Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Feature Selection using Extra Trees for Breast Cancer Prediction

Shahad Abdelmuniem Mohamed Awadelkarim
muniemshahd@gmail.com
Department of Computer Science, National Ribat University, Khartoum, Sudan

Article Information

Submitted : 28 Mar 2024
Reviewed: 31 Mar 2024
Accepted : 20 Apr 2024

Keywords

Breast Cancer, Machine Learning, Feature Selection, Extra Trees, SVM.

Abstract

Breast cancer is a disease that seriously threatens women's health. It is considering a common death cause in women. Machine learning has made significant progress in recent years to improve the effectiveness of early diagnosis of various diseases. Accurate predication and detection help decrease the death rate of breast cancer. This paper aims to predict breast cancer using several machine-learning techniques. The idea is to lower the number of features in the Wisconsin Breast Cancer Dataset (WCDB) and use it for prediction. The study used the extra trees method for feature selection and Random forest, Logistic regression, and Support Vector Machine (SVM) for testing the dataset. According to the results, SVM achieved the best performance among the other models with 98% accuracy. The proposed method in this study proved its effectiveness in breast cancer prediction.

A. Introduction

Breast cancer is the most commonly occurring cancer in women and the most common cancer overall, with more than 2.26 million new cases of breast cancer in women since 2020 [1]. Breast cancer occurs in breast cells, the fleshy tissue or the stringy connective tissue within the breast. Breast cancer is a dangerous tumor that grows rapidly and eventually causes death in critical cases [2]. Although this type of cancer is more common in females, it can rarely occur in males [2]. Factors such as age and a family history of breast cancer can increase the risk of breast cancer [2, 3]. There are two types of tumors Benign and Malignant [3] Benign is not critical for a human body and rarely lead to human death. This type of tumor grows in one part (spot) of the body and has limited growth. Malignant is seriously dangerous and can lead to death; this type is called breast cancer. The malignant tumor appears when cells in the breast tissue grow abnormally. The treatment options for breast cancer are based on the patient age, and the cancer stage and type. Treatment can be one therapy or a collection of several therapies such as Chemotherapy, Radiotherapy, Surgery, and others [3, 4].

Data mining and Machine learning are automatic methods used to teach models how to handle the data more efficiently. They are used to uncover correlations between factors and are used widely in disease prediction. Therefore, these types of research helped with decision-making [5]. So, the study aims to predict breast cancer using several machine-learning algorithms with the least number of features possible.

This paper is structured as follows: Section II reviews recent research on the detection and prognosis of breast cancer. Section III explains the algorithms and tools of data mining and machine learning used for breast cancer prediction. Section IV discusses the results, and section V concludes the research.

B. Literature Review

In this research, researchers used Computer-Aided Diagnosis or Detection (CAD) systems to predict breast cancer. Researchers used several machine learning algorithms to train CAD systems. The algorithms included the Random Forest algorithm, K Nearest Neighbor algorithm, Support Vector Machine, and Gradient Boosting. According to the trained dataset, the most accurate algorithm among the previous algorithms was the Random Forest algorithm, which used both classifying and regression methods. It gave the highest accuracy rate of more than 70% [6]. This research used the genetic programming (GP) technique to select the best features and perfect parameter values for the machine learning classifiers. GP aimed to resolve the hyper parameters problems, which are those parameters that cannot estimate from the data. Breast cancer detection was the study of the research. The present technique proceeded in different experiments using the breast cancer dataset. By combining feature preprocessing methods and classifier algorithms, GP found the best model with the highest accuracy rate among eleven algorithms: K neighbors, Decision tree, Random forest AdaBoost classifier, Gradient boosting, Gaussian NB, Linear discriminant analysis, Quadratic discriminant analysis, Logistic regression, and Extra trees. The most accurate algorithm was the Extra Tree classifier 97.34% accurate [7].

Researchers in this study attempted to develop a Hierarchical Clustering Random Forest (HCRF) model. This model measures the similarities among decision trees and clusters them hierarchically. It constructs the hierarchical clustering random forest with low similarity and high accuracy. The Variable Importance Measure (VIM) method has optimized the selected feature number for breast cancer prediction. Researchers tested the HCRF algorithm on two different datasets, and results showed the highest accuracy rate with 97.05% and 97.76% compared to Decision Tree, Adaboost, and Random Forest algorithms [8]. A novel model has been used in this study to predict breast cancer. This model combined K-means and the Gaussian mixture model (GMM). It was a hybrid combination of segmentation and detection models. Several kinds of breast images, including Normal, Benign, and Malignant, were segmented and classified using this technique. The proposed model approved its effectiveness among several algorithms. The hybrid model had the highest accuracy rate of 95.5% [9].

