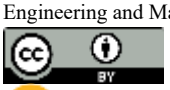
Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and their Analysis

[1] Mr. SK. Sharif , [2] G. ANUSHA, [3] G. NITHISH, [4] P. SUMANJALI
[1]Assistant Professor, [234]Student
[1234]Department of CSE (AI & ML)
[1234]CMR Technical Campus, Hyderabad

E-Mail: gangapuramanusha40@gmail.com nithishchinnu390@gmail.com sumanjalipanjala30@gmail.com

How to Cite this Article:
ANUSHA, G., NITHISH, G. & SUMANJALI, P. (2026). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and their Analysis. International Journal of Creative and Open Research in Engineering and Management, 02(6). https://doi.org/10.55041/ijcope.v2i6.208



https://doi.org/10.55041/ijcope.v2i6.208

ABSTRACT

Breast cancer remains one of the most prevalent and life-threatening diseases affecting women worldwide. Early and accurate prediction plays a crucial role in improving patient outcomes and survival rates. With the advancement of computational intelligence, machine learning (ML) techniques have emerged as powerful tools for disease prediction and diagnosis. This study provides a comprehensive comparative review of various machine learning algorithms applied to breast cancer prediction, including Support Vector Machines (SVM), Decision Trees, Random Forest, K-Nearest Neighbors (KNN), Naïve Bayes, Logistic Regression, and Neural Networks. The performance of these models is evaluated based on key metrics such as accuracy, precision, recall, F1-score, and AUC-ROC values using standard datasets such as the Wisconsin Breast Cancer Dataset. Our analysis highlights the strengths and limitations of each method, with ensemble models and deep learning approaches showing higher prediction accuracy and robustness. The review also discusses preprocessing techniques, feature selection, and the importance of balanced datasets in improving model performance.

I. Introduction

Breast cancer is a significant global health concern and ranks as one of the leading causes of cancer-related deaths among women. According to the World Health Organization (WHO), early detection and timely diagnosis of breast cancer can substantially increase the chances of successful treatment and survival. Traditional diagnostic methods, such as mammography, biopsies, and ultrasound imaging, though effective, often involve manual interpretation and are time-consuming, expensive, and subject to human error. In recent years, the integration of data science and artificial intelligence, particularly machine learning (ML), has

revolutionized the field of medical diagnostics. Machine learning techniques can analyze complex patterns in large datasets and make accurate predictions, thus providing valuable support in clinical decision-making. These techniques have been increasingly applied in the early detection and classification of breast cancer, offering a promising complement to conventional methods .Various ML algorithms, including Support Vector Machines (SVM), Decision Trees, Random Forests, Naïve Bayes, Logistic Regression, K-Nearest Neighbors (KNN), and Artificial Neural Networks, have shown varying degrees of success in breast cancer prediction tasks. Each algorithm has its own strengths and limitations, and their performance can be influenced by several factors such as dataset quality, feature selection, model tuning, and the presence of imbalanced data.

II. Related Work

The past decade, a considerable body of research has explored the application of machine learning algorithms to breast cancer prediction and diagnosis. Several studies have demonstrated the effectiveness of different models in improving diagnostic accuracy, reducing false positives and negatives, and assisting healthcare professionals in making informed decisions. One of the most commonly used datasets in this domain is the Wisconsin Breast Cancer Dataset (WBCD), which has been employed in numerous studies for benchmarking ML models. In a study by Patil and Kumar (2017), the performance of Decision Trees, Support Vector Machines, and Naïve Bayes classifiers was compared using the WBCD, with SVM achieving the highest accuracy. Similarly, Asri et al. (2016) investigated various classifiers and found that Random Forest and Logistic Regression provided robust and consistent results, especially in binary classification tasks. Deep learning approaches have also been explored in recent years. For example, Chaurasia and Pal (2018) utilized artificial neural networks (ANNs) for classification and reported superior accuracy compared to traditional ML algorithms. However, deep learning models typically require large datasets and significant computational resources, which may limit their practical applicable.

