



Enhancing Breast Cancer Detection Accuracy through Combined Classification Algorithms

Niloofer Fayazi^{1*}, Nayereh Zaghari²

¹ Department of Computer Engineering, West Tehran Branch, Islamic Azad University, Tehran, Iran.

² Department of Cognitive Computational Science, CT.C, Islamic Azad University, Tehran, Iran.

ARTICLE INFO

Article Type:

Original Research

Received: 2024.12.24

Revised: 2025.04.15

Accepted: 2025.04.21

Keyword:

Breast Cancer
Machine Learning
Random Tree
J48
Hybrid Models
Diagnostic Accuracy
Classification

*Corresponding Author:

Niloofer Fayazi

Email:

nilofayazi1990@gmail.com

ABSTRACT

Breast cancer is the most commonly identified cancer among women and a major cause of death worldwide. Early detection significantly improves survival rates, but challenges remain in accurately differentiating between benign and malignant tumors. This study proposes a hybrid method combining machine learning algorithms, specifically Random Tree and J48, to improve diagnostic accuracy in breast cancer detection. The research used two datasets: the Ljubljana dataset, consisting of 286 patient samples with ten essential features such as tumor size, malignancy grade, and lymph node involvement, and the Wisconsin (WDBC) dataset, which contains 569 samples. Data preprocessing steps, including standardization, duplicate removal, and feature filtering, were applied to ensure data quality and relevance. The study compared the performance of multiple classifiers, including Naive Bayes, Support Vector Machines (SVM), and Random Tree, using metrics such as accuracy, precision, and recall. Results showed that the Random Tree algorithm achieved a remarkable accuracy of 97.9% on the Ljubljana dataset and 100% on the Wisconsin (WDBC) dataset, outperforming other single algorithms. Combined algorithms significantly increased accuracy. The hybrid methods Random Tree + KNN and Random Tree + J48 achieved 96% and 99% accuracy, respectively, across the datasets. These combinations, with higher processing speed and improved accuracy, enhanced diagnostic efficiency and demonstrated greater suitability for clinical implementation and early breast cancer detection. This study highlights the potential of machine learning in advancing early breast cancer detection and sets a benchmark for scalable, accurate, and interpretable diagnostic tools in healthcare.



Introduction

Breast cancer is one of the earliest documented diseases in human history. Ancient civilizations have the first recorded medical accounts of breast cancer. Medical writings from ancient Egypt, dated around 1600 BCE, noted occurrences of breast tumors, which were untreatable at the time and often regarded as fatal conditions. Breast cancer is the most prevalent cancer among women worldwide. According to the World Health Organization (WHO), in 2020, about 2.3 million new cases of breast cancer were identified globally, making it the most frequently diagnosed cancer among women and, in some countries, the most common cancer overall.[1] Breast cancer is the second most frequent cause of cancer-related deaths, following lung cancer, accounting for approximately 685,000 deaths in the same year. [2, 3] In Iran, breast cancer is the most common type of cancer and the primary cancer affecting women, with an incidence rate of approximately 30 cases per 100,000 women. Thanks to growing public awareness and advancements in the healthcare sector, currently, 55% of breast cancer diagnoses are made at early stages, while 45% are still being detected at advanced stages.[4] The increasing global incidence of breast cancer highlights the essential role of early detection in decreasing mortality. Among women, breast cancer and fibroids are among the most frequently occurring masses. When detected early, the prognosis and treatment outcomes are significantly improved.[5] In recent decades, advances in machine learning and data mining have significantly enhanced breast cancer diagnosis, making them powerful tools for improving accuracy and reducing mortality. However, early detection remains challenging due to tumor complexity, patient data variability, and limitations of traditional diagnostic methods.[6] This study aims to address these issues by applying machine learning techniques—particularly ensemble models combining algorithms such as Random Tree, KNN, J48, and Logistic Regression—to improve classification performance. These models are capable of processing large volumes of clinical, imaging, and laboratory data, providing effective decision support in medical environments.[7] Classification, as a core component of machine learning, plays a vital role in medical diagnosis. Extensive research on breast cancer datasets has demonstrated the effectiveness of combined algorithms in achieving higher accuracy compared to individual models. Our experiments, based on real patient datasets, highlight the potential of hybrid approaches in building intelligent diagnostic systems.[8]

Literature Review

Research on machine learning for breast cancer diagnosis began in the late 1980s. With advancements in computational power, intelligent algorithms have become integral to medical diagnosis. In the 1990s, the Wisconsin Breast Cancer Dataset (WBCD) emerged, marking a key milestone. W.H. Wolberg and O.L. Mangasarian introduced a linear programming technique in 1990 to distinguish benign and malignant patterns in breast cytology, significantly improving early detection accuracy. Their work demonstrated the potential of computational methods to enhance diagnostic precision.[9] In 1995, W.N. Street, O.L. Mangasarian, and W.H. Wolberg developed an inductive learning approach for predicting breast cancer prognosis. Their method used diagnostic features to identify high-risk patients, aiding clinicians in making treatment decisions. This approach demonstrated effectiveness in prognostic predictions for breast cancer patients.[10] By 2018, research on breast cancer diagnosis expanded with Hongwei Chen, Annette Payne, and Xiaohui Liu using ML techniques such as ANNs, SVMs, DTs, and k-NNs on the Wisconsin Breast Cancer Database (WBCD), improving classification accuracy. They introduced Particle Swarm Optimization (PSO) algorithms, like SPSO and SDPSO, to enhance ML performance and explored ML applications for analyzing Friedreich Ataxia (FRDA) datasets, aiming to develop an intelligent healthcare system.[11] In 2018, Dai, B., Chen, R.-C., Zhu, S.-Z., and Zhang applied the Random Forest (RF) algorithm to diagnose breast cancer using medical data. RF, known for handling complex datasets, accurately classified breast cancer cases as benign or malignant, outperforming simpler algorithms. Their research highlighted RF's potential in medical diagnostics, particularly in breast cancer, where precise classification is crucial for effective treatment.[12] In 2022, Ali Bou Nassif et al. reviewed breast cancer detection using AI, focusing on genetic sequencing and histopathological imaging. Histopathological imaging achieved 99.3% accuracy, while genetic sequencing performed slightly better at 99.8%. Despite the higher accuracy of genetic sequencing, histopathological imaging is more practical due to cost and complexity. The study emphasized the effectiveness of machine learning and deep learning techniques in early detection and diagnosis, highlighting their potential to improve survival rates. [13] In 2022, Adyasha Sahu et al. reviewed the use of machine learning (ML) and deep learning (DL) techniques, especially convolutional neural networks (CNNs), for detecting breast cancer from mammograms. They noted significant improvements in detection accuracy, with DL models achieving up to 97%, though challenges like data quality and the need for larger datasets remain.[14] In 2022, Mohammed Amine et al. tested five machine learning algorithms on the WBCD dataset. SVM

outperformed the others with 97.2% accuracy, 97.5% precision, and an AUC of 96.6%. The study suggested that these results were specific to the WBCD dataset and recommended validating the algorithms with other datasets.[15] In 2022, Varsha Nemade and Vishal Fegade explored various machine learning techniques for predicting breast cancer, including decision trees, support vector machines (SVM), and k-nearest neighbors (K-NN). The study highlighted their potential in early detection, with some models achieving high accuracy. It also emphasized the importance of feature selection, dataset quality, and algorithm optimization to enhance model performance in breast cancer prediction.[16] In 2022, P. Mounika and Pitchai R explored the use of machine learning algorithms like decision trees, SVM, and random forests for detecting breast cancer. They highlighted the importance of feature extraction and selection in improving model performance, achieving high accuracy rates. The study also emphasized the need for high-quality datasets and preprocessing to train reliable models for early and accurate diagnosis.[17] This review highlights the progress in applying machine learning for breast cancer diagnosis, emphasizing the growing use of hybrid models and deep learning techniques. These approaches are designed to improve accuracy and robustness, reflecting a trend towards integrating various methods for enhanced diagnostic performance.

