International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 1-12 March 2026-Sept 2026 www.csjournals.com

Random Forest Based Breast Cancer Classification

Ravina[1], Dr. Vishal Verma[2], and Mrs. Neetu[3]
[1]MCA Student, [2]Professor, [3]P.hd Scholar
Deptt. of Comp. Sc. and Application, Chaudhary Ranbir Singh University, Jind, Haryana, India

Abstract

It is well known that the problem of breast cancer is increasing rapidly. Breast cancer is important to analysis at initial stage for better survival rates. The suggested framework is executed with several Exploratory Data Analysis (EDA) techniques and class balancing based on SMOTE for better classification outcomes. The presented RF model shows accuracy 97.20% on training and 94.44% on testing. Evaluation metrics provides good classification ability with ROC-AUC value of 0.9876.

Keywords: ML, Random Forest, Breast Cancer, ROC Curve, AUC, SMOTE

Introduction

According to WHO, in 2022 nearly 670,000 deaths were reported world-wide due to breast cancer[13]. Many countries which are in developing phase suffering from the uncontrolled growth of abnormal cells stays highly concern in the women[9]. Yet, this technology is capable to process complex patient details as well as find subtle irregularities which is highly challenging for manual analysis[3][10].

Random Forest has emerged as a leading Machine Learning framework for tumor classification[1][4][5]. It works by combining multiple decision trees and producing a final result based on majority voting. This ensemble strategy reduces overfitting, where the algorithm simply stores the training instances instead of generalizing new cases [1][5][13]. Since Random Forest maintains the concept of bias and variance using averaging and randomness to improve balance because both of these terms are inversely proportional to each other. We require low bias and low variance and it is achieved by the concept of Bagging.

$$\text{Bias} \propto \frac{1}{\text{Variance}}$$

Bagging stands for Bootstrapped Aggregation and Random Forest comes under this technique. In this technique all base trees are decision trees means base algorithm is same but these are trained on different data like rows sampling, feature sampling after that average or random occurring data is used which helps to manage the bias and variance which are opposite to each other[1][4][5].

Previous work also supports this approach. Few algorithms combine PCA to handle impurities and maintain only the important biological features [7]. Rest of the scientist integrate RF classifications with specialist clinical details to improve security and accuracy[13]. Moreover, Random Forest can approximate the role of attributes, assisting healthcare providers understand which factors contribute strongly to classify [12].

Machine Learning in Healthcare

ML is a branch of AI that allows systems to automatically discover data patterns and make decisions. It identifies similarity in the previous data and continuous improvement with time. Today, Machine Learning is utilized in every sector like finance, education, transportation and especially in healthcare [8].

It is divided into three major parts:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Regression and Classification categorized under supervised learning, while data segmentation, feature compression, discovering hidden relationships, and outliers belong to unsupervised learning.Particularly, supervised learning is commonly applied in healthcare because medical datasets often contain well-organized outcomes. It trains on labeled datasets where the correct answers are provided earlier. The system uncovers the connection among input variables and outputs, then applies which knowledge to predict unseen cases [8].

This model implements medical information such as cell radius, texture, smoothness, or shape. Once trained on previous patient cases with confirmed outcomes then model easily differentiating between benign and malignant.

Abbreviations

Acronyms	Full Form
AI	Artificial Intelligence
CAD	Computer-Aided Diagnosis
IQR	Interquartile Range
ML	Machine Learning
PCA	Principal Component Analysis
RF	Random Forest
ROC-AUC	Receiver Operating Characteristic – Area Under Curve
SMOTE	Synthetic Minority Oversampling Technique
WBCD	Wisconsin Breast Cancer Dataset
WHO	World Health Organization
SVM	Support Vector Machine
LR	Logistic Regression

Research Gap

- Past researchers mainly focus on performance of Random Forest by calculating only accuracy rather than the use of other evaluation metrics.[2][8][13].
- Previous studies do not focus on class imbalance dataset [5][12].
- The analyzed literature depend on either raw features or techniques like Principal Component Analysis, which highlights essential requirement of extracting highly correlated key attributes[7][12].

