Journal of Harbin Engineering University
ISSN: 1006-7043
Vol 45 No. 11
November 2024

Effect of Subset on Classification Accuracy of Breast Cancer Detection

[1]Manmohansahoo, [2]Amalendu Bag, [3]Swasthee Priya Mallick, [4]Aswini kumarMohanty
[1]Research scholar,BPUT,Odisha,
for_manumohan@yahoomail.co.in
[2]Research Scholar ,BPUT
amalendu.bag@gmail.com
[3]Asst. Prof. , KIIT Polytechnic
mallick.swasthee@gmail.com
[4]SIT, Bhubaneswar,Odisha
asw_moh@yahoo.com

Abstract
Data mining is the process of discovering valid, novel, useful, and understandable patterns in data. It involves extracting information from large databases and plays a crucial role in various fields, including business, education, government, health care and engineering .In health care ,data mining is particular useful for disease predictions. Techniques such as classification, clustering, association rules, summarization, and regression are commonly used.
Breast cancer is a serious illness that affects many women worldwide. Early detection significantly increases the chances of successful treatment; with success rates reaching up to 80%.Analyzing existing data for early detection is therefore essential. In our study, we used data from cancer patients provided by the Wisconsin dataset from the UCI learning Repository, which includes 35 different features.
We applied the Ant Colony Optimization (ACO) feature selection algorithm to reduce the number of features. The selected features were then used as input for various classification algorithms. We compared the accuracies of these algorithms before and after applying ACO to assess the improvement in performance .Our results showed that ACO significantly enhanced classification accuracy, with the Random Forest algorithm achieving the highest accuracy of 99.02%.

