iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75793

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

Comparative Study of Feature Selection Techniques for Breast Cancer Prediction

Anagha P[1], Sajana T[2]
Assistant Professor, Department of Computer Science, Chinmaya Institute of Technology Kannur University, Kerala, India

Abstract: ***This study presents a comparative analysis of five feature selection methods—Chi-Square, Mutual Information, RFE, LASSO, and Random Forest Importance—applied to the Breast Cancer Wisconsin Diagnostic dataset. Their effectiveness was evaluated using Logistic Regression, SVM, and Random Forest classifiers based on accuracy, F1-score, ROC-AUC, runtime, and Jaccard-based stability. RFE achieved the highest predictive performance, whereas Chi-Square and Mutual Information provided the strongest stability and fastest computation. Random Forest Importance offered a balanced trade-off, while LASSO showed reduced stability due to aggressive regularization. The results highlight clear performance–stability trade-offs and provide practical guidelines for selecting reliable feature selection techniques in breast cancer prediction.***
Keywords: Feature Selection, Breast Cancer Prediction, Machine Learning, Classification Algorithms, Model Evaluation

I. INTRODUCTION

Breast cancer remains one of the most prevalent and life-threatening diseases affecting women worldwide, accounting for a significant proportion of cancer-related morbidity and mortality [1], [2]. Early and accurate diagnosis plays a crucial role in improving survival rates, and with the increasing availability of biomedical datasets, machine learning (ML) has become a powerful tool for developing predictive models to support clinical decision-making [3], [4]. However, high-dimensional medical data often lead to challenges such as overfitting, reduced interpretability, and increased computational cost, highlighting the importance of effective dimensionality reduction and feature selection techniques [5], [6].

Feature selection aims to identify the most informative attributes that significantly contribute to prediction performance while eliminating redundant or irrelevant features [5], [7]. These techniques are generally categorized into three groups—filter, wrapper, and embedded methods—each differing in methodology, computational requirements, and dependency on learning algorithms.

Although several studies focus on improving classification accuracy in breast cancer prediction, many overlook key factors such as feature stability and runtime efficiency. Stability refers to the consistency of selected features across multiple data samples or folds, which is essential for ensuring reproducibility and trustworthiness of ML models in clinical settings [8], [9].

In this study, a comparative analysis of five feature selection methods—Chi-Square, Mutual Information, Recursive Feature Elimination (RFE), LASSO regularization, and Random Forest Importance—is carried out using multiple classifiers. The evaluation considers accuracy, F1-score, ROC-AUC, feature stability (via Jaccard similarity), and computational time, providing a comprehensive understanding of the trade-offs between predictive performance, robustness, and efficiency in breast cancer prediction.

II. METHODOLOGY

The methodology includes data preprocessing, application of feature selection methods, model training using multiple classifiers, and evaluation through predictive accuracy, feature stability, and runtime analysis.

A. Dataset Description

The experiments were conducted using the publicly available Wisconsin Breast Cancer Diagnostic (WBCD) Dataset, originally introduced by Wolberg et al. [10] and later made accessible via the UCI Machine Learning Repository [11]. The dataset contains 569 samples with 30 features. 0 indicates Benign and 1 indicated Malignant tumour. The features represent computed attributes of cell nuclei extracted from digitized images, including mean, standard error (SE), and worst values of radius, texture, perimeter, smoothness, concavity, and related morphological descriptors [10], [11].

B. Data Preprocessing

The following preprocessing steps were applied:

1) *Handling missing values:* The WBCD dataset contains no missing values; hence no imputation was required [11].
2) *Feature standardization:* All features were scaled using StandardScaler, which improves performance of models such as SVM and Logistic Regression [12].
3) *Train-Test split & Cross validation:* A stratified 5-fold cross-validation approach was used to preserve class distribution across folds, aligning with best practices for medical datasets [13].

C. Feature Selection Techniques

Five feature selection methods from filter, wrapper, and embedded categories were evaluated.