Literature Review

Table 1: Overview of different research studies

S. No.	Authors	Year	Dataset	Used Algorithms	Accuracy	Best Model	Limitations
1.	T. L. Octaviani, Z. Rustam[1]	2019	Wisconsin Breast Cancer Dataset	Random Forest	~96–98%	Random Forest	Missing ROC-AUC evaluation, No cross-validation strategy
2.	M. M. Islam et al.[2]	2020	Wisconsin Breast Cancer Dataset	RF, SVM, KNN, DT, LR, NB	~98.2%	SVM / RF	Focused mainly on benchmark dataset; limited real-world validation
3.	B. O. Macaulay et al.[3]	2021	African women breast cancer risk data	Random Forest	~90%	Random Forest	Regional dataset only; limited sample diversity
4.	M. Minnoor, V. Baths[4]	2023	Breast Cancer Dataset	Random Forest	~97%	Random Forest	Only RF emphasized; comparison limited
5.	A. Batool, Y.-C. Byun[5]	2023	Wisconsin Dataset	Random Forest	~98%	Random Forest	No imbalance handling discussed
6.	E. S. Mohamed et al.[6]	2023	Mental health dataset	SVM, MLP, RF	~94%	Hybrid Model	Not focused on breast cancer

7.	S. Laghmati et al.[7]	2024	Breast Cancer Dataset	PCA + ML (RF, SVM, others)	~99%	PCA + RF	Complex preprocessing; interpretability reduced
8.	M. Darwich, M. Bayoumi[8]	2024	Multiple breast cancer datasets	Various ML models	~96–99%	Comparative Study	No single deployable model proposed
9.	A. Mehak et al.[9]	2025	Breast Cancer Dataset	RF, LR, SVM, KNN	~98%	Random Forest	Small dataset; no feature explainability
10.	A. T. Garba, H. S. Hamza[10]	2025	Breast Cancer Dataset	Interpretable ML, RF, Tree-based models	~97%	Explainable RF Model	Accuracy slightly lower than black-box models
11.	A. Bilal et al.[11]	2025	Breast cancer imaging dataset	SqueezeNet + SVM	~99%	Quantum-optimized SqueezeNet-SVM	High computational complexity
12.	A. Yaqoob et al.[12]	2025	Breast Cancer Dataset	SGA + RF	~99.3%	SGA + Random Forest	Complex feature selection process
13.	Sharifah Nurulhikmah Syed Yasin et al. [13]	2025	Wisconsin Breast Cancer Dataset	RF	91.23%	RF	Missing ROC-AUC evaluation, No cross-validation strategy

Proposed Methodology

The presented methodology targeting tumor classification using ML consists multiple stages including dataset gathering, cleaning, variable analysis, handling imbalance, model development, as well as validation. The provided architecture is designed to increase accuracy and reduce redundancy while maintaining model efficiency.