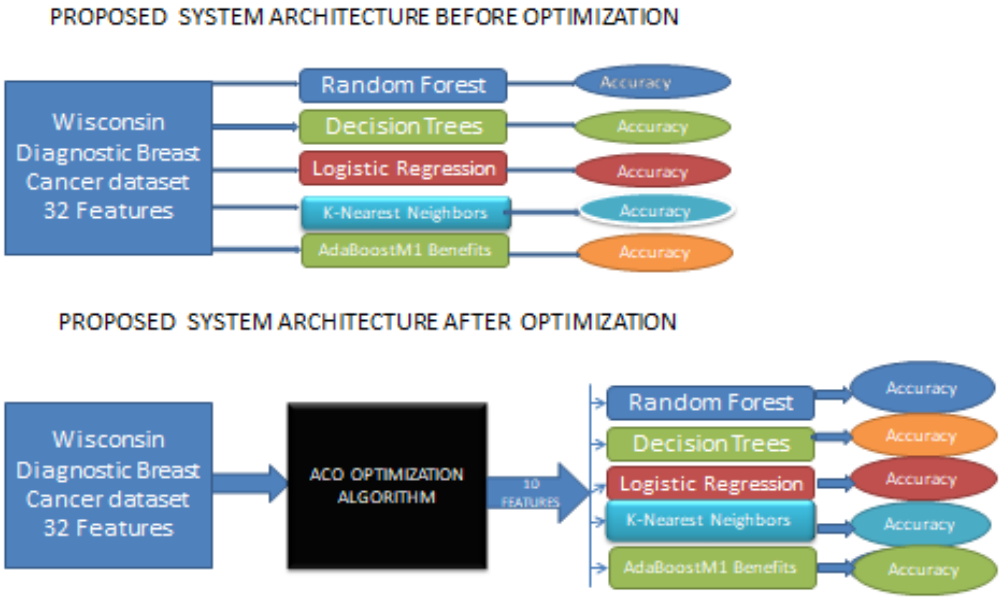
INTRODUCTION

According to the world health Organization's Globocan 2012 report[1], the breast cancer is the most common cancer among women globally. Indian women are particularly affected by this disease report highlights. which making it a leading cause of death in this population. The survival chances of patients increases significantly through Early Detection. Allowing for timely preventive measures, various biological techniques can be utilized for early breast cancer detection.In this paper, we employ different data mining algorithms to predict recurrent cases of breast cancer using the Wisconsin Prognostic Breast Cancer (WPBC) dataset from the UCI machine learning repository[2]. We evaluate the performance of five classification algorithms: Random_forest_tree, Decision Trees
Logistic Regression, K-means and AdaBoostM1 Benefits. The implementation is carried out using Python, which facilitates comprehensive data analysis [3]. The goal is to identify the most effective data mining algorithm for predicting recurrent breast cancer cases and to determine the key attributes that significantly influence recurrence prediction.
This paper is structured into two main parts. Firstly, by using Wisconsin dataset we analyze the accuracy of several data mining classification algorithms. Second, we apply the feature selection algorithm, Ant Colony Optimization (ACO) to identify important features from the dataset. To analyze and compare the accuracy before and after feature selection, these selected features are used with the data mining classification algorithms. This complete analysis aims to improve prediction accuracy and identify critical attributes for breast cancer recurrence.

PURPOSE OF THIS PAPER

In today's world, many women globally are seriously affected by breast cancer.Due to Early

detection, the chances of successful treatment significantly increases, with 80% success rate. Therefore, analyzing existing data is crucial for early detection. In our study, we used the Wisconsin dataset from the UCI Machine Learning Repository [10], which contains 32 different attributes. To reduce the number of features, we applied the Ant Colony Optimization (ACO) feature selection algorithm. Subsequently, we tested various data mining classification algorithms and compared their accuracy before and after feature reduction to determine the most effective approach.



CLASIFICATION ALGORITHMS

- RANDOM_FOREST_TREE
- DECISION TREES:
- LOGISTIC REGRESSION:
- K-NEAREST NEIGHBORS (K-NN):
- ADABOOSTM1 BENEFITS:

RANDOM FOREST TREE:

Random Forest is a supervised model that implements both decision trees and bagging method. The idea is that the training dataset is resampled according to a procedure called “bootstrap”. Each sample contains a random subset of the original columns and is used to fit a decision tree. The number of models and the number of columns are hyperparameters to be optimized. Two hyperparameters that need to be optimized are the quantity of models and columns. Lastly, the combined predictions from the trees are used to determine the mean value (for regression) or, for classification, by soft voting. The concept of bagging is based on the bias-variance tradeoff, which states that by averaging the outputs of individual decision trees, the standard error and model variance both decrease. For this reason, Random Forest has gained a lot of notoriety in recent years.[4].

DECISION TREES:

Decision Trees constitute a crucial and comprehensible machine learning algorithm applicable to both classification and regression tasks. This hierarchical model organizes decisions through a sequence of questions and outcomes, where each internal node signifies a test on a particular feature, each branch represents the potential result of the test, and each leaf node signifies the ultimate decision or prediction.[3]. This algorithm constructs the tree in an iteratively selecting feature that provides the most information at each node based on criteria like Gini impurity or information gain . The process

continues until a specified stopping condition is met, resulting in a tree that can be navigated to generate predictions for new data points. Decision Trees[5] offer transparency and ease of interpretation, making them valuable in scenarios where understanding the reasoning behind predictions is crucial. They excel in capturing complex decision boundaries and interactions within the data. However, there is a risk of over fitting, especially when the tree becomes too deep and specific to the training data. Techniques like pruning are employed to address this issue, preventing the tree from becoming overly complex and enhancing its generalization to new data. Decision Trees are employed across various domains including finance, healthcare, and marketing, owing to their capability to process both numerical and categorical data. While standalone Decision Trees are powerful, they also serve as the building blocks for ensemble methods like Random Forests, combining multiple trees to improve overall predictive performance.

LOGISTIC REGRESSION

Logistic Regression is an essential algorithm in data mining, frequently utilized for both binary and multi-class classification problems. Contrary to what its name might suggest, it is designed for classification rather than regression tasks. The algorithm estimates the probability that a given instance falls into a specific class by using the logistic function [6] to ensure that the output is restricted within the range of 0 to 1.The algorithm estimates coefficients for input features, and the decision boundary is determined by a threshold. Logistic Regression is computationally efficient, interpretable, and suitable for linearly separable data. It finds applications in diverse fields, including healthcare and finance, where predicting binary outcomes or multiple classes is essential.