Furthermore, this research also used a novel model to classify breast cancer patients based on their subtypes and survival rates. Researchers created a multiplatform network called the Multimodal Auto encoders (MAE) classifier. DNA methylation, gene expression (GE), and miRNA expression were the features to classify breast cancer. Testing results showed that the proposed model scored the highest rates on predicting breast cancer subtypes and survival rates among several models, but the top three models after MAE were support vector machine (SVM), Gradient Boosting Trees (GBT), and Random Forest (RF). The accuracy results of the model were 91% and 86% for subtypes prediction and survival prediction, respectively [10]. This study [11] presented five machine-learning techniques for predicting breast cancer. The algorithms were support vector machine (SVM), K-nearest neighbors, random forests, artificial neural networks (ANN), and logistic regression. The highest accuracy rate was by ANN of 98.57%. The Random Forest and logistic regression were the second-best models by 95.71%.

On the other hand, the authors of this study [12] used several methods, such as K-means and Spectral Clustering (SC) algorithms, to cluster two different datasets. Then, they used Support Vector Machines (SVM), Decision trees, and Random tree algorithms for prediction. SVM scored the highest accuracy rate of 96.5% on the WCDB dataset and 78.7% on the WPBC. In this research [13], the authors applied a class weight function to balance the dataset. After that, they developed a model using the logistic regression algorithm for prediction. The results showed that the model scored a high accuracy rate of 98.2%.

According to the review, the previous works used the Wisconsin breast cancer dataset with all attributes. This research proposes a method to reduce the features used for prediction in the WCDB dataset.

C. Research Method

This section discusses the processes of selecting features from the dataset and using several models for testing it. Figure 1 displays the flowchart outlining the steps in this research.

Data preprocessing

This research used the Wisconsin Diagnosis of Breast Cancer (WDBC) dataset from the machine learning repository of UCI [14]. The WDBC database contains 569 instances. Each instance consists of 30 real-value attributes and a class label. The dataset features were from a digitized image of an FNA of a breast mass, which describes the traits of the cell nuclei [15]. Table 1 shows the dataset description.

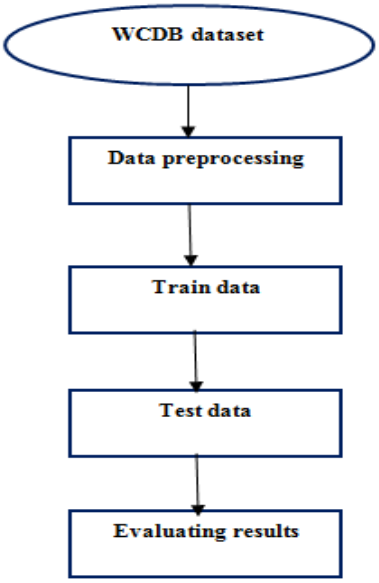

Figure1. Flow diagram for the breast cancer prediction

Table1. WDBC Dataset Description

Attribute	Description
Radius	Mean, standard Error, worst area
Texture	Mean, standard Error, worst area
Area	Mean, standard Error, worst area
Perimeter	Mean, standard Error, worst area
Smoothness	Mean, standard Error, worst area
Compactness	Mean, standard Error, worst area
Concavity	Mean, standard Error, worst area
Symmetry	Mean, standard Error, worst area
Fractal dimension	Mean, standard Error, worst area
Concave points	Mean, standard Error, worst area

Selecting the most featured attributes is critical in breast cancer prediction because it provides clinical information that can help decision-making. Therefore, this research used the Extra Trees method to calculate the importance of all attributes and rank them according to their weight of importance [16]. The Extra trees generate multiple individual decision trees from the whole training dataset. It selects a random split to divide the parent node into two random child nodes. This process repeats in each child node until reaching the leaf node [16, 17]. The predictions of all the trees are combined to set the final prediction through a

majority vote [18]. For feature selection, for each feature, the Gini importance is computed. Gini Index shows the probability of category inconsistency of two samples randomly selected from the subset after node split [19]. The smaller Gini index, the higher the purity of the subset is. The mathematical formula of the Gini index is:

$$G_m = \sum_{C=1}^{C} P_{mc}(1 - P_{mc}) \quad (1)$$

C represents the number of categories on the training set, and P represents the probability of a classification c at node m.
The feature importance of the N feature at node m is calculated by:

$$I_{jm=G_m - w_L\ G_L - w_R\ G_R} \quad (2)$$

G_L and G_R are the "Gini Index'' of the left and right nodes after node m split, respectively. w_L and w_R represent the number of weighted samples reaching the left and right nodes after node m split, respectively.
Each feature displays in descending order according to the Gini importance of each attribute. Finally, the user selects the top k features according to his; or her choice as input for the classification model.
Three models will test the modified dataset for prediction: Random Forest, Logistic Regression, and Support Vector Machine (SVM).

