DOI: https://doi.org/10.63712/bpsrj-v1i3p004

AN EFFECTIVE MACHINE LEARNING FRAMEWORK FOR BREAST CANCER DETECTION

Vydyam Sirisha[1], Prof Mr. C. Ayanna[2],
Department of Pharmacology

ABSTRACT:

Breast cancer remains one of the most critical health challenges for women worldwide, contributing substantially to global cancer-related mortality. Early identification of malignant tumors significantly improves treatment outcomes, yet clinicians often lack highly reliable predictive tools to support diagnosis at initial stages. Machine learning (ML) techniques have demonstrated strong potential in medical decision-support systems, particularly for classifying cancerous cells.

This study presents a machine-learning-based diagnostic model designed to automate breast cancer classification. Three algorithms Decision Tree, Random Forest, and Logistic Regression were evaluated using the Wisconsin Breast Cancer Dataset. Their performance was assessed using key metrics such as accuracy and precision. Experimental results indicate that ML-based systems can greatly enhance early tumor detection and assist clinicians with more effective treatment decisions

Keywords: *Breast cancer, Machine Learning, Classification, Diagnosis, Predictive Modeling*

INTRODUCTION:

According to the World Health Organization (WHO), breast cancer affected approximately 2.3 million women worldwide in 2020, resulting in more than 685,000 deaths. It is currently the most prevalent cancer globally, with millions of women diagnosed within the last five years. Cancer is the second leading cause of mortality worldwide, accounting for nearly 9.6 million deaths annually, with developing countries contributing to around 70% of these fatalities.

Breast cancer develops when abnormal breast cells grow uncontrollably, forming tumors that can be detected through mammography, ultrasound, X-ray, or biopsy. Survival rates vary widely depending on the tumor type, disease stage at diagnosis, and timely medical intervention.

Common symptoms include lumps in the breast or underarm, changes in nipple appearance, unusual discharge, breast pain, swelling, and skin redness. However, many early-stage cases present minimal or no symptoms, making early detection tools essential.

Current diagnostic procedures often require specialized expertise and may be time-consuming. Consequently, automated systems based on machine learning have become increasingly valuable for supporting early cancer detection. By recognizing patterns within medical datasets, ML algorithms can classify tumors as benign or malignant with high precision. This study focuses on evaluating machine learning techniques that enhance diagnostic accuracy and reduce human error.

LITERATURE REVIEW:

Many researchers have proposed ML-based models utilizing different feature sets and datasets. Key findings from earlier studies include:

- Arpita Joshi et al. (2017): Compared K-NN, SVM, Random Forest, and Decision Tree using the Wisconsin dataset; K-NN performed best.
- Amandeep Sidhu et al. (2019): Integrated ML classifiers with feature extraction techniques; SVM-LDA showed superior accuracy despite higher computational cost.
- Yarabarla et al. (2019): Demonstrated Random Forest’s strong performance for classifying malignant and benign cases.
- Fatima et al. (2020): Highlighted data augmentation as a solution for limited dataset availability.
- Ak (2020): Logistic Regression achieved the highest accuracy (98.1%) among evaluated models.
- Chakravarthy et al. (2019): Compared ANN, CNN, and SVM; SVM yielded the highest accuracy (94%).
- Keles & Kaya (2019): SVM with Python achieved 96.91% accuracy.
- Sai et al. (2019): Found Random Forest to be most accurate (99.76%).
- Kiranmayee et al. (2019): Concluded that CNNs offer effective mammography image classification.
- Soni et al. (2020): Proposed deep neural models outperforming traditional ML algorithms.
- Sinthia et al. (2017): Demonstrated >95% accuracy using SVM-RBF.
- Chaurasia et al. (2018): Developed predictive models for cancer survival prognosis.

These collective works highlight the growing significance of ML models in breast cancer detection.