1) *Chi-Square Test (Filter):* The Chi-Square test measures statistical dependence between each feature and the target variable. Features with the highest χ^2 scores were selected. For each feature, a Chi-Square statistic is calculated by comparing the observed frequency of feature-class combinations with the expected frequency if the feature and class were independent. Features with higher Chi-Square values indicate stronger association with the target class and are considered more informative for classification [14].
2) *Mutual Information (Filter):* Mutual Information (MI) measures both linear and nonlinear dependency using entropy-based information gain. Mutual Information measures the amount of information shared between a feature and the class label, quantifying how much knowing the feature reduces uncertainty about the class. Features with higher mutual information values are more relevant, as they provide greater predictive power for classification. This method can handle both categorical and continuous variables (with discretization) and is widely used to select informative features while reducing dimensionality [15].
3) *Recursive Feature Elimination — RFE (Wrapper):* Recursive Feature Elimination is a wrapper-based feature selection method that recursively removes the least important features based on a model's weight or importance scores. At each iteration, the model is trained, feature rankings are computed, and the least significant features are eliminated until a desired number of features remain. RFE is effective for identifying a subset of features that maximizes model performance while reducing dimensionality [5].
4) *LASSO — L1 Regularization (Embedded):* LASSO is a regularization technique that performs both feature selection and regression by adding an L1 penalty to the loss function. The L1 penalty forces some feature coefficients to shrink to zero, effectively removing less important features from the model. This approach helps in selecting a subset of relevant features while preventing overfitting and improving model interpretability [16].
5) Random Forest Feature Importance (Embedded): *Random Forest ranks feature by their contribution to impurity reduction (Gini importance), enabling selection of the most influential attributes.* it calculates feature importance during model training by measuring how much each feature contributes to reducing impurity (e.g., Gini impurity) or affects prediction accuracy when permuted. Features with higher importance scores are more relevant and are automatically selected by the model, making Random Forest effective for identifying informative features while building a predictive model *[17*].

D. Classification Methods

Three supervised learning classifiers were selected:

1) *Logistic Regression (LR):* A widely used linear classifier suitable for medical diagnostic tasks [12].
2) *Support Vector Machine (SVM):* Used with an RBF kernel to capture nonlinear relationships in biomedical data [18].
3) *Random Forest (RF):* A robust ensemble classifier capable of modelling complex feature interactions [17].

Each classifier was trained using the feature subsets generated by the five feature selection methods.

E. Evaluation Metrics

1) *Predictive Performance:* Performance was assessed using widely adopted metrics such as Accuracy, F1-score and ROC-AUC. These metrics measure correctness, handling of class imbalance, and discriminative ability in medical classification tasks [12], [18].
2) *Feature Stability:* Feature stability across cross-validation folds was computed using the Jaccard Similarity Index, a standard metric for assessing reproducibility of selected features [8].

$$J(A, B) = |A \cup B| / |A \cap B|$$

3) *Runtime Efficiency: Execution time for each feature selection method was recorded to compare computational cost, following recommendations for FS evaluation in high-dimensional datasets [9].*

F. Experimental Procedure:

1) Pre-processed the dataset
2) Applied each feature selection technique
3) Extracted selected feature subsets
4) Trained classifiers using 5-fold cross-validation
5) Calculated feature stability via Jaccard similarity
6) Measured runtime for each method
7) Compiled comparative results for interpretation and analysis

III. RESULTS

This section presents the comparative evaluation of five feature selection techniques—Chi-Square, Mutual Information, RFE, LASSO, and Random Forest Importance—combined with three classifiers: Logistic Regression, Random Forest, and SVM. The results were analysed based on accuracy, F1-score, ROC-AUC, runtime, and feature selection stability.

A. Classification Performance:

TABLE 1 summarizes the performance metrics across all combinations of feature selection and classifiers. All models achieved high predictive performance, with ROC-AUC values ranging from 0.98 to 0.99, confirming the strong separability of the Wisconsin Breast Cancer dataset.

The highest performance was achieved by:

RFE + SVM: Accuracy: 0.9666**,** F1-Score: 0.9742, ROC-AUC: 0.9934

RFE + Logistic Regression: Accuracy: 0.9666, F1-Score: 0.9743

These results demonstrate that Recursive Feature Elimination consistently produced the most discriminative subset of features, likely because RFE iteratively removes least-important predictors based on model feedback.

LASSO achieved comparatively lower performance, with stability and AUC values lower than other methods. Filter methods like Chi-Square and Mutual Information performed reasonably well with lower computational cost.

TABLE I

Feature Selection	Classifier	Accuracy Mean	F1 Mean	ROC_AUC Mean	Runtime (s)
Chi-Square	Logistic Regression	0.9332	0.9484	0.9866	0.6992
Chi-Square	Random Forest	0.9455	0.9569	0.9848	3.5395
Chi-Square	SVM	0.9402	0.9531	0.9876	0.4828
Mutual Information	Logistic Regression	0.9332	0.9484	0.9882	2.0243
Mutual Information	Random Forest	0.9473	0.9581	0.9855	4.4263
Mutual Information	SVM	0.9420	0.9545	0.9889	3.3076
RFE	Logistic Regression	0.9666	0.9743	0.9932	2.5077
RFE	Random Forest	0.9613	0.9693	0.9864	5.1094
RFE	SVM	0.9666	0.9742	0.9934	1.1530
LASSO	Logistic Regression	0.9314	0.9473	0.9865	1.3304

Feature Selection	Classifier	Accuracy Mean	F1 Mean	ROC_AUC Mean	Runtime (s)
LASSO	Random Forest	0.9490	0.9598	0.9857	2.1189
LASSO	SVM	0.9385	0.9524	0.9900	0.8757
Random Forest Importance	Logistic Regression	0.9385	0.9526	0.9879	1.9020
Random Forest Importance	Random Forest	0.9455	0.9568	0.9839	3.0143
Random Forest Importance	SVM	0.9437	0.9560	0.9883	2.2718

B. Runtime Analysis

Runtime results indicate the computational efficiency of each feature selection method:

1) Chi-Square and SVM combinations were the fastest, often completing within 0.5–0.7 seconds.

