© NOV 2025 | IRE Journals | Volume 9 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1712217

Prediction of Breast Cancer Using Machine Learning Techniques

YUSUF MAGAJI MADA[1], ABDULMALIK ABDULLAHI[2]
[1]*Department of Computer Engineering, Abdu Gusau Polytechnic Talata Mafara*
[2]*Department of Computer science, Abdu Gusua Polytechnic Talata Mafara*

Abstract- ***Breast cancer is a highly dangerous disease affecting women worldwide. Its prevalence is rapidly increasing, emphasizing the importance of early detection and preventive measures. This study focuses on utilizing various machine learning classification algorithms, including Support Vector Machine (SVM), Logistic Regression (LR), K Nearest Neighbor (KNN), Random Forest Classifier (RFC), and Decision Tree Classifier, to predict breast cancer in women. We will assess and compare these classifiers using metrics like accuracy, precision, recall, and f1-Score. We will utilize the breast cancer dataset from the UCI Machine Learning Repository, which is accessible to the public. The data will be split, with 80% allocated for training and 20% for testing purposes. Based on the outcomes, it is evident that the Random Forest Classifier outperformed the other classifiers across all evaluation criteria.***

Index Terms Machine Learning, Breast Cancer, Classification, Accuracy, and Precision

I. INTRODUCTION

Women should be aware of their breast cancer risk factors and discuss any concerns with their healthcare provider. These risk factors include age, an individual's family background, genetic variations, reproductive experiences, and lifestyle choices are all factors that can influence their health and well-being. like alcohol consumption and physical activity levels. Breast cancer is a common cause of death worldwide and primarily affects women. The causes of breast cancer are multifaceted, involving factors such as family genetics, hormonal influences, radiation therapy, and reproductive factors. Every year, the World Health Organization (WHO) releases a yearly publication known as the World Health Report. that provides information on various health issues, including breast cancer. (WHO, n.d.) The 2021 World Health Report states that breast cancer globally, is the most prevalent form of cancer identified in women, with approximately 2.3 million new cases reported. identified solely in 2021. The report emphasizes the significance of early detection and screening for breast cancer, as timely diagnosis can enhance treatment effectiveness and increase survival rates. It also underscores the importance of ensuring that women have access to high-quality healthcare and treatment, regardless of their socioeconomic status or geographical location. Numerous imaging techniques have been developed to detect and treat breast cancer at an early stage, aiming to reduce mortality rates. Assisted breast cancer diagnosis methods are employed to improve diagnostic accuracy. (Singh, 2020) Machine Learning algorithms play a significant role in intelligent healthcare systems. This research paper compares various machine learning algorithms, such as k Nearest Neighbors, Support Vector Machine, Logistic Regression, Random Forest classification, for diagnosing and predicting outcomes of breast cancer.

II. MACHINE LEARNING

The breast cancer classification model utilizes machine learning algorithms to analyze a breast cancer dataset. The model extracts relevant features and undergoes training to predict whether a case is benign or malignant. Benign cases are deemed noncancerous and pose no significant threat. On the other hand, malignant cancer originates from abnormal cell growth and has the potential to rapidly spread or invade nearby tissues, making it dangerous.

A. K NEAREST NEIGHBOR (KNN)

The k Nearest Neighbors algorithm utilizes the idea of "similarity in features" to make predictions for the values of new data points. It determines the value of a new data point by comparing its similarity to the points in the training set. The closer the match, the more likely the new data point will be assigned a similar value as those in the training set. (L. T. S. M. S. A. Yu, 2021)

B. LOGISTIC REGRESSION (LR)

Logistic Regression is a well-known classification technique in the field of machine learning (Rajendran, 2020) that falls under the category of linear classifiers and shares similarities with polynomial and statistical regression. It is known for its speed, simplicity, and ease of interpretation. While it is primarily used for binary classification, it can also be applied to multi-class problems. Unlike statistical regression, which deals with the prediction of continuous values, logistic regression models the probability of a response falling into a specific category. By utilizing the Sigmoid function, a logistic regression model addresses scenarios where the output can only take two values, 0 or 1.

