International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 212-221 March 2026-Sept 2026 www.csjournals.com

Feature Selection Based Breast Cancer Prediction Model

Sonia, Dr. Vishal Verma and Ms. Neetu
[1]MCA Student, [2]Professor, [3]Asst Professor
Deptt. of Comp. Sc. and Appl, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract

Breast cancer among the more widely recognized and a very serious diseases of women around the world. Early detection and precise diagnosis is essential for increasing the survival rate of patients. The origins of this medical condition that could contribute for the occurrence of breast cancer are not well known. But not all features are effective to use in a dataset some features are useless and can even reduce accuracy. This study presents a comparative analysis based upon (ML) machine learning breast cancer classifying with feature selection using WBCD Dataset. We presented a four-stage hybrid feature selection pipeline, consisting of Chi-Square filtering, Mutual Information scoring, random forest embedded significance ranking, recursive feature elimination (RFE), upon reduce from nine attributes to only three such as Cell Shape, Cell Size, and Bare nuclei. Two models are implemented, including logistic regression, random forest. Standard model performance have been measured with precision, recall, F1-Score, confusion matrix (CM) and ROC AUC curve. Logistic Regression obtained accuracy 92% and Random Forest obtained an accuracy of 94% on WBCD dataset. Random forest performs best and effective in breast cancer prediction.

Keywords: Breast Cancer, Classification, Random Forest, Logistic Regression, Feature Selection, ROC-AUC, WBCD Dataset, Machine Learning (ML).

1. Introduction

Breast cancer is a major factors that lead to death of females worldwide. The cancer is primary and secondary most common leading cause of death in humans in the world, based on data obtained in 2019 from the World Health Organization [1]. In addition, there might be 2.3 million diagnoses and 685000 dying around the globe in 2020. By the end of 2020, 7.8 million females who were living over the last five years had been identified as having breast cancer, which is considered the world's most malignant tumors common [2]. Breast cancer is the uncontrolled growth of some of the cells in the breast. [1]. The prognosis is significantly improved with early diagnosis, and research shows that the earliest diagnosis can result in five-year survival rates up to 90% [3]. However, conventional diagnostic procedures such as mammography, biopsy, histopathological analysis, magnetic resonance imaging (MRI) and ultrasound, although effective in detecting breast cancer, are expensive, time-consuming and require specialized expertise [3,6]. Things like age and having a beloved ones history of breast cancers can increase a woman's threat for breast cancer.

Here different kinds of tumor Benign and Malignant. Not benign is harmful for human beings or does not usually cause human death. The tumor of cancer occurs when breast tissue cells grow improperly. Malignant is very dangerous and can cause death. The breast cancer treatment options depend on the individual age and stage of cancer and category. Therapy can be a combination of different therapies such as Chemotherapy, Radiotherapy, Surgery etc. [5]. Classification of breast cancer has been done using different ML techniques. Machine learning (ML) techniques achieved popularity in the medical care industry for their ability to identify and categorize various diseases, particularly in the presence of breast cancers. But many ML algorithms do not work well due to the availability of irrelevant features. The importance of proper data organization and decreased dimensionality is emphasized before using ML algorithms as it can influence the results. Precise, consistent and noise-free data make detection and classification of disease more efficient [8]. This paper used Wisconsin Breast Cancer Dataset and we applied a hybrid FS pipeline (Chi-square, Mutual Information, Random Forest Importance, RFE) to minimize nine features to three. LR and RF classifiers were trained and evaluated. Random Forest achieved best performance.