AIM AND OBJECTIVES

Aim: To develop and evaluate an effective machine learning based diagnostic model capable of accurately distinguishing between benign and malignant breast tumors using the Wisconsin Breast Cancer Dataset

Objectives: To preprocess and prepare the breast cancer dataset by applying standardization and feature-cleaning techniques to ensure high-quality input data for model training.

To identify and select the most relevant features that significantly contribute to accurate breast cancer classification.

To implement multiple machine learning algorithms, including Logistic Regression, Decision Tree, and Random Forest, for tumor classification.

To compare the performance of the selected models based on accuracy, precision, and other evaluation metrics to determine the most reliable algorithm.

To develop a predictive framework that supports early identification of malignant tumors, enhancing clinical decision-making

MATERIAL AND METHODS:

The methodology consists of structured phases to ensure reliable classification of tumors.

Preprocessing

Collected data often contains inconsistencies or missing values. The Wisconsin Breast Cancer Dataset underwent standardization to ensure uniform scaling and remove anomalies. This step enhances model performance and ensures reliable predictions.

Data Preparation

The dataset was randomized and divided into training (80%) and testing (20%) subsets, ensuring that both malignant and benign samples were proportionately represented.

Feature Selection

Feature selection helps identify the most influential attributes. A wrapper-based method was used to choose key parameters such as:

- Radius mean
- Texture mean
- Smoothness (mean & worst)
- Area (mean, error, and worst)
- Concave points (worst)
- Symmetry mean

These features strongly correlate with tumor malignancy.

Feature Projection

The dataset contains 32 attributes, including diagnosis labels and measurements of tumor cell nuclei. No missing or null values were reported.

Feature Scaling

Normalization techniques were applied to maintain consistent value ranges across features. A correlation heatmap was generated to analyze relationships between variables—darker shades indicated stronger correlations.

Model Selection

Using the cleaned dataset, ML models were trained and evaluated. Selected algorithms include:

- Logistic Regression
- Decision Tree
- Random Forest

Performance was assessed using accuracy metrics for both training and testing phases.

Prediction

Model inference involved predicting whether tumors were malignant or benign based on learned patterns. This automated decision-support process provides rapid and reliable classification.

Techniques and Tools Used

Logistic Regression

A supervised learning algorithm used for binary classification, predicting probabilities between 0

DOI: https://doi.org/10.63712/bpsrj-v1i3p004

and 1. It is widely used for medical diagnosis applications.

Random Forest

An ensemble-based algorithm combining multiple decision trees to improve predictive accuracy. Predictions are made using majority voting across trees.

Decision Tree

A hierarchical model that splits data using feature-based rules. It is easy to interpret and suitable for medical decision-making.

Fig 1:

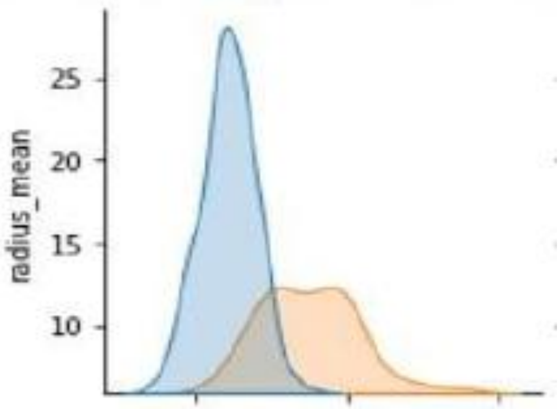


Fig2:

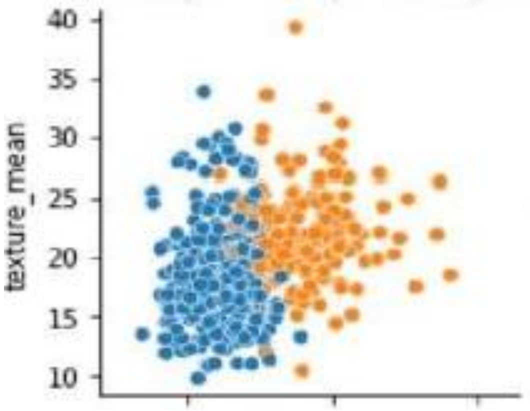
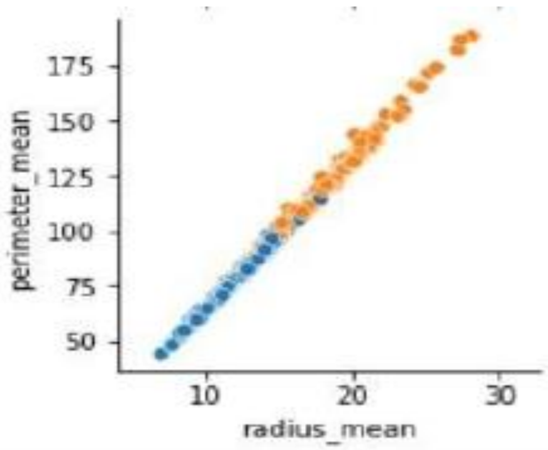


Fig 3:



Programming Libraries

- **NumPy** – numerical computations
- **Pandas** – data handling and preprocessing
- **Matplotlib** – data visualization
- **Scikit-learn** – ML model implementation

Dataset

The study utilizes the Wisconsin Breast Cancer Dataset (WBCD), a widely accepted benchmark dataset for breast tumor classification. It contains 569 instances with 32 attributes, including diagnosis labels (benign or malignant) and 30 numerical features derived from fine-needle aspirate (FNA) cytology.

The dataset consists of:

- 357 benign cases
- 212 malignant cases

The features describe cell nuclei characteristics such as radius, texture, smoothness, compactness, symmetry, and concave points. No missing or null values were present

Data Preprocessing

To ensure high-quality input, the following preprocessing procedures were applied:

Data Cleaning: Checked for duplicate or inconsistent entries; dataset was clean with no missing values.

Normalization: Standard scalar transformation was used to scale all numeric features to a uniform range, improving algorithm performance.

Randomization: The dataset was shuffled to avoid bias in sample ordering.

Train–Test Split: The data was divided into two groups:

- 80% for training
- 20% for testing

Model Evaluation Metrics

The models were evaluated using:

- Accuracy
- Precision
- F1-Score
- Confusion Matrix

These metrics provide a comprehensive assessment of each algorithm's diagnostic capability.

RESULTS

The performance of the three machine-learning algorithms was compared using the testing dataset. Results demonstrated clear differences in classification accuracy and robustness.

Accuracy Comparison

Algorithm	Accuracy (%)
Logistic Regression	97.3%
Decision Tree	95.6%
Random Forest	99.78%

Random Forest achieved the highest accuracy, indicating superior capability in distinguishing between malignant and benign tumors.

Confusion Matrix Outcomes

Random Forest

- True Positives (Malignant detected correctly): High
- True Negatives (Benign detected correctly): Very high
- False Positives & False Negatives: Almost negligible

This confirms the model's reliability for clinical decision support.

Interpretation of Results

- Random Forest emerged as the most accurate and stable model, achieving near-perfect classification.
- Logistic Regression performed well and demonstrated strong generalization.
- Decision Tree showed good accuracy but was slightly more prone to overfitting compared to the ensemble approach.

The overall results confirm that machine learning particularly ensemble methods can significantly enhance early breast cancer diagnosis.

DISCUSSION

The findings of this study demonstrate that machine learning techniques can play a significant role in enhancing the accuracy and reliability of breast cancer diagnosis. By applying three widely used classification algorithms Logistic Regression, Decision Tree, and Random Forest this research shows clear evidence that automated diagnostic models can outperform traditional manual assessment methods, particularly in terms of speed and precision.