C. RANDOM FOREST CLASSIFIER (RFC)

Random Forest Classifier is a classification technique based on random forest' Theorem, (Hu, 2018) A Random Forest Classifier is a machine learning algorithm that belongs to the ensemble learning family. It combines multiple decision trees to create a powerful and robust model for classification tasks. Every decision tree within the random forest is taught using a distinct portion of the training data and produces its own set of predictions. The ultimate classification is arrived at by combining the predictions from all trees, which can be done through voting or averaging. This technique aids in diminishing overfitting, leading to enhanced overall accuracy and better generalization of the model.

D. SUPPORT VECTOR MACHINE (SVM)

Support Vector Machines (SVMs) are supervised AI models that create a hyperplane when classifying items. The hyperplane acts as a line on a plane to distinguish between two classes. SVM is a non-probabilistic, linear classifier. By using a training algorithm, an SVM generates a model that assigns new instances to one or more categories or vice versa, based on a set of training samples, each of which is labeled as belonging to one or more classes. (L. T. L. L. X. C. Z. Y. and T. S. K. Yu, 2021)

E. DECISION TREE CLASSIFIER

The Decision Tree Classifier is a machine learning technique that utilizes a hierarchical arrangement to make decisions or forecasts. It is an algorithm used mainly for classification purposes in supervised learning, although it can also be employed for regression tasks. The decision tree is formed by dividing the input data repeatedly, considering various features and their corresponding values. Each internal node within the tree signifies a decision made using a particular feature, while each leaf node represents a class label or a prediction.

III. ABOUT THE DATA SET

This paper utilizes a publicly available dataset sourced The UCI Machine Learning Repository provides a dataset comprising numerous records of human cell tests., each containing measurements of a set of cell characteristics.

The ID Number attribute holds the unique identifiers for patients. The attributes Clump Thickness to Mitoses encompass the characteristics of cell tests for each patient. These values range from 1 to 10, with 1 indicating the closest to normal. The class field indicates the diagnosis based on clinical procedures. In this dataset, a value of 2 corresponds to a benign condition, while a value of 4 indicates a malignant condition.

IV. LITERATURE REVIEW

The authors of this study (Nassif et al., 2022) provided a summary of various techniques employed in classifying breast cancer using histopathological image analysis (HIA) and different architectures of Artificial Neural Networks (ANN). They organized their work based on the datasets used, arranging them in chronological order. Their analysis revealed that ANNs were first utilized in HIA around 2012, with ANNs and PNNs being the most commonly applied algorithms. However, most studies focused on extracting textural and morphological features for

feature extraction. Deep CNNs were found to be highly effective in early detection and diagnosis of breast cancer, leading to improved treatment outcomes. Additionally, the prediction of Non-Communicable Diseases (NCDs) was explored using various algorithms.

In another study, (Chen et al., 2023) researchers developed an intelligent method for breast cancer detection using a support vector machine (SVM) classifier and an artificial neural network (ANN). They employed the Wisconsin Diagnostic dataset to build an SVM model capable of distinguishing between benign and malignant breast clusters obtained from needle aspirates (FNA). Extensive research has been conducted to compare traditional statistical methods with standard machine learning (ML) classification approaches to assess the reliability and potential of ML. The results showed that ML methods exhibited the highest reliability due to advancements in AI techniques and the increasing volume and complexity of data.

In a different approach described in (Jia et al., 2022), a team-based strategy was employed, combining multiple models to achieve improved accuracy across different types of classes. This approach utilized SVM, naive Bayes, classifiers with a democratic classifier methodology, resulting in an accuracy of 89.13%, surpassing the performance of individual classifiers.