Table 1: Comparison of My Study with Recently Published Papers

References	Journal Name	Method	Advantages	Weaknesses	Accuracy
Mostafa Shanbehzadeh, et al [18]	(Elsevier) (2022)	Various machine learning techniques (e.g., SVM, Decision Trees)	Effective classification of breast cancer, high prediction reliability	Dependent on dataset quality and preprocessing	79.002%
Yudi Jin, et al [19]	(European Journal of Medical Research) (2023)	Random Forest (RF), Logistic Regression	High specificity and sensitivity in predicting tumor recurrence and events	Logistic regression has lower performance compared to RF	AUC: 82.9%
Ganjar Alfian, et al [20]	(International Journal of Computer Science and Network Security)	SVM with Extra-Trees for feature selection	Improved feature relevance and prediction accuracy	Small dataset size, limited generalizability	80.23%

Mohammed Amine Naji , Sanaa El Filali[15]	(Elsevier) (2021)	SVM, Random Forest, Logistic Regression, C4.5, KNN	High accuracy in classification , effective early detection	Performance depends on data quality and preprocessing	97.2%
Sara Laghmati, et al [21]	(Springer) (2024)	The method uses PCA and machine learning.	High accuracy and recall, especially with the XGboost and S-LR models.	No comparison with other advanced ensemble techniques.	97.37%
Farnaz Hoseini1* , Hamed Sepehrzadeh [22]	(Journal of Engineering and Applied Research (JEAR)) (2024)	Frequency transformations, Naive Bayes	High accuracy (91%), high precision (98%) and recall (98.7%), effective feature extraction	May require extensive research for optimization, future developments needed for cost reduction in digital mammography	91%
Proposed method (self)	(Journal of Engineering and Applied Research (JEAR)) (2024)	Combination of machine learning algorithms like Random Tree and other classification methods	High accuracy in distinguishing benign and malignant tumors	Requires careful data preprocessing and advanced tools	100%

Theoretical Background

- 1. ZeroR:** ZeroR is a simple baseline classification algorithm that predicts the most frequent class of the target variable without using any input features. It helps establish a baseline for evaluating more complex models by providing a minimal benchmark, ensuring that advanced models truly improve classification performance. ZeroR is computationally efficient, requires no training time, and is useful for analyzing class distribution. In this study, it was used to establish baseline accuracy for comparing more sophisticated models in breast cancer detection.[23]
- 2. J48:** J48, based on the C4.5 decision tree algorithm, is ideal for classification tasks due to its transparent decision-making, effective handling of missing data, and ability to manage varying feature costs. Its pruning mechanism reduces overfitting, improving generalizability, and it is computationally efficient for large datasets. In this study, J48 was

selected for its strong performance, robustness to noisy data, and interpretability, making it a suitable choice for breast cancer diagnosis.[24]

3. **Naive Bayes:** Naïve Bayes (NB) is a simple and efficient probabilistic classification algorithm based on Bayes' Theorem and the assumption of conditional independence between features. Despite this assumption often not holding in real-world data, it remains effective due to its speed and computational efficiency. Variations of the algorithm have been developed to handle more complex data, improving accuracy. Its simplicity makes it easy to implement, even for high-dimensional data or limited sample sizes, making it a robust choice for various machine learning tasks.[25]
4. **Random Tree:** A Random Tree is a decision tree algorithm that incorporates randomness in both feature and data subset selection when building the tree. Unlike traditional decision trees that consider all features at each split, Random Tree randomly selects a subset of features, helping to reduce overfitting and enhance generalization, especially in noisy or complex datasets. Its robustness to noise makes it particularly suitable for medical data, such as in breast cancer detection, where variability and irrelevant features are common. Additionally, Random Tree can offer insights into feature importance, aiding in the identification of key predictive factors.

Key advantages include:

- Reduced overfitting through random feature selection
- Improved generalization on noisy data
- Capability to identify influential features in classification tasks

These properties make Random Tree an effective and interpretable model for medical diagnosis applications.

5. **Random Forest:** Random Forest is an ensemble learning technique that builds multiple decision trees using bootstrapped data samples and random feature subsets, combining their outputs through voting (classification) or averaging (regression). This approach enhances accuracy, reduces overfitting, and improves model robustness. In the context of breast cancer detection, Random Forest offers several advantages:

- Generalization: Randomization prevents overfitting to training data.

- Noise tolerance: Effective in handling variability and noise inherent in medical datasets.
- Feature importance: Highlights key predictors, aiding interpretability.

These strengths make Random Forest a reliable and high-performing model for medical classification tasks such as cancer detection.

6. SVM: Support Vector Machine (SVM) is a supervised learning algorithm that constructs an optimal hyperplane to separate data classes with maximum margin. For nonlinear data, SVM uses kernel functions (e.g., RBF) to project data into higher dimensions, enabling effective classification.

SVM is particularly suitable for high-dimensional, nonlinear, and small to medium-sized datasets, offering:

- Strong generalization with limited training data
- High accuracy and resistance to overfitting
- Efficient decision boundaries defined by a subset of data (support vectors)

In this study, SVM was selected for breast cancer detection due to its robustness, precision, and suitability for complex medical datasets.[26]

7. K-Nearest Neighbors: K-Nearest Neighbors (KNN) is a simple algorithm that classifies or predicts outcomes based on the closest data points. It selects the K nearest neighbors and predicts the most common label or average value. KNN uses distance metrics like Euclidean distance to determine proximity. Its strengths include simplicity and ease of implementation, though performance depends on the choice of K and feature scaling. In breast cancer detection, KNN helps classify tumors based on proximity to known cases, with preprocessing techniques like normalization used to improve accuracy.

8. Logistic regression: Logistic Regression is a supervised learning algorithm used for binary classification by predicting the likelihood of an instance belonging to a category based on its features. It assumes a linear relationship between predictors and the log odds of the outcome. Regularization techniques like Lasso and Ridge help prevent overfitting and improve generalization. Its simplicity and interpretability make it suitable for breast cancer detection in this study.

9. Precision: The formula for Precision in machine learning and classification model evaluation is as follows:

(1)

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

Where:

- True Positives (TP): The number of samples that the model correctly predicted as positive.
- False Negatives (FN): The number of samples that the model incorrectly predicted as negative.

Recall, also known as sensitivity, indicates the model's effectiveness in correctly identifying actual positive samples. This metric is particularly crucial in cases where it is important to identify all positive cases, such as in rare disease detection, where missing even a small number of true positive cases can be critical, even if some false positives occur.

10. Recall (or Sensitivity): The formula for Recall is as follows:

(2)

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

Where:

- True Positives (TP): The number of samples that the model correctly predicted as positive.
- False Negatives (FN): The number of samples that the model incorrectly predicted as negative.

Recall, also known as sensitivity, indicates the model's effectiveness in correctly identifying actual positive samples. This metric is particularly crucial in cases where it is important to identify all positive cases, such as in rare disease detection, where missing even a small number of true positive cases can be critical, even if some false positives occur.

11. F-Measure (or F1-Score):

The F1-Score, also known as the F-Measure, is a combination of both Precision and Recall and is used to evaluate the overall performance of a model. The formula for the F1-Score is as follows:

(3)

$$F1 - Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where:

- Precision: The model's accuracy in predicting positive outcomes.
- Recall: The model's sensitivity in identifying true positive samples.

The F1-Score is particularly useful when there is a need to balance Precision and Recall. It is especially valuable when dealing with imbalanced datasets, where the number of positive and negative samples is significantly different. The F1-Score is typically used in scenarios where a balanced combination of accuracy and sensitivity is required.

12. AUC (Area Under the ROC Curve):

The AUC (Area Under the Receiver Operating Characteristic Curve) is a metric that measures the area under the ROC curve and ranges from 0 to 1:

- AUC = 1: The model performs flawlessly and can completely separate the two classes.
- AUC = 0.5: The model cannot differentiate and is equivalent to random prediction.
- AUC < 0.5: The model performs worse than a random classifier.