2. Literature Review

Hasan et.al, The preliminary and early work has shown that diagnostic performance has been reliable with the use of ML classifiers with structured feature selection on Wisconsin breast cancer datasets. Applied a Wrapper based feature selection method using a WEKA 's Classifier Attribute Eval with Best First search, applying Logistic Regression, LSVM, and QSVM on three Wisconsin datasets (WBCO, WDBC and WPBC). LR with feature selection got 97.1% accuracy upon WBCO dataset, while QSVM without feature selection got the best accuracy 97.9% on the WDBC dataset, which indicates that the effects of feature selection are different for different classifiers and datasets [1]. Maccido et. al, Similarly, when it employed a set of standard classifiers such as SVM, KNN, Decision Tree or Naïve Bayes on WDBC dataset, SVM achieved a very good accuracy of 97% [4]. Silaich et.al, A Systematic analysis of Four attribute selection Methods – Correlation -based selection, Information Gain, Sequential Feature Selection (FS), Recursive Feature Elimination (RFE) — paired with SVM, RF, Logistic Regression, ANN, XGBoost upon WDBC dataset. The maximum accuracy 97% was achieved by SVM combined with RFE, and all models obtained accuracies above 90%, which confirms that wrapper-type methods like RFE which assess subsets of features using classifier feedback, always provide better discriminative performance than simple filter methods [3]. Garba et.al, performed RFE and PCA on WDBC datasets with SVM, Random Forest, Logistic Regression, Random Forest achieved the highest accuracy 98% [5]. Awadelkarim utilized the Extra Trees method for feature selection on the WBCD Dataset, with SVM getting best accuracy of 98%, which shows that the tree-based importance classification simultaneously reduces redundancy and improves the predictive accuracy [7]. Shaban presented a New Hybrid Feature Selection Method that combines Information Gain as a fast pre-selection filter and a hybrid Bat Algorithm and Particle Swarm Optimization as a wrapper refinement stage. Proposed strategy achieved 97% with a low error rate of 0.03 on a Kaggle mammography dataset [2]. Karimi suggested two wrapper attribute selection frameworks upon WDBC dataset based on Bat Algorithm (WFSB) and Imperialist

Competitive Algorithm (WFSIC) combined with nine ML classifiers. The BA-based method has a maximum accuracy of 99.12% with a reduction of feature dimension up to 90% (only 10 out of 30 features are selected), while WFSIC with AdaBoost achieved 98.95% accuracy [6].

3. Methodology

3.1 Data Collection: This study used WBCD that was extracted from UCI Machine Learning Repository. Original version was developed by Dr. William H. Wolberg, University Hospital, Madison, University of Wisconsin among 1989 and 1991, The data set consists of clinical information from Fine Needle Aspirates (FNA) lumps in the breast [6]. The breast cancer dataset corresponds to 682 instance , the benign cases are 443 (64.96%) and 239 (35.04%) are malignant cases. The dataset contains ten numerical features with nine diagnostic features and one target class. These features characterizes the properties of cell nuclei.

3.2 Data Preprocessing:

Accurate or consistent data is critical for efficient model training. Raw datasets, like the WBCD are often imperfect, with inconsistent information and noisy data that can reduce classification accuracy [6]. In the WBCD ,234 Duplicate rows were found. Dataset size changed from 682 to 448 rows. This would lead to in duplicate rows in, causing the model to see the same data multiple times during training. This would artificially increase the training accuracy and create overfitting (the model learns the duplicated records too well, instead of learning general patterns).

In this dataset, The target attribute ‘class’ had string classifications (Benign and Malignant). Then, they were transformed into binary integers values with the Label Encoding: Benign, 0 (negative class) Malignant, 1 (positive class). This is necessary because all the ML algorithms require numerical labels to measure the errors and update the weights in training.

3.3 Feature Selection Techniques:

Feature selection aims at reducing the numbers of attributes by eliminating irrelevant, unreliable features and enhancing a performance of the classification [1].

1. Chi-Square (Chi2) Filter Method: A Chi-Square filter method is a FS technique used in order to test a statistical independence of classification features with respect to a target variable.

2. Mutual Information (MI) Filter Method:

Mutual Information filter method is not-parametric feature selection technique which is widely used in data science.ML that determines the statistical dependence between an input feature and the target variable.