The study utilized a dataset consisting of 912 ultrasound images, with a total of 185 features extracted from these images. The images were classified into two categories: malignant and benign tumors. The extracted features were saved as tabular data in CSV format. Additionally, a framework for computer-aided diagnosis (CAD) was presented with the aim of supporting radiologists in the classification of breast ultrasound images as either benign or malignant tumors.

To assess the performance of the proposed framework, five classifiers were employed: k-NN, SVM, RF, XG Boost, and Light GBM. The framework's performance was evaluated using 10-fold cross-validation, and Bayesian optimization with a tree-structured Parzen estimator was applied. Among the five classifiers, the Light GBM classifier achieved the highest accuracy, precision, recall, and F1-score. The experiment results showed that the Light GBM classifier outperformed the other classifiers, with accuracy, precision, recall, and F1-score values of 99.86%, 100.00%, 99.60%, and 99.80%, respectively. (Jia et al., 2022)

According to the provided methodology in (K & L, 2021), the numerical dataset was analyzed using four distinct algorithms: SVM, KNN, Decision tree (CART), and Naïve Bayes. The results of this analysis yielded the following outcomes. Initially, the baseline algorithm analysis was conducted on the dataset, which resulted in accuracy values obtained by these classifier algorithms. Subsequently, to enhance the accuracy further, standard scaling was applied to center and scale the data. Additionally, regularization techniques were employed in conjunction with the algorithm.

V. METHODOLOGY

The breast cancer classification model utilizes machine learning algorithms to analyze and predict the nature of breast tumors. In this process, a breast cancer dataset is first loaded, and relevant features are extracted from the data. These features are then used to train the classification model, which is capable of distinguishing between benign and malignant tumors. Benign tumors are non-cancerous and generally considered harmless, as they do not spread to other parts of the body. On the other hand, malignant tumors are cancerous and arise from abnormal cell growth. Malignant cancer cells have the potential to grow uncontrollably and spread to surrounding tissues or other parts of the body, making them dangerous and often life-threatening. By training the model on such data, it becomes capable of accurately predicting whether a tumor is benign or malignant, which is crucial for early detection and appropriate

© NOV 2025 | IRE Journals | Volume 9 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1712217

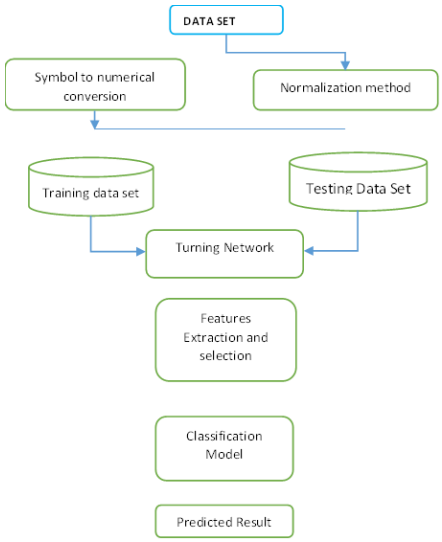

Figure 1 model flow chart

VI. ACCURRACY

The classifier's accuracy assesses its ability to correctly predict classifications for cases, and it is computed by dividing the number of accurate predictions by the total number of instances in the dataset. It is crucial to acknowledge that accuracy is influenced by the chosen threshold of the classifier and can differ across various testing sets. Hence, although it may not be the ideal approach to compare different classifiers, it can still provide a summary of their classification performance

The table provided presents the accuracy values corresponding to each of the four machine learning algorithms.

	model_name	score	accuracy_score	accuracy_percentage
0	LogisticRegression	0.916010	0.909574	90.96%
1	RandomForestClassifier	0.992126	0.925532	92.55%
2	DecisionTreeClassifier	1.000000	0.909574	90.96%
3	SVC	0.923885	0.914894	91.49%

Table 1.1

6.1 CONFUSION MATRIX

The confusion matrix is a widely utilized table that showcases the performance of a classification model on a test dataset where the true values are already known.