13. Dataset Description : In this research, two datasets are used for breast cancer data mining and classification:

- Breast Cancer Dataset (CSV format): The Breast Cancer Dataset, collected in the 1980s at the University Medical Center Ljubljana, Slovenia, contains 286 instances with nine key attributes related to tumor characteristics and patient condition. It is clinically significant, with features like tumor size and lymph node status, crucial for diagnosis and prognosis. Its historical perspective and widespread use enhance its credibility for developing predictive models in breast cancer research.
- The Breast Cancer Wisconsin (Diagnostic) Dataset, compiled in 1995 by Dr. William H. Wolberg and his team, contains 569 samples and 32 features describing tumor properties like texture, perimeter, and symmetry. A benchmark in medical machine learning, it is widely used to evaluate classification algorithms, making it a reliable resource for developing predictive models for early breast cancer diagnosis and treatment planning.[27]

Features of the Datasets:

Table 2: Features of the Ljubljana dataset from Slovenia

Feature	Description
Age	Patient's age, categorized into ranges like 10–19, 20–29, etc. It is important for determining disease progression and treatment.
Menopause	Menopausal status, categorized as premenopausal, perimenopausal, or postmenopausal. It affects cancer progression and treatment.
Tumor Size	Size of the tumor in millimeters, categorized into ranges. It helps determine cancer stage and treatment planning.
Inv Nodes	Number of involved lymph nodes, categorized into ranges. It indicates cancer spread within the body.
Node Caps	Indicates if the lymph node capsule is involved (yes/no). It can influence disease progression and survival prediction.
Deg Malign	Degree of malignancy, rated on a scale from 1 to 3. It's crucial for treatment decisions and predicting treatment response.
Breast	Indicates which breast is affected (left or right). It's useful for medical information and further analysis.
Breast Quad	Specific quadrant of the breast affected (upper-left, upper-right, lower-left, lower-right, or central). Important for spatial analysis of cancer.

Table 3: Features of the Wisconsin dataset

Feature	Description
id	Unique identifier for each tumor sample.
diagnosis	Indicates whether the tumor is malignant (M) or benign (B).
radius mean	Average distance from the center to the edge of the tumor.
Texture mean	Average variance in the grayscale intensity.
Perimeter mean	Average perimeter of the tumor.
Area mean	Average area of the tumor.
Smoothness mean	Average variation in the tumor's surface.
Compactness mean	Ratio of the tumor's perimeter to its area.
Concavity mean	Average concavity of the tumor.
Concave points mean	Average number of concave points in the tumor boundary.
Symmetry mean	Average symmetry of the tumor.
Fractal dimension mean	Average fractal dimension of the tumor boundary.
Radius se	The standard error of the tumor's radius, indicating how much the radius value varies
Texture se	The standard error of the tumor's texture, representing the variability in the tumor's surface texture.
Perimeter se	The standard error of the tumor's perimeter, showing how much the perimeter value varies.
Area se	The standard error of the tumor's area, indicating the variation in the tumor's size.
Smoothness se	The standard error of the tumor's surface smoothness, indicating its variability.
Compactness se	The standard error of the tumor's compactness, indicating variability in its shape and size.
Concavity se	The standard error of the tumor's concavity, showing the variability in concave features of the tumor
Concave points se	The standard error of the number of concave points on the tumor's surface.
Symmetry se	The standard error of the tumor's symmetry, reflecting how much the symmetry changes.

Fractal dimension se	The standard error of the fractal dimension, showing shape complexity variability.
Radius worst	The worst (largest) radius of the tumor, reflecting the biggest size of the tumor in the images.
Texture worst	The worst (most irregular) texture of the tumor, showing the highest level of texture variation.
Perimeter worst	The worst (largest) perimeter of the tumor, showing the largest boundary measurement.
Area worst	The worst (largest) area of the tumor, reflecting the biggest tumor size.
Smoothness worst	The worst (roughest) smoothness of the tumor surface.
Compactness worst	The worst (least compact) compactness of the tumor, showing the least regularity in shape
Concavity worst	The worst concavity of the tumor, indicating the most inward curves in its surface.
Concave points worst	The worst (most) concave points on the tumor's surface.
Symmetry worst	The worst (least) symmetry of the tumor, showing the highest degree of asymmetry.
Fractal dimension worst	The worst fractal dimension of the tumor, indicating the highest shape complexity.
id	Unique identifier for each tumor sample.

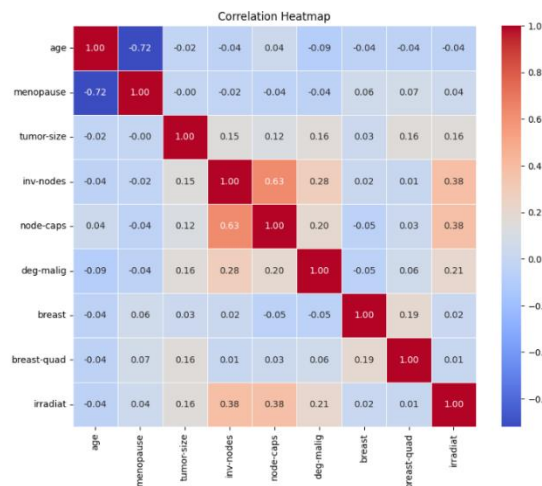


Figure 1: The heatmap of the Ljubljana Dataset

The heatmap shows the highest correlation (0.63) between inv-nodes and node-caps, suggesting that more involved lymph nodes tend to be encapsulated. The weakest correlation (0.01) is between breast and breast-quad, indicating

little to no relationship. Red colors represent strong positive correlations, while blue colors indicate weaker or negative correlations.

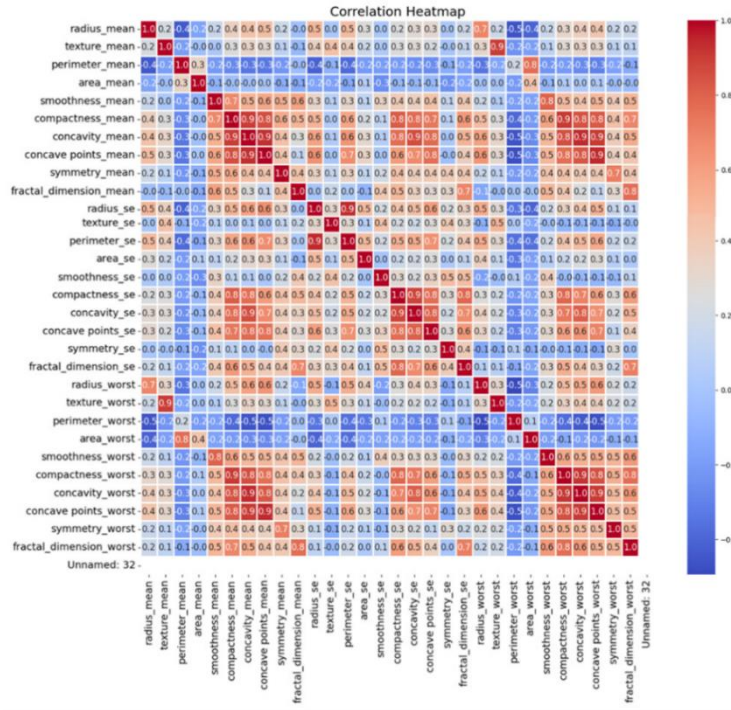


Figure 2: The heatmap of the Wisconsin dataset

In the provided heatmap, strong linear relationships between various features of the breast cancer dataset are visible. For example, the features Compactness mean and Concavity mean have a high correlation of 0.9, and the features Concavity mean and Concave points mean also exhibit a strong correlation of 0.9. Additionally, a strong correlation is observed between Texture mean and Texture worst (0.9). The features radius_se and Perimeter_se also have a close relationship with a correlation of 0.9. Furthermore, a moderate correlation is observed between area_mean and area_worst (0.4).

Methodology

Data Mining Techniques (Classification): In data mining, various methods are used to discover patterns and insights from datasets. The categorization process follows structured phases that help extract useful

information for predictive modeling. In this study, the categorization steps are specifically designed to support prediction-focused applications.

