proposed a method that significantly boosts predictive accuracy by integrating multiple learning algorithms through a meta-learner. This approach has had a lasting impact on the evolution of ensemble learning techniques within the machine learning field. Furthermore, ensemble methods significantly enhance classification tasks by combining the strengths of various algorithms, demonstrating their effectiveness in fields like medical diagnostics [20].

The aim of this research is to develop a model capable of predicting breast cancer using 31 features. In this study, we compared several ensemble models and proposed a new model that outperforms others not only in terms of accuracy but also with higher precision, recall, and F1 scores. To build this model, we utilized six algorithms: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Random Forest, Gradient Boosting, and XGBoost. These algorithms were then combined using multiple ensemble techniques, with Hard Voting playing a crucial role in the final outcome. Through this process, we successfully identified an optimal ensemble model capable of classifying breast cancer data into malignant and benign categories.

In the following sections, we discuss each step of our work in detail, with a particular emphasis on improving precision.

2. Related Study

In recent years, the use of machine learning (ML) classifiers has expanded rapidly within various domains. Utilizing a range of algorithms on medical datasets has led to better results in diagnosing diseases, with researchers demonstrating how diverse methodologies can improve healthcare predictions [21]. This section will explore notable research efforts that utilized machine learning to analyze breast cancer datasets.

In the article [22], the authors presented an ensemble approach for breast cancer classification using a bagging technique that combined Decision Trees and KNN. Impressively, this model achieved 100% accuracy, with performance evaluated using accuracy, confusion matrices, and classification reports. Notably, they relied on the Coimbra dataset from UCI, splitting the data with 90% for training

DOI: https://doi.org/10.70774/ijist.v2i2.22

and 10% for testing, demonstrating the robustness of the approach. Similarly, in [23], researchers proposed a two-layer nested ensemble model, named SV-Naïve Bayes-3-MetaClassifier, which demonstrated remarkable performance with an accuracy of 98.07%. The model's effectiveness was highlighted through metrics such as precision, recall, F1 score, and ROC curve, with k-fold cross-validation ensuring reliable results. Furthermore, in [24], the authors compared various classification algorithms and determined that the AdaBoost ensemble method stood out, achieving an impressive accuracy of 98.77% using 10-fold cross-validation. This demonstrated the method's strong predictive capability in breast cancer diagnosis.

Moving forward, an ensemble model for breast cancer detection is proposed in [25], combining decision trees, SVM, and KNN, achieving 78% accuracy—better than individual classifiers and models like Naïve Bayes, ANN, and logistic regression. Additionally, [26] showed the Gradient Boosting model as the top performer in breast cancer classification, with an impressive F1 score of 96.77%. Meanwhile, [27] highlighted XGBoost as the top performer among five machine learning models for breast cancer classification, with 95.42% accuracy, 98.5% sensitivity, 97.5% specificity, and 99% F1 score. The authors used data cleaning to handle missing values and oversampled the data but their 80-30 train-test split is unfeasible in machine learning. Innovatively, [28] employed snapshot ensembling, achieving 86.6% accuracy, opening new avenues for breast cancer classification methods. Moreover, in [29], the authors evaluated their ensemble model against five individual base models and achieved an impressive accuracy of 98.14%. This outcome highlighted the enhanced performance of the ensemble approach over the individual models.

In [30], researchers utilized 60% of the WBCD dataset for training and 40% for testing, applying 5-fold cross-validation to ensure model reliability. They identified Extreme Gradient Boosting and Extremely Randomized Trees as the top-performing classifiers. However, they acknowledged the absence of feature selection in their model, identifying it as a potential area for future work. In a recent study [31], researchers introduced the ELRL-E approach, which combines four classifiers to achieve an accuracy of 97.6%. The model's evaluation was conducted using a dataset split, with 70% allocated for training and 30% for testing. Finally, [32] demonstrated the effectiveness of a stacked ensemble learning framework to classify breast cancer, where they recorded 95% accuracy on the BreakHis dataset and an even higher 99% accuracy on the WBCD dataset.

3. Methodology Framework

In this study, we set out to develop an efficient and precise breast cancer classification model using machine learning techniques, with the primary goal of distinguishing between malignant and benign cancer cells by leveraging ensemble modeling to optimize performance. Our approach followed a systematic process, beginning with data collection and preprocessing, and progressing through model training and evaluation, as illustrated in Figure-1. To enhance the dataset and improve classification accuracy, we employed various techniques, including feature selection, normalization, and data balancing. Multiple ensemble models were used to capture underlying patterns and ensure precise predictions, with a focus on both accuracy and precision.

This section provides a detailed overview of the key steps in our methodology, covering dataset preparation, model design, and validation strategies.