Fig 1.1 Confusion matrix For Logistic Regression

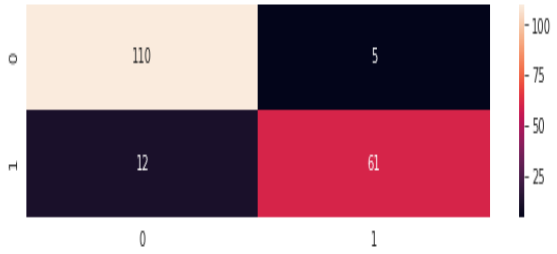

Fig 1.2 Confusion matrix For Random Forest Classifier

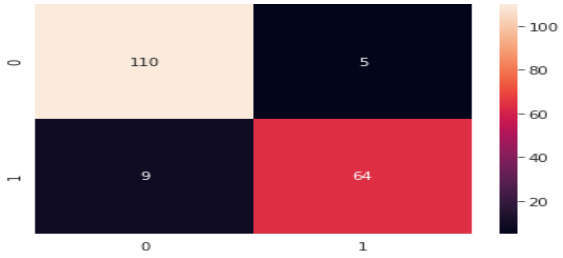

Fig 1.3 Confusion matrix For Decision Tree Classifier.

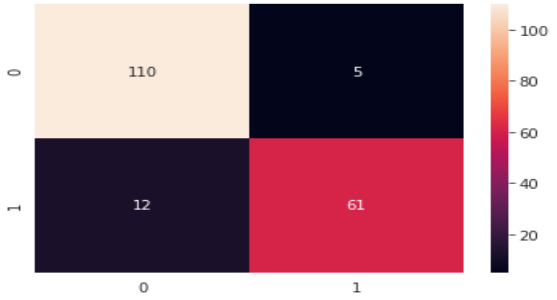

Fig 1.4 Confusion matrix For Support Vector Machine.

© NOV 2025 | IRE Journals | Volume 9 Issue 5 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I5-1712217

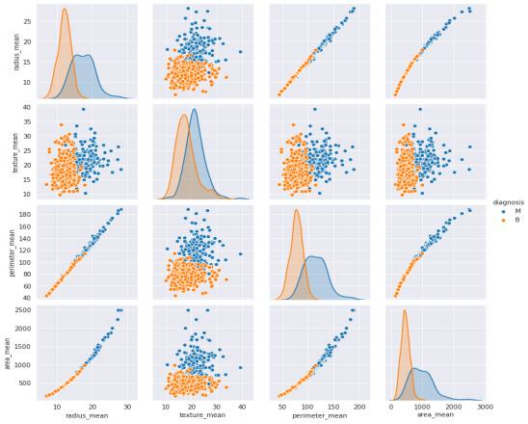

Fig 1.5 An over view of Confusion matrix

Matrices of Evaluation figure 1.5 the confusion matrix (CM) serves as an evaluative tool for AI classification models, providing an overview of their performance. It assesses how often the models correctly predict outcomes and how often they make incorrect predictions. The CM assigns false positives and false negatives to incorrectly predicted values, while true positives and true negatives are assigned to accurately predicted values. To evaluate the model's performance, metrics such as accuracy, precision, recall, and area under the curve (AUC) are calculated after organizing the predicted values in the matrix.

6.2 CLASSIFICATION REPORT

Classification report of various machine learning techniques covering Accuracy, precision, recall fi-score support

Classification Report of 'LogisticRegression '

	precision	recall	f1-score	support
0	0.90	0.96	0.93	115
1	0.92	0.84	0.88	73
accuracy			0.91	188
macro avg	0.91	0.90	0.90	188
weighted avg	0.91	0.91	0.91	188

Classification Report of 'RandomForestClassifier '

	precision	recall	f1-score	support
0	0.92	0.96	0.94	115
1	0.93	0.88	0.90	73
accuracy			0.93	188
macro avg	0.93	0.92	0.92	188
weighted avg	0.93	0.93	0.93	188