FEATURE SELECTION ALGORITHM:

Ant Colony Optimization:

Ant Colony Optimization (ACO) feature selection algorithms offer a powerful and flexible approach to identifying relevant features, thereby enhancing the performance of machine learning models through improved accuracy, reduced over fitting, and decreased computational complexity. Inspired by the foraging behavior of ants, the ACO algorithm mimics the way ants find the shortest path to food sources. In the context of feature selection, a colony of artificial ants explores the feature space, depositing pheromones on features that contribute to better model performance. The pheromone intensity influences the likelihood of a feature being selected. Through iterative refinements based on pheromone concentrations, the algorithm identifies an optimal subset of features. This approach is particularly effective for high-dimensional datasets, optimizing feature sets

K-NEAREST NEIGHBORS (K-NN):

K-Nearest Neighbors (K-NN) is a classification algorithm in data mining that assigns a class to a data point based on the majority class of its k closest neighbors within the feature space. The algorithm assesses proximity using distance metrics such as Euclidean or Manhattan distance.

K-NN[7] is characterized as non-parametric and instance-based, which allows it to adapt to various data distributions. Its simplicity and effectiveness make it particularly useful in scenarios where local patterns are significant. Nevertheless, the algorithm can be sensitive to outliers and necessitates a meticulous selection of the parameter k.

Due to its intuitive concept and adaptability, K-NN is utilized in a wide range of applications, including recommendation systems, pattern recognition, and anomaly detection.

ADABOOSTM1 BENEFITS:

AdaBoost (Adaptive Boosting) is an influential ensemble classification algorithm that amalgamates multiple weak learners to form a robust classifier. Its benefits include robustness against over fitting, as it focuses on misclassified instances, continuously improving model performance. AdaBoost[8]can handle various types of data and is not limited to specific types of classifiers, making it versatile. Additionally, it automatically selects relevant features, reducing the need for extensive feature engineering. Its simplicity in implementation and ability to work well with both classification and regression tasks make it widely used in real-world applications, offering high accuracy and interpretability.

to enhance machine learning model accuracy by leveraging swarm intelligence principles. The success of ACO underscores the importance of swarm-based optimization techniques for improving other optimization algorithms, showcasing the potential for bio-inspired methods to address complex computational problems.[9].

ACCURACY , Correctly classified , Incorrectly classified ANALYSIS BEFORE FEATURES SELECTION									
SL_NO	ALGORITHM	NO OF FEATURES	ACCURACY	Sensitivity	Precision	F-Measure	Correctly classified	Incorrectly classified	CONFUSION_MATRIX
1	Random forest tree	32	95.61%	0.95	0.97	0.96	196	9	[[129 3] [6 67]]
2	Decision Trees	32	92.20%	0.95	0.93	0.94	189	16	[[123 9] [7 66]]
3	Logistic Regression	32	96.10%	0.96	0.98	0.97	197	8	[[129 3] [5 68]]
4	K-Nearest Neighbors (k-NN):	32	95.61%	0.97	0.96	0.96	196	9	[[127 5] [4 69]]
5	AdaBoostM1 Benefits:	32	95.61%	0.97	0.96	0.96	196	9	[[127 5] [4 69]]