Random Forests Classifier

Random forests are groups of classification and regression trees, which are simple models using binary splits on predictor variables to locate outcome predictions. Many classification and regression trees are formed in the random forest setting using randomly chosen training datasets and random selections of predictor variables for modeling outcomes. The scores from each tree are combined to create a prediction. The random forest has a significant advantage in prediction modeling; it can handle datasets with multiple predictor variables [20].

Logistic Regression

This method predicts a categorical dependent variable's output. Therefore, the result must be a categorist or discrete value. True or false, 0 or 1, or Yes or No are all possible outcomes. It gives the probabilistic values which lie between 0 and 1. Except for how they are applied, it is similar to linear regression. Linear regression is utilized for solving Regression problems, while Logistic regression is for solving classification problems [21].

Support Vector Machine (SVM)

It is a supervised learning technique employed in problems involving classification and regression. It consists of theoretical and numeric functions to solve most regression problems. It is a powerful machine-learning technique based on 3D and 2D models [16, 20]. The classification formula of SVM is:

$$max_{\alpha} \left[\sum_{i=1}^{n} - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i\, \alpha_j y_i y_j K(x_i x_j) \right] \quad (3)$$

With restrictions:

$$\sum_{i=1}^{n} \alpha_i y_i = 0{,}0 \ \leq \ \alpha_i \ \leq C, i = 1{,}2, \ldots, n \quad (4)$$

Where α presents the parameter vector for the classifier hyperplane, C is the penalty parameter that controls the number of misclassifications, xi is the real-valued *n*-dimensional input vector, and $\boldsymbol{yi}$ presents the class label associated with the training vector [22].

Evaluation measures

The confusion matrix evaluates the performance of the classifiers. It divides the samples into two categories: Positive and Negative, according to the model prediction and the fact. True positive (TP) and true negative (TN) represent data that are correctly classified, whereas false positive (FP) and false negative (FN) represent data that are incorrect in classification. These measures can be collected to analyze the accuracy, precision, recall, and F1 score, which will be the metrics of this paper [23]. Table 2 shows the model performance measures.

Table2. Models Performance Measures

Measure	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + TN}$
F1 score	$2 * \frac{precision * recall}{precision + recall}$

The Random Forest, SVM, and Logistic Regression are set in training and testing the dataset of WCDB using Python, utilizing the Sci-kit learn libraries of the models in Jupyter Lab editor.

D. Result and Discussion

In the research experiments, firstly, the implementation of feature selection using the Extra tree method. Figure 2 displays the attribute ranking for the breast cancer dataset using the extra trees model. According to the diagram, the attributes:(smoothness_se,smoothness_mean,smoothness_worst,symmetry_se,symmetry_worst,symmetry_mean,fractal_dimension_se,fractal_dimension_mean, fractal_dimension_worst) removed from the dataset. Therefore, the dataset now includes 569 instances and 24 features; one of them is the label that includes two classes: Malignant (M) and Benign (B). Secondly, I used the standard Scaler function to scale the numeric values of the selected features into the same range and remove outliers [24, 25].

SVM, Random forest, and Logistic Regression trained and tested the dataset with the selected features. Table 3 displays the testing experiments of the models on the original dataset. Then, I divided the new dataset into 70% for training and 30% for testing. According to the testing results, SVM performed the best among the other proposed models in precision, recall, F1 score, and accuracy with 97%, 99%, 98%, and 98%, respectively. Furthermore, SVM performed better compared to its performance on the original dataset, as shown in Table 4.

On the other hand, the Random forest performed the same on both datasets with a 96% accuracy rate. Linear regression had little better results on the original dataset by 0.5% compared to its performance on the selected features. Also, some

previous studies tested the WCDB dataset using different models; Table 5 displays the details.

