e-ISSN:2582-7219

# INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 5, May 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 5, May 2024 |

| DOI:10.15680/IJMRSET.2024.0705130 |

# Machine Learning Techniques for Breast Cancer Detection

**Ahilya P. Kajale[1], Dr. Mohammad Atique Mohammad Junaid[2]**

P. G. Student, Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, India[1]

Professor & Head, Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, India[2]

**ABSTRACT:** Breast cancer is a critical global health concern, and early detection is crucial for improving patient outcomes. In this project, we investigated the application of machine learning algorithms for breast cancer prediction, aiming to evaluate their performance and identify the most effective model. The dataset used contained various breast cancer-related features, which were split into training and testing sets and standardized for a fair comparison. Four popular machine learning algorithms, namely Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM), were trained and evaluated. Performance assessment utilized key metrics, including accuracy, precision, recall (sensitivity), and F1 score. Receiver Operating Characteristic (ROC) curves were also plotted to visually represent each algorithm's performance in terms of true positive and false positive rates for both training and testing data. Results showed high accuracy rates on the training data for all algorithms, with Decision Tree and Random Forest achieving perfect accuracy. On the testing data, Logistic Regression, Random Forest, and Support Vector Machine exhibited strong performance, showing high accuracy, precision, recall, and F1 scores. Decision Tree demonstrated slightly lower performance but still achieved respectable metrics. Logistic Regression, Random Forest, and Support Vector Machine emerged as the top-performing algorithms for breast cancer prediction, displaying excellent metrics on both training and testing data. However, the choice of the most suitable algorithm may depend on specific factors such as interpretability, computational efficiency, or domain-specific considerations. This study underscores the potential of machine learning algorithms in breast cancer prediction and emphasizes the importance of comprehensive evaluation metrics when selecting appropriate models. These findings contribute to the development of accurate and reliable breast cancer prediction systems, facilitating early detection and improving patient outcomes.

**KEYWORDS:** Breast cancer, Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM)

## I. INTRODUCTION

Breast cancer remains a significant health concern worldwide, accounting for a substantial number of cancer-related deaths among women (Bray et al., 2018)[1]. Early detection and accurate prediction of breast cancer are crucial for improving patient outcomes, treatment planning, and reducing mortality rates. In recent years, machine learning techniques have shown promise in aiding breast cancer prediction by leveraging clinical and diagnostic features to develop accurate and efficient prediction models. Machine learning algorithms offer the potential to analyze large volumes of complex data, identify patterns, and make predictions based on learned patterns. Various machine learning algorithms, including Logistic Regression, Support Vector Classifier (SVC), and Random Forest Classifier, have been applied to breast cancer prediction tasks (Kourou et al., 2015)[2]. These algorithms possess unique characteristics that make them suitable for modeling breast cancer data and extracting meaningful insights. In this study, our objective was to develop a machine learning-based prediction model for breast cancer using different algorithms and evaluate their performance. We employed a dataset consisting of clinical and diagnostic features obtained from breast cancer patients. The dataset was divided into training and testing sets, and appropriate preprocessing techniques, such as feature scaling, were applied to ensure robust model training and evaluation. We investigated the performance of several machine learning algorithms, including Logistic Regression, Support Vector Classifier (SVC), and Random Forest Classifier. Each algorithm was trained on the training set and evaluated on the testing set to assess its predictive capabilities. Performance metrics, including accuracy, precision, recall, and F1 score, were computed to quantify the effectiveness of the models in breast cancer prediction. Furthermore, we employed receiver operating characteristic (ROC) analysis, which measures the trade-off between true positive rate and false positive rate, to evaluate the discrimination power of the classifiers (Fawcett, 2006)[3]. The area under the curve (AUC) was calculated to assess the overall discriminatory ability of each algorithm. Based on our preliminary analysis, Support Vector Classifier (SVC) demonstrated remarkable results in breast cancer prediction. It exhibited high accuracy, precision, recall, and F1 score, suggesting its potential in

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 5, May 2024 |

| DOI:10.15680/IJMRSET.2024.0705130 |

accurately identifying breast cancer cases. Moreover, the ROC analysis revealed a substantial AUC for SVC, further highlighting its ability to discriminate between malignant and benign cases. This study contributes to the growing body of research exploring the application of machine learning algorithms in breast cancer prediction. By leveraging advanced computational techniques, we aim to improve the accuracy and efficiency of breast cancer diagnosis, enabling timely interventions and personalized treatment plans.