Table-1

ACCURACY , CORRECTLY CLASSIFIED , INCORRECTLY CLASSIFIED ANALYSIS IN GRAPH

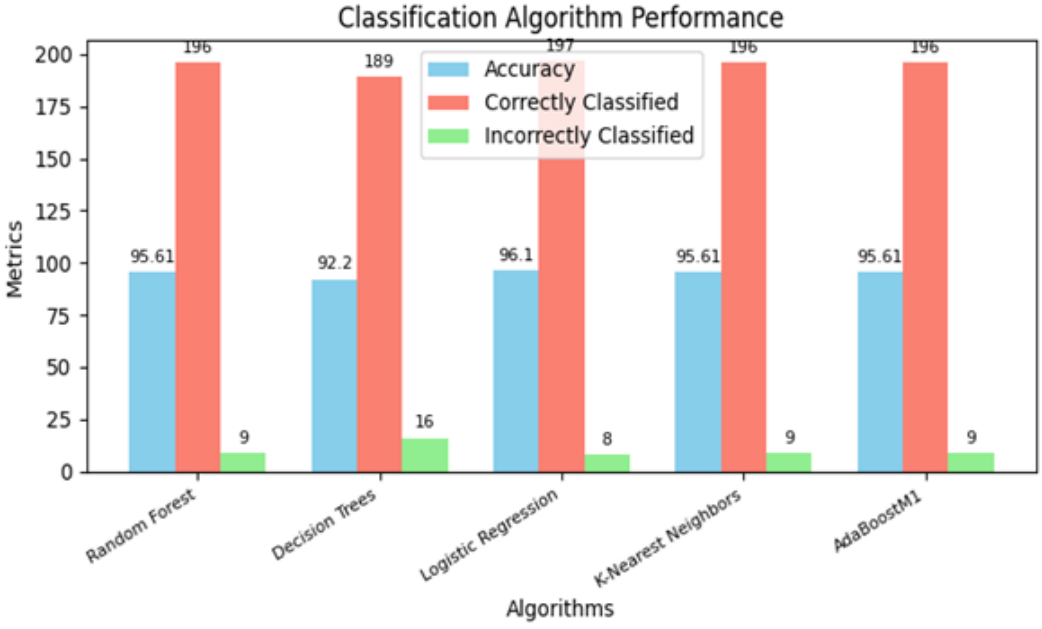


Figure-1

ACCURACY ANALYSIS AFTER FEATURES SELECTION(Ant Colony Optimization (ACO) is a metaheuristic optimization algorithm inspired by the foraging behavior of ants)									
SL_NO	**ALGORITHM**	**NO OF FEATURES**	**ACCURACY**	Sensitivity	Precision	F-Measure	Correctly classified	Incorrectly classified	CONFUSION_MATRIX
1	Random_forest_tree	**10**	**99.02%**	1.0	0.98	0.99	203	2	[[130 2] [0 73]]
2	Decision Trees	**10**	95.12%	0.97	0.94	0.96	195	10	[[125 7] [3 70]]
3	Logistic Regression	**10**	96.59%	0.96	0.98	0.97	198	7	[[130 2] [5 68]]
4	K-Nearest Neighbors (k-NN):	**10**	96.10%	0.96	0.97	0.97	197	8	[[129 3] [5 68]]
5	AdaBoostM1 Benefits:	**10**	96.10%	0.98	0.95	0.97	197	8	[[126 6] [2 71]]

Table-2

ACCURACY , CORRECTLY CLASSIFIED , INCORRECTLY CLASSIFIED ANALYSIS AFTER FEATURES SELECTION IN GRAPH

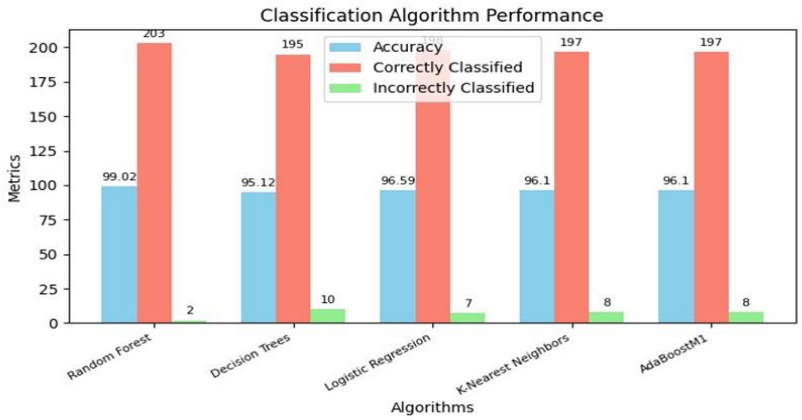


Figure -2

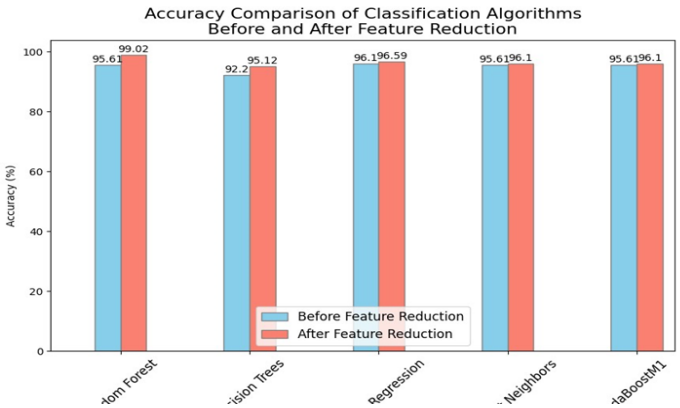


Figure -3

EXPERIMENT AND RESULT:

Accuracy

Accuracy is the ratio of correctly classified instances to the total number of instances in the dataset.

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

True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN)

- **TP Rate (Sensitivity)**: Measures the ability to correctly identify true positive instances.

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

Precision

Precision is the ratio of correctly classified fault-prone modules to the total number of modules classified as fault-prone. It indicates the proportion of units accurately predicted as faulty.

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

F-Measure

The F-Measure combines precision and sensitivity into a single metric, providing a balance between them.

$$\text{F} - \text{Measure} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

DATA SET-The dataset comes from the University of California-Irvine's machine learning repository and is called the WBDC (Wisconsin Diagnostic Breast Cancer) dataset [10].

It includes information on 569 cases. Among them, 357 cases are identified as benign (not harmful)and 212 cases are identifiedas malignant (potentially harmful). There are 32 different pieces of information, or attributes, for each case.

DATA SELECTION- Data selection has made research in detection and pattern recognition better. Its goal is to turn complex data into simpler forms by picking out the most important parts.

This is done using methods like feature selection and extraction. Feature selection focuses on choosing the important bits and leaving out the less important ones. This paper uses a method called ACO (Ant Colony Optimization) for feature selection.

EXPERIMENT SETUP

To start, we're using a method called Ant Colony Optimization (ACO) to make the data less complex. The goal is to pick out the important parts from the dataset. In this paper, we'll use various classification methods like RF, DT, LR, K-NN, and AdaBoostM1. We allocated 64% of our dataset for training purposes, while the remaining 36% was designated for testing.

Then, we'll compare how accurate they are when using the simplified data.

EXPERIMENTAL RESULTS

We started by testing five different classification algorithms without doing any feature selection. Logistic Regression turned out to be the best, with 96.10% accuracy. The others—Random Forest, Decision Trees, K-Nearest Neighbors, and AdaBoostM1—had accuracies of 95.61%, 92.20%, 95.61%, and 95.61% respectively.

We thought we could improve these results by choosing the most important features and getting rid of the less useful ones. So, we selected the 'N best attributes' for each classifier, where N could range from 1 to 32, since we had 32 features in our dataset.

After some testing, we found that Random Forest achieved the highest accuracy of 99.02% when we usedthe top 10 features. We then removed the other 22 less important features and ran Random Forest again, which increased its accuracy by 3.41%. For Decision Trees, Logistic Regression, K-Nearest Neighbors, and AdaBoostM1, the accuracies improved by 2.92%, 0.49%, 0.49%, and 0.49% respectively compared to before we did any feature selection[10].

The table summarizes these results without using the ACO feature selection algorithm. It's clear that Random Forest performed the best for breast cancer accuracy, andLogistic Regression was also strong without feature selection. Table 2 shows that every evaluation criteria improved when we used appropriate feature selection. This suggests that a dataset with many features can benefit from selecting only the most important ones for classification. Also fig-1 and fig-2 showing the comparison graph of accuracy ,correctly classified and incorrectly classified among data mining classification algorithms between before and after feature reduction .Also Fig-3 only show the accuracy comparison before and after feature selection.

CONCLUSION AND FUTURE SCOPE

In our study, we emphasized the significance of feature selection in predicting breast cancer outcomes. By choosing the right attributes, any classification method can see a big boost in performance. Features that don't contribute much can lead to inaccurate predictions. We found that Random Forest performed well both before and after feature selection, and other algorithms also improved with this technique. Moving forward, we plan to explore newer algorithms with improved feature selection methods.

REFERENCE

[1] "J. Ferlay, Globocan 2012 v1.0 Cancer Incidence and Mortality Worldwide: Iarc Cancerbase no. 11, 2014, [online]Available: http://globocan.iarc.fr.".

[2] "A. Frank and A. Asuncion, 'UCI machine learning repository,' 2010. [online] Available: http://archive.ics.uci.edu/ml".

[3] "R Core Team2013, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, [online] Available: http://www.R-project.org".

[4] "Random forest for breast cancer prediction".

[5] "Performance analysis of decision tree algorithms for breast cancer classification".

[6] "Predicting breast cancer using logistic regression and multi-class classifiers".

[7] "Breast cancer classification using k-nearest neighbors algorithm".

[8] X. Shu and P. Wang, "An Improved Adaboost Algorithm Based on Uncertain Functions," in *2015 International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration*, Wuhan: IEEE, Dec. 2015, pp. 136–139. doi: 10.1109/ICIICII.2015.117.

[9] M. A. H. Dalfi, S. Chaabouni, and A. Fakhfakh, "Breast Cancer Detection Using Random Forest Supported by Feature Selection," *Int. J. Intell. Syst. Appl. Eng.*.

[10] H. Peng, C. Ying, S. Tan, B. Hu, and Z. Sun, "An Improved Feature Selection Algorithm Based on Ant Colony Optimization," *IEEE Access*, vol. 6, pp. 69203–69209, 2018, doi: 10.1109/ACCESS.2018.2879583.

[11] P. A. Wingo, T. Tong and S. Bolden, "Cancer statistics," CA. J. Clin.,vol. 45, pp. 8–30, 1995.

[12] W. K. Pratt, Digital Image Processing: 3rd Edition, Wiley-Interscience; 2001.

[13] R. C. Gonzalez and Richard E. Woods Digital Image Processing, Prentice Hall; 2nd Edition, 2002.

[14]J.N. Kapur, P.K. Sahoo and A.K.C. Wong, "A new method of gray level picture thresholding using entropy of the histogram", Computer Vision & Graphics Image processing, Vol. 29, 1985, pp. 273-285.

[15] Li E.H., Lee C.K. "Minimum cross-entropy thresholding", Pattern Recognition, pp. 617-625,1992.

[16] A.S. Abutaleb, "Automatic thresholding of gray level pictures using two dimensional entropy", Computer Vision and Graphics Image Processing, Vol. 47, 1989, pp. 22-32.

[16] M. Hanmandlu, V.K. Madasu and S. Vasikarla, "A fuzzy Approach to texture segmentation", Proc. International Conference on Information Technology: Coding and Computing, vol-1, pp.636-642, 2004.

[17] V. Chalana and Y. Kim, "A Methodology for Evaluation of Boundary algorithms on Medical Images," IEEE Trans. on Medical Imaging, vol. 16(5), pp. 642-52, 1997.

[18] R Gupta et al., "The use of texture analysis to delineate suspicious masses in mammography", Phys. Med.Biol., vol 40 835- 855, 1995.

[19] N.R. Pal and S.K. Pal, "Entropic Thresholding", Signal Processing, vol.16, pp.97-108,1989.

[20] R. Gupta and P.E. Undrill, "The use of texture analysis to identify suspicious masses inmammography", Phys. Med. Bio., vol 15. 835-855, 1997.

[21] J. S. Weska et al., "A survey of threshold selection techniques", Computer Graphics &Image Processing,vol. 7. pp. 259-265, 1978.

[22] W. A. Perkins, "Area segmentation of images using edge points." IEEE Trans. on Pattern Analysis & Machine Intelligence, vol. PAMI-2, pp. 8-15. 1980.

[23] S. Zucker. "Region growing: Childhood and adolescence", Computer Graphics and ImageProcessing,vol. 5. pp. 382-399. 1976.