1. **Data Collection:** The dataset used comes from reputable, widely recognized sources in medical research and machine learning, ensuring credibility and relevance for predictive tasks. Selecting such datasets is essential for achieving accurate and reliable results, as they are comprehensive and representative of the problem being studied.
2. **Data Preprocessing:** Data preprocessing is essential for preparing raw data for machine learning by cleaning, handling missing values, reducing noise, and correcting inconsistencies. These steps improve data quality, ensuring better model performance and more reliable predictions:
 - 2.1. **Handling Missing Values:** Handling missing values is vital for machine learning models like SVM, KNN, Naïve Bayes, and Random Forest. Simple Imputer techniques (mean, median, or mode) are commonly used to fill missing data. If missing values are extensive, affected rows or columns may be removed to maintain data quality.
 - 2.2. **Encoding Categorical Data:** Categorical variables must be converted into numerical formats for machine learning. Label Encoding is used for ordinal data with a natural order, while One-Hot Encoding is applied to nominal data, creating separate binary columns for each category.
 - 2.3. **Normalization and Standardization:** Algorithms like KNN and SVM require data to be on a consistent scale. StandardScaler standardizes features around a mean of zero and a standard deviation of one, while MinMaxScaler normalizes values between 0 and 1. These scaling techniques are critical for improving the accuracy of distance-based algorithms.
 - 2.4. **Outlier Detection and Removal:** Outliers can distort model results. Techniques such as Z-Score or IQR can be used to identify and remove outliers. This process may involve removing data points that fall outside a specified range.
 - 2.5. **Feature Engineering and Feature Selection:** To improve model efficiency, irrelevant features should be removed. Techniques like Principal Component Analysis (PCA) and Correlation Analysis help reduce dimensions while keeping important information. For example, in the Wisconsin dataset, the "id" column was excluded because it was not relevant to the target variable.
 - 2.6. **Feature Transformation and Use of New Data:** For models like Naïve Bayes (GaussianNB), which assume that data is normally

distributed, it may be necessary to apply transformations like logarithmic scaling or other techniques to better align with the model's assumptions.

2.7. Handling Combined Models: When using combined algorithms like RF + KNN, RF + KNN + J48, and others, it's important to ensure that preprocessing is done separately for each algorithm, and then their results are combined using methods like Ensemble or Voting Classifiers.

3. **Data Splitting:** After preprocessing, the dataset is divided into training and testing sets, commonly with an 80/20 split. This ensures that the model is evaluated on unseen data, allowing for an unbiased assessment of its generalization ability.
4. **Model Selection:** Choosing the right classification algorithm is crucial for obtaining the best performance. The algorithm selected depends on the dataset's nature and the problem being solved. Commonly used classification algorithms in this study include: SMO (SVM), IBK (KNN), Naïve Bayes, Random Tree, J48 (C4.5), Zero R, Random Forest. Additionally, combined algorithms can be used to improve classification performance by leveraging the strengths of multiple algorithms. These include: RF + KNN, RF + KNN + J48, Logistic Regression + J48 + RF, J48 + RF, Naive Bayes + RF + KNN
5. **Model Training:** This step involves using the training data to "teach" the algorithm how to recognize patterns and relationships within the data. The model iteratively adjusts its internal parameters to minimize prediction errors and better fit the data.
6. **Model Evaluation:** After training, the model is evaluated using metrics like accuracy, precision, recall, F1 score, and the confusion matrix to assess its performance and prediction effectiveness on unseen data.
7. **Hyperparameter Tuning:** Optimizing model performance requires fine-tuning hyperparameters (e.g., learning rate, tree depth, regularization). Techniques like Grid Search and Randomized Search are used to explore and find the best hyperparameter values.
8. **Model Deployment:** After achieving satisfactory results, the model is deployed to make predictions on new, unseen data. This step involves integrating the trained model into a real-world application or system, enabling it to support decision-making processes and solve the problem at hand.

This study focused on using classification methods for breast cancer detection, specifically distinguishing between benign and malignant tumors. Machine learning algorithms, such as Random Tree and J48, were utilized due to their high accuracy, aiding in early and accurate detection of breast cancer.[28]

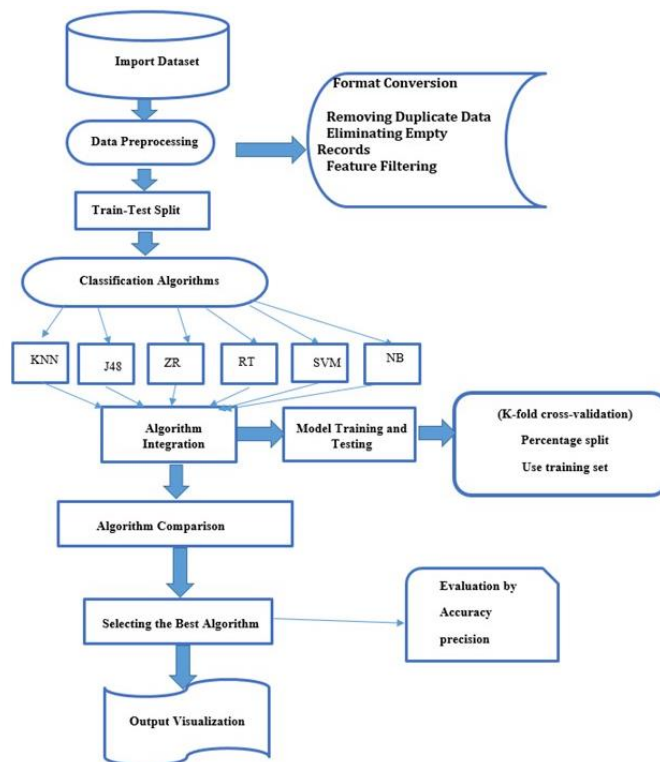


Figure 3: This flowchart illustrates the process applied to a breast cancer dataset for classification and prediction using data mining techniques.

The importance of research: Early and accurate diagnosis of breast cancer significantly improves treatment outcomes and reduces the physical and emotional burden on patients. This research aims to explore how machine learning algorithms, through the analysis of clinical and diagnostic data, can support healthcare professionals in making timely decisions. By identifying hidden patterns within patient data, the study enhances the precision of cancer risk prediction. The insights gained from this project can contribute to the development of decision support systems in medical settings, ultimately

improving the quality of care and optimizing treatment planning for women at risk.

Result and discussion

Scenario 1

In this stage of the study, various algorithms were used individually for breast cancer detection. The algorithms applied in this stage included SMO (SVM), IBK (KNN), Naïve Bayes, Random Tree, J48 (C4.5), ZeroR, and Random Forest. The objective of this scenario was to evaluate each algorithm separately in order to observe and compare their performance. After running each algorithm on the data, the results obtained from them included metrics such as Accuracy, Precision, Recall, and F-Measure.

Final Comparison of Algorithms using the Ljubljana Dataset.

Table 4: Comparison of Algorithms Using the Ljubljana Dataset.

Algorithm Name	SMO(SVM)	IBK(KN N)	Naïve Bayes	RANDO M TREE	J48 (C4.5)	ZERO R	Random Forest
Correctly Classified Instances	77.2727	79.0210	73.0769	97.9021	97.9021	70.2956	97.9021
Detection Accuracy (%)	77.2727	97.9021	73.0769	97.9021	97.9021	70.2956	97.9021
TP Rate (Sensitivity)	0.3294	79.0210	0.5059	0.9412	0.9412	0.0000	0.9529
FP Rate	0.0398	0.0945	0.1741	0.0050	0.0050	0.0000	0.0100
Precision	0.7737	0.7812	0.7247	0.9792	0.9792	0.7914	0.9790
Recall	0.7727	0.7902	0.7308	0.9790	0.9792	0.7030	0.9790
F1-Measure	0.7391	0.7801	0.7273	0.9789	0.9789	0.5804	0.9789
MCC	0.3990	0.4666	0.3404	0.9496	0.9496	0.0000	0.9496
ROC Area	0.7794	0.8349	0.7389	0.9989	0.9989	0.5000	0.9982
Training Time (sec)	0.0794	0.0044	0.0053	0.0058	0.0048	0.0011	0.1982
Prediction Time (sec)	0.0156	0.0382	0.0037	0.0035	0.0025	0.0007	0.0135
Memory Usage (MB)	285.54	271.95	270.62	290.70	265.46	266.35	276.88

CPU Usage Before Training	3.5%	4.1%	64.4%	7.5%	4.5%	4.1%	4.0%
CPU Usage After Training	3.0%	3.0%	8.0%	24.7%	4.0%	2.5%	3.5%
Estimated Energy Consumption(watt-seconds) ¹	0.06	0.01	0.18	0.31	0.02	0.00	0.86
Estimated FLOPS (million operations per second) ²	0.2659	3.7667	0.5050	0.0689	0.0785	0.11	3.0880
Model Parameters (Number of internal parameters) ³	136	5	9	199	199	0.2597	100
Method	Use training set	Use training set	Use training set	Use training set	Use training set	10-fold cross-validation	Use training set

¹ "Estimated Energy Consumption" represents the approximate energy used during model training, calculated based on CPU usage and training time.