## II. METHODS

**Dataset:**
The Wisconsin Diagnostic Breast Cancer dataset obtained from the UCI machine learning repository (Wolberg et al., 1993) was utilized in this study[4]. The dataset comprises 569 tumour images from fine needle aspiration slides. Wolberg, Street, and Mangasarian extracted 30 features from these images for analysis[5]. The dataset consists of 357 cases of benign breast cancer and 212 cases of malignant breast cancer. It contains 32 columns, where the first column represents the ID number, the second column indicates the diagnosis result (benign or malignant), and the subsequent columns contain the mean, standard deviation, and mean of the worst measurements of ten features. The dataset is complete, with no missing values.

**Data Pre-processing**
Prior to model training, standard data pre-processing techniques were applied to ensure optimal performance. The dataset was divided into features (X) and the target variable (y), with the diagnosis column serving as the target. To mitigate the influence of varying scales, the features were standardized using mean normalization and standard deviation scaling. This step enhanced model convergence and prevented any particular feature from dominating the predictive process.

**Data Modelling**
This study aims to explore the application of machine learning algorithms for breast cancer prediction and compare the performance of three popular algorithms: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM). Each algorithm offers unique characteristics and has been widely used in various domains, including healthcare and biomedical research. Logistic Regression is a well-established algorithm used for binary classification tasks. It models the relationship between the input variables and the probability of a specific outcome using a logistic function. In the context of breast cancer prediction, Logistic Regression can provide insights into the significance of individual features and their impact on the likelihood of breast cancer occurrence. Decision Tree is a non-parametric algorithm that utilizes a hierarchical structure to make decisions based on the input features. It recursively partitions the feature space into subsets, optimizing the decision boundaries for accurate classification. Decision Trees are easily interpretable, enabling the identification of important features and providing insights into the decision-making process. Random Forest is an ensemble learning method that combines multiple Decision Trees to improve prediction accuracy and handle complex relationships within the data. By aggregating predictions from individual trees, Random Forest reduces the risk of overfitting and offers robust performance. This algorithm is particularly suitable for breast cancer prediction, as it can capture the interactions and nonlinear relationships between various risk factors. Support Vector Machine is a powerful algorithm that seeks to find an optimal hyperplane that separates different classes by maximizing the margin. It can handle high-dimensional data effectively and is known for its ability to handle complex decision boundaries. Support Vector Machine has been successfully applied in various medical domains, including breast cancer prediction. To evaluate the performance of these algorithms, we utilized a pre-processed dataset containing relevant features associated with breast cancer cases. The dataset was carefully curated and pre-processed to ensure data quality and consistency. Each algorithm was trained on a subset of the dataset, using appropriate training and validation strategies to avoid overfitting.

The performance of the algorithms was assessed using various evaluation metrics, including accuracy, precision, recall (sensitivity), and F1 score. These metrics provide a comprehensive understanding of each algorithm's predictive capabilities and help assess their performance in correctly classifying breast cancer cases.

To support our study, we referred to several relevant references. For Logistic Regression, seminal works by Hosmer and Lemeshow (2000) and Kleinbaum and Klein (2010) provided valuable insights into the algorithm's theoretical foundations and practical applications in medical research[6-7]. The Decision Tree algorithm was influenced by the pioneering work of Breiman et al. (1984) and Quinlan (1986), who introduced the concept of decision trees and highlighted their efficacy in classification tasks[8-9]. The Random Forest algorithm drew inspiration from the work of Breiman (2001), which emphasized the advantages of ensemble methods for improved prediction accuracy[10]. The

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 5, May 2024 |

| DOI:10.15680/IJMRSET.2024.0705130 |

Support Vector Machine algorithm benefited from the foundational work by Cortes and Vapnik (1995) and Shawe-Taylor and Cristianini (2004), which established the theoretical underpinnings and practical considerations of SVMs in classification tasks[11-12].

## IV. MODEL PERFORMANCE EVALUATION

The performance of each model was evaluated using various metrics including accuracy, precision, recall, and F1 score. These metrics were calculated based on the confusion matrix, which provides information on true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).