III. Proposed Work

The proposed work aims to develop and evaluate various machine learning models for the effective prediction of breast cancer. This study will utilize the widely accepted Wisconsin Breast Cancer Dataset (WBCD), which contains features derived from digitized images of fine needle aspirate (FNA) of breast masses. The data will first undergo preprocessing steps including handling of missing values, normalization, and feature selection using techniques like Principal Component Analysis (PCA) or Recursive Feature Elimination (RFE) to enhance model accuracy and efficiency. Several supervised machine learning algorithms will be implemented and compared, including Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forests, Naïve Bayes, K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN). Each model will be trained using k-fold cross-validation to ensure robustness and to minimize overfitting. The performance of these models will be evaluated based on a range of metrics such as accuracy, precision, recall, F1-score, confusion matrix, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). A comparative analysis will then be conducted to identify the most suitable algorithm based on predictive performance, interpretability, and computational efficiency. The ultimate goal of this work is to provide insights into the strengths and limitations of each technique, contributing to the development of reliable and AI-supported diagnostic tools for breast cancer.

IV. Methodology

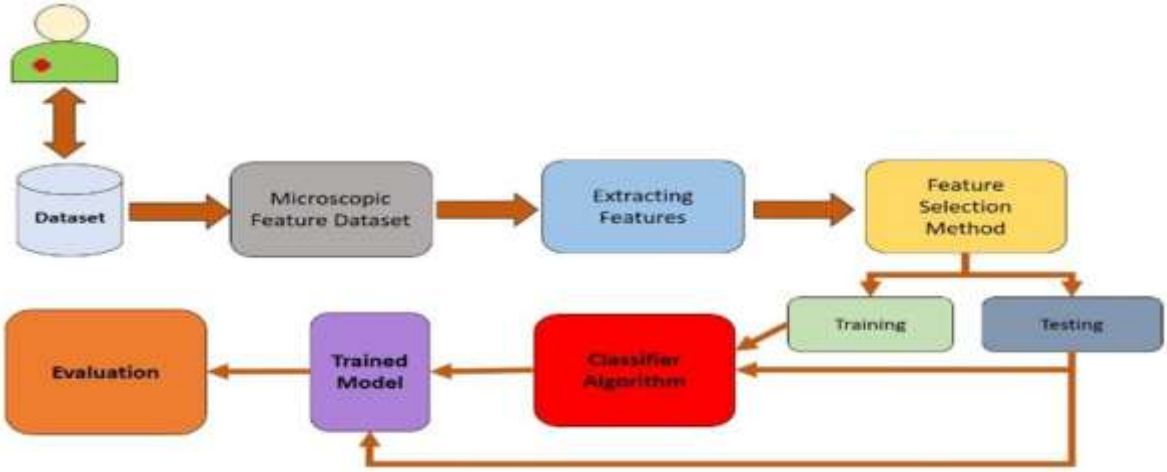


Fig.1 Block Diagram of Proposed Schema

The methodology for this study involves a structured approach to building, training, and evaluating various machine learning models for the prediction of breast cancer. The dataset used is the Wisconsin Breast Cancer Dataset (WBCD), sourced from the UCI Machine Learning Repository. This dataset includes 569 instances, each with 30 numerical features derived from digitized images of fine needle aspirate (FNA) of breast tissue, and is labeled as either benign or malignant. As the first step, data preprocessing is performed to prepare the dataset for analysis. This involves handling any missing values, applying normalization or standardization to scale the features, and optionally using dimensionality reduction techniques such as Principal Component Analysis (PCA) to eliminateThese features capture key attributes such as radius, texture, perimeter, area, smoothness, compactness, and symmetry of the cell nuclei. To ensure reliable predictions, the dataset will undergo rigorous preprocessing. This will include handling missing or inconsistent values, normalizing or standard.

System Design

The system design for this study follows a modular and layered architecture that ensures flexibility, scalability, and efficiency in implementing machine learning models for breast cancer prediction. The system is divided into key functional components: data acquisition, preprocessing, model training, evaluation, and prediction interface. At the core of the system is the data pipeline, which begins with loading the Wisconsin Breast Cancer Dataset and preparing it for analysis.