² "Estimated FLOPS" is a measure of computational complexity that estimates the number of floating-point operations performed per second.

³ "Model Parameters" refers to the number of internal parameters, such as tree nodes in decision trees, that determine the model's structure and complexity.

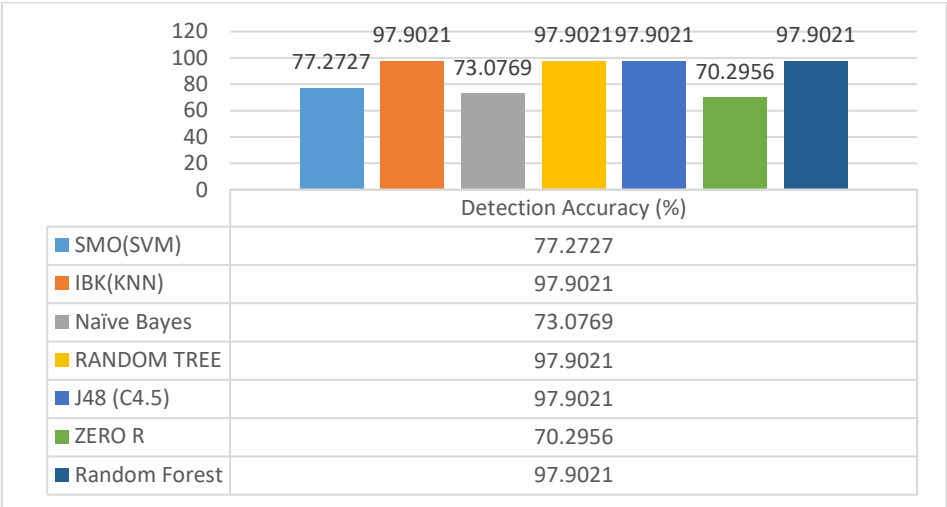


Figure 4: Comparison of Algorithms Using the Ljubljana Dataset.

In the Ljubljana dataset, Random Tree, J48 (C4.5), and Random Forest showed high classification accuracy (97.90%) in distinguishing malignant and benign cases. Random Forest was the most reliable, with a sensitivity of 0.9529, precision of 0.9790, and an F1-measure and MCC of 0.9789 and 0.9496, respectively. Although J48 and Random Tree had faster training times, Random Forest provided superior performance, making it the most effective model despite longer training time.

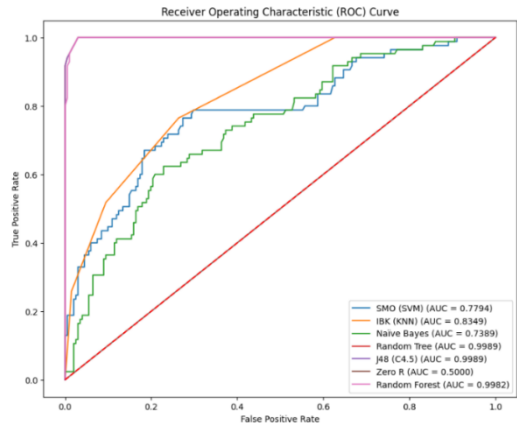


Figure 5: ROC Curve Using the Ljubljana Dataset

The ROC curve illustrates the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) at different thresholds. A higher ROC area indicates better model performance. Random Forest, Random Tree, and J48 (C4.5) achieved the highest ROC area of 0.9989, showing excellent distinction between positive and negative cases. In contrast, Zero R had a ROC area of 0.5000, indicating poor performance, similar to random guessing.

Final Comparison of Algorithms Using the Wisconsin Dataset.

Table 5: Comparison of Algorithms Using the Wisconsin Dataset

Algorithm Name	SMO(SVM)	IBK(KNN)	Naïve Bayes	RANDOM TREE	J48 (C4.5)	ZERO R	Random Forest
Correctly Classified Instances	96.6608	94.7276	94.2004	100	100	62.7417	100
Detection Accuracy (%)	96.6608	94.7276	94.2004	100	100	62.7417	100
TP Rate (Sensitivity)	0.9481	0.9009	0.8915	1.0000	1.0000	0.0000	1.0000
FP Rate	0.0224	0.0252	0.0280	0.0000	0.0000	0.0000	0.0000
Precision	0.9666	0.9475		1.0000	1.0000	0.7662	1.0000
Recall	0.9666	0.9473	0.9420	1.0000	1.0000	0.6274	1.0000
F1-Measure	0.9666	0.9469	0.9416	1.0000	1.0000	0.4838	1.0000
MCC	0.9284	0.8868	0.8755	1.0000	1.0000	0.0000	1.0000
ROC Area	0.9952	0.9917	0.9890	1.0000	1.0000	0.5000	1.0000
Training Time (sec)	13.8418	0.0039	0.0045	0.0050	0.0158	0.0012	0.2917
Prediction Time (sec)	0.0052	0.1968	0.0035	0.0019	0.0019	0.0005	0.0162
Memory Usage (MB)	283.79	282.52	271.28	256.52	253.18	248.93	262.93
CPU Usage Before Training	3.0%	2.0	20.5	3.0%	3.5	2.0	3.0
CPU Usage After Training	42.7	2.0	56.7	3.0%	2.0	2.0	3.0
Estimated Energy	0	0.56	0.56	0.01	0.04	0.11	0.88

Consumption(watt-seconds)							
Estimated FLOPS (million operations per second)	18.4544	0.2821	0.2821	0.5093	39	0.0008	0.0322
Model Parameters (Number of internal parameters)	57	65	65	65	0.0054	1	4696
Method	Use training set	Use training set	Use training set	Use training set- (Decision Tree with Random Splitter)	Use training set	10-fold cross- validati on	Use training set

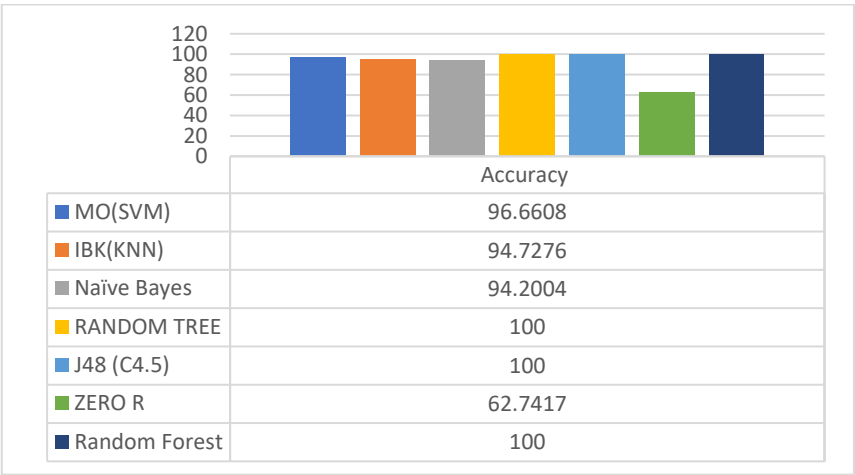


Figure 6: Comparison of Algorithms Using the Wisconsin Dataset.

In the Wisconsin dataset, the best-performing algorithm in this study was Random Forest, achieving 100% classification accuracy and perfect sensitivity, precision, F1-Measure, and MCC (1.0000). Despite a longer training time, it delivered superior overall performance. J48 (C4.5) also achieved 100% accuracy with similar metrics but had a slightly faster training time. However, Random Forest outperformed J48 in terms of stability and overall effectiveness.

The ROC area values indicate that Random Forest, Random Tree, and J48 (C4.5) all achieved perfect scores of 1.0000, showing their strong ability to distinguish between positive and negative cases. Zero R, with a ROC area of 0.5000, performed poorly, essentially equivalent to random guessing.

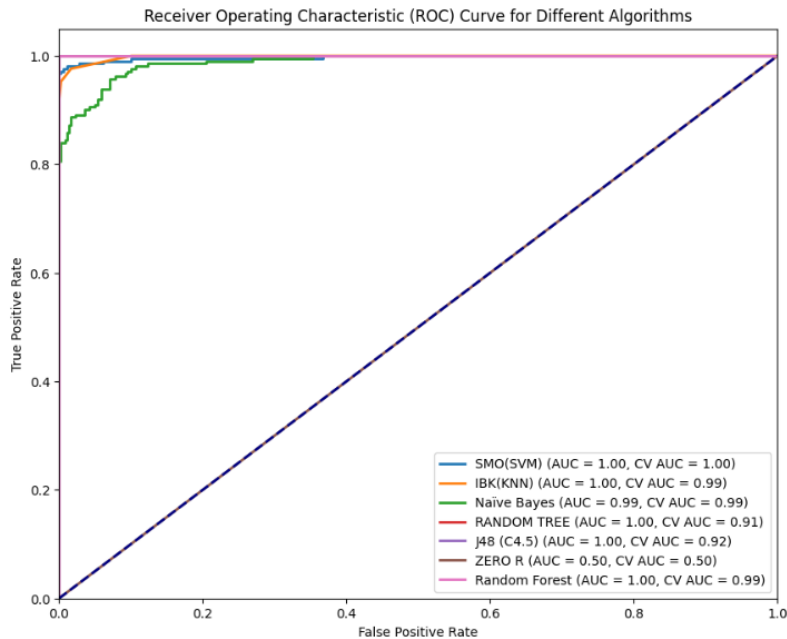


Figure 7: ROC Curve Using the Wisconsin Dataset

Scenario 2: Using Combined Algorithms

In this scenario, instead of using individual algorithms, a combination of multiple algorithms was employed to enhance the accuracy and performance of the breast cancer detection system. This approach is particularly effective in cases where the data is complex and contains diverse features. In this study, combinations of algorithms such as RF + KNN, RF + KNN + J48, Logistic Regression + J48 + RF, J48 + RF, and Naïve Bayes + RF + KNN were utilized. The goal of these combinations was to improve model accuracy and achieve better results compared to using individual algorithms. By leveraging the strengths of different algorithms and addressing the weaknesses of each, these combined models enhance the overall performance of the detection system.

The final comparison of combined algorithms using the Ljubljana Dataset.

Table 6: Comparison of Algorithms Using the Ljubljana Dataset

Combined algorithms	RF+KNN	RF + KNN + J48	Logistic Regression+J48+RF	J48+RF	Naive Bayes + RF + KNN
Correctly Classified Instances (%)	96	99	98.68	99	93
Accuracy (%)	96	99	98.68	99	93.42
TP Rate (Sensitivity)	0.94	0.99	99	0.99	0.93
FP Rate	0.01	0.01	0.01	0.01	0.07
Precision	0.96	0.99	0.99	0.99	0.94
Recall	0.96	0.99	0.99	0.99	0.93
F1-Measure	0.96	0.99	0.99	0.99	0.93
MCC	0.90	0.97	0.97	0.97	0.84
ROC Area	0.77	0.78	0.78	0.77	0.83
Training Time (sec)	0.1622	0.45	0.79	0.92	0.60
Prediction Time (sec)	0.0257	0.03	0.41	0.03	0.02
Memory Usage (MB)	1282.03	292.45	302.41	299.00	299.16
CPU Usage Before Training (%)	12.0	3.50	5.50	4.5	9.6
CPU Usage After Training	56.6	2.00	2.50	4.5	3.0
Estimated Energy Consumption(watt-seconds)	13.5681	355.93	385.08	753.31	333.97
Model Parameters (Number of internal parameters)	0	0	44	0	0
Method	Hard Voting	Voting Classifier	Voting Classifier	Voting Classifier	Voting Classifier

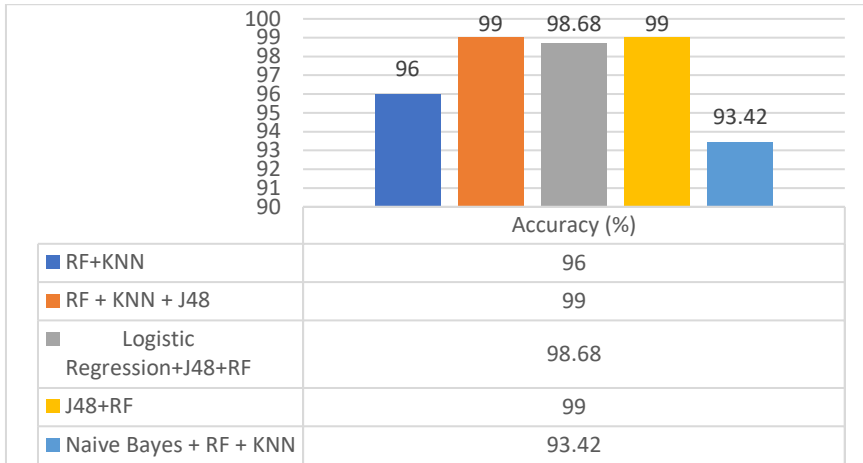


Figure 8: The final comparison of combined algorithms using the Ljubljana Dataset.

In the Ljubljana dataset, combined algorithms like RF+KNN, RF+KNN+J48, Logistic Regression+J48+RF, and J48+RF generally perform better than individual algorithms. Accuracy, sensitivity, and F1-Measure are higher in the combinations, especially RF+KNN+J48 and J48+RF, which achieve 99% accuracy and 0.99 sensitivity. While training time is higher for the combinations, the higher accuracy and better performance in identifying positive and negative cases compensate for it. MCC is significantly higher in the combined models compared to individual algorithms. Although energy and memory usage are higher in some combinations, overall, the combined algorithms perform better in terms of accuracy and overall performance.

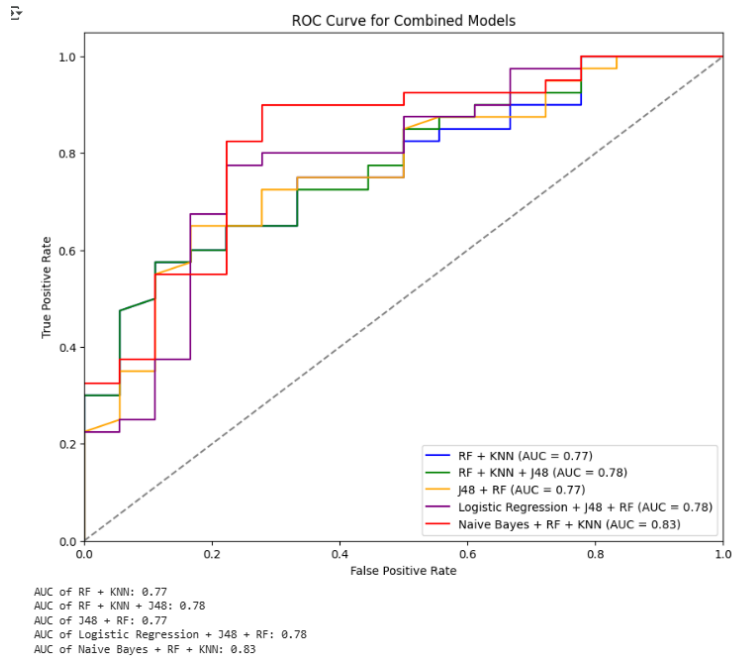


Figure 9: ROC Curve Using the Ljubljana Dataset.

In the Ljubljana dataset, when we look at the ROC area for the combined algorithms, we can see that these algorithms generally perform well in distinguishing between malignant and benign cases. Among the combinations, Naive Bayes + RF + KNN stands out with the highest ROC area of 0.83, indicating the best ability to differentiate between the two classes. Other combinations, such as RF + KNN, RF + KNN + J48, and Logistic Regression + J48 + RF, also perform well, with ROC areas ranging from 0.77 to 0.78.

The final comparison of combined algorithms using the Wisconsin Dataset.

Table7: comparison of combined algorithms using the Wisconsin Dataset.