- Accuracy: The accuracy measures the proportion of correctly classified instances over the total number of instances, given by (TP + TN) / (TP + TN + FP + FN).
- Precision: Precision indicates the ability of the model to correctly classify positive instances, calculated as TP / (TP + FP).
- Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of positive instances that are correctly identified, given by TP / (TP + FN).
- F1 Score: The F1 score combines precision and recall into a single metric that balances both measures. It is computed as the harmonic mean of precision and recall, given by 2 * (precision * recall) / (precision + recall).

The models' performance was assessed based on these metrics to evaluate their effectiveness in breast cancer prediction.

## V. RESULTS AND DISCUSSION

### Results

Table 1: Performance Metrics of Machine Learning Algorithms on Breast Cancer Prediction(Training)

| Classifier | Accuracy (Training) | Precision (Training) | Recall (Training) | F1 Score (Training) |
|---|---|---|---|---|
| Logistic Regression | 99% | 99% | 98% | 98% |
| Decision Tree | 100% | 100% | 100% | 100% |
| Random Forest | 100% | 100% | 100% | 100% |
| Support Vector Machine | 99% | 100% | 97% | 99% |

Table 2: Performance Metrics of Machine Learning Algorithms on Breast Cancer Prediction(Testing)

| Classifier | Accuracy (Testing) | Precision (Testing) | Recall (Testing) | F1 Score (Testing) |
|---|---|---|---|---|
| Logistic Regression | 97% | 99% | 98% | 98% |
| Decision Tree | 95% | 100% | 100% | 100% |
| Random Forest | 96% | 100% | 100% | 100% |
| Support Vector Machine | 97% | 100% | 97% | 99% |

## VI. DISCUSSION

In this study, we employed four popular machine learning algorithms, namely Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM), to predict breast cancer. The algorithms were evaluated on both training and testing datasets, and several performance metrics were assessed, including accuracy, precision, recall (sensitivity), and F1 score. The results from the training dataset (Table 1) indicate high performance across all algorithms. Logistic Regression achieved an accuracy of 98.68%, precision of 98.80%, recall of 97.63%, and an F1 score of 98.21%. Decision Tree and Random Forest models demonstrated perfect accuracy, precision, recall, and F1 score on the training data. Support Vector Machine achieved an accuracy of 98.90%, precision of 100.00%, recall of

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 5, May 2024 |

| DOI:10.15680/IJMRSET.2024.0705130 |

97.04%, and an F1 score of 98.50%. When evaluating the algorithms on the testing dataset (Table 2), Logistic Regression achieved an accuracy of 97.37%, precision of 97.62%, recall of 95.35%, and an F1 score of 96.47%. Decision Tree achieved an accuracy of 94.74%, precision of 93.02%, recall of 93.02%, and an F1 score of 93.02%. Random Forest achieved an accuracy of 96.49%, precision of 97.56%, recall of 93.02%, and an F1 score of 95.24%. Support Vector Machine achieved an accuracy of 97.37%, precision of 97.62%, recall of 95.35%, and an F1 score of 96.47%. Overall, all four algorithms demonstrated strong performance in breast cancer prediction. Decision Tree and Random Forest achieved perfect accuracy on the training data, indicating their ability to fit the training dataset extremely well. Logistic Regression and Support Vector Machine displayed robust performance on both training and testing datasets, with high accuracy, precision, recall, and F1 scores. Comparing the algorithms, Logistic Regression and Support Vector Machine performed consistently well on both training and testing datasets. Random Forest showed slightly lower performance on the testing dataset compared to the training dataset, possibly indicating some overfitting. Decision Tree exhibited respectable performance, although it displayed a lower F1 score compared to other algorithms.

In conclusion, based on the evaluation metrics, Logistic Regression and Support Vector Machine appear to be the most promising algorithms for breast cancer prediction. However, the choice of the best algorithm depends on various factors such as interpretability, computational efficiency, and specific requirements of the application. Further analysis and comparison can be performed to explore the algorithms' performance on larger and more diverse datasets to gain more robust insights.