2) Random Forest–based methods required the longest runtime due to tree-based training overhead.

3) RFE had moderate runtime, balancing computational cost with high accuracy.

This shows that Chi-Square and Mutual Information are suitable when quick model development is required, whereas feature elimination (RFE) is preferred for maximizing accuracy.

C. Feature Selection Stability

Stability was assessed using Jaccard similarity, measuring the consistency of selected features across data folds.

TABLE II

Feature Selection	Jaccard Stability
Chi-Square	1.00
Mutual Information	0.927
Random Forest Importance	0.945
RFE	0.794
LASSO	0.507

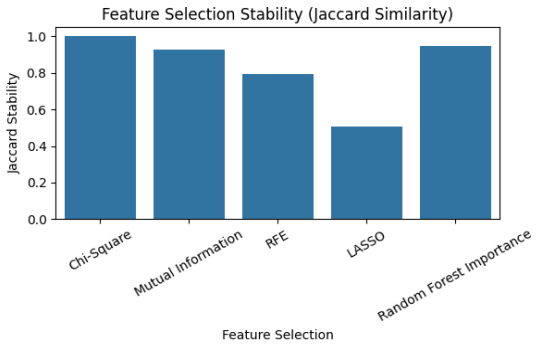


Fig 1. Stability Graph

Fig 1. Shows:

1) Chi-Square achieved perfect stability, consistently selecting the same features in each run.

2) Random Forest Importance and Mutual Information were also highly stable, with Jaccard scores above 0.90.

3) RFE showed moderate stability, likely due to its wrapper-based nature where small data variations affect feature rankings.

4) LASSO had the lowest stability, confirming findings from previous studies that L1-penalized models tend to be sensitive to dataset variations

D. Interpretation of Results

1) RFE offers the best classification performance. It yields the highest accuracy and ROC-AUC, making it ideal where predictive performance is the top priority.
2) Chi-Square is the most stable and computationally efficient**.** Perfect stability + fast execution makes it suitable for clinical settings where consistency is essential.
3) Random Forest Importance balances accuracy, stability, and interpretability. A highly stable method with strong AUC values.
4) LASSO is not recommended alone for this dataset. Lower feature stability and moderate performance suggest it may exclude important correlated features.

