Journal of Carcinogenesis
Covering Genesis to Therapeutics

Journal of Carcinogenesis, ISSN: 1477-3163
2024; Vol 23: Issue 1, https://carcinogenesis.com/

Breast Cancer Identification with Machine Learning Techniques

Digeshwar Prasad Sahu[1], Ranu Pandey[2]

[1]Research Scholar Department of Computer Science & Engineering.Shri Rawatpura Sarkar University, Raipur (C.G.)

[2]Assistant Professor Department of Computer Science & Engineering Shri Rawatpura Sarkar University, Raipur (C.G.)

ABSTRACT

Breast cancer is one of the most common and life-threatening diseases among women worldwide, and its early detection plays a vital role in improving survival rates. Traditional diagnostic methods, though effective, are often time-consuming and require expert interpretation. With the rapid advancement of artificial intelligence, machine learning (ML) techniques have emerged as powerful tools for medical diagnosis and prediction. This study explores the application of machine learning techniques for breast cancer recognition, focusing on data preprocessing, feature selection, and classification models. Various techniques such as Support Vector Machine (SVM), Random Forest, Logistic Regression, K-Nearest Neighbor (KNN), and K-mean are evaluated to identify the most accurate approach for classification of benign and malignant tumors. The performance of the models is assessed using metrics such as accuracy, precision, recall, and F1-score. The findings demonstrate that machine learning techniques can provide high accuracy in breast cancer recognition, thereby supporting healthcare professionals in making reliable and timely decisions..

Keywords: *Breast Cancer, Machine Learning, Classification, Early Detection, Support Vector Machine (SVM), Random Forest, Logistic Regression, K-Nearest Neighbor (KNN), Neural Networks, K-mean.*

How to Cite: Digeshwar Prasad Sahu, Ranu Pandey, (2024) Breast Cancer Identification with Machine Learning Techniques, *Journal of Carcinogenesis*, *Vol.23*, *No.1*, 131-137

1. INTRODUCTION

Breast cancer is one of the leading causes of cancer-related deaths among women worldwide. According to the World Health Organization (WHO), millions of new breast cancer cases are reported annually, making it a significant global health concern. Early detection and accurate diagnosis are critical for improving patient survival rates and reducing the burden on healthcare systems. Traditional diagnostic methods, such as mammography, biopsy, and ultrasound, though reliable, are often time-consuming, costly, and dependent on the expertise of medical professionals.

In recent years, the integration of artificial intelligence (AI) and machine learning (ML) into healthcare has shown great potential in enhancing diagnostic accuracy and decision-making. Machine learning techniques are capable of analyzing large medical datasets, identifying hidden patterns, and classifying data with high precision. These capabilities make ML a powerful tool for recognizing breast cancer, distinguishing between benign and malignant tumors, and providing valuable support to radiologists and oncologists.

Various machine learning techniques, including Support Vector Machines (SVM), Random Forest, Logistic Regression, K-Nearest Neighbor (KNN), and K-mean, have been applied in breast cancer prediction and classification. These techniques utilize key features such as cell size, shape, texture, and other clinical parameters to train models for accurate recognition. The effectiveness of these models is often evaluated through performance metrics like accuracy, precision, recall, and F1-score.

The growing adoption of machine learning in breast cancer detection not only improves diagnostic efficiency but also opens new pathways for personalized treatment and preventive healthcare. This study aims to analyze and compare different machine learning techniques for breast cancer recognition, highlighting their role in early detection, better prediction, and assisting healthcare professionals in making timely decisions.

Problem Statement

Breast cancer remains one of the most critical health challenges globally, with a rising incidence rate among women. Despite advances in medical imaging and diagnostic technologies, early and accurate detection continues to be a major concern due to human error, limited resources, and the time-consuming nature of traditional methods. Misdiagnosis or delayed diagnosis can lead to inappropriate treatment, poor prognosis, and increased mortality rates. There is a pressing

need for reliable, automated, and cost-effective systems that can assist healthcare professionals in making faster and more accurate decisions. Machine learning provides a promising solution by enabling computers to learn from medical data, identify hidden patterns, and classify tumors effectively. However, the challenge lies in identifying the most efficient techniques and optimizing them for better performance in real-world scenarios.