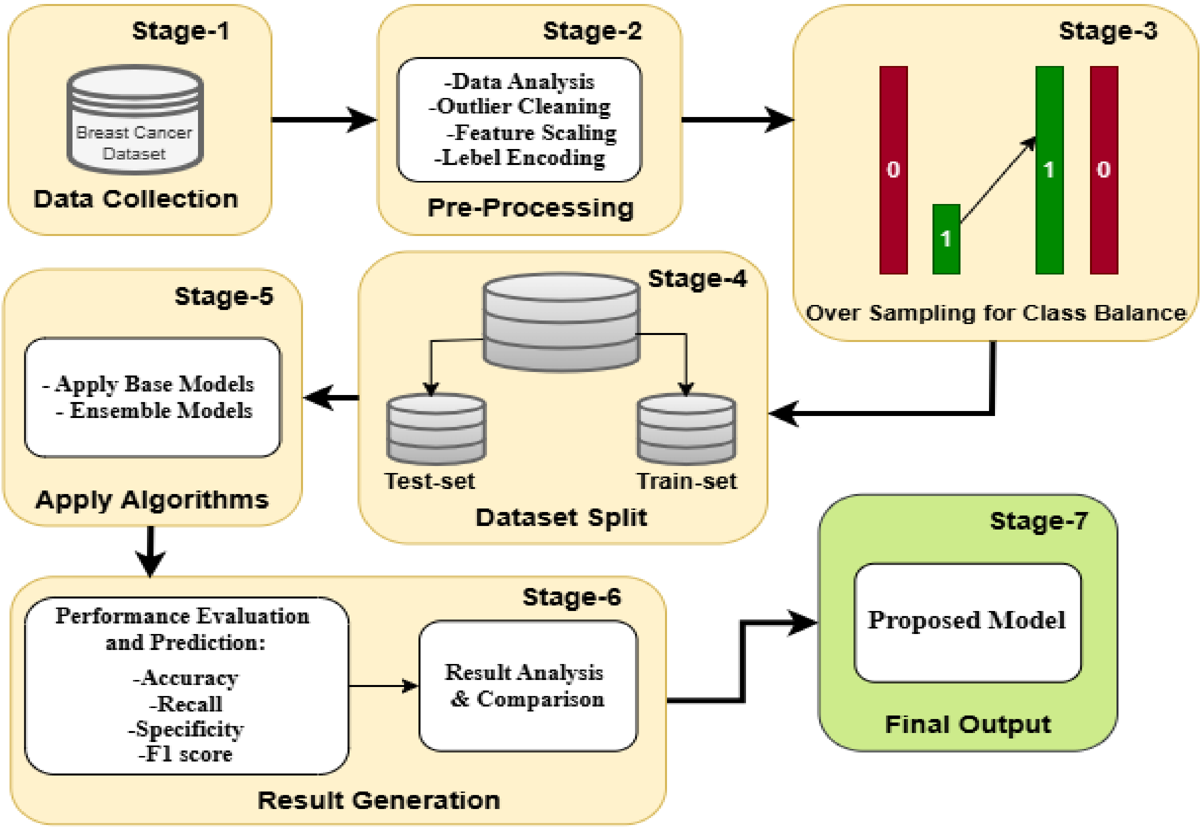

Figure-1: Work Stages

3.1 Dataset and Data Preparation

3.1.1 Dataset Collection:

In the initial phase, we selected and acquired the Breast Cancer dataset from the University of Wisconsin (WBC) [33] and collected from kaggle, which includes 31 attributes and 569 instances—212 malignant and 357 benign. After data collection, we proceeded with a series of preprocessing steps to ensure the dataset was ready for further analysis.

3.1.2 Preprocessing steps:

To prepare the dataset for analysis, the first step is to identify and remove any missing values. This is essential, as missing data can introduce inaccuracies and biases in model predictions. Following this, we focused on detecting and eliminating outliers—data points that significantly differ from the rest. Outliers can skew model training, resulting in misleading outcomes and diminished performance. Once the data is prepared, the next critical step is feature scaling, which ensures that all features in the breast cancer dataset have an equal influence on the model. Effective machine learning models depend on proper scaling, as the ability to adjust feature values can significantly impact performance. Techniques such as normalization and standardization are commonly applied to this process. Ultimately, effective feature scaling leads to a more reliable and accurate model. After performing feature scaling, we labeled our data, assigning 0 to benign cases and

INTERNATIONAL JOURNAL OF IMMINENT SCIENCE & TECHNOLOGY
Journal homepage: www.ijisnt.com
ISSN.3006-3116
IJIST 2023 Vol-2 Issue-1

1 to malignant cases. Following this labeling process, we encoded the dataset to prepare it for model training.

3.1.3 Oversampling and Dataset split:

In this stage, we observed a significant imbalance between the two classes in our dataset: benign instances totaled 300, while malignant instances were only 98. To address this disparity, we employed an oversampling technique to enhance the representation of the malignant cases in our dataset. To prepare our dataset for analysis, we divided it into two parts: a training set and a testing set. We utilized a 70:30 ratio for this split, allocating 70% of the data for training the model and 30% for testing its performance.

To enhance our analysis of the Wisconsin Breast Cancer (WBC) dataset, it's important to look at how the features are related to one another. In Figure-2, we presented the correlation matrix, which effectively shows the strong positive correlations between various features. This visual representation not only highlights the mathematical relationships among the features but also plays a vital role in distinguishing between benign and malignant cases, enhancing our ability to spot significant patterns. By using a correlation matrix, we can identify key feature interactions that ultimately improve the model's accuracy and effectiveness. Such data visualization techniques are essential for extracting valuable insights and making informed decisions in our analysis [34] [35].

3.1.4 Data visualization:

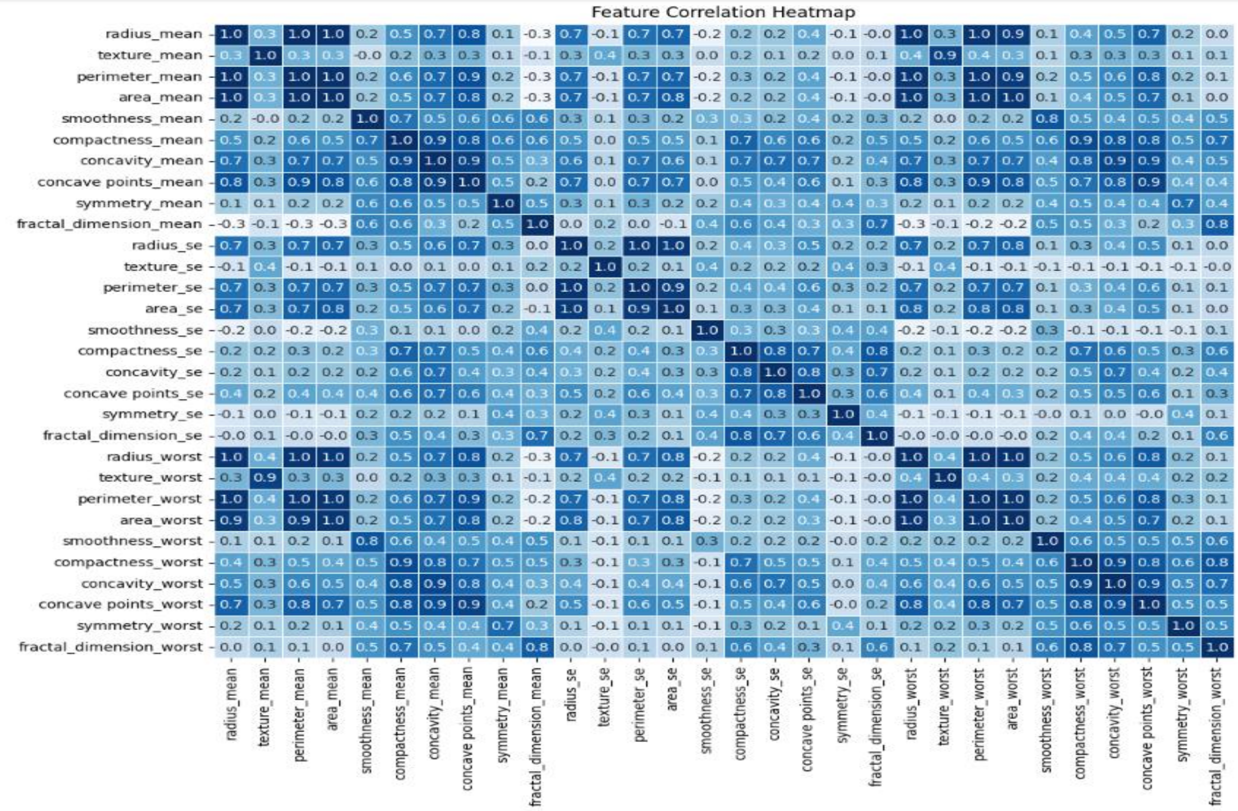

Figure-2: Correlation Matrix

3.2 Algorithms

3.2.1 Support Vector Machine (SVM):

SVM is a supervised machine learning algorithm that is frequently used for tasks involving regression and classification. It works by determining which hyperplane in a high-dimensional feature space is best for dividing data points of various classes. To improve the model's resilience and capacity for generalization, SVM's main objective is to maximize the margin between the hyperplane and the closest data points, or support vectors. SVM uses kernel functions to map the data into higher dimensions when the data is not linearly separable, allowing for the successful separation of classes. The equation of the hyperplane can be represented as, $w^T x + b = 0$ where w^T is the weight vector and b is the bias term [36].

3.2.2 K-Nearest Neighbors:

A straightforward, non-parametric supervised learning algorithm for classification and regression problems is K-Nearest Neighbors (KNN). Based on the similarity principle, the algorithm finds the kkk data points that are closest to a given input point in the feature space. These neighbors are then categorized by the algorithm based on the majority class. Euclidean distance is commonly used to measure the distance between points, and the k nearest neighbors vote to determine the final classification [37].

The general equation to calculate the distance d between two points x_i and x_j in an n-dimensional space is:

$$d(x_i, x_j) = \sum_{p=1}^{n} (x_{ip} - x_{jp})^2$$

Where x_i and x_j are two points in the feature space, p denotes each feature dimension.

3.2.3 Logistic Regression:

For binary classification tasks, logistic regression (LR) is a popular statistical technique that predicts the likelihood that a given input falls into a particular category. Using the logistic function, logistic regression transforms its output into a probability score between 0 and 1, in contrast to linear regression, which forecasts continuous outcomes. Using the logistic function, which is represented by the following equation, the model creates a relationship between the input features and the binary outcome:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}$$

P stands for the positive class probability, Y for the binary outcome, X for the input features, and β for the dataset-estimated coefficients in this equation [6].

3.2.4 Random Forest:

Random Forest (RF) is an ensemble learning technique that uses the strength of several decision trees to increase prediction accuracy and manage overfitting. It is applicable to both regression and classification tasks. It works by building a large number of decision trees during training, each of which is constructed using a random subset of features and data. By combining the predictions of every single tree, usually by majority voting for classification or averaging for regression, the final prediction is produced. The Random Forest algorithm can be represented by the equation:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} h_i(x)$$

iJIST OPEN ACCESS
INTERNATIONAL JOURNAL OF IMMINENT SCIENCE & TECHNOLOGY
ISSN.3006-3116-Lead Journal Publisher
INTERNATIONAL JOURNAL OF IMMINENT SCIENCE & TECHNOLOGY
Journal homepage: www.ijisnt.com
ISSN.3006-3116
iJIST 2023 Vol-2 Issue-1

In this equation, $\hat{□}$ is the predicted output, T is the total number of trees in the forest, and $h_{□}(□)$ is the prediction made by the t-th tree for the input features X [38].

3.2.5 Naïve Bayes:

Bayes' theorem is the foundation of the probabilistic classification algorithm known as Naïve Bayes (NB), which makes the strong (naïve) assumption that features are independent of one another. For tasks involving text classification and other applications where the independence assumption might roughly hold, it works especially well. The algorithm predicts the class $□_{□}$ for a given input **X=($□_{□}$,$□_{□}$,...,$□_{□}$)** by computing the posterior probability using the following equation:

$$□(□_{□}|□) = \frac{□(□_{□})\prod_{□=1}^{□} □(□_{□}|□_{□})}{□(□)}$$

Where P($□_{□}$ | □) is the posterior probability of the class $□_{□}$ given the features □, $□(□_{□})$ is the prior probability of the class, $□(□_{□}|□_{□})$ is the likelihood of feature $□_{□}$ given class $□_{□}$, and P(X) is the evidence (constant for all classes). Despite its simplicity, Naïve Bayes performs well in many real-world applications [39].

3.2.6 Gradient Boosting Machine:

The ensemble learning technique known as Gradient Boosting Machine (GBM) builds models one after the other, with each new model aiming to fix the mistakes of the ones that came before it. A strong predictive model is created by combining the outputs of multiple weak learners, usually decision trees. By including additional trees that forecast the model's residual errors, the main idea is to minimize a differentiable loss function. The prediction at each stage $□_{□}(□)$ is updated as:

$$□_{□+1}(□) = □_{□}(□) + \eta \, . \, h_{□}(□)$$

Where $□_{□}(□)$ is the current model, $h_{□}(□)$ is the new weak learner (typically a decision tree), and η\etaη is the learning rate, which controls the contribution of the new learner. This process is repeated until a desired level of accuracy is achieved or a stopping criterion is met, making GBM a powerful and flexible method for both regression and classification tasks [40].

3.2.7 XGboost:

Extreme Gradient Boosting, or XGboost, is a version of gradient boosting that has been optimized for speed and performance. With an emphasis on both classification and regression issues, it employs a group of decision trees to increase accuracy. XGBoost employs regularization strategies to avoid overfitting and second-order derivatives of the loss function to give the model more precise updates.

3.3 Model Evaluation Metrics:

3.3.1 Confusion Matrix:

A confusion matrix is a table that compares the results of predictions with the actual results in order to assess how well a classification model is performing. It provides a clear breakdown of the model's correct and incorrect predictions for each class in a classification problem.

- The confusion matrix typically contains four key components for binary classification:
- **True Positives (TP):** The number of instances that were correctly predicted as positive.
- **False Positives (FP):** The number of instances that were incorrectly predicted as positive (i.e.,

actual negatives that were predicted as positives).

- **True Negatives (TN):** The number of instances that were correctly predicted as negative.
- **False Negatives (FN):** The number of instances that were incorrectly predicted as negative (i.e., actual positives that were predicted as negatives).

3.3.2 Classification Metrics:

One tool for assessing a classification model's performance is a classification matrix, also known as the confusion matrix. However, if you're searching for confusion matrix-derived metrics that offer a more thorough understanding of a model's performance, the following classification metrics are included:

3.3.3 Accuracy:

The ratio of correctly predicted instances to the total instances. It measures the overall effectiveness of the model.

$$\square\square\square\square\square\square\square\square = \frac{\square\square + \square\square}{\square\square + \square\square + \square\square + \square\square}$$

3.3.4 Precision:

The ratio of correctly predicted positive instances to the total predicted positive instances. It measures the accuracy of the positive predictions.

$$\square\square\square\square\square\square\square\square = \frac{\square\square}{\square\square + \square\square}$$

3.3.5 Recall:

The ratio of correctly predicted positive instances to the actual positive instances. It measures the model's ability to find all the relevant cases (true positives).

$$\square\square\square\square\square\square = \frac{\square\square}{\square\square + \square\square}$$

3.3.6 F1-Score:

The harmonic mean of precision and recall. It provides a balance between the two metrics, especially in cases where there is an uneven class distribution.

$$\square 1\ \square\square\square\square\square = 2 * \frac{\square\square\square\square\square\square\square\square\square * \square\square\square\square\square\square}{\square\square\square\square\square\square\square\square\square + \square\square\square\square\square\square}$$

4. Findings

This section provides an in-depth analysis of the performance of each model, including both individual methods and ensemble combinations applied to our dataset using a 70:30 train-test split ratio. Each model was analyzed using four primary metrics: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive view of each model's strengths and weaknesses of performance, particularly in distinguishing between classes effectively.

We present the results for the base models in Table 4.1, followed by the ensemble models in Table 4.2, which show the differences and improvements among the model performances. The base model results are initially discussed, followed by the performance of different ensemble combinations.

4.1 Base Models:

Table 4.1 illustrates the the performance metrics for six base models, which are K-nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Gradient Boosting Machine (GBM), and XGBoost. Among all of these models, Gradient Boosting Machine (GBM) achieved the highest accuracy at 98.33%, with a perfect recall of 100%, resulting in an F1-score of 98.55%. This performance indicates that GBM was particularly effective in identifying positive cases without any false negatives.

XGBoost demonstrated strong performance, achieving an accuracy of 97.22% and an F1-score of 97.61%, showcasing a balanced approach between precision and recall. Similarly, Logistic Regression (LR) performed commendably, achieving 97.22% across accuracy, recall, F1-score, and a slightly higher precision of 97.23%. Furthermore, both SVM and Random Forest (RF) also performed reliably, with accuracies close to 96%, suggesting that they too are effective classifiers within this context. Figure-3 shows the performance of different machine learning (ML) models.

Table 4.1: Performance of base models

ML Algorithms	Accuracy	Precision	Recall	F1-score
K-nearest neighbor (KNN):	96.11%	96.15%	96.11%	96.10%
Support Vector Machine (SVM):	96.67%	96.67%	96.67%	96.67%
Logistic Regression (LR):	97.22%	97.23%	97.22%	97.22%
Random Forest (RF):	96.67%	96.85%	96.67%	96.65%
Gradient Boosting Machine (GBM):	98.33%	97.14%	100%	98.55%
XGBoost:	97.22%	95.33%	100%	97.61%

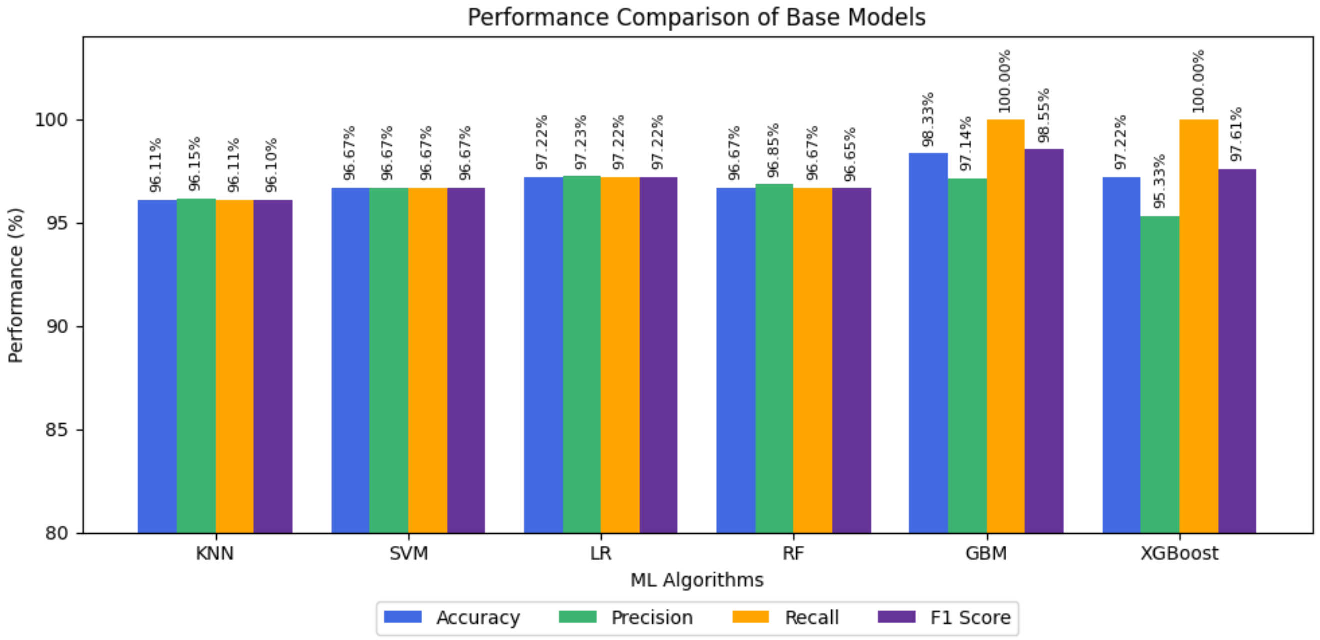

Figure-3: Performance of six base models

4.2 Ensemble Models:

To enhance our results, we combined our base models into multiple sets and analyzed the performance of the new ensemble models. Consequently, these ensembles demonstrated significant improvements over the base models, highlighting the advantages of integrating multiple algorithms for better predictive performance. Moreover, certain combinations consistently outperformed others, thus emphasizing effective algorithm pairings. Table 4.2 presents the performance metrics of these ensembles.

Notably, the ensemble model combining Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Machine (GBM) demonstrates superior performance across all metrics. It achieves an impressive accuracy of 98.89%, along with a precision of 98.08%, a recall of 100%, and an F1-score of 99.03%. This remarkable performance underscores the effectiveness of ensemble methods in leveraging the strengths of individual models, resulting in a balanced and highly accurate output that outperforms each model used independently. As illustrated in **Figure-4**, this combination stands out as a leading solution in predictive modeling, with Figure-5 depicting the corresponding confusion matrix.

Other ensembles, such as GBM + XGBoost and LR + RF + GBM, performed well with 98.33% accuracy and an F1-score of 98.55%, further demonstrating ensemble superiority. However, certain combinations, like SVM + KNN, showed comparatively lower results, indicating the importance of appropriate model pairings for optimal performance.

Table 4.2: Performance of ensemble models

Algorithms	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM + KNN	94.44	100.0	90.20	94.85
SVM + LR	93.89	97.89	91.18	94.42
(SVM + RF), (SVM+GBM), (SVM+XGBoost)	95.56	100	92.16	95.92
KNN + LR	93.33	98.91	89.22	93.81
(KNN + RF),	96.67	97.06	97.06	97.06
KNN+GBM	97.22	98.02	97.06	97.54
KNN+XGBoost	96.11	96.12	97.06	96.59
LR + RF	94.44	98.94	91.18	94.90
LR + GBM	93.89	98.92	90.20	94.36
LR + XGBoost	94.44	98.94	91.18	94.90
RF+GBM	97.78	97.12	99.02	98.06
RF+XGBoost	97.78	96.23	100	98.08

GBM+XGBoost	98.33	97.14	100	98.55
SVM+KNN+LR	93.89	96.91	92.16	94.47
SVM+KNN+RF	98.33	98.06	99.02	98.54
SVM+KNN+GBM	97.78	98.04	98.04	98.04
SVM+KNN+XGBoost	97.22	96.19	99.02	97.58
SVM+LR+RF	93.89	96.91	92.16	94.47
SVM+LR+GBM	93.89	96.91	92.16	94.47
SVM+LR+XGBoost	93.89	96.91	92.16	94.47
SVM+RF+GBM	**98.89**	**98.08**	**100**	**99.03**
SVM+RF+XGBoost	97.78	97.12	99.02	98.06
SVM+GBM+XGBoost	98.33	97.14	100	98.55
KNN+LR+RF	97.78	97.12	99.02	98.06
KNN+LR+GBM	98.33	98.06	99.02	98.54
KNN+LR+XGBoost	97.22	96.19	99.02	97.58
KNN+RF+GBM	97.78	96.23	100	98.08
KNN+RF+XGBoost	96.67	94.44	100	97.14
KNN+GBM+XGBoost	97.22	95.33	100	97.67
LR+RF+GBM	98.33	97.14	100	98.55
LR+RF+XGBoost	97.78	96.23	100	98.08
LR+GBM+XGBoost	98.33	97.14	100	98.55
RF+GBM+XGBoost	97.78	96.23	100	98.08
SVM+KNN+LR+RF	95	98.95	92.16	95.43
SVM+KNN+LR+GBM	94.44	98.94	91.18	94.90
SVM+KNN+LR+XGBoost	95	98.95	92.16	95.43
SVM+KNN+RF+GBM	98.33	98.06	99.02	98.54
SVM+KNN+RF+XGBoost	97.78	97.12	99.02	98.06
SVM+KNN+GBM+XGBoost	98.33	98.06	99.02	98.54
SVM+LR+RF+GBM	95	98.95	92.16	95.43
SVM+LR+RF+XGBoost	95	98.95	92.16	95.43
SVM+LR+GBM+XGBoost	95	98.95	92.16	95.43
SVM+RF+GBM+XGBoost	98.33	97.14	100	98.55
KNN+LR+RF+GBM	98.33	98.06	99.02	98.54
KNN+LR+RF+XGBoost	97.78	97.12	99.02	98.06
KNN+LR+GBM+XGBoost	98.33	98.06	99.02	98.54
KNN+RF+GBM+XGBoost	97.22	95.33	100	97.61
LR+RF+GBM+XGBoost	98.33	97.14	100	98.55

SVM+KNN+LR+RF+GBM	98.33	98.06	99.02	98.54
SVM+KNN+LR+RF+XGBoost	98.33	98.06	99.02	98.54
SVM+KNN+LR+GBM+XGBoost	98.33	98.06	99.02	98.54
SVM+KNN+RF+GBM+XGBoost	97.78	96.23	100	98.08
SVM+LR+RF+GBM+XGBoost	98.33	97.14	100	98.55
KNN+LR+RF+GBM+XGBoost	97.78	96.23	100	98.08
SVM+KNN+LR+RF+GBM+XGBoost	98.33	98.06	99.02	98.54

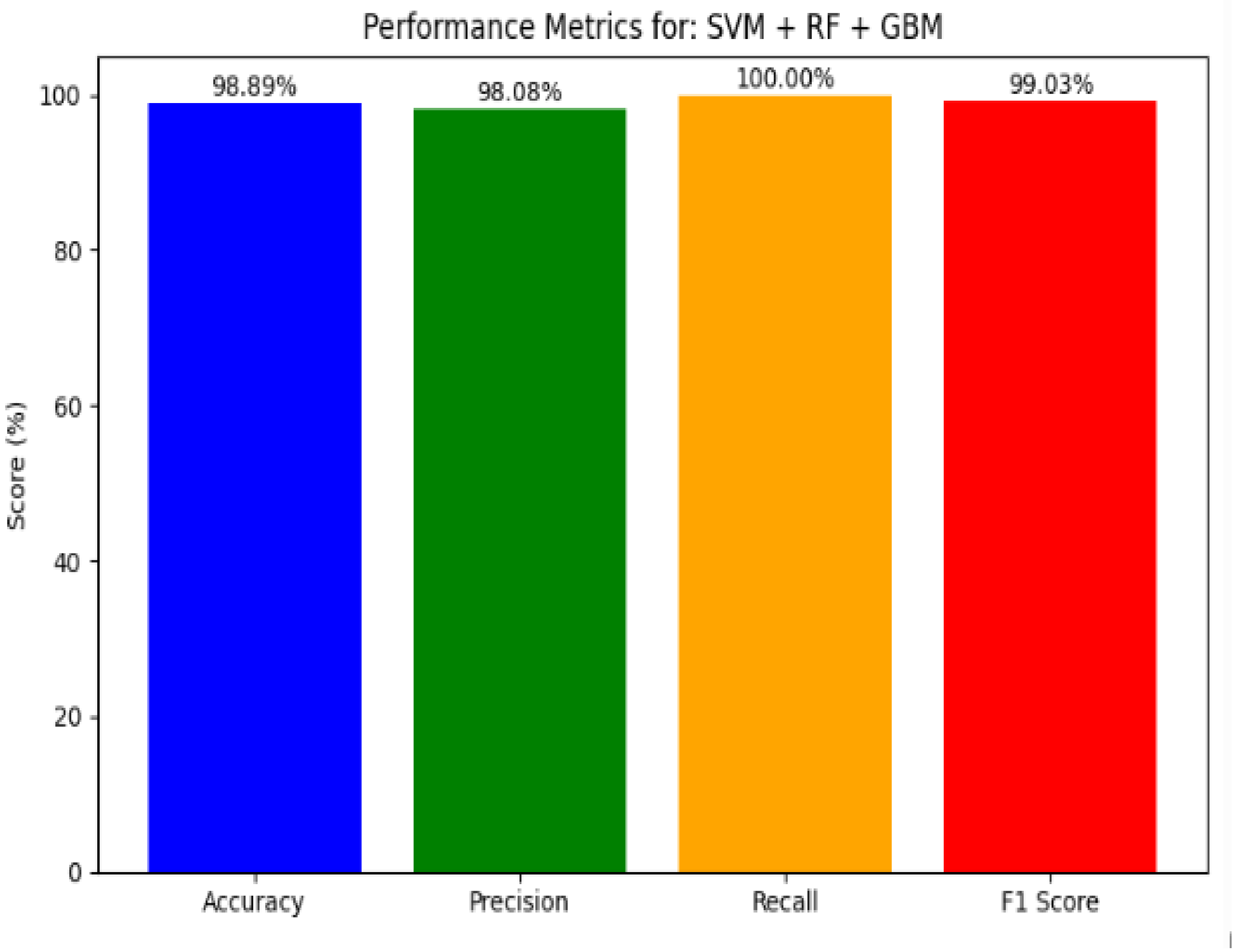

Figure-4: Performance of SVM+RF+GBM

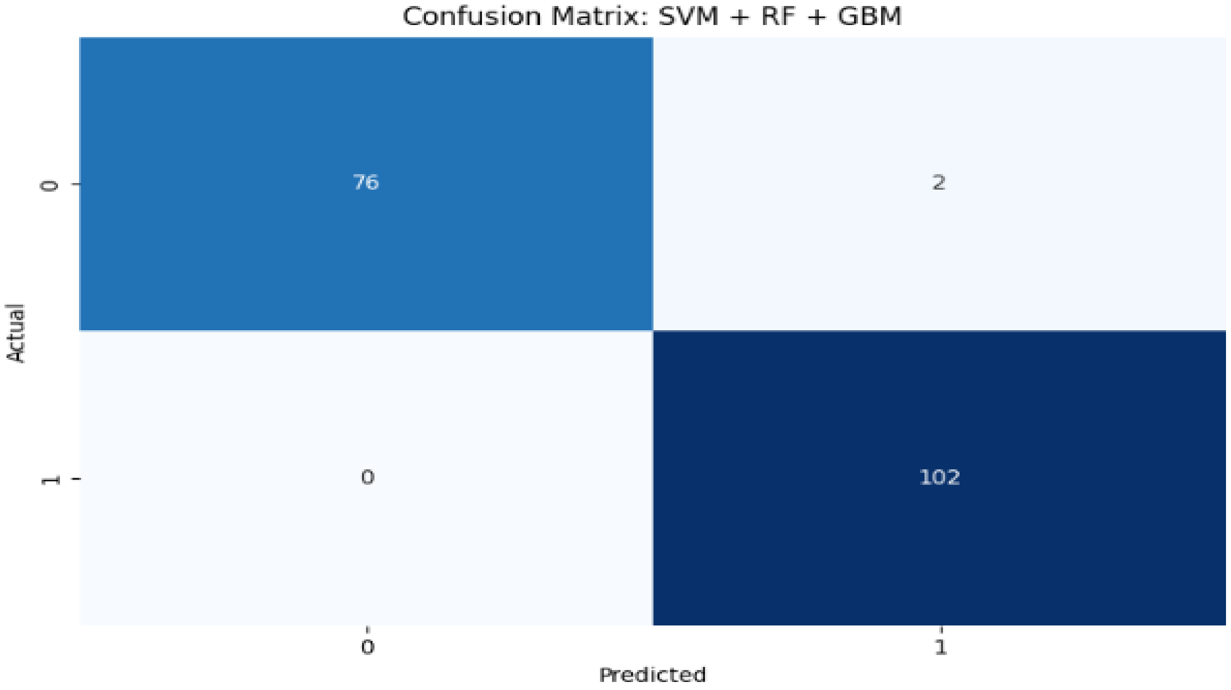

Figure-5: Confusion matrix of SVM+RF+GBM

5. Discussion and Future Work

5.1 Discussion

The findings of this study underscore the potential of machine learning, specifically ensemble models, in enhancing breast cancer diagnosis accuracy. By leveraging multiple ML algorithms—such as K-Nearest Neighbors, Support Vector Machines, and Gradient Boosting—combined through Hard Voting, the ensemble model achieved notable improvements in classification metrics. This integration demonstrated the advantages of ensemble methods, which balance individual model biases and increase robustness, ultimately leading to superior predictive power compared to single models. However, the results also revealed some limitations: while ensemble models generally boost performance, they can be computationally intensive and may demand more resources for both training and evaluation. Additionally, the dataset's imbalanced nature required oversampling techniques, which, while effective, may introduce potential biases that could affect the model's generalizability to other datasets or populations.

5.2 Limitation

The limitations of this study include a relatively small sample size and reliance on a single dataset (WBC), which may limit the generalizability of the model to broader populations. Additionally, while we implemented oversampling to address class imbalance, more sophisticated balancing techniques, such as SMOTE, could further enhance model performance. Future research could expand upon this study by incorporating a larger, more diverse dataset and employing advanced ensemble approaches such as stacking with deep learning models to potentially improve accuracy and generalizability. Incorporating feature selection algorithms could also refine input features, minimizing noise and improving model interpretability.

5.3 Future Work

Future research could focus on several directions to expand and refine the current study. One promising avenue is exploring advanced data augmentation and synthesis techniques, such as synthetic minority oversampling (SMOTE) and generative adversarial networks (GANs), to better address class imbalance without potentially introducing bias. Furthermore, incorporating deep learning architectures like Convolutional Neural Networks (CNNs) for image data or transformer-based models for feature-rich datasets could offer enhanced pattern recognition capabilities, potentially improving the model's accuracy and generalizability. Additionally, applying feature selection methods or dimensionality reduction techniques could refine the model by isolating the most predictive features, reducing computational demands, and mitigating overfitting risks. Lastly, deploying and validating the model in clinical settings or on diverse, real-world datasets could provide insights into its effectiveness and reliability in practical applications, potentially guiding future developments toward a clinically viable diagnostic tool for breast cancer.