Intersection of Chi2 and MI Selections:

Features were selected based on the intersection of Chi2 and MI. Features common to both selections: Cell Size, Cell Shape, Bare Nuclei, Marginal Adhesion, Bland, Nucleoli (6 features). Features that were present in only one of the two were excluded. This double verification provides better proof that these 6 features are really important.

3. Random Forest Feature Importance

It determines the impact of each attribute on the accuracy of the prediction system's. It is useful to identify the most influential input variables, improve performance, interpretability and computational efficiency.

4. Recursive Feature Elimination (RFE)

RFE is a feature selection method in which an base estimator is iteratively trained, feature importances or coefficients are extracted and the feature with the smallest importance is removed at each recursion. This process is repeated until the right number of features is obtained.

3.4 Machine Learning Classification Algorithms

ML is a subfield of artificial intelligence that's focus on methods which may “learn” the patterns Using training data and then make right inferences on the new data. The pattern recognizing capability allows machine learning models to take standard decisions or make predictions without having to be specifically configured to do so. Breast cancer prediction is done using two major methods of classification of machine learning namely LR and RF classifier.

3.4.1 Logistic Regression

LR is a robust supervised machine learning approach. This describes a relation between a output variables as well as all the variables which affect it features. The contribution of each variable to the last fitting is readily got it and the results of back-fitting when the data can be directly assessed the corresponding probabilities [1]. It uses sigmoid function to map input characteristics corresponding to the probability of being in a certain class. In this study, it was used as an initial linear classifier to assess the effect of selecting features on the accuracy of standardized classification in breast cancer detection [6].

3.4.2 Random Forest

RF is a collections of decision trees and regression trees. They are basic models that use binarized split-tings on predictive parameters to predict an result. In the setting for the random forest, many categorization and regression tress are built by using a randomly selected training dataset and a random choice of predictor variables to model the outcome. The scores for each tree are added together to give the prediction. The random forest has great advantage in predictive modeling in general. It can process the data set with multiple predictors [5].

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 212-221 March 2026-Sept 2026 www.csjournals.com

3.4.3 Evaluation Metrics

The models were validated by a combination of classification measures and computational efficiency ones in order to confirm their predictive quality and practical applicability [3]. The metrics used for evaluation were as follows:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

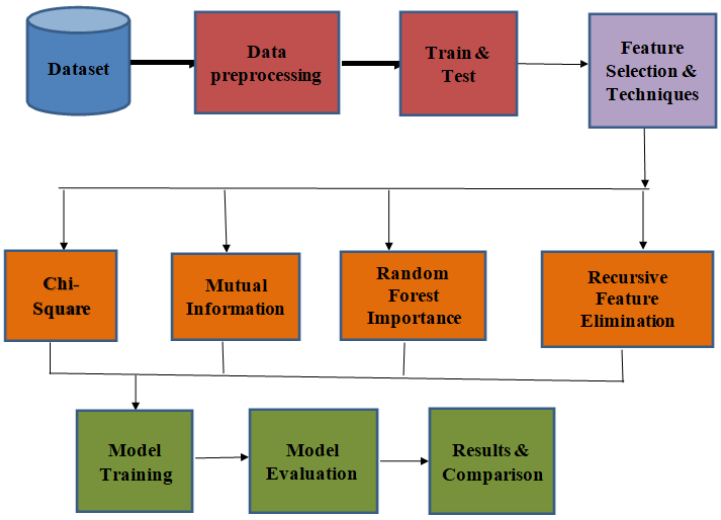
$$\text{Precision} = \frac{TP}{(TP + FP)}$$

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

$$\text{F1- Score} = \frac{2^{*}\ (recall * precision)}{(recall + precision)}$$

$$\text{Specificity} = \frac{TN}{(TN+FP)}$$

3.5 Work Flow Chart



4. Results and Discussion

A comparative standard evaluation of two different models is done in this work. To check the model performance, confusion matrix, Feature importance and ROC-AUC Curve are used.

Metric	Logistic Regression	Random Forest
Accuracy	92%	94%
Precision	94%	94%
Recall	91%	94%
F1-Score	92%	94%

Table1: Model Performance Comparison

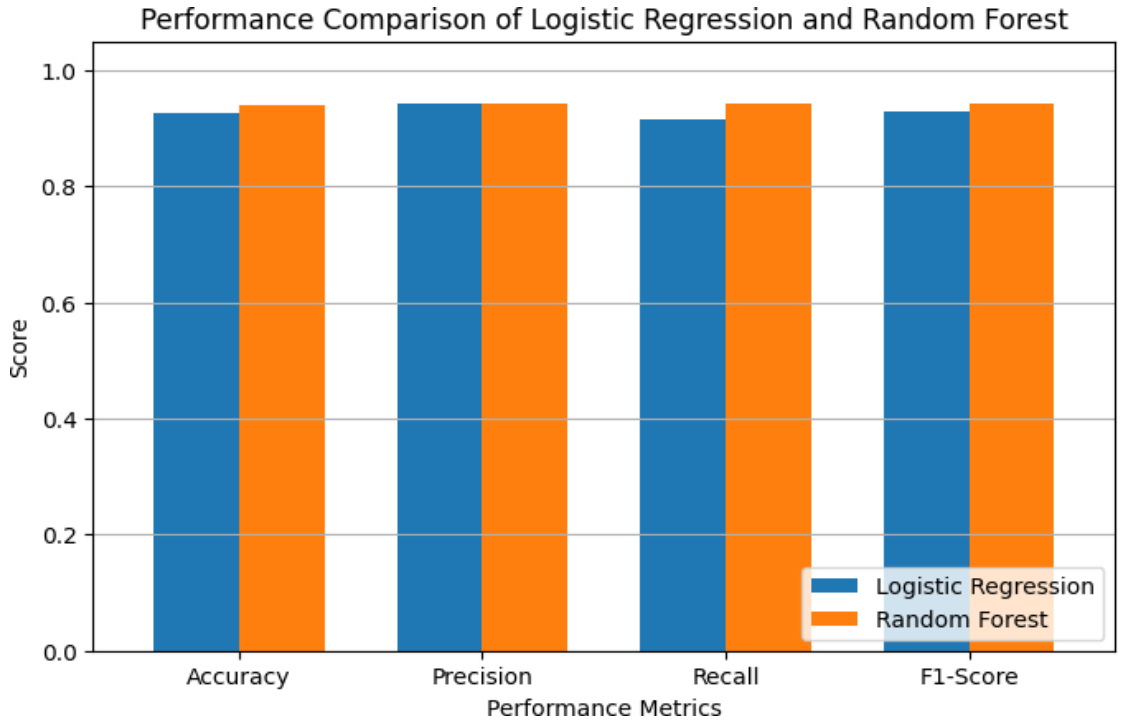


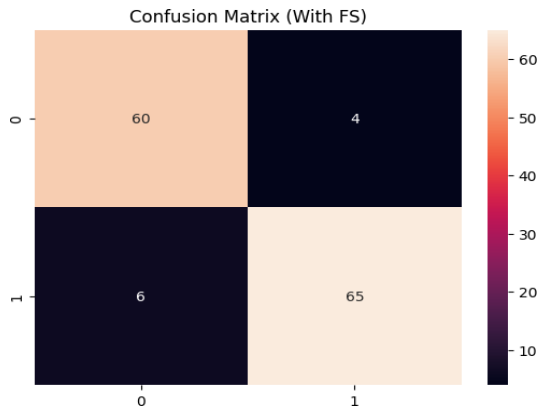
Fig2. Model performance graph

Figure 2. shows a comparison of the performance for Logistic Regression and Random Forest based on the four evaluation metrics Accuracy, Precision, Recall ,F1-Score. Random Forest always performed better than Logistic Regression in all metrics. Random Forest model provided accuracy 94%, precision 94%, recall 94% and F1-score is also 94%, respectively. Logistic Regression model showed accuracy 92%, precision 94%, recall 91% and F1-score 92%, respectively. The biggest