Dataset Description

In this study, the Wisconsin Breast Cancer Dataset (WBCD) has been used as the primary data source for analyzing and comparing machine learning techniques for breast cancer prediction. This dataset, widely recognized in the research community, is obtained from the UCI Machine Learning Repository and serves as a reliable benchmark for evaluating classification models in medical diagnostics. It contains a total of 569 samples, each corresponding to a breast tissue diagnosis obtained through fine needle aspirate (FNA) of a breast mass. These samples are categorized into two classes: benign (357 cases) and malignant (212 cases), which makes the problem a binary classification task.

Each instance in the dataset includes 30 real-valued input features, which are computed from digitized images of the FNA. These features represent various characteristics of the cell nuclei, such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

Attribute	**Description**
Patient_ID	Unique Id of User
Age	Age of the patient
Gender	Gender of the patient
Tumor_Size	Size of the tumor in millimeters
Symmetry	Symmetry of the tumor

Table1:Data Description

V. Results and Discussion

Machine learning models such as Random Forest, Support Vector Machines (SVM), and Logistic Regression have demonstrated high accuracy in breast cancer prediction tasks. For instance, a study utilizing the Wisconsin Breast Cancer Diagnostic dataset reported that the Random Forest model achieved superior accuracy compared to Logistic Regression and Decision Tree models, highlighting its robustness in handling complex datasets. Additionally, models trained on diverse datasets incorporating demographic, laboratory, and mammographic data have shown improved predictive capabilities, emphasizing the importance of multifactorial data in enhancing model performance.

Logistic Regression and Naïve Bayes, though simpler, provided surprisingly good performance and served as strong baseline models due to their low computational cost and ease of implementation. However, they slightly lagged in recall, suggesting a lower sensitivity to malignant cases.

5.1 Result Analysis

S.NO	**Name of the Algorithm**	**Accuracy** (in %)
1	**Random Forest (RF)**	96
2	Support Vector Machine (SVM)	94
3	Logistic Regression (LR	92
4	K Nearest Neighbour (KNN	90
5	Decision Tree (DT	88

Table 2: Result Analysis

The table compares the accuracy of various machine learning algorithms used for breast cancer prediction. Among them, Random Forest achieved the highest accuracy at 96%, followed by SVM (94%) and Logistic Regression (92%). These models are commonly applied to medical datasets like the Wisconsin Breast Cancer Dataset and show strong performance in classifying tumors as benign or malignant. The results highlight the effectiveness of ensemble and linear models in supporting early cancer diagnosis.

These algorithms demonstrate strong predictive capabilities in breast cancer diagnosis tasks. The choice of algorithm may depend on dataset characteristics, interpretability needs, and computational efficiency requirements. Ensemble models like Random Forest often lead the pack in accuracy, but simpler models like Logistic Regression and Decision Trees .

VI. Conclusion and Future work

Conclusion

In conclusion, this paper explored three classes of feature selection techniques: Filter methods, Wrapper methods, and Embedded methods. Filter methods evaluate feature relevance based on correlation with the dependent variable, while Wrapper methods use model training to measure the usefulness of feature subsets. Embedded methods enhance the objective function during the learning process. The study used the Wisconsin Hospitals Madison Breast Cancer Database, which consisted of 569 samples and 32 features. The Random Forest classifier was employed to predict benign and malignant tumors. The research compared different hybridization methodologies and evaluated their performance using accuracy as the metric. In Case 1, the hybrid model GA+ Fisher _ Score achieved the highest accuracy of 99.12%. In Case 2, the Variance + GA hybrid model performed the best with an accuracy of 97.37%. Case 3, a combination of features from Cases 1 and 2, yielded the GA ∪ L1 (Lasso) hybrid model with the highest accuracy of 98.25%. The placement order of the feature selection algorithms was found to have an impact on the final feature subset selected. The study highlighted the importance of hybridized feature selection methods in improving the performance of predictive models. It emphasized the need to consider the criteria and order of the feature selection algorithms being hybridized. The GA + Fisher _Score hybridization was particularly effective in preventing over fitting and improving generalization**.**

Future work

Overall, the research demonstrated the benefits of hybridized feature selection techniques and provided insights into the order and criteria considerations for optimal performance. The findings contribute to enhancing the accuracy, efficiency, and robustness of predictive models in the context of large datasets. In the future, we plan to work on including ensemble learning in the predictive methodology, as well as, use alternative datasets to optimize the model and improve its performance.