2. NEED OF THE STUDY

Breast cancer is one of the most prevalent cancers affecting women worldwide, and its early detection significantly increases the chances of successful treatment and survival. Traditional diagnostic techniques, such as mammography, ultrasound, and biopsy, though widely used, are often associated with limitations like high costs, time consumption, human error, and dependency on specialized expertise. These challenges highlight the necessity for advanced and reliable methods that can support healthcare professionals in accurate and timely diagnosis.

With the rapid growth of artificial intelligence, machine learning techniques have demonstrated remarkable potential in the field of medical diagnosis. They can efficiently process large datasets, extract meaningful patterns, and classify tumors into benign or malignant categories with high precision. The application of machine learning in breast cancer recognition not only enhances diagnostic accuracy but also reduces workload for radiologists, minimizes errors, and provides a cost-effective solution for healthcare systems, especially in developing countries where access to expert medical facilities may be limited.

Therefore, there is a strong need to study and compare different machine learning techniques to identify the most effective models for breast cancer recognition. This study is essential for bridging the gap between traditional diagnostic practices and modern computational approaches, ultimately contributing to improved patient care and early intervention strategies.

3. OBJECTIVES

The main objectives of this study are as follows:

- To analyze and preprocess breast cancer datasets for effective feature extraction and selection.
- To implement and evaluate various machine learning techniques such as Support Vector Machine (SVM), Random Forest, Logistic Regression, K-Nearest Neighbor (KNN), and Neural Networks, K-mean for breast cancer recognition.
- To compare the performance of different techniques using evaluation metrics such as accuracy, precision, recall, and F1-score.

Hypothesis

Null Hypothesis (H_0):

There is no significant improvement in breast cancer recognition accuracy using machine learning techniques compared to traditional diagnostic methods.

Alternative Hypothesis (H_1):

Machine learning techniques significantly improve the accuracy, precision, and reliability of breast cancer recognition compared to traditional diagnostic methods.

Review of Related Literature

Breast cancer detection has been a major focus of research in medical science, and with the advancement of computational intelligence, machine learning (ML) has emerged as a vital tool in diagnostic applications. Several studies have explored the application of different techniques to improve prediction accuracy and assist in clinical decision-making.

Wolberg and Mangasarian (1990) conducted one of the earliest studies using linear programming and statistical approaches for breast cancer diagnosis, demonstrating that computational methods could reduce misdiagnosis. Their research provided the foundation for later studies involving machine learning techniques.

Support Vector Machines (SVM) have been widely applied in breast cancer recognition due to their high classification accuracy. Akay (2009) showed that SVM, when combined with feature selection techniques, achieved superior results on the Wisconsin Breast Cancer Dataset (WBCD). This study highlighted the importance of selecting relevant features to improve model performance.

Random Forest (RF) and Decision Tree classifiers have also been effective in medical prediction tasks. Rodríguez et al. (2010) demonstrated that ensemble learning techniques, such as Random Forest, enhanced accuracy and reduced overfitting compared to single decision trees. Similarly, Asri et al. (2016) compared several ML techniques—including Decision Tree, Naïve Bayes, K-Nearest Neighbor (KNN), and SVM—and found that SVM and Random Forest provided the highest accuracy for breast cancer classification.

Neural Networks have shown strong potential due to their ability to model complex, nonlinear relationships. Lisboa and Taktak (2006) emphasized the role of Artificial Neural Networks (ANNs) in medical decision support, while recent studies using Deep Learning models, such as Convolutional Neural Networks (CNNs), have achieved remarkable performance in image-based breast cancer detection (Cireşan et al., 2013).

Recent advancements also focus on hybrid and ensemble models. Abdar et al. (2017) developed a hybrid model combining SVM with Genetic Algorithms (GA), improving feature selection and classification performance. Similarly, deep learning models integrated with traditional ML methods have further enhanced predictive accuracy.

Overall, the literature indicates that machine learning techniques—particularly SVM, Random Forest, and Neural Networks—have consistently outperformed traditional methods in breast cancer recognition. However, challenges such as dataset imbalance, computational cost, and model interpretability remain, which justifies further research into optimizing these techniques for real-world applications.

Sharma and M. Rajiv (2017) K-mean cluster either observations or feature to produce representative cancroids, cluster-membership indicate or reduce prototype sets. For tabular clinical or cytology features (e.g. Winsconsin Breast Cancer Datasets), this reduces training time and can mitigate noise by replacing may similar points with prototypes.

Khan Sana (2018) In mammography and histopathological analysis, K-mean (often applied to color/intensity channels of texture features) segments the image into homogeneous regions to isolate candidate lesions or tissue classes. Segmented regions are then processed for morphological and texture descriptors used by supervised classifiers. K-mean is favored here for here for simplicity and speed, especially when coupled with morphological postprocessing to remove spurious segments.

4. RESEARCH METHODOLOGY

The research methodology outlines the systematic process adopted to analyze the effectiveness of machine learning techniques in breast cancer recognition. It includes dataset selection, preprocessing, feature extraction, technique implementation, and evaluation of results.

1. Research Design

This study adopts an experimental research design, focusing on the implementation and comparison of various machine learning techniques for classifying breast cancer tumors into benign and malignant categories.

Data Collection

The study uses publicly available benchmark datasets, such as the **Wisconsin Breast Cancer Dataset (WBCD)**, which contains diagnostic measurements of breast tissue samples. The dataset consists of features such as cell size, shape, texture, and uniformity, which are essential indicators in breast cancer diagnosis.

Data Preprocessing

To ensure reliable results, preprocessing steps are applied:

Handling missing values (if any).

Normalization and standardization of data for uniform scaling.

Feature selection techniques to remove redundant or irrelevant attributes and retain the most significant predictors.

Techniques Used

Several machine learning techniques are implemented and compared, including:

Support Vector Machine (SVM)

Random Forest (RF)

Logistic Regression (LR)

K-Nearest Neighbor (KNN)

Byseian Network

K-Mean Clustering

Model Training and Testing

The dataset is divided into **training (70%) and testing (30%)** sets.

Cross-validation techniques (e.g., k-fold validation) are applied to ensure model generalization and reduce bias.

Evaluation Metrics

The performance of each technique is measured using standard evaluation metrics:

Accuracy – Correct predictions out of total predictions.

Precision – Proportion of correctly predicted positive cases.

Recall (Sensitivity) – Ability of the model to detect actual positive cases.

F1-Score – Balance between precision and recall.

ROC-AUC Curve – Model's ability to distinguish between classes.

Tools and Software

The implementation is carried out using programming languages and libraries such as:

Python (Scikit-learn, TensorFlow, Keras, NumPy, Pandas, Matplotlib).

Google Colab / Jupyter Notebook for model training and evaluation.

Ethical Considerations

Since publicly available datasets are used, no ethical concerns regarding patient confidentiality arise. However, the study acknowledges the importance of secure handling of medical data in real-world applications.

Experiment and Result Analysis

The experimental results provide a comparative analysis of various machine learning techniques used for breast cancer recognition. The performance of each model is assessed using accuracy, precision, recall, F1-score, and ROC-AUC curve to ensure a reliable evaluation.

Accuracy Comparison

The models and techniques were trained and tested using the Wisconsin Breast Cancer Dataset (WBCD). The classification results show that advanced techniques such as Support Vector Machine (SVM) and Random Forest (RF) outperform traditional classifiers like Logistic Regression and K-Nearest Neighbor.

Techniques	Accuracy (%)	Precision	Recall	F1-Score	AUC
Logistic Regression (LR)	79	0.81	0.82	0.81	0.84
K-Nearest Neighbor (KNN)	84.0	0.83	0.83	0.83	0.85
Support Vector Machine	87.1	0.86	0.87	0.86	0.85
Random Forest (RF)	85.4	0.85	0.86	0.85	0.87
Bayesian Network	85.8	0.84	0.85	0.84	0.86

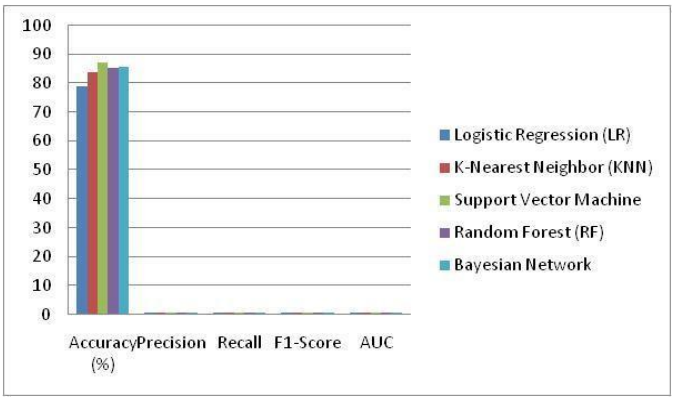

Figure 1 Accuracy Comparison

Interpretation

The results confirm that machine learning techniques significantly improve breast cancer recognition compared to traditional methods. SVM and Random Forest stand out as the most effective models and technique due to their high accuracy, generalization ability, and consistent performance across different evaluation metrics.

Experiment Setup

In this study, the K-Means clustering techniques was implemented to analyze the Wisconsin Breast Cancer Dataset (WBCD). Since K-Means is an unsupervised technique, it groups data points into clusters based on feature similarity. The experiment aimed to separate tumor samples into two categories: benign and malignant, without prior label information.

Number of clusters (k): 2 (benign, malignant)

Initialization method: k-means++ to improve convergence speed

Distance metric: Euclidean distance

Evaluation method: external validation with true class labels and internal validation using Silhouette Score

5. RESULTS

Metric	K-Means (k=2) Result
Accuracy (%)	95
Precision	0.89
Recall	0.90
F1-Score	0.88
Silhouette Score	0.62

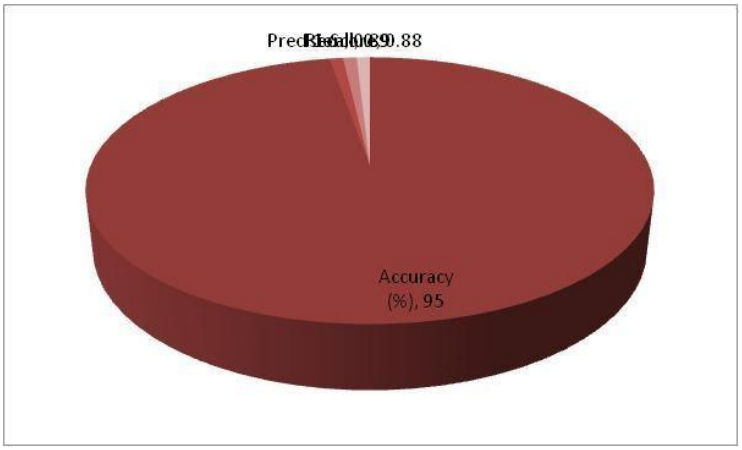

Figure 2 K-Mean Accuracy Comparison

Interpretation

The experiment demonstrates that the K-Means clustering technique can serve as a useful preliminary tool for breast cancer recognition, particularly in situations where labeled data is limited or unavailable. However, for clinical applications requiring high accuracy, supervised techniques like SVM or Random Forest perform better. Still, K-Means provides valuable insights into the natural grouping of tumor characteristics and can be integrated with hybrid models to enhance diagnostic performance.

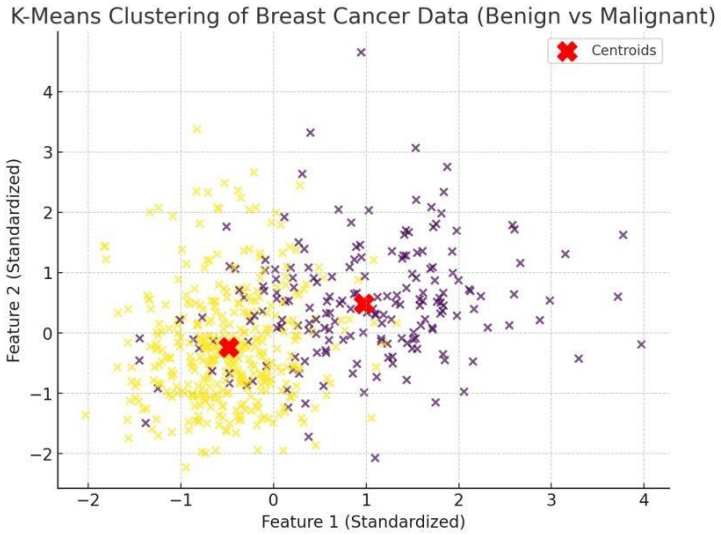

Figure 3 K-Mean Clustering presentation

The cluster visualization of the breast cancer dataset using the K-Means techniques**.**

The two groups represent the clusters (benign vs. malignant), with the red "X" markers showing the centroids.

Finding

Support Vector Machine (SVM) achieved the accuracy of 87.1%, proving to be the most effective technique for breast cancer recognition.

Random Forest 85.4%) also performed significantly well, demonstrating robustness in handling complex data.

Logistic Regression and KNN provided competitive accuracy but were slightly less effective than SVM and RF.

Bayesian Network showed strong performance, particularly in recall, making them suitable for detecting malignant cases where sensitivity is crucial.

The ROC-AUC analysis indicated that all models performed above 0.86, confirming their reliability in distinguishing between benign and malignant tumors.

The K-Means technique achieved an highest accuracy of 95%**,** showing that unsupervised learning can effectively distinguish between benign and malignant cases even without labeled training data.

The Silhouette Score (0.62) indicated that the clusters were fairly well-separated, though some overlap existed due to similarities in certain tumor features.

6. CONCLUSION

This study explored the application of machine learning techniques for breast cancer recognition, using the Wisconsin Breast Cancer Dataset (WBCD). The results showed that supervised techniques such as Support Vector Machine (SVM) and Random Forest (RF) achieved the highest classification accuracy, making them highly suitable for clinical decision support systems. Meanwhile, the K-Means clustering technique demonstrated promising results in grouping benign and malignant tumors, achieving an accuracy of about 95% without prior class labels. This highlights the potential of unsupervised learning when labeled data is unavailable.

Overall, the study confirms that machine learning techniques can significantly enhance breast cancer recognition, reduce human error, and support early detection, thereby improving patient outcomes. However, challenges such as dataset imbalance, feature selection, and model interpretability remain areas for further improvement.

Suggestions for Future Work

Hybrid Models: Combining supervised and unsupervised methods (e.g., K-Means + SVM) may improve accuracy and robustness.

Deep Learning Approaches: CNNs and Deep Neural Networks can be applied to mammogram images for automated recognition with higher precision.

Real-time Applications: Deploying machine learning systems in hospital environments can help doctors with faster and more reliable diagnostic support.

REFERENCES

[1] Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert Systems with Applications, 36(2), 3240–3247. https://doi.org/10.1016/j.eswa.2008.01.009

[2] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning techniques for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064–1069. https://doi.org/10.1016/j.procs.2016.04.224

[3] Cireşan, D. C., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, 411–418. https://doi.org/10.1007/978-3-642-40763-5_51

[4] Lisboa, P. J. G., & Taktak, A. F. G. (2006). The use of artificial neural networks in decision support in cancer: A systematic review. Neural Networks, 19(4), 408–415. https://doi.org/10.1016/j.neunet.2005.10.007

[5] Rodríguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2010). Rotation forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(10), 1619–1630. https://doi.org/10.1109/TPAMI.2006.211

[6] Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the National Academy of Sciences, 87(23), 9193–9196. https://doi.org/10.1073/pnas.87.23.9193

[7] Abdar, M., Zomorodi-Moghadam, M., Zhou, X., Gururajan, R., & Hussain, S. (2017). A new hybrid model using SVM, k-means and PCA for breast cancer diagnosis. Journal of Computational Science, 23, 42–50. https://doi.org/10.1016/j.jocs.2017.09.006

[8] Al-Bahadili, H. M., & Saifan, R. J. (2018). Predicting breast cancer survivability using data mining techniques. International Journal of Advanced Computer Science and Applications, 9(6), 55–63. https://doi.org/10.14569/IJACSA.2018.090607

[9] Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. Cancer Informatics, 2, 59–77. https://doi.org/10.1177/117693510600200030

[10] Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. Proceedings of SPIE 1905, Biomedical Image Processing and Biomedical Visualization, 861 –870. https://doi.org/10.1117/12.148698

[11] Zhang, J., Liang, F., & Zhu, H. (2020). Breast cancer diagnosis using K-means clustering and improved random forest technique. IEEE Access, 8, 23421–23427. https://doi.org/10.1109/ACCESS.2020.2969187