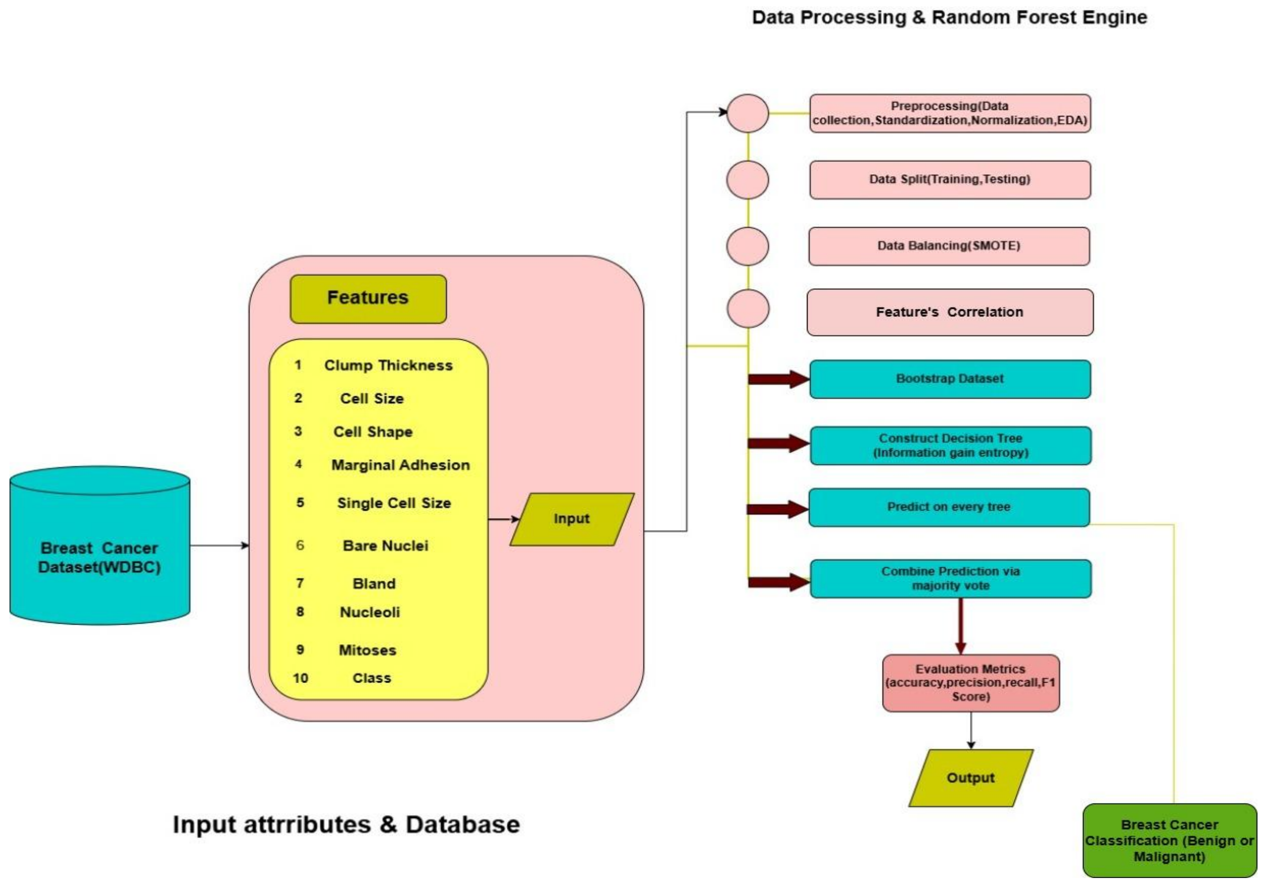

Figure 1: Machine learning Flowchart

1. Data Acquisition

WBCD is used in this research collected from Kaggle having 682 rows and 10 columns which includes various medical attributes like cell shape, cell size, clump thickness and many more.

This dataset is imported using Pandas library of Python for preprocessing and further analysis.

2. Libraries and Tools Used

Several Python computational utilities were used during this work for datapreparation, visualization, classifier training, imbalance handling, and performance evaluation.

Library / Tool Purpose

Pandas Data loading and preprocessing

NumPy Numerical computations

Matplotlib Graph plotting and visualization

Seaborn Correlation heatmap visualization

Scikit-learn Machine learning model development

RandomForestClassifier Breast cancer classification

LabelEncoder Converting categorical labels into numerical values

SMOTE Handling imbalanced dataset

Train-Test Split Splitting training and testing data

Accuracy Score Measuring overall model accuracy

Precision Score Measuring correctness of positive predictions

Recall Score Measuring sensitivity

F1-Score Balancing precision and recall

Confusion Matrix Classification performance analysis

ROC Curve & AUC Evaluating classification capability

3. Data Preprocessing

It is an important step because in this we handle null, missing and duplicate values by mean, mode and median. But, for this purpose firstly we have to understand the dataset and apply suitable operations to improve the quality of dataset.

a.) Data Understanding and Exploration

Exploratory Data Analysis is performed to analyze architecture and statistical properties of the data. The following insights were conducted:

- Checking dataset dimensions
- Viewing feature names and data types
- Generating statistical summaries
- Analyzing class distribution

This step helped in identifying feature behavior and understanding the overall dataset structure.

b.) Missing and Null Value Analysis

To check missing or null values in the dataset below given method is used.