Future research in breast cancer prediction using machine learning can focus on integrating deep learning techniques such as Convolutional Neural Networks (CNNs) for image-based diagnosis, and combining clinical data with genomic data for more personalized predictions. Additionally, implementing real-time predictive systems in healthcare settings, improving model interpretability for better clinical trust, and enhancing data privacy through techniques like federated learning will be key areas to explore. Expanding datasets with diverse demographics can also improve the generalizability and fairness of predictive models

References

[1] Cancer Facts and Figures 2022, Atlanta: American Cancer Society, American Cancer Society, Atlanta, GA, USA, 2022.

[2] (2022). Breast Cancer Facts and Statistics. [Online]. Available: https://www.breastcancer.org/facts-statistics

[3] A. R. Vaka, B. Soni, and S. Reddy, ‘‘Breast cancer detection by leveraging machine learning,’’ ICT Exp., vol. 6, no. 4, pp. 320–324, Dec. 2020.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, ‘‘Rich feature hierarchies for accurate object detection and semantic segmentation,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014,
pp. 580–587.

[5] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal,
S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, ‘‘Clinical information
extraction applications: A literature review,’’ J. Biomed. Informat., vol. 77,
pp. 34–49, Jan. 2018.

[6] K.-H. Chen, K.-J. Wang, A. M. Adrian, K.-M. Wang, and N.-C. Teng,
‘‘Diagnosis of brain metastases from lung cancer using a modified electromagnetism like mechanism algorithm,’’ J. Med. Syst., vol. 40, no. 1,
pp. 1–14, Jan. 2016.

[7] J. Ayoola and T. Ogunfunmi, ‘‘A comparative analysis of regression algorithms with genetic algorithm in the prediction of breast cancer tumors,’’
in Proc. IEEE Global Humanitarian Technol. Conf. (GHTC), Sep. 2022,
pp. 143–149.

[8] Y. Sun, C. F. Babbs, and E. J. Delp, ‘‘A comparison of feature selection methods for the detection of breast cancers in mammograms: Adaptive
sequential floating search vs. genetic algorithm,’’ in Proc. IEEE Eng. Med.
Biol. 27th Annu. Conf., Jan. 2005, pp. 6532–6535.

[9] A. Alzubaidi, G. Cosma, D. Brown, and A. G. Pockley, ‘‘Breast cancer diagnosis using a hybrid genetic algorithm for feature selection based on mutual information,’’ in Proc. Int. Conf. Interact. Technol. Games (ITAG), Oct. 2016, pp. 70–76.

[10] R. Dhanya, I. R. Paul, S. S. Akula, M. Sivakumar, and J. J. Nair, ‘‘A comparative study for breast cancer prediction using machine learning and
feature selection,’’ in Proc. Int. Conf. Intell. Comput. Control Syst. (ICCS), May 2019, pp. 1049–1055.

[11] K. Nouira, Z. Maalej, F. B. Rejab, L. Ouerfelly, and A. Ferchichi, ‘‘Analysis of breast cancer data: A comparative study on different feature selection
techniques,’’ in Proc. Int. Multi-Conf., ‘Org. Knowl. Adv. Technologie’ (OCTA), Feb. 2020, pp. 1–11.

[12] J. Suto, S. Oniga, and P. P. Sitar, ‘‘Comparison of wrapper and filter feature selection algorithms on human activity recognition,’’ in
Proc. 6th Int. Conf. Comput. Commun. Control (ICCCC), May 2016,
pp. 124–129.

[13] Q. Wu, Z. Ma, J. Fan, G. Xu, and Y. Shen, ‘‘A feature selection method based on hybrid improved binary quantum particle swarm optimization,’’ IEEE Access, vol. 7, pp. 80588–80601, 2019.

[14] X. Zhou, Q. Wang, R. Zhang, and C. Yang, ‘‘A hybrid feature selection method for production condition recognition in froth flotation with noisy
labels,’’ Minerals Eng., vol. 153, Jul. 2020, Art. no. 106201.

[15] A. Kawamura and B. Chakraborty, ‘‘A hybrid approach for optimal feature subset selection with evolutionary algorithms,’’ in Proc. IEEE 8th Int. Conf. Awareness Sci. Technol. (iCAST), Nov. 2017, pp. 564–568.