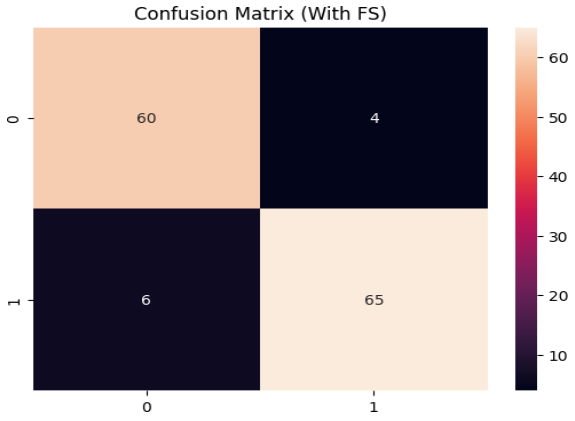
difference was in Recall (2.6 percentage points), with Random Forest being better at correctly identifying malignant cases, which is an important property for cancer diagnosis."

Confusion Matrix Analysis:

Confusion matrix plays important role in evaluating classification problem. There are 443 benign cases and 239 malignant cases in the test dataset.



Logistic Regression



Random Forest

The assessment of the confusion matrix has confirmed that the minimum classification error rate of all the configurations tested was obtained with the Random Forest model with feature selection. Matrix showed a high true-positive rate for the detection of malignant tumours (high sensitivity/recall) which is of great importance in the clinical setting of screening for breast cancer as the consequence of missing a malignancy is a more serious error than a false-positive result. Logistic Regression was also highly sensitive, and had a similar false negative rate.

Feature Selection Importance

Random Forest model a feature importance scores after feature selection are shown in Figure 3. The selected features ranked according to their effect on breast cancer classification. Within the selected features, Cell Size had highest importance score ~0.30, showing that it was the most effective predictor in differentiating between benign and malignant breast tumors. Cell Shape was second with an importance score of ~0.22, indicating a strong effect on the model's predictive accuracy. The other features like Bare Nuclei 0.17, Bland Chromatin 0.16 were also important in the classification process. The lowest importance scores based on the selected features were for Nucleoli (around 0.08) and Marginal Adhesion (around 0.07). This indicates that while these features provide useful diagnostic information, their impact to the final prediction is relatively limited. Cell Size and Cell Shape are the most selective characteristics for breast cancer prediction. The results shows that the RF model heavily depends on a few most relevant features for precise

classification of breast cancer.

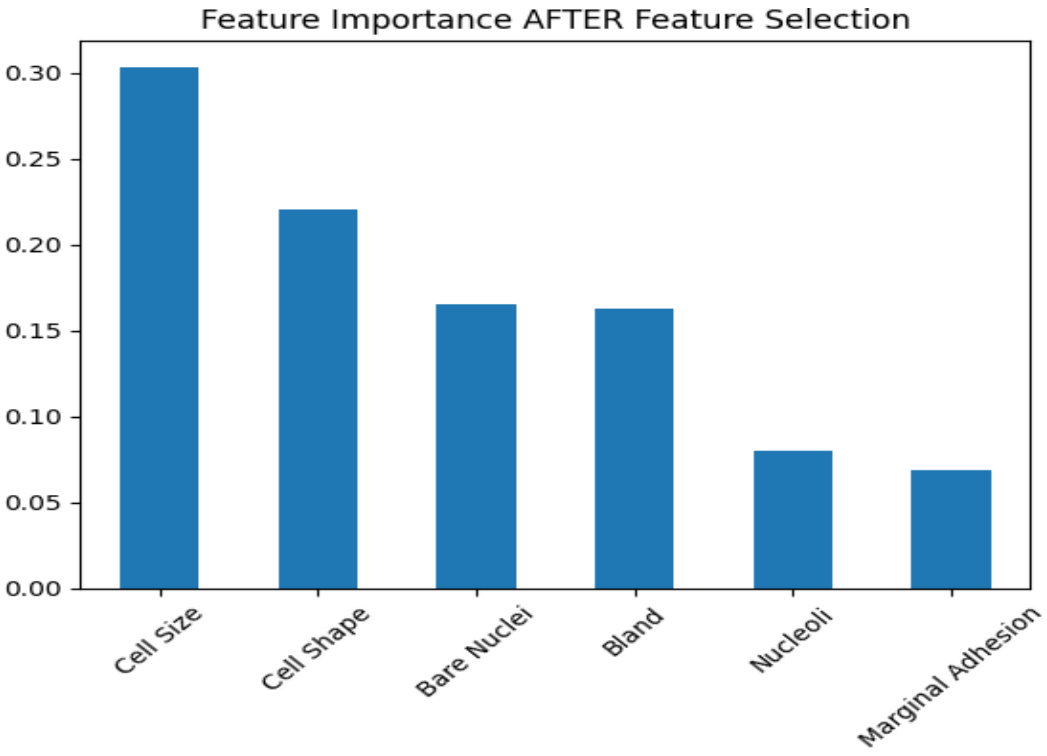
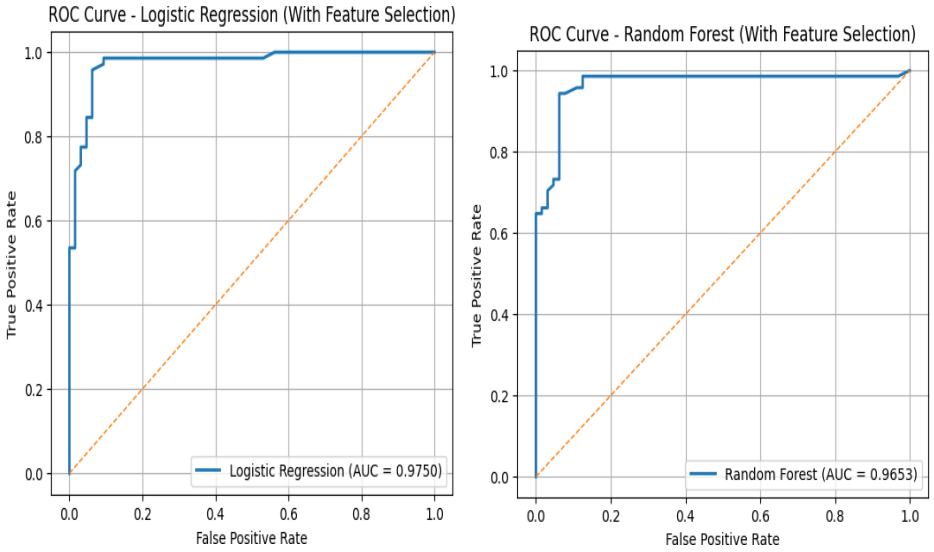


Fig3. Feature Importance

AUC-ROC

Area under the Receiver Operating Characteristic curves for LR and RF classification are shown respectively. The classifiers were analyzed after applying the hybrid feature selection pipeline. The AUC score of Logistic Regression was 0.9750 and of Random Forest was 0.9653, both significantly higher than 0.50 score of a random classifier, indicating good classification capability to distinguish in between the benign and malignant tumour classes. The ROC curve of Logistic Regression shows a steep slope to the upper left corner for very low false positive rates, which means that the model has high true positive rates as keeping false alarms to a minimum. The ROC curve of Random Forest is also characterized by a quick rise at the beginning, reaching a true positive rate of about 0.95 at a false positive rate of 0.15, which indicates a high sensitivity for a range of classification limits. In fact, both models can be considered as outstanding classifiers according to the standard criteria for AUC interpretation (AUC > 0.95), which further confirms the performance of the proposed feature selection pipeline in maintaining class-discriminative information in a limited feature subset.



Conclusion and Future Scope

This study, have a comparison analysis of feature selection based breast cancer prediction using Logistic Regression and RF classifiers with a hybrid feature selection procedure is performed. Feature selection was done using Chi-Square, Mutual Information, Random Forest feature importance and Recursive Feature Elimination to get the most applicable features. Both models were evaluated in terms of accuracy, precision, recall, F1-score, ROC-AUC curve, confusion matrix, and feature importance evaluation. Logistic Regression performed well in classification but Random Forest classification outperformed with higher accuracy and better predictive ability in all cases**.**

Results indicate that the proposed hybrid feature selection technique enhances the working of model by reducing dimensionality of data and maintaining most relevant diagnostic features. The study was performed based on a single breast cancer dataset which might limit the generalization of the model to different medical datasets.

In the future work, the dataset could be augmented with bigger and more distinct clinical datasets from various healthcare organizations to enhance model accuracy and standardization. We may explore more advanced ML and deep learning methods like XGBoost, LightGBM, CatBoost, Artificial Neural Networks and Convolutional Neural Networks to improve the accuracy of predictions. The transparency and interpretability of the prediction process can be improved by utilizing Explainable Artificial Intelligence (XAI) such as SHAP and LIME.

References

[1] R. Hasan and A. S. M. Shafi, "Feature selection based breast cancer prediction," Int. J. Image Graph. Signal Process. (IJIGSP), vol. 15, no. 2, pp. 13–23, Apr. 2023, doi: 10.5815/ijigsp.2023.02.02.

[2] K. Karimi, A. Ghodratnama, and R. Tavakkoli-Moghaddam, "Two new feature selection methods based on learn-heuristic techniques for breast cancer prediction: A comprehensive analysis," Kharazmi University,
Tehran, Iran. [Online]. Available: https://github.com/kamyab24/Machine-learning-Algorithms.git

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

[4] S. Silaich and R. Yadav, "Performance analysis of breast cancer classification using feature selection and machine learning," NeuroQuantology, vol. 20, no. 19, pp. 5043–5048, Nov. 2022, doi: 10.48047/nq.2022.20.19.nq99467.

[5] S. A. M. Awadelkarim, "Feature selection using extra trees for breast cancer prediction," Indones. J. Comput. Sci., vol. 13, no. 3, pp. 1808–1817, 2024.

[6] H. S. Maccido, "Non-linear feature selection and stacked ensemble learning for breast cancer diagnosis: A fusion of statistical and neural classifiers," J. Stat. Sci. Comput. Intell., vol. 1, no. 2, pp. 85–105, Jul. 2025, doi: 10.64497/jssci.21.

[7] W. H. Wolberg, "Breast Cancer Wisconsin (Original) Dataset," UCI Machine Learning Repository, 1992. [Online].
Available: https://www.kaggle.com/datasets/choupeiyuan/breast-cancer-wisconsin-original

[8] W. M. Shaban, "Insight into breast cancer detection: new hybrid feature selection method," *Neural Computing and Applications*, vol. 35, pp. 6831–6853, Dec. 2022, doi: 10.1007/s00521-022-08062-y.

[9] J. Zhu, Z. Zhao, B. Yin, C. Wu, C. Yin, R. Chen, and Y. Ding, "An integrated approach of feature selection and machine learning for early detection of breast cancer," *Scientific Reports*, vol. 15, no. 1, p. 13015, Apr. 2025, doi: 10.1038/s41598-025-97685-x.

[10] V. Nanda Gopal, F. Al-Turjman, R. Kumar, L. Anand, and M. Rajesh, "Feature selection and classification in breast cancer prediction using IoT and machine learning," *Measurement*, vol. 178,
p. 109442, 2021, doi: 10.1016/j.measurement.2021.109442.