5.4 Ethical Consideration

In conducting this research, we considered and addressed several ethical issues, specifically related to data privacy, model bias, and patient impact. We utilized publicly available, anonymized datasets to protect patient identities and adhered to regulatory standards like HIPAA and GDPR. Recognizing the potential for bias, we validated the model with diverse data to ensure equitable performance across demographic groups, aiming to reduce disparities in predictive accuracy. We also prioritized model interpretability, enabling healthcare professionals to understand and validate the machine learning outputs rather than relying solely on algorithmic predictions. This transparency supported patient trust and respected the clinical judgment of healthcare providers. Additionally, we minimized the likelihood of false positives and negatives, understanding that accurate early predictions have critical implications for patient outcomes and care. Throughout the study, we continuously evaluated and refined the model to uphold ethical standards and ensure it remained aligned with best practices in clinical settings.

6. Conclusion

In conclusion, this research underscores the significant potential of machine learning, particularly ensemble techniques, in the realm of breast cancer diagnostics. By integrating six diverse algorithms—K-Nearest Neighbors, Support Vector Machine, Logistic Regression, Random Forest, Gradient Boosting, and XGBoost—through a Hard Voting ensemble, the model effectively classified breast cancer cases with enhanced accuracy, precision, recall, and F1 scores. These metrics are crucial in clinical diagnostics, where accurate early detection of malignancy can dramatically improve treatment outcomes and patient survival rates. Ensemble methods proved advantageous by capitalizing on the unique strengths of each algorithm, thereby delivering a more robust classification than individual models.

Despite the notable results, the study also highlights certain limitations, particularly the challenges posed by class imbalance in the dataset and the increased computational resources required for ensemble learning. The use of oversampling techniques to balance classes helped mitigate these issues; however, it may introduce a degree of bias that could affect the model's performance on entirely new datasets. Furthermore, computational intensity could pose barriers to practical application in clinical settings where speed and efficiency are critical.

Overall, this study contributes to the growing body of research advocating for machine learning-based diagnostic tools in healthcare, providing a foundation for further exploration. Future studies could extend these findings by validating the model across larger, more diverse datasets and exploring deep learning architectures or advanced data augmentation techniques. By building on this foundation, future developments could lead to even more accurate, resource-efficient, and clinically viable models, ultimately supporting timely and precise breast cancer diagnostics on a wider scale.