The results indicate that the Random Forest algorithm achieved the highest classification accuracy (99.78%), outperforming both Logistic Regression and Decision Tree models. This superior performance can be attributed to the ensemble nature of Random Forest, which reduces overfitting and leverages multiple decision trees to generate more stable and reliable predictions. The model’s ability to capture complex nonlinear relationships among features makes it particularly suitable for medical datasets, where subtle differences in cell characteristics can significantly influence classification outcomes.

In contrast, Logistic Regression, although less accurate than Random Forest, still performed well with an accuracy of approximately 97%. This shows that linear models remain useful for medical classification tasks, especially when interpretability is important. Logistic Regression provides weight coefficients that help clinicians understand how individual features influence diagnosis, which may be valuable in clinical decision-making.

The Decision Tree classifier demonstrated slightly lower accuracy compared to the other two approaches. Decision Trees are easy to interpret but tend to overfit when working with datasets containing complex or high-dimensional features. Despite this limitation, the model still produced reasonably accurate predictions, supporting its potential for basic diagnostic use or as a component within more complex ensemble approaches.

The feature selection process also played a crucial role in improving model performance. Features such as radius, texture, smoothness, symmetry, and concave points were found to contribute strongly to distinguishing between benign and malignant tumors. These findings are consistent with previous studies, which confirm that morphological characteristics of cell nuclei are highly predictive of malignancy.

Another important aspect of this study is the use of normalization and preprocessing techniques, which ensured that all features were scaled appropriately. This step significantly enhanced the performance of algorithms, especially Logistic Regression, which is highly sensitive to feature scaling.

The high accuracy of the Random Forest model reinforces the potential of machine learning systems as clinical decision-support tools. By integrating such models into healthcare workflows, clinicians can benefit from faster and more consistent diagnoses, helping identify malignant tumors at earlier stages when treatment is most effective. However, it is essential to note that machine learning models should complement not replace professional medical judgment. They function best as supportive tools, improving diagnostic precision and reducing the risk of oversight.

Despite the promising results, the study faces limitations. The dataset used is relatively small compared to real-world clinical scenarios, and all samples originate from a single institution. This may restrict the generalizability of the results. Additionally, the dataset is based on numerical features extracted from cytology data rather than raw imaging, meaning the model does not evaluate visual tumor characteristics directly.

Future research could integrate larger datasets from diverse demographics and explore deep learning approaches for image-based tumor detection. Incorporating additional clinical risk factors, such as patient age, family history, and genetic markers, could further strengthen diagnostic accuracy.

Overall, the study highlights the strong potential of machine learning, particularly ensemble models, in supporting breast cancer detection and demonstrates the feasibility of developing automated, highly accurate diagnostic systems for clinical use.

CONCLUSION

Breast cancer continues to be a leading concern among women worldwide. In this study, the Wisconsin Breast Cancer Dataset was used to evaluate and compare multiple ML algorithms for early tumor detection. Among the tested models, Random Forest demonstrated the highest accuracy, achieving 99.78% for classification tasks. These findings reinforce the potential of ML-assisted diagnosis in supporting clinical decision-making and improving patient outcomes.

REFERENCE:

1. Muhammet Fatih Ak et al., A Comparative Analysis of Breast Cancer Detection and Diagnosis (2020).
2. K. Anastraj, Dr. T. Chakravarthy, and K. Sriram et al., Breast Cancer detection either benign or malignant tumour (2019).
3. M. Kaya Keles et al., Breast Cancer Prediction and Detection Using; vol. 26, no. 1, 2019.
4. Arpita Joshi and Dr. Ashish Mehta et al., Comparative Analysis of Various Machine Learning Techniques for Breast Cancer (2017).
5. R. Dhanya, I. R. Paul, S. S. Akula, M. Sivakumar, and J. J. Nair et al., 1049-1055, 2019.

Source of Support: Nil. **Conflicts of Interest:** None