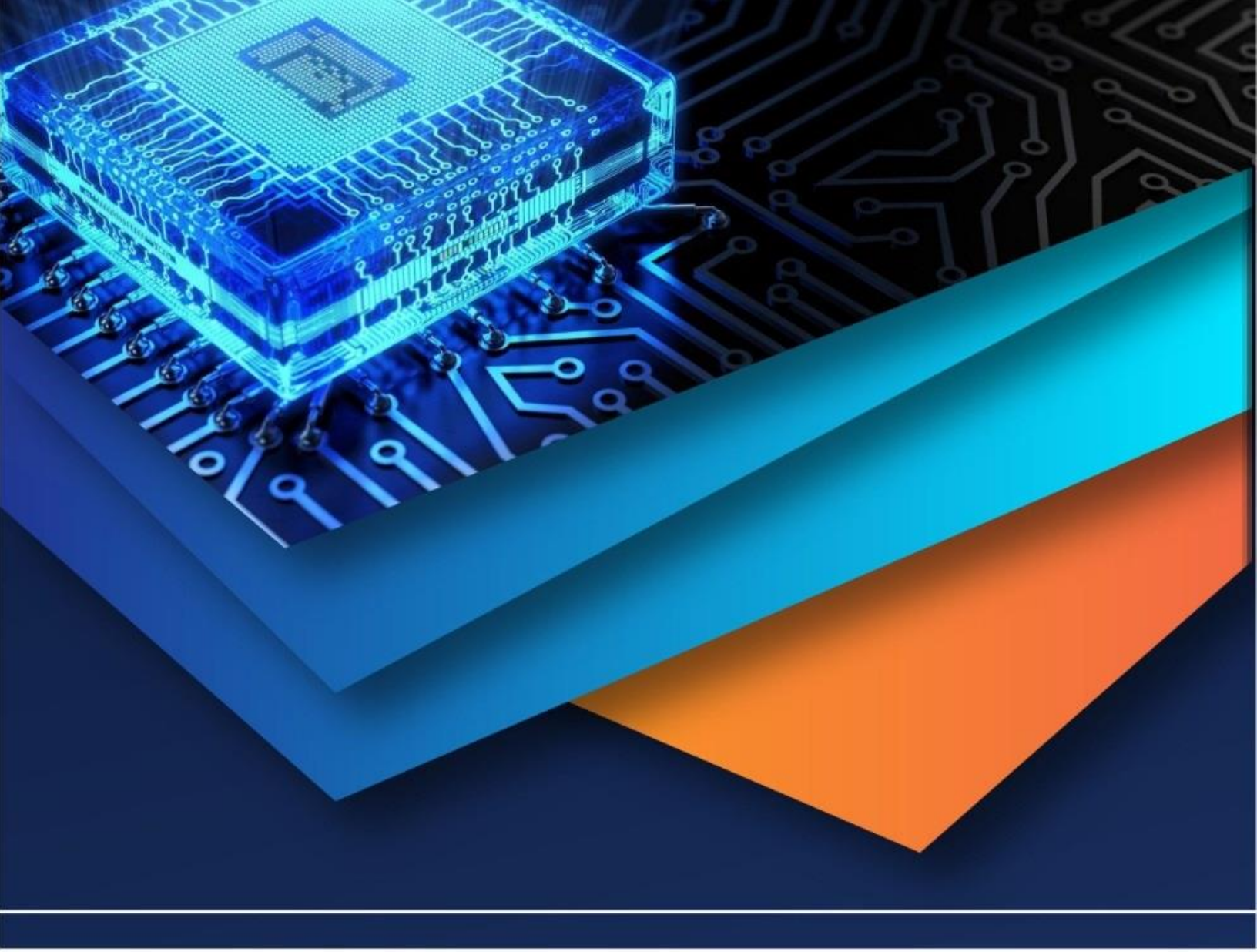
IV. CONCLUSIONS

This study provides a comprehensive evaluation of five feature selection techniques—Chi-Square, Mutual Information, RFE, LASSO, and Random Forest Importance—for breast cancer prediction using Logistic Regression, Random Forest, and SVM**.** The results reveal a clear trade-off between predictive accuracy, feature stability, and computational efficiency**.** RFE achieved the highest accuracy and ROC-AUC, demonstrating the strength of wrapper-based methods, while Chi-Square and Mutual Information offered the greatest stability and fastest computation, highlighting the reliability of filter-based approaches. Random Forest Importance provided a balanced compromise, whereas LASSO showed limited stability on small biomedical datasets.

These findings suggest that RFE with SVM or LR is optimal for maximizing predictive performance, whereas Chi-Square or Mutual Information is ideal when reproducibility and interpretability are prioritized. By integrating accuracy, stability, and runtime considerations**,** this study offers practical guidance for developing robust, interpretable, and clinically relevant machine learning models for breast cancer prediction.

REFERENCES

[1] World Health Organization, *Breast Cancer: Key Facts*. WHO, 2024.
[2] F. Bray et al., “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.
[3] L. Cruz and D. Wishart, “Applications of machine learning in cancer prediction and prognosis,” *Cancer Informatics*, vol. 2, pp. 59–77, 2007.
[4] S. Abdar et al., “A review of computational intelligence methods for breast cancer diagnosis,” *Neural Computing and Applications*, vol. 32, pp. 643–680, 2020.
[5] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
[6] H. Liu and H. Motoda, *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Springer, 1998.
[7] M. Robnik-Šikonja and I. Kononenko, “Theoretical and empirical analysis of ReliefF and RReliefF,” *Machine Learning*, vol. 53, pp. 23–69, 2003.
[8] M. Nogueira, K. Sechidis, and G. Brown, “On the stability of feature selection algorithms,” *Journal of Machine Learning Research*, vol. 18, no. 174, pp. 1–54, 2018.
[9] K. Somol and P. Pudil, “Feature selection toolbox and stability issues in high-dimensional data,” in *Proceedings of the 16th International Conference on Pattern Recognition*, vol. 2, pp. 556–559, 2002.
[10] W. Wolberg, W. Street, and O. Mangasarian, “Breast cancer diagnosis and prognosis via linear programming,” *Operations Research*, vol. 43, no. 4, pp. 570–577, 1995.
[11] D. Dua and C. Graff, “UCI Machine Learning Repository,” University of California, Irvine, 2019.
[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.
[13] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation,” in *IJCAI*, 1995.
[14] A. Duda, P. Hart, and D. Stork, *Pattern Classification*, Wiley, 2001.
[15] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, 1991.
[16] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” *Journal of the Royal Statistical Society: Series B*, 1996.
[17] L. Breiman, “Random forests,” *Machine Learning*, 45(1), 5–32, 2001.
[18] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, 20, pp. 273–297, 1995.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY