

Performance Analysis of Different Data Mining Algorithms Using Combined Features from Two Optimization Algorithms for Breast Cancer Detection.

Manmohan Sahoo¹, Aswini Kumar Mohanty²

¹Research Scholar, Biju Patnaik University of Technology, Rourkela, Odisha, India

Email: for_manumohan@yahoo.co.in

²Principal, Synergy Institute of Technology, Phulnakhara, Bhubaneswar, India

Email: asw_moh@yahoo.com

Abstract

Globally, breast cancer ranks as the second leading cause of cancer-related mortality among women, with a significant impact on middle-aged populations. Early detection and prevention play a vital role in reducing mortality, and accurate prognosis—along with the ability to predict recurrence risk—is essential for effective disease management.

In this study, we focus on improving breast cancer prediction and detection accuracy by applying a set of classification algorithms after optimizing the dataset through a combined feature selection strategy. This research is based on the Wisconsin Breast Cancer Dataset (WBCD), which was accessed via the UCI Machine Learning Repository and used as the principal dataset. A total of five classification models—Random Forest, Decision Tree, Logistic Regression, k-NN, and AdaBoost—were implemented to analyze 35 variables and measure their predictive capability.

To enhance these models, two complementary feature selection algorithms were applied separately to identify the most informative features. The top-ranked features from both methods were then merged into a unified set, while low-importance attributes were removed. Eliminating less relevant features not only streamlined the data but also reduced noise, ultimately leading to notable improvements in accuracy across all five classification approaches.

Keywords: Breast cancer detection, Data mining algorithms, Feature selection, Wisconsin Breast Cancer Dataset, Random Forest, Decision Tree, Logistic Regression, k-NN, AdaBoost, Optimization algorithms.

How to cite this article: Sahoo M, Mohanty AK. Performance Analysis of Different Data Mining Algorithms Using Combined Features from Two Optimization Algorithms for Breast Cancer Detection. *Int J Drug Deliv Technol.* 2026;16(7s): 764-770; DOI: 10.25258/ijddt.16.7s.82

Introduction

Breast cancer ranks among the leading cancer types affecting women globally, although men can also be diagnosed with the disease. Breast cancer usually begins in the tissue of the ducts or lobules and can manifest through signs like swelling, changes in breast contour, skin puckering, or nipple secretion. If left untreated, malignant tumors can spread to nearby lymph nodes and other organs.

The World Health Organization reported that in 2018, breast cancer represented 25.4% of all new cancer cases among women. By 2025, the worldwide cancer incidence is projected to rise to 19.3 million, with breast cancer continuing to play a major role. Although incidence rates continue to rise, progress in medical imaging, image processing, and machine learning has

helped reduce mortality through earlier detection and improved treatment planning. When detected in its early stages, treatment success rates can be as high as 96%, substantially improving survival and quality of life.

It's crucial to remember that not all breast tumors are cancerous. Benign tumors stay in one place and don't spread. However, differentiating malignant from benign instances is essential for sound clinical decision-making. In this context, biomedical data analytics and machine learning approaches present significant potential for early diagnosis and categorization.

This research introduces a Combined Hybrid Feature Selection approach that merges Correlation-Based Feature Selection (CFS) with Random Forest importance scoring to pinpoint the most relevant

Performance Analysis of Different Data Mining Algorithms Using Combined Features from Two Optimization Algorithms for Breast Cancer Detection

predictors in the Wisconsin Breast Cancer Dataset (WBCD). The integration of these two selection methods results in a cleaner set of features that improves the accuracy and stability of predictive models. In the second phase, this optimized feature set is evaluated using multiple classification algorithms to improve predictive performance and achieve more reliable breast cancer detection.

3. Related Work

Several researchers have investigated different approaches to breast cancer detection and diagnosis, often integrating classification algorithms with feature selection methods to improve predictive accuracy. Iranpour (2007) applied a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel and reported an accuracy of 98.1%, outperforming linear SVM (94%), a fluorescence-based classifier (95.8%), and an edited nearest neighbor method with pure filtering (95%). Veerabhadrapa et al. (2010) extended this work by analyzing dimensionality reduction methods on the WDBC dataset, as well as on the wine and zoo datasets. A hierarchical reduction strategy was used, where Level-1 relied on mutual similarity-based feature selection, and Level-2 integrated PCA and LPP for further dimensionality reduction. Their findings showed that LPP achieved average F-measure scores of 95.148, 91.898, and 89.752 across the three datasets, whereas PCA produced slightly lower scores of 92.950, 85.146, and 87.073. Similarly, Sarvestan Soltani et al. (2010) compared several data mining techniques and found that the decision table classifier provided the highest accuracy of 93.62%, surpassing Neural Networks, Naïve Bayes, SVM, Bayesian Networks, and Logistic Regression across multiple datasets, including SEER. Feature selection (FS) has remained a key focus in medical machine learning, as it eliminates irrelevant attributes and retains the most predictive features. For instance, Fu (2005) used Sequential Forward Selection (SFS) with SVM and a General Neural Regression Network (GRNN), achieving high AUC values of 0.9800 (SVM) and 0.9780 (GRNN). In another study, Osman Hegazy and Omar (2015) designed five bio-inspired optimization algorithms—Algorithms inspired by swarm intelligence, namely ABC and MCS, Bat Algorithm (BA), and Flower Pollination Algorithm (FPA)—to train and optimize Least Squares SVM (LS-SVM) models. Their approach resulted in faster convergence and improved accuracy while avoiding

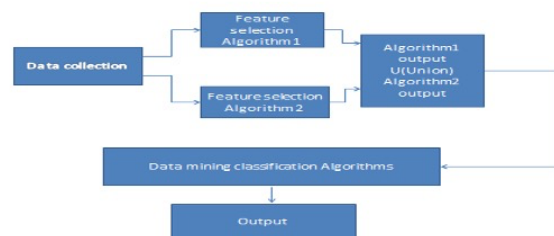
local minima. Other notable contributions include Pawar and Talbar's (2016) genetic fuzzy feature selection method, which achieved an accuracy of 89.47%, and Mohanty et al.'s (2018) Forest Optimization Algorithm, which was tested with multiple classifiers such as SVM, k-NN, and Naïve Bayes to distinguish between normal and abnormal mammographic lesions.

4. Proposed Work

This study employs the WBCD from the UCI Machine Learning Repository, which contains 569 records with 32 attributes, including the diagnosis label (Dua & Graff, 2019). Measurements captured from fine needle aspiration (FNA) breast tissue images form the basis of each dataset instance. Attributes in the dataset reflect various structural traits of cell nuclei, such as size (radius, area, perimeter), surface qualities (texture, smoothness, compactness), and shape characteristics (concavity, symmetry).

The workflow of the planned system, presented in Figure 1, is divided into two primary stages. The first stage applies a Composite Hybrid Feature Selection approach. By integrating Random Forest-based ranking with Correlation-Based Feature Selection (CFS), the method retains only the most relevant and non-redundant features. The second stage utilizes the optimized set of features as input for various classification algorithms, including Random Forest. By comparing the predictive performance of these models on the optimized dataset, the aim is to determine the classifier that yields the most accurate and reliable breast cancer predictions.

This integrated strategy takes advantage of both statistical correlation analysis and ensemble-based ranking methods, enabling the final models to maintain high classification accuracy while reducing computational load. The overall workflow of the proposed approach is illustrated in Figure 1.



4.1 Data Collection

Performance Analysis of Different Data Mining Algorithms Using Combined Features from Two Optimization Algorithms for Breast Cancer Detection

The dataset was generated using the *Multisurface Method-Tree (MSM-T)* algorithm in conjunction with linear programming for decision tree creation. Each data entry contains:

- Unique ID number
- .Diagnosis outcome (M: Malignant, B: Benign)
- .Thirty quantitative attributes computed from digitized images that describe nuclear morphology:

Feature	Description
Radius	Average distance from the center of the mass to its perimeter.
Texture	Variation in grayscale pixel intensities within the image.
Perimeter	The complete outline length of the tumor.
Area	Overall size of the tumor region.
Compactness	Calculated as $(\text{Perimeter}^2 / \text{Area}) - 1.0$, representing shape compactness.
Concavity	Extent to which the contour contains inward-curving sections.
Concave Points	Count of distinct concave segments along the contour.
Symmetry	Measurement of how symmetrical the tumor shape appears.
Fractal Dimension	Approximation of the tumor boundary's complexity

Each property is recorded in three variants:

1. *Mean* – Average value for the feature.
2. *Standard Error* – Measure of variability.
3. *Worst* – Mean of the three largest values observed.

Relation to the Original Dataset

While the Diagnostic dataset contains more refined features than the *Original Breast Cancer Wisconsin dataset*, the core attributes map conceptually:

- **Clump Thickness** → Smoothness, Radius
- **Uniformity of Cell Size/Shape** → Radius, Perimeter, Area
- **Marginal Adhesion** → Compactness, Concavity
- **Single Epithelial Cell Size** → Radius, Area
- **Bare Nuclei, Bland Chromatin, Normal Nucleoli** → Texture, Symmetry, Fractal Dimension
- **Mitoses** → Indirectly represented via multiple morphological features

4.2 Feature Selection

High-dimensional datasets often contain **irrelevant or redundant variables** that reduce model efficiency and accuracy. **Feature Selection (FS)** is therefore a critical preprocessing step in medical AI applications, serving to:

- Reduce computational complexity
- Improve interpretability of the model
- Increase predictive performance

FS strategies can be categorized into:

1. **Optimal Subset Selection** – Choose the subset that maximizes evaluation metrics.
2. **Minimal Subset Meeting Criteria** – Select the smallest feature set meeting performance thresholds.
3. **Trade-off Optimization** – Balance subset size and performance.

4.2.1 Feature Selection Using Correlation-Based Strategy (CFS)

Through CFS, features with strong relevance to the outcome are chosen, while redundancy among them is reduced.

Algorithm Steps:

Algorithm: CFS (Correlation-based Feature Selection)[13]

Input:

$S(F_1, F_2, \dots, F_N, C)$

δ

Output:

S_{best}

Steps:

Begin

Performance Analysis of Different Data Mining Algorithms Using Combined Features from Two Optimization Algorithms for Breast Cancer Detection

```
For i=1 to N do:
  Calculate SU_(i,c) for Fi
  If SU_(i,c)≥δ:
    Append F_i to S_"list" ^
  Order S_"list" ^ in descending SU_(i,c) value.
  Fp← getNextElement(S_"list" ^)
  Do:
    Fq← getNextElement(S_"list" ^,Fp)
    If F_q≠NULL:
      Do:
        Fq^←Fq
        If SU_(p,q)≥SU_(q,c):
          Remove Fq from S_"list" ^
          Fq← getNextElement(S_"list" ^,Fq^)
        Else:
          Fq← getNextElement(S_"list" ^,Fq)
        Until Fq==NULL
        F_p← getNextElement(S_"list" ^,Fp)
        Until F_p==NULL
      S_"best" ←S_"list" ^
End
```

4.2.2 Random Forest (RF) is an ensemble-based method that builds a collection of decision trees in the training phase to improve predictive accuracy and robustness.

It incorporates two key strategies:

- **Bootstrap Sampling (Bagging):** Training subsets are generated by sampling the original dataset with replacement.

Randomized Feature Selection: At every decision node, the algorithm considers a randomly chosen subset of features to determine the most effective split.

RF computes two **importance measures**:

1. **%IncMSE** – The percentage increase in Mean Squared Error when a feature's values are permuted in out-of-bag (OOB) samples.
2. **IncNodePurity** – The cumulative improvement in node purity attributed to a feature across all trees.

5. Methodology

Five supervised machine learning algorithms—**Random Forest, Decision Trees, Logistic Regression, K-Nearest Neighbors (k-NN), and AdaBoostM1**—were implemented for classification. Each method is briefly described below.

5.1 Random Forest

Random Forest is an ensemble learning technique that constructs multiple decision trees using bootstrapped samples of the training data. At each node, a random

subset of features is evaluated, which enhances diversity and reduces correlation among trees.. Final classification is determined through majority voting, improving generalization and mitigating overfitting.

5.2 Decision Trees

Decision Trees partition the dataset step by step into smaller subsets according to the values of the input features. At each node, the best split is selected using measures such as Gini impurity or information gain. This recursive process continues until a termination condition is reached, after which the leaf nodes correspond to the predicted class labels.

5.3 Logistic Regression predicts the probability of a binary class by using the sigmoid (logistic) function, defined as

$$\sigma(z) = 1 / (1 + e^{(-z)}),$$

where z represents the linear combination of input features and their weights.

Parameters are estimated using maximum likelihood, typically via gradient descent. Predictions above a threshold of 0.5 are classified as positive, while those below are classified as negative.

5.4 K-Nearest Neighbors (k-NN)

The k-Nearest Neighbors algorithm classifies an unseen data point by locating the k neighboring examples in the training dataset, using a chosen distance measure—commonly Euclidean distance. The new instance is then assigned the class most frequently represented among these neighbors. As a non-parametric approach, k-NN does not rely on prior assumptions about the distribution of the data.

5.5 AdaBoostM1

AdaBoostM1 is a boosting technique that merges multiple simple classifiers—often decision stumps—into a single, more accurate model. At the start, all training instances are assigned equal weights. After each training round, the weights of incorrectly predicted samples are increased so that subsequent classifiers focus more on these harder-to-classify cases. The final decision is produced through a weighted vote, giving greater influence to models with higher accuracy.

6. Experiment Analysis In this study, two different methods were applied to identify the most significant features from the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which consists of 32 attributes.

Performance Analysis of Different Data Mining Algorithms Using Combined Features from Two Optimization Algorithms for Breast Cancer Detection

The first method, Correlation-Based Feature Selection (CFS), identified 26 features strongly related to the diagnosis label while minimizing redundancy among variables. The second, Random Forest Feature Importance, ranked features by their contribution to model accuracy, selecting 17 high-impact variables. By taking the union of the outputs from both methods, we arrived at a combined set of 28 distinct features.

These selected features were then used as inputs to five classification algorithms:

- Random Forest Classifier
- Decision Tree Classifier
- Logistic Regression
- k-Nearest Neighbors (k-NN)
- AdaBoostM1 Classifier

For model evaluation, the dataset was divided into 66% training data and 34% testing data. All implementations were carried out in Python. The effectiveness of each classifier was evaluated using:

- Training Accuracy – Proportion of correctly classified samples in the training set.

- Cohen’s Kappa – Measures agreement between predicted and actual classifications, adjusted for chance.
- Mean Absolute Error (MAE) – Average absolute difference between predictions and true values.
- Root Mean Squared Error (RMSE) – Square root of the mean squared prediction errors.
- Relative Absolute Error (RAE) – MAE expressed as a percentage relative to a baseline model.
- Root Relative Squared Error (RRSE) – RMSE compared to the RMSE of a baseline model.

These metrics were calculated for both training and testing phases to assess not only accuracy but also the ability of the models to generalize to new, unseen data. A combined evaluation was then performed for an overall comparison, as shown in Table 1.

BREAST CANCER DATA ANALYSIS										
SL_NO	ALGORITHM	NO OF ATTRIBUTES	TRAINING & TESTING DATA	ACCURACY %	CONFUSION MATRIX	Kappa	MAE	RMSE	RAE	RRSE
1	RandomForestClassifier	28	66 & 34	99.99	$\begin{bmatrix} 713 & 1 \\ 0 & 424 \end{bmatrix}$	0.99	0.00	0.03	0.00	0.06
2	DecisionTreeClassifier	28	66 & 34	99.21	$\begin{bmatrix} 709 & 5 \\ 4 & 420 \end{bmatrix}$	0.98	0.01	0.09	0.02	0.18
3	Logistic Regression	28	66 & 34	95.34	$\begin{bmatrix} 694 & 20 \\ 33 & 391 \end{bmatrix}$	0.9	0.05	0.22	0.1	0.45
4	K-Nearest Neighbors	28	66 & 34	95.08	$\begin{bmatrix} 700 & 14 \\ 42 & 382 \end{bmatrix}$	0.89	0.05	0.22	0.11	0.46
5	AdaBoostClassifier	28	66 & 34	99.65	$\begin{bmatrix} 714 & 0 \\ 4 & 420 \end{bmatrix}$	0.99	0	0.06	0.01	0.12

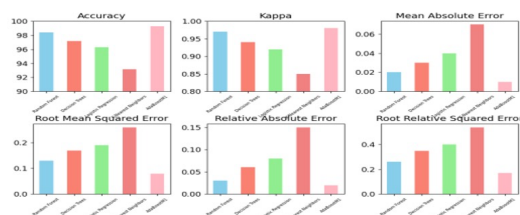


Fig-2

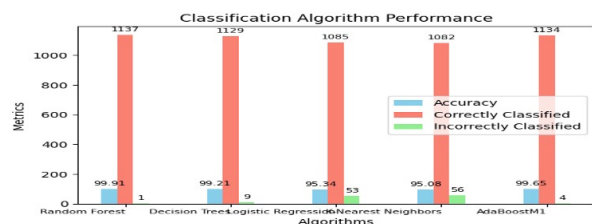


Fig-3

Performance Analysis of Different Data Mining Algorithms Using Combined Features from Two Optimization Algorithms for Breast Cancer Detection

Performance Comparison

Table 1 The results indicate that the RF Classifier delivered the best overall performance, achieving an outstanding accuracy of 99.99% with very low error rates. Its effectiveness can be explained by its capability to manage high-dimensional data, mitigate overfitting through bootstrapped sampling, and ensure stability in classification tasks.

AdaBoostM1 ranked second, with only marginally lower accuracy, followed by the **Decision Tree Classifier**. **Logistic Regression** and **K-Nearest Neighbors** achieved slightly lower accuracies, ranking fourth and fifth, respectively.

Figure 2 (not shown here) illustrates the comparative performance of the algorithms, highlighting Random Forest's clear advantage. Figure 3 further visualizes accuracy rates alongside counts of correctly and incorrectly classified instances for each classifier.

7. Conclusion

In this study, two feature selection techniques—Correlation-Based Feature Selection (CFS) and Random Forest importance ranking—were applied to the WDBC dataset to identify the most relevant attributes for classification. The union of the features selected by both methods was used as input for five classification algorithms: Random Forest, Decision Tree, Logistic Regression, K-Nearest Neighbors, and AdaBoostM1.

The target variable was dual, where “**1**” denoted **M** cases and “**0**” denoted **B** cases. By applying both feature selection methods prior to model training, we reduced potential bias and improved the interpretability of the models.

The experimental results demonstrated that the **Random Forest Classifier** achieved the highest accuracy (**99.99%**), outperforming all other algorithms. The **AdaBoostM1 Classifier** ranked second with **98.97% accuracy**, followed closely by the **Decision Tree Classifier** with **99.21% accuracy**. Logistic Regression and K-Nearest Neighbors produced slightly lower accuracies but still performed competitively.

The results demonstrate that ensemble approaches—especially Random Forest—are highly reliable for breast cancer classification, as they can model complex feature interactions, maintain robustness, and effectively prevent over fitting.

8. REFERENCE:

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A, “Global Cancer Statistics 2018,” GLOBOCAN estimates of incidence and mortality

worldwide for 36 cancers in 185 countries. CA Cancer J Clin, in press.

[2] Dina A. Ragab, Maha Sharks and Omneya Attallah, “Breast Cancer Diagnosis Using an Efficient CAD System Based on Multiple Classifiers,” *Diagnostics* 2019, 9, 165; DOI:10.3390/diagnostics9040165.

[3] Kalló, Gergő, Miklós Emri, et al. “Changes in the Chemical Barrier Composition of Tears in Alzheimer’s Disease Reveal Potential Tear Diagnostic Biomarkers.” *PLoS One*, vol. 11, no. 6, Public Library of Science, June 2016, p. e0158000.

[4] Veerabhadrapa, Lalitha Rangarajan, “Bi-level dimensionality reduction methods using feature Selection and feature extraction,” *International Journal of Computer Applications (0975 – 8887)* vol. 4, No.2 July 2010.

[5] Sarvestan Soltani A, Safavi A A, Parandeh M N, and Salehi M, “Predicting Breast Cancer Survivability using Data Mining Techniques,” *Software Technology and Engineering (ICSTE), 2nd International Conference*, Vol. 2, pp. 227-231, 2010.

[6] Fu, J.C.; Lee, S.K.; Wong, Wang, A.H.; Wu, H.K.; introduce Image segmentation feature selection and pattern classification for microcalcifications. *Comput. Med. Imaging Graph.* 2005, 29, 419–429.

[7] Osman Hegazy, Omar S. Soliman, and Mustafa Abdul Salam, “Comparative Study between FPA, BA, MCS, ABC, and PSO Algorithms in Training and Optimizing of LS-SVM for Stock Market Prediction,” *International Journal of Advanced Computer Research* vol.5, Issue. 18, pp.35-45, March-2015.

[8] Pawar, M.M.; Talbar, S.N. Genetic Fuzzy System (GFS) based wavelet co-occurrence feature selection in mammogram classification for breast cancer diagnosis. *Perspect. Sci.* 2016, 8, 247– 250

[9] Mohanty, F.; Rup, S.; Dash, B.; Majhi, B.; Swamy, M.N.S. Mammogram classification using contourlet features with forest optimization-based feature selection approach. *Multimed. Tools Appl.* 2018, 1–30.

[10] Dua, D. and Graff, C. (2019). [<http://archive.ics.uci.edu/ml>]. UIC Machine Learning Repository.

[11] A. Kalousis, J. Prados, M. Hilario, “Stability of Feature Selection Algorithms: a study on high dimensional spaces,” *Knowledge and information System*, vol. 12, no. 1, pp. 95-116, 2007. Article (CrossRef Link)

[12] Luis Carlos Molina, Lluís Belanche, Àngela Nebot, “Feature Selection Algorithms: A Survey and

Performance Analysis of Different Data Mining Algorithms Using Combined Features from Two Optimization Algorithms for Breast Cancer Detection

Experimental Evaluation,” in Proc.of ICDM, pp. 306-313, 2002.

[13] L Yu, H Liu - Proceedings of the 20th international conference on ..., 2003 - cdn.aaai.org

[14] V. N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[15] K. B. Jyothi, K. H. Bindu and D. Suryanarayana, A comparative study of random forest & knearest neighbors on the HAR dataset using Caret, International Journal of Innovative Research in Technology, vol.3, no.9, pp.6-9, 2017.

[16] U. Grömping, Variable importance assessment in regression: Linear regression versus random forest, Amer. Statistician, vol.63, no.4, pp.308-319, 2009.

[17] H. Ishwaran, Variable importance in binary regression trees and forests, Electr. J. Stats, no.1, pp.519- 537, 2007.

[18] C. Strobl, A. Boulesteix, T. Kneib, T. Augustin and A. Zeileis, Conditional variable importance for random forests, BMC Bioinf., vol.9, p.307, 2008.