To further enhance model performance and applicability, it is recommended to explore advanced techniques for addressing class imbalance, such as SMOTE or GAN-based synthetic data generation. Future implementations should also consider optimizing feature selection to streamline model performance without sacrificing accuracy. Finally, validating this model with larger and more diverse datasets would support its generalizability and clinical utility, contributing to more accurate, efficient, and accessible breast cancer diagnostic tools.

Reference

[1] World Cancer Research Fund. “Breast Cancer Statistics.” World Cancer Research Fund, 2023, www.wcrf.org/cancer-trends/breast-cancer-statistics/.

[2] World Health Organization. “Breast Cancer.” World Health Organization, 5 Apr. 2023, www.who.int/news-room/fact-sheets/detail/breast-cancer.

[3] Cancer Health. “Breast Cancer Mortality Continues Three-Decade Decline, Disparities Remain.” Cancer Health, 10 Oct. 2023, www.cancerhealth.com/article/breast-cancer-mortality-continues-three-decade-decline-disparities-remain.

[4] MurtiRawat, Ram, et al. "Breast Cancer detection using K-nearest neighbors, logistic regression and ensemble learning." 2020 international conference on electronics and sustainable communication systems (ICESC). IEEE, 2020.

[5] Gupta, Manoj Kumar, and Pravin Chandra. "A comprehensive survey of data mining." International Journal of Information Technology 12.4 (2020): 1243-1257.

[6] Esteva, Andre, et al. "Dermatologist-level classification of skin cancer with deep neural networks." nature 542.7639 (2017): 115-118.

[7] Cortes, Corinna. "Support-Vector Networks." Machine Learning (1995).

[8] Cover, Thomas, and Peter Hart. "Nearest neighbor pattern classification." IEEE transactions on information theory 13.1 (1967): 21-27.

[9] Langley, Pat, Wayne Iba, and Kevin Thompson. "An analysis of Bayesian classifiers." Aaai. Vol. 90. 1992.

[10] Breiman, Leo. Classification and regression trees. Routledge, 2017.

[11] Cox, David R. "The regression analysis of binary sequences." Journal of the Royal Statistical Society Series B: Statistical Methodology 20.2 (1958): 215-232.

[12] Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.

[13] Freund, Yoav, and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." Journal of computer and system sciences 55.1 (1997): 119-139.

[14] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[15] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." nature 323.6088 (1986): 533-536.

[16] Jiang, Fei, et al. "Artificial intelligence in healthcare: past, present and future." Stroke and vascular neurology 2.4 (2017).

[17] Bauer, Eric, and Ron Kohavi. "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants." Machine learning 36 (1999): 105-139.

[18] Kittler, Josef, et al. "On combining classifiers." IEEE transactions on pattern analysis and machine intelligence 20.3 (1998): 226-239.

[19] Wolpert, David H. "Stacked generalization." Neural networks 5.2 (1992): 241-259.

[20] Dietterich, Thomas G. "Ensemble methods in machine learning." International workshop on multiple classifier systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000.

[21] Jiang, Chaoqun, et al. "Automatic facial paralysis assessment via computational image analysis." Journal of Healthcare Engineering 2020.1 (2020): 2398542.

[22] Sharma, R. K., and Anil Ramachandran Nair. "Efficient breast cancer prediction using ensemble machine learning models." 2019 4th International conference on recent trends on electronics, information, communication & technology (RTEICT). IEEE, 2019.

[23] Abdar, Moloud, et al. "A new nested ensemble technique for automated diagnosis of breast cancer." Pattern Recognition Letters 132 (2020): 123-131.

[24] Mashudi, Nurul Amirah, et al. "Comparison on some machine learning techniques in breast cancer classification." 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES). IEEE, 2021.

[25] Nanglia, S., et al. "An enhanced Predictive heterogeneous ensemble model for breast cancer prediction." Biomedical Signal Processing and Control 72 (2022): 103279.

[26] Kadhim, Rania R., and Mohammed Y. Kamil. "Comparison of machine learning models for breast cancer diagnosis." IAES International Journal of Artificial Intelligence 12.1 (2023): 415.

[27] Khan, Razib Hayat, et al. "A comparative study of machine learning algorithms for detecting breast cancer." 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2023.

[28] Sharma, Nonita, et al. "Breast cancer classification using snapshot ensemble deep learning model and t-distributed stochastic neighbor embedding." Multimedia Tools and Applications 82.3 (2023): 4011-4029.

[29] Mahesh, T. R., et al. "Early predictive model for breast cancer classification using blended ensemble learning." International Journal of System Assurance Engineering and Management 15.1 (2024): 188-197.

[30] ALABI, OI, et al. "Performance Assessment of Ensemble-Tree Learning Models on Breast Cancer Dataset." Journal of Information Sciences 23.1 (2024): 90-104.

[31] Batool, Amreen, and Yung-Cheol Byun. "Towards Improving Breast Cancer Classification

using an Adaptive Voting Ensemble Learning Algorithm." IEEE Access (2024).

[32] Jakhar, Amit Kumar, Aman Gupta, and Mrityunjay Singh. "SELF: a stacked-based ensemble learning framework for breast cancer classification." Evolutionary Intelligence 17.3 (2024): 1341-1356.

[33] William, H., W. Nick Street, and Olvi L. Mangasarian. "Breast cancer wisconsin (diagnostic) data set." UCI Machine Learning Repository (1995).

[34] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine 17.3 (1996): 37-37.

[35] Hastie, Trevor, et al. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. New York: springer, 2009.

[36] Cortes, Corinna. "Support-Vector Networks." Machine Learning (1995). https://doi.org/10.1023/A:1022627411411

[37] Cover, Thomas, and Peter Hart. "Nearest neighbor pattern classification." IEEE transactions on information theory 13.1 (1967): 21-27. https://doi.org/10.1109/TIT.1967.1053964

[38] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. John Wiley & Sons, 2013.

[39] Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.

[40] Zhang, Harry. "The optimality of naive Bayes." Aa 1.2 (2004): 3.

[41] Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." Annals of statistics (2001): 1189-1232. https://doi.org/10.1214/aos/1013203451

[42] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. https://doi.org/10.1145/2939672.2939785