Combined algorithms	RF+KNN	RF + KNN + J48	Logistic Regression+J48+RF	J48+RF	Naive Bayes + RF + KNN
Correctly Classified Instances (%)	96.49	95.61	96.49	94.74	96.49

Accuracy (%)	96.49	95.61	96.49	94.74	96.49
TP Rate (Sensitivity)	0.93	0.91	0.93	0.88	0.93
FP Rate	0.01	0.01	0.01	0.01	0.01
Precision	96.52	95.69	96.52	94.88	96.52
Recall	96.49	95.61	96.49	94.74	96.49
F1-Measure	96.47	95.58	96.47	94.68	96.47
MCC	0.93	0.91	0.93	0.89	0.93
ROC Area	1.00	0.99	1.00	0.99	0.99
Training Time (sec)	0.6936	2.6403	2.2317	4.5564	33.9877
Prediction Time (sec)	0.010801	0.040966	0.022824	0.032563	0.039326
Memory Usage (MB)	0.27	0.37	0.37	0.39	0.57
CPU Usage Before Training (%)	5.50	3.00	3.00	2.50	3.00
CPU Usage After Training	2.00	3.00	2.50	52.50	47.50
Estimated Energy Consumption(watt-seconds)	1.39	4.17	4.00	77.07	1614.41
Model Parameters (Number of internal parameters)	50	100	100	100	200
Method	Grid Search + Weighted Voting	Voting Classifier	Voting Classifier	Voting Classifier	Voting Classifier

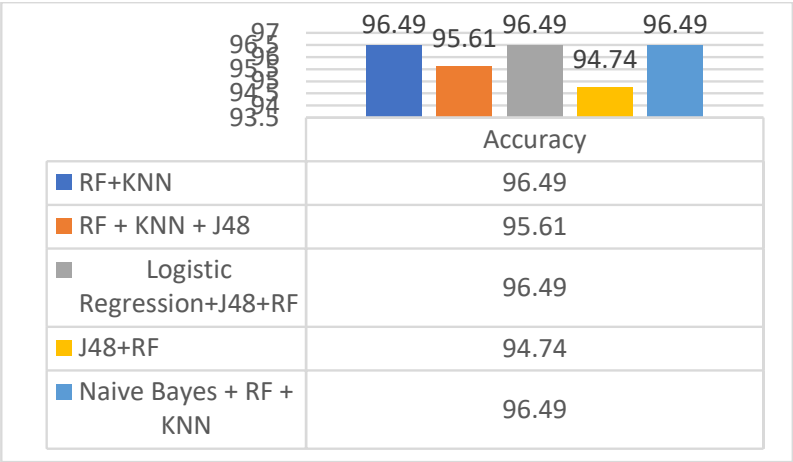


Figure10: The final comparison of combined algorithms using the Wisconsin Dataset.

In the Wisconsin models perform well in F1-Measure and MCC, with most combinations showing values around 0.93. However, they come with increased training times, especially Naive Bayes + RF + KNN, which takes 33.99 seconds. While prediction time remains fast, memory and energy consumption are higher, with Naive Bayes + RF + KNN consuming significantly more energy. Overall, combined algorithms improve classification performance but at the cost of higher computational requirements. On the dataset, the combined algorithms, such as RF+KNN and Logistic Regression+J48+RF, generally maintain high accuracy (around 96.5%) and precision, with slight reductions in sensitivity (around 0.88–0.93). These

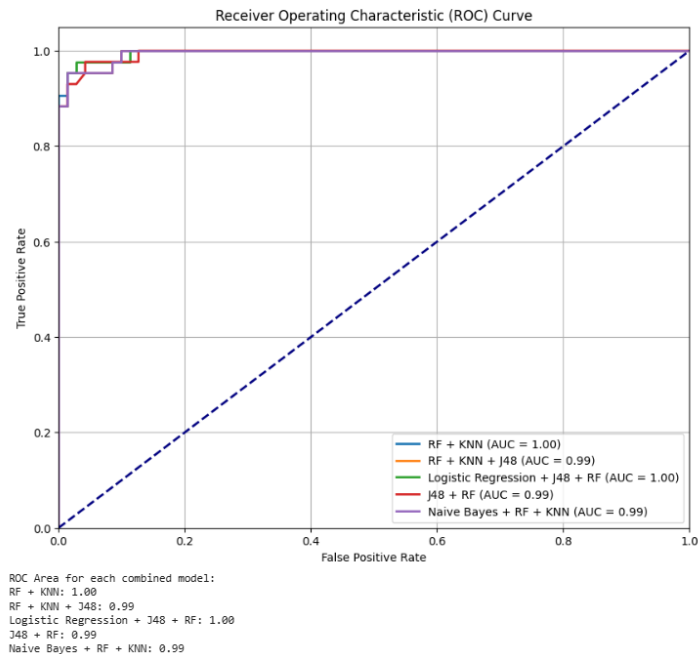


Figure 11: The ROC Curve Using the Wisconsin Dataset.

In the Wisconsin dataset, the combined models demonstrated very high performance, with ROC area values ranging from 0.99 to 1.00. Models such as RF+KNN, RF+KNN+J48, and Logistic Regression+J48+RF achieved ROC scores of 1.00 or very close to it, indicating excellent capability in distinguishing between benign and malignant tumors. These high ROC values confirm the superior accuracy and effectiveness of combined models compared to individual algorithms in breast cancer detection.

Performance of the Random Forest Algorithm on the Ljubljana and Wisconsin Datasets

Based on the analysis of the four tables and the datasets, the findings of the study are as follows: Breast cancer, a leading cause of death among women, was studied using machine learning algorithms to classify tumors as benign or malignant. In this study, two datasets—Ljubljana and the Wisconsin Breast Cancer Dataset (WBCD)—were utilized. The Ljubljana dataset comprises 286 patient samples, while the WBCD consists of 569 samples.

- For the Ljubljana Dataset: The best-performing individual algorithm was Random Forest, with an accuracy of 97.9%, achieving a high TP rate of 94.12% and a low FP rate of 0.0050. The performance of Random Tree and J48 (C4.5) also showed promising results, both achieving an accuracy of 97.9%.

- For the Wisconsin Breast Cancer Dataset (WBCD): The highest accuracy was observed with Random Forest, achieving an accuracy of 100%, with an impressive 1.0000 TP rate and 0.0000 FP rate. This demonstrated the algorithm's high effectiveness in distinguishing between benign and malignant tumors.
- Combined Algorithms: Combining Random Forest with KNN and J48 (such as RF+KNN and RF+KNN+J48) further improved accuracy. The RF+KNN+J48 combination achieved an accuracy of 99%, offering a balanced combination of sensitivity and specificity. Additionally, models such as Logistic Regression+J48+RF and J48+RF demonstrated similar performance, achieving an accuracy of 98.68% and 99%, respectively.
- Conclusion: Based on the results, Random Forest emerged as the most accurate individual algorithm across both datasets, achieving 100% accuracy in the WBCD. The combined algorithms also demonstrated a significant improvement in diagnostic performance, with combinations like RF+KNN+J48 achieving a 99% accuracy, which is a notable enhancement compared to individual algorithms. These results highlight the potential of machine learning techniques in the early and accurate detection of breast cancer

Conclusion

In this study, the accuracy, precision, recall, and performance of various machine learning algorithms for breast cancer detection were examined. Using two datasets, Ljubljana and Wisconsin Breast Cancer Dataset (WBCD), the results showed that combined algorithms outperformed individual ones. For the Ljubljana dataset, the best performance was achieved by the Random Forest algorithm, which attained an accuracy of 97.9%, with a high TP rate of 94.12% and a low FP rate of 0.0050. Additionally, Random Forest demonstrated high precision and recall, indicating its ability to identify malignant tumors with minimal false positives. The AUC for Random Forest was also impressive, suggesting strong classification ability. Moreover, the Random Tree and J48 algorithms also yielded satisfactory results with an accuracy of 97.9%, though slightly lower than Random Forest in terms of precision and recall. For the WBCD dataset, Random Forest demonstrated the best performance with 100% accuracy, a TP rate of 1, and an FP rate of 0, showcasing its high precision and ability to distinguish between benign and malignant tumors. The AUC for Random Forest in this dataset was also optimal, reinforcing its effectiveness. In the combined algorithms, the RF+KNN+J48 combination achieved an accuracy of 99%, offering a good balance of sensitivity and specificity. The precision,

recall, and AUC values for this combination were all significantly high, making it one of the most reliable approaches. Models like Logistic Regression+J48+RF and J48+RF also showed similar performance, achieving accuracies of 98.68% and 99%, respectively, with comparable precision, recall, and AUC values. The final conclusion is that Random Forest was the best individual algorithm across both datasets, and the combination of algorithms, especially RF+KNN+J48, increased breast cancer detection accuracy to 99%. The strong precision, recall, and AUC values for the combined models demonstrate the high efficacy of machine learning in the early detection of breast cancer and the potential of these methods for providing more precise models in the future.