- isnull()

But, this dataset does not contain any null and missing value, we do not need to apply any operation for handling null values.

c.) Duplicate Record Verification

This method is used to ensure data consistency and avoid to learning same pattern during model training.

However, this WBCD dataset have 234 row's duplicate value after removing these, it contains 448 rows and 10 coulmns.

d.) Data Encoding

It changes categorical value into numeric which is required by machine learning algorithms. In this dataset Class column contains categorical data, we have to change it into numeric data using label encoder such as :

- Benign → 0
- Malignant → 1

e.) Outlier Detection and Treatment

Values outside the acceptable lower and upper limits were treated as outliers. These are identified using Boxplot. Outliers in this dataset are:

These are handled using IQR (Interquartile Range) method using below formula's

$IQR = Q3 - Q1$

$Lower = Q1 - 1.5(IQR)$

$Upper = Q3 + 1.5(IQR)$

4.) Training Testing procedure

Provided patient database is categorized into:

- 80% training data
- 20% testing data

5.) Class Balancing (SMOTE)

It contained irregularities in class, SMOTE is used for training.

SMOTE creates artificial observations for the underrepresented class, helping the model to establish balanced boundary and improving classification performance.

6.) Feature's Correlation

This is used to show the relation of two features and tells which features are highly connected. This visualization is done in this study using heatmap.

7.) Bootstrap Sampling

Bootstrap sampling was applied by randomly selecting observations with replacement from the training dataset. Random feature sampling was also performed to generate diverse feature subsets.

This technique helps simulate the ensemble learning mechanism used in Random Forest classification.

8.) Random Forest Model Development

The final classification architecture constructed using RF model. It aggregates numerous decision trees to optimize:

- Classification
- Generalization capability
- Robustness against overfitting

8.1) Hyperparameter Optimization

To optimize the predictive efficacy and robustness capability of the Random Forest classifier, hyperparameter optimization was performed during architecture implementation.

- Number of estimators = 1000
- Maximum tree depth = 5

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 1-12 March 2026-Sept 2026 www.csjournals.com

- Minimum samples required to split a node = 50
- Minimum samples required at leaf node = 10

9.) Results and Discussion

9.1) Model Performance Analysis

- **Training and Testing Accuracy**

Table 2: Training and Testing Accuracy

Accuracy Type	Values (in %)
Training Accuracy	97.20
Testing Accuracy	94.44

- **Cross Validation Results**

Table 3: Cross validation analysis

Validation Method	Mean Accuracy (in %)
20-Fold Cross Validation	95.51

9.2) Attribute Relevance

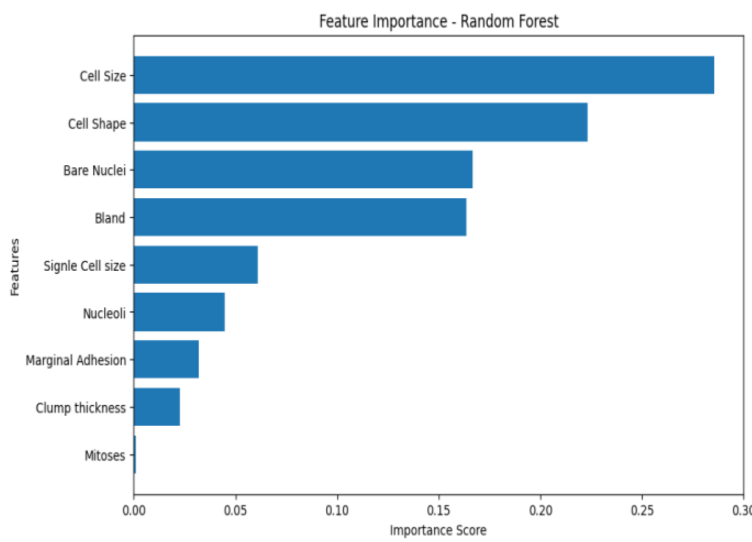

Figure 2: Feature's Contribution

9.3) Evaluation Metrics

a.) Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

b.) Precision = $\frac{TP}{TP+FP}$

c.) Recall = $\frac{TP}{TP+FN}$

d.) F1-Score = $2\times\frac{Precision+Recall}{Precision\times Recall}$

Table 4: Random Forest Evaluation Metrics

Random Forest Evaluation Metrics	Benign (in %)	Malignant (in %)
Accuracy	94.44	94.44
Precision	97	92
Recall	91	98
F1-Score	94	95
Support	43	47

9.5) Confusion Matrix Analysis

Following confusion matrix shows:

- **39** cases known as **True Negatives**.
- **46** cases known as **True Positives**.
- **4** cases known as **False Positives**.
- **1** case known as **False Negatives**.

