International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Breast Cancer Detection Using Machine Learning

FAHIM ZAHIR SHAIKH

Department of BACHELOR OF VOCATIONAL IN ARTIFICIAL AND DATA SCIENCE ,
Anjuman I Islam's Abdul Razzaq Kalsekar Polytechnic, New Panvel , Maharashtra , India

---

Abstract -Breast cancer is the most prevalent cancer among women worldwide and early diagnosis is crucial for increasing survival rates. This research proposes a machine learning-based approach for the classification of breast cancer as benign or malignant using the Wisconsin Breast Cancer Dataset (WBCD). The Random Forest Classifier was selected for its robustness and accuracy in handling structured data. The trained model was deployed through a user-friendly Python GUI built with Tkinter, enabling real-time predictions. The system achieved an accuracy of over 96%, demonstrating its potential as a low-cost, accessible, and reliable decision support tool in breast cancer screening and diagnosis.

Key Words: Breast Cancer, Machine Learning, Random Forest, GUI, Classification, Detection, WBCD

1. INTRODUCTION

Breast cancer remains a significant health threat for women globally, with millions of cases diagnosed every year. Despite advancements in medical imaging and diagnostics, access to such resources remains limited in rural and underdeveloped regions. To address this gap, machine learning (ML) provides a promising avenue for fast, cost-effective diagnosis through computational models trained on medical data.

This study aims to develop an intelligent system that classifies breast tumors using clinical data. The system leverages supervised learning algorithms, specifically Random Forest, and is accessible via a Tkinter-based GUI. The core goal is to offer a supportive diagnostic tool that empowers early detection.

2. LITERATURE SURVEY

Previous works in the field have employed various classifiers like Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees for breast cancer classification. Studies have shown that ensemble methods like Random Forest improve performance through feature aggregation and internal cross-validation. Tools such as Scikit-learn, Python Tkinter, and datasets like WBCD have become standard in academic research, enhancing model training and deployment.

3. DATASET DESCRIPTION

The Wisconsin Breast Cancer Dataset (WBCD), sourced from the UCI Machine Learning Repository, comprises 699 instances with 9 numerical features and a target class indicating whether the tumor is benign (2) or malignant (4). Features include:

- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape
- Marginal Adhesion
- Single Epithelial Cell Size
- Bare Nuclei
- Bland Chromatin
- Normal Nucleoli
- Mitoses

The dataset is clean, well-structured, and ideal for binary classification tasks.

4. SYSTEM ARCHITECTURE

The system comprises:

1. ***Data Preprocessing*** *– Handling missing values, normalizing features.*
2. ***Model Training*** *– 80/20 train-test split; classifier trained on training data.*
3. ***Model Selection*** *– Random Forest outperformed other models like SVM and KNN.*

4. ***GUI Development*** *– Developed using Tkinter for real-time input and output.*

5. ***Model Deployment*** *– Trained model saved with Joblib and integrated into the GUI.*

5. METHODOLOGY

5.1 Preprocessing

- Missing values replaced with mode
- All input features scaled using MinMaxScaler
- Class labels mapped (Benign → 0, Malignant → 1)

5.2 Training and Evaluation

- Random Forest Classifier trained using 80% of data
- Evaluated using metrics: accuracy, precision, recall, F1-score

Confusion Matrix Results

	Predicted Benign	Predicted Malignant
Actual Benign	86	4
Actual Malignant	2	108

6. RESULTS

- **Accuracy**: 96.3%
- **Precision**: 95.8%
- **Recall**: 96.0%
- **F1-Score**: 95.9%

The system showed consistent performance across test cases and demonstrated reliability during live testing via the GUI.

7. SAMPLE OUTPUTS

Test Case 1
Input: [9, 8, 8, 7, 6, 10, 7, 8, 1] → **Malignant**

Test Case 2
Input: [2, 1, 2, 1, 2, 1, 2, 1, 1] → **Benign**

Test Case 3
Input: [10, 10, 10, 8, 7, 10, 9, 8, 2] → **Malignant**

Test Case 4
Input: [3, 2, 2, 2, 3, 1, 2, 1, 1] → **Benign**

Test Case 5
Input: [7, 7, 8, 5, 6, 9, 7, 7, 3] → **Malignant**

8. DISCUSSION

The high performance of the Random Forest model indicates the dataset is well-suited for ensemble learning. The GUI ensures accessibility and usability even for users without technical backgrounds. However, the model's reliance on numeric clinical inputs makes it less effective for image-based or real-world unstructured data without prior preprocessing.

9. CONCLUSION

This research demonstrates that machine learning can be used to create reliable diagnostic support systems. The developed tool is capable of providing early breast cancer predictions with high accuracy and minimal resource requirements. It is not a replacement for professional medical evaluation but can aid in preliminary screening, especially in underserved regions.

10. FUTURE SCOPE

- Extend input support to image-based data using CNNs
- Web or mobile app deployment for remote usage
- Integration with cloud-based databases for real-time analytics
- Clinical testing for certification and validation

11. ACKNOWLEDGEMENT

I sincerely thank the Department of Artificial Intelligence and Data Science, Anjuman-I-Islam's Abdul Razzaq Kalsekar Polytechnic, and my project guide Mr. Ali Karim Sayed for their continuous support, mentorship, and encouragement.

12. REFERENCES

1. UCI Machine Learning Repository – Breast Cancer Dataset
2. Scikit-learn Documentation – https://scikit-learn.org
3. Python Tkinter Documentation – https://docs.python.org/3/library/tkinter.html
4. Nahato D. et al., "Rule Mining from Breast Cancer Data Using Improved Genetic Algorithm", Journal of Biomedical Informatics
5. Breiman, L. "Random Forests", Machine Learning Journal, 2001.