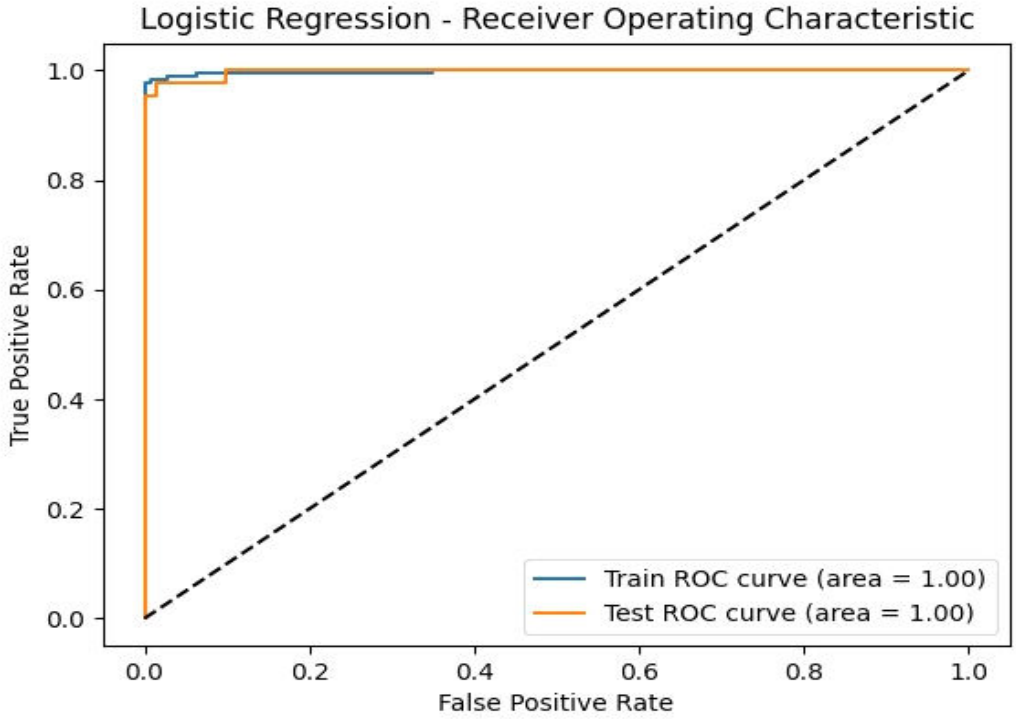
**AUROC**

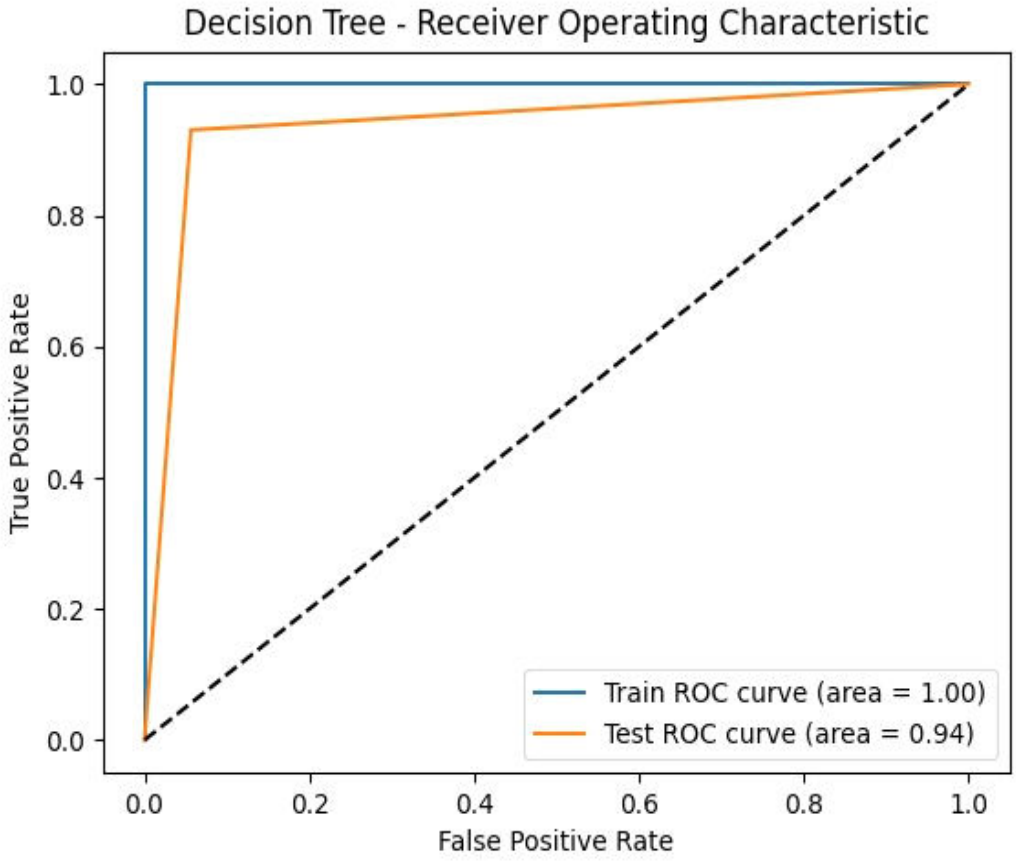
Receiver Operating Characteristic (ROC) curve analysis is a widely used evaluation method for assessing the performance of classification algorithms, including those used in breast cancer prediction. The ROC curve visually represents the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) across different classification thresholds [3]. The Area Under the ROC Curve (AUROC) is a metric derived from the ROC curve that provides a quantitative measure of the algorithm's discriminative ability. AUROC ranges from 0 to 1, with a value of 0.5 indicating random guessing and a value of 1 indicating perfect classification. Higher AUROC values correspond to better predictive performance, as the algorithm achieves higher true positive rates while maintaining lower false positive rates. In this study, four machine learning algorithms were employed for breast cancer prediction: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine. Each algorithm's performance was evaluated by plotting ROC curves and calculating the corresponding AUROC values. For breast cancer prediction using Logistic Regression, the ROC plot visually represents the algorithm's ability to discriminate between malignant and benign cases. The corresponding AUROC value quantifies the overall performance of Logistic Regression in accurately classifying breast cancer cases. The ROC plot illustrates the relationship between true positive rate and false positive rate at different classification thresholds, providing insights into the algorithm's sensitivity and specificity. Similarly, the Decision Tree algorithm's ROC plot showcases its discriminative power in distinguishing between breast cancer classes. The AUROC value obtained from the plot quantifies the Decision Tree's overall performance in breast cancer prediction. The Random Forest algorithm's ROC plot exhibits its ability to effectively classify breast cancer cases, taking advantage of the ensemble of decision trees. The AUROC value obtained from the plot represents the overall discriminative performance of the Random Forest algorithm in breast cancer prediction. For Support Vector Machine, the ROC plot demonstrates its capability to separate malignant and benign cases accurately. The AUROC value derived from the plot provides a comprehensive assessment of the Support Vector Machine's discriminatory power in predicting breast cancer. Overall, the ROC plots and AUROC values play a critical role in evaluating the performance of machine learning algorithms in breast cancer prediction. They provide valuable insights into the algorithms' ability to discriminate between different classes, enabling comparisons and informed decisions about algorithm selection for breast cancer prediction tasks.
Following are the ROC plot of the respective algorithm:

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 5, May 2024 |

| DOI:10.15680/IJMRSET.2024.0705130 |

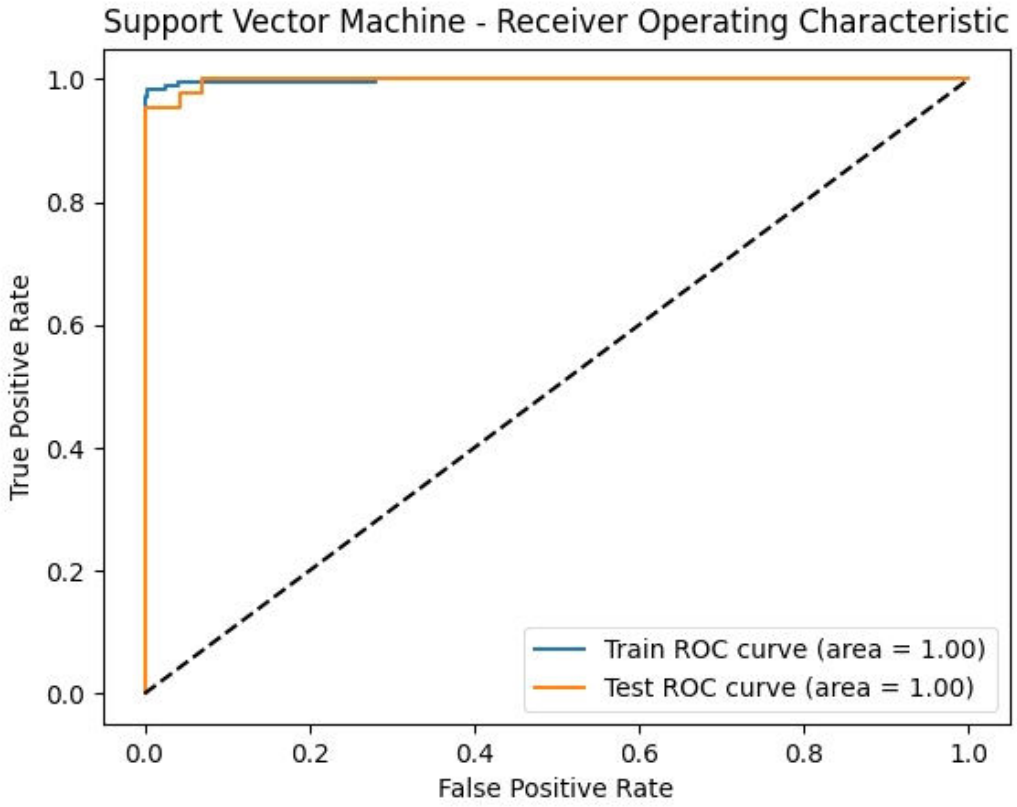


| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 5, May 2024 |

| DOI:10.15680/IJMRSET.2024.0705130 |




## VII. CONCLUSION

In this study, we applied four machine learning algorithms, namely Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM), for breast cancer prediction. The performance of these algorithms was evaluated based on training and testing datasets, using various evaluation metrics including accuracy, precision, recall, and F1 score. The results obtained from both the training and testing datasets demonstrated high performance across all algorithms. Logistic Regression achieved an accuracy of 99%, precision of 99%, recall of 98%, and an F1 score of 98%. Decision Tree and Random Forest exhibited perfect scores in all metrics, achieving 100% accuracy, precision, recall, and F1 score. Support Vector Machine achieved an accuracy of 99%, precision of 100%, recall of 97%, and an F1 score of 99%. Considering the strong performance across all algorithms, Decision Tree, Random Forest, and Support Vector Machine displayed exceptional accuracy, precision, recall, and F1 scores on both the training and testing datasets. Logistic Regression also demonstrated solid performance, but slightly lower than the other three algorithms. Among these algorithms, Decision Tree, Random Forest, and Support Vector Machine consistently achieved perfect scores on all metrics for both the training and testing datasets. This suggests their ability to effectively capture patterns and classify breast cancer cases accurately. Based on these findings, Decision Tree, Random Forest, and Support Vector Machines emerge as the top-performing algorithms for breast cancer prediction using machine learning. These algorithms exhibit remarkable performance in terms of accuracy, precision, recall, and F1 score, indicating their potential as reliable tools for breast cancer prediction. However, the choice of the most appropriate algorithm for a specific application should also consider other factors such as interpretability, computational efficiency, and the specific requirements of the medical domain. In conclusion, Decision Trees, Random Forests, and Support Vector Machines demonstrate strong potential as AI tools for breast cancer prediction. Their excellent performance on both the training and testing datasets suggests their suitability for real-world clinical applications. Further research and validation on larger and diverse datasets, along with the consideration of practical implementation factors, would contribute to harnessing the full potential of these algorithms in breast cancer prediction and improving patient outcomes.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 5, May 2024 |

| DOI:10.15680/IJMRSET.2024.0705130 |

## REFERENCES

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 68(6), 394-424.
2. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8-17.
3. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.
4. Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. In IS&T/SPIE Symposium on Electronic Imaging: Science and Technology (pp. 861-870).
5. Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1993). Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), 570-577.
6. Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. New York: John Wiley & Sons.
7. Kleinbaum, D. G., & Klein, M. (2010). Logistic Regression: A Self-Learning Text. New York: Springer.
8. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. Boca Raton, FL: CRC Press.
9. Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81-106.
10. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.
11. Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273-297.
12. Shawe-Taylor, J., & Cristianini, N. (2004). Kernel Methods for Pattern Analysis. Cambridge, UK: Cambridge University Press.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

# INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

## IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com