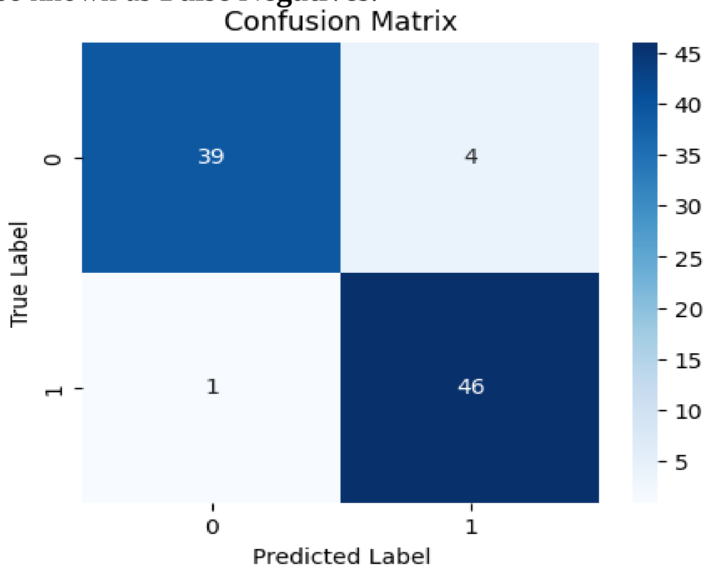

Figure 3: Confusion Matrix

9.5) ROC Curve and AUC Analysis

The algorithm which are used in this framework provides an AUC score of 0.9876 which highlight the best performance of classification.

- TPR = $\frac{TP}{TP+FN}$
- FPR = $\frac{FP}{FP+TN}$

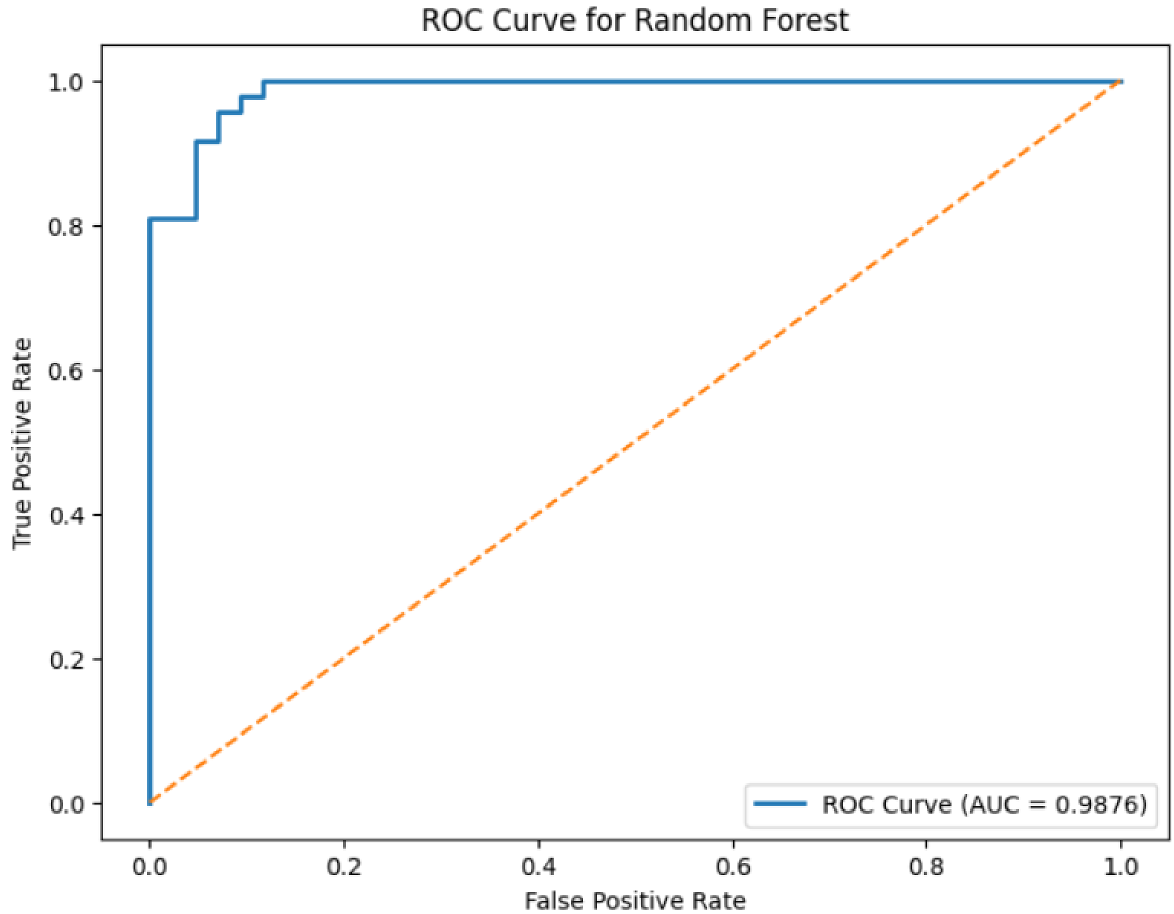

Figure 4: ROC-AUC Curve

9.6) Actual Value vs Predicted Value

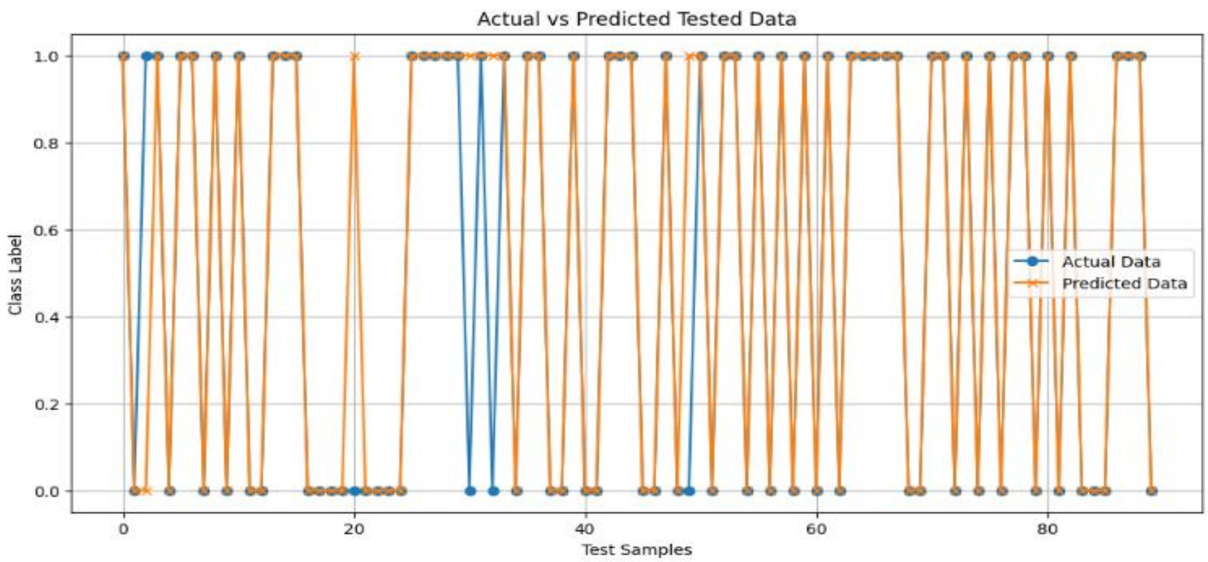

Figure 5: Actual vs Predicted test data visualization

Conclusion and Future Scope

The used Random Forest model gives 97.20% training, 94.44% testing and 95.51% cross validation accuracies. All in all, these evaluation metrics show strong capability for prediction of breast cancer. Feature Selection also plays crucial role in tumor classification.

The above used area under curve defines difference between benign and malignant tumors.

In future, the provided system can be enhanced by utilizing various deep learning techniques and ensemble ML models. In addition, this system can operate on distinct data.

References

[1] T. L. Octaviani and Z. Rustam, "Random forest for breast cancer prediction," AIP Conference Proceedings, vol. 2168, no. 1, Art. no. 020050, Nov. 2019, doi: 10.1063/1.5132477.

[2] M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, and M. N. Kabir, "Breast cancer prediction: A comparative study using machine learning techniques," SN Computer Science, vol. 1, no. 290, pp. 1–14, 2020, doi: 10.1007/s42979-020-00305-w.

[3] B. O. Macaulay, B. S. Aribisala, S. A. Akande, B. A. Akinnuwesi, and O. A. Olabanjo, "Breast cancer risk prediction in African women using Random Forest Classifier," Cancer Treatment and Research Communications, vol. 28, Art. no. 100396, 2021, doi: 10.1016/j.ctarc.2021.100396.

[4] M. Minnoor and V. Baths, "Diagnosis of breast cancer using random forests," Procedia Computer Science, vol. 218, pp. 429–437, 2023, doi: 10.1016/j.procs.2023.01.025.

[5] A. Batool and Y.-C. Byun, "Breast cancer classification using Random Forest algorithm," Journal of Physics: Conference Series, vol. 2559, no. 1, Art. no. 012002, 2023, doi: 10.1088/1742-6596/2559/1/012002.

[6] E. S. Mohamed, T. A. Naqishbandi, S. A. C. Bukhari, I. Rauf, V. Sawrikar, and A. Hussain, "A hybrid mental health prediction model using Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms," Healthcare Analytics, vol. 3, Art. no. 100185, 2023, doi: 10.1016/j.health.2023.100185.

[7] S. Laghmati, S. Hamida, K. Hicham, B. Cherradi, and A. Tmiri, "An improved breast cancer disease prediction system using ML and PCA," Multimedia Tools and Applications, vol. 83, pp. 33785–33821, 2024, doi: 10.1007/s11042-023-16874-w.

[8] M. Darwich and M. Bayoumi, "An evaluation of the effectiveness of machine learning prediction models in assessing breast cancer risk," Information Medicine Unlocked, vol. 52, Art. no. 101550, 2024, doi: 10.1016/j.imu.2024.101550.

[9] A. Mehak, T. M. Ali, A. Nawaz, and L. S. Parwati, "Breast cancer predictions using machine learning algorithms," in Proc. 7th Int. Global Conf. Series on ICT Integration in Technical Education & Smart Society, Aizuwakamatsu City, Japan, Jan. 20–26, 2025, Engineering Proceedings, vol. 107, p. 129, 2025, doi: 10.3390/engproc2025107129.

[10] A. T. Garba and H. S. Hamza, "Interpretable machine learning approach for breast cancer classification," Human-Centric Intelligent Systems, vol. 5, pp. 308–322, 2025, doi: 10.1007/s44230-025-00111-8.

[11] A. Bilal, A. Alkhathlan, F. A. Kateb, A. Tahir, M. Shafiq, and H. Long, "A quantum-optimized approach for breast cancer detection using SqueezeNet-SVM," Scientific Reports, vol. 15, art. no. 3254, 2025, doi: 10.1038/s41598-025-86671-y.

[12] A. Yaqoob, N. K. Verma, M. A. Mir, G. G. Tejani, N. H. B. Eisa, H. M. H. Osman, and M. A. Shah, “SGA-driven feature selection and random forest classification for enhanced breast cancer diagnosis: A comparative study,” Scientific Reports, vol. 15, art. no. 10944, 2025, doi: 10.1038/s41598-025-95786-1.

[13] Sharifah Nurulhikmah Syed Yasin, Aiman Azhar, and Rajeswari Raju, “Breast Cancer Prediction: A Random Forest-based System with Expert Validation,” Pertanika Journal of Science & Technology, vol. 33, no. S3, pp. 157–180, 2025, doi: 10.47836/PJST.33.S3.10.