References

- [1] W. H. Organization. "World Health Organization." <https://www.who.int/> (accessed).
- [2] W. H. Organization. "Breast Cancer." <https://www.who.int/news-room/fact-sheets/detail/breast-cancer> (accessed ٢٠٢٤, ٢٠٢٤ دسامبر).
- [3] L. Feng, L. Guo, Y. Tanaka, and L. Su, "Tumor-derived small extracellular vesicles involved in breast cancer progression and drug resistance," *International Journal of Molecular Sciences*, vol. 23, no. 23, p. 15236, 2022. <https://doi.org/10.3390/ijms232315236>
- [4] S. M. Mousavi *et al.*, "Breast cancer in Iran: an epidemiological review," *The breast journal*, vol. 13, no. 4, pp. 383–391, 2007. <https://doi.org/10.1155/2017/2481021>
- [5] C. B. Abdallah and A. N. S. Moh, "Machine Learning-Based Algorithms for Breast Cancer Prediction," in *2023 IEEE International Conference on Networking, Sensing and Control (ICNSC)*, 2023, vol. 1: IEEE, pp. 1–6. <https://doi.org/10.1109/ICNSC58704.2023.10319051>
- [6] R. K. Yadav, P. Singh, and P. Kashtriya, "Diagnosis of breast cancer using machine learning techniques-a survey," *Procedia Computer Science*, vol. 218, pp. 1434–1443, 2023. <https://doi.org/10.1016/j.procs.2023.01.122>
- [7] T. Mahesh, O. Geman, M. Margala, and M. Guduri, "The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthcare Analytics*, vol. 4, p. 100247, 2023. <https://doi.org/10.1016/j.health.2023.100247>
- [8] S. K. Maliha, R. R. Ema, S. K. Ghosh, H. Ahmed, M. R. J. Mollick, and T. Islam, "Cancer disease prediction using naive bayes, K-nearest neighbor and J48 algorithm," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2019: IEEE, pp. 1–7. <https://doi.org/10.1109/ICCCNT45670.2019.8944686>
- [9] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," ed: July, 2009. <https://doi.org/10.1287/opre.43.4.570>
- [10] W. H. Wolberg, W. N. Street, D. M. Heisey, and O. L. Mangasarian, "Computerized breast cancer diagnosis and prognosis from fine-needle aspirates," *Archives of*

- Surgery*, vol. 130, no. 5, pp. 511–516, 1995.
<https://doi.org/10.1001/archsurg.1995.01430050061010>
- [11] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, "Machine learning with applications in breast cancer diagnosis and prognosis," *Designs*, vol. 2, no. 2, p. 13, 2018. <https://doi.org/10.3390/designs2020013>
- [12] B. Dai, R.-C. Chen, S.-Z. Zhu, and W.-W. Zhang, "Using random forest algorithm for breast cancer diagnosis," in *2018 International symposium on computer, consumer and control (IS3C)*, 2018: IEEE, pp. 449–452.
<https://doi.org/10.1109/IS3C.2018.00119>
- [13] A. B. Nassif, M. A. Talib, Q. Nasir, Y. Afadar, and O. Elgendy, "Breast cancer detection using artificial intelligence techniques: A systematic literature review," *Artificial intelligence in medicine*, vol. 127, p. 102276, 2022.
<https://doi.org/10.1016/j.artmed.2022.102276>
- [14] A. Sahu, P. K. Das, S. Meher, R. Panda, and A. Abraham, "An efficient deep learning-based breast cancer detection scheme with small datasets," in *International Conference on Intelligent Systems Design and Applications*, 2022: Springer, pp. 39–48. https://doi.org/10.1007/978-3-031-35510-3_5
- [15] M. A. Naji, S. El Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouahid, and O. Debauche, "Machine learning algorithms for breast cancer prediction and diagnosis," *Procedia Computer Science*, vol. 191, pp. 487–492, 2021.
<https://doi.org/10.1016/j.procs.2021.07.062>
- [16] V. Nemade and V. Fegade, "Machine learning techniques for breast cancer prediction," *Procedia Computer Science*, vol. 218, pp. 1314–1320, 2023.
<https://doi.org/10.1016/j.procs.2023.01.110>
- [17] P. Mounika and R. Pitchai, "Breast Cancer Detection Using Machine Learning Algorithms," in *2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN)*, 2023: IEEE, pp. 231–236.
<https://doi.org/10.1109/CICN59264.2023.10402232>
- [18] M. Shanbehzadeh, H. Kazemi-Arpanahi, M. B. Ghalibaf, and A. Orooji, "Performance evaluation of machine learning for breast cancer diagnosis: A case study," *Informatics in Medicine Unlocked*, vol. 31, p. 101009, 2022.
<https://doi.org/10.1016/j.imu.2022.101009>
- [19] Y. Jin, A. Lan, Y. Dai, L. Jiang, and S. Liu, "Development and testing of a random forest-based machine learning model for predicting events among breast cancer patients with a poor response to neoadjuvant chemotherapy," *European Journal of Medical Research*, vol. 28, no. 1, p. 394, 2023.
<https://doi.org/10.1186/s40001-023-01361-7>
- [20] G. Alfian *et al.*, "Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method," *Computers*, vol. 11, no. 9, p. 136, 2022.
<https://doi.org/10.3390/computers11090136>
- [21] S. Laghmati, S. Hamida, K. Hicham, B. Cherradi, and A. Tmiri, "An improved breast cancer disease prediction system using ML and PCA," *Multimedia Tools and Applications*, vol. 83, no. 11, pp. 33785–33821, 2024.
<https://doi.org/10.1177/20552076251332738>
- [22] F. Hoseini and H. Sepehrzadeh, "Breast Cancer Diagnosis based on Frequency Converters and Extraction of Effective Features," *Journal of Engineering and*

- Applied Research*, vol. 1, no. 2, pp. 65–78, 2024.
<https://doi.org/10.48301/jear.2024.458787.1026>
- [23] R. Naveenkumar, V. K. Joshi, B. Pramanik, and R. Sarkar, "An exploratory research and useful information on enhancing the effectiveness of Zero-R and One-R algorithms in data classification methods." <https://doi.org/10.48047/AFJBS.6.14.2024.11255-11272>
- [24] M. Raj and J. R. Saini, "A Novel Comparison of Charotar Region Wheat Variety Classification Techniques using Purely Tree-based Data Mining Algorithms," *Procedia Computer Science*, vol. 235, pp. 568–577, 2024.
<https://doi.org/10.1016/j.procs.2024.04.056>
- [25] I. Wickramasinghe and H. Kalutarage, "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation," *Soft Computing*, vol. 25, no. 3, pp. 2277–2293, 2021.
<https://doi.org/10.1177/20552076251332738>
- [26] D. M. Abdullah and A. M. Abdulazeez, "Machine learning applications based on SVM classification a review," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 81–90, 2021. <https://doi.org/10.48161/qaj.v1n2a50>
- [27] t. U. M. C. Ljubljana, 1980 1980. [Online]. Available: <https://www.kaggle.com/datasets/faysalmiah1721758/breast-cancer-data>.
- [28] E. Elbasi and A. I. Zreikat, "Efficient early prediction and diagnosis of diseases using machine learning algorithms for IoMT data," in *2021 IEEE World AI IoT Congress (AllIoT)*, 2021: IEEE, pp. 0155–0159.
<https://doi.org/10.1109/AllIoT52608.2021.9454231>