Classification Report of 'DecisionTreeClassifier '

	precision	recall	f1-score	support
0	0.90	0.96	0.93	115
1	0.92	0.84	0.88	73
accuracy			0.91	188
macro avg	0.91	0.90	0.90	188
weighted avg	0.91	0.91	0.91	188

Classification Report of 'SVC '

	precision	recall	f1-score	support
0	0.90	0.97	0.93	115
1	0.94	0.84	0.88	73
accuracy			0.91	188
macro avg	0.92	0.90	0.91	188
weighted avg	0.92	0.91	0.91	188

VII. RESULT AND DISCUSSION

From the result obtained Random Forest classifier Achieves the crucial aspect of performance in terms of accuracy, accomplishes the Support Vector Machine become second in performance as far as accuracy, the third and fourth are the Logistic Regression and Decision Tree become third as their all have the same result

The analysis is constrained by the relatively small size of the dataset employed for training and testing purposes. In order to obtain more dependable outcomes in clinical settings, it becomes essential to perform the analysis using a larger dataset.

VIII. CONCLUSION

In this study, four Machine Learning algorithms, specifically Support Vector Machine, Logistic Regression, and Random Forest classifier, Decision tree classifier, are compared. The comparison is carried out using the Wisconsin breast cancer dataset accessible from the UCI Machine Learning Repository. The primary goal of this analysis is to determine the most precise algorithm that can serve as a diagnostic tool for breast cancer. According to the prediction outcomes, the Random Forest classifier demonstrates the highest accuracy among the algorithms considered.

REFERENCES

[1] Chen, H., Wang, N., Du, X., Mei, K., Zhou, Y., & Cai, G. (2023). Classification Prediction of Breast Cancer Based on Machine Learning. *Computational Intelligence and Neuroscience*, *2023*, 1–9. https://doi.org/10.1155/2023/6530719

[2] Hu, L. N.-T. N. W. T. M. C. L. X. F. L. M. R. S. and S. N. A. S. (2018). Modeling of cloud-based digital twins for smart manufacturing with MT connect. *Procedia Manufacturing*, 26–34.

[3] Jia, X., Sun, X., & Zhang, X. (2022). Breast Cancer Identification Using Machine Learning. *Mathematical Problems in Engineering*, *2022*. https://doi.org/10.1155/2022/8122895

[4] K, Y. H., & L, C. M. (2021). *Breast Cancer Prediction Using Machine Learning Techniques*.

[5] K. Yu, L. T. L. L. X. C. Z. Y. and T. S. (2021). Deep-Learning-Empowered Breast Cancer Auxiliary Diagnosis for 5GB Remote E-Health. *IEEE Wireless Communications*, *Vol28*(No3), 54–61.

[6] K. Yu, L. T. S. M. S. A.-R. A. A.-D. A. K. B. F. A. K. "Securing C. I. (2021). Securing Critical Infrastructures: Deep Learning-based Threat Detection in the IIoT. *IEEE Communications Magazine,*

[7] Nassif, A. B., Talib, A., Nasir, Q., Afadar, Y., & Elgendy, O. (2022). Breast cancer detection using artificial intelligence techniques: A systematic literature review. In *Artificial Intelligence in Medicine* (Vol. 127). Elsevier. https://www.cancer.org/

[8] Rajendran, G. B, U. M. K. C. Z. P. B. D. and S. L. Ullo. (2020). Land-use and land-cover classification using a human group-based particle swarm optimization algorithm with an LSTM Classifier on hybrid pre-processing remote-sensing images. *Remote Sensing*, *Vol 24*.

[9] Singh, G. (2020). Breast Cancer Prediction Using Machine Learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 278–284. https://doi.org/10.32628/CSEIT206457

[10] WHO. (n.d.). *Breast Cancer https://www.who.int/cancer/*.