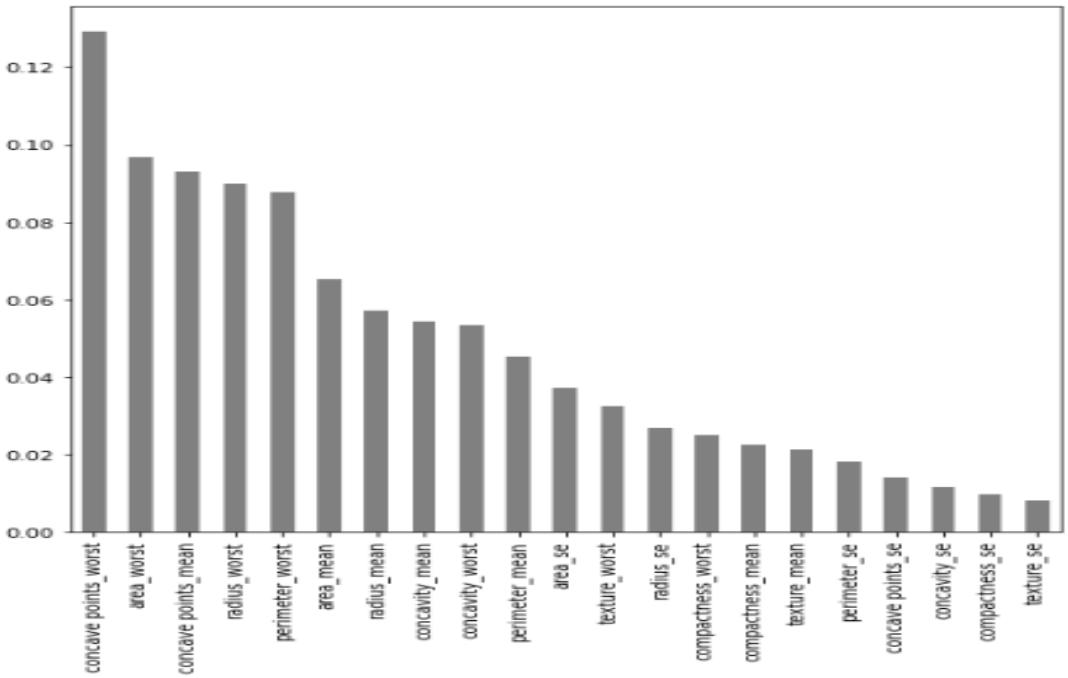

Figure2. Feature importance using Extra Trees on breast cancer dataset

Table3. The Performance of Different Models on The Original WCDB Dataset

Model	Precision	Recall	F1 score	Accuracy
SVM	97%	98%	98%	97%
Random Forest	98%	96%	97%	96.4%
Logistic Regrisssion	97%	97%	97%	96%

Table4. The Performance of Different Models on The WCDB Dataset with The Selected Features

Model	Precision	Recall	F1 score	Accuracy
SVM	97%	99%	98%	98%
Random Forest	98%	96%	97%	96.4%
Logistic Regrisssion	95%	98%	97%	95.9%

Table5. Comparison of Accuracy with other Models on WCDB Dataset

Reference	Year	Model	Accuracy rate
[6]	2019	Random Forest	70%
[7]	2019	Extra Tree	97%

[10]	2019	MAE	91%
[11]	2020	ANN	98%
[19]	2021	HCRF	97%
[9]	2021	K-means and GMM	95.5%
[12]	2022	SVM	96.5%
[13]	2023	Logistic regression (class weight function was used)	98%
This work	2023	SVM (Extra Tree was used to select features)	98%

From the table, it is clear that our suggested models, which used Extra trees as a feature selection method, performed the best.

E. Conclusion

This research predicted breast cancer from the WCDB dataset using several models. The study aimed to reduce the features on the dataset to predict the disease; the Extra tree was the method for feature selection. Several models tested the dataset before and after feature selection. The models included SVM, Random Forest, and Logistic Regression. The testing results showed that the feature selection process improved the performance of the models. SVM achieved the highest rates among the other models, and also, compared to its performance on the original dataset. Since the study used the WCDB dataset, the results may not apply to other cases. Studies should look at other clinical datasets, prediction models, and feature selection methods in the future.

References

[1] "Breast cancer statistics." (2023) WCRF International. https://www.wcrf.org/cancer-trends/breast-cancer-statistics/#:~:text=Breast%20cancer%20is%20the%20most%20commonly%20occurring%20cancer%20in%20women,cancer%20in%20women%20in%202020.

[2] "World Health Organization. Cardiovascular diseases (CVDs)." https://www.who.int/cardiovascular_diseases/en/ (accessed 25 June, 2023).

[3] "Mayo Clinic. Breast Cancer: Symptoms and causes." (2023) https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

[4] "NHS. Breast cancer in women: Treatment - NHS." (2023) https://www.nhs.uk/conditions/breast-cancer/treatment/

[5] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, "Machine learning classification techniques for breast cancer diagnosis," in *IOP Conference Series: Materials Science and Engineering*, 2019, vol. 495, no. 1: IOP Publishing, p. 012033.

[6] M. S. Yarabarla, L. K. Ravi, and A. Sivasangari, "Breast cancer prediction via machine learning," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2019: IEEE, pp. 121-124.

[7] H. Dhahri, E. Al Maghayreh, A. Mahmood, W. Elkilani, and M. Faisal Nagi, "Automated breast cancer diagnosis based on machine learning algorithms," *Journal of healthcare engineering,* vol. 2019, 2019.

[8] Z. HUANG and D. CHEN, "A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering

Random Forest Algorithm," *IEEE Access,* vol. 10, 2022, doi: 10.1109/ACCESS.2021.3139595.

[9] P. E. JEBARANI, N. UMADEVI, I. HIEN DANG (Member, and M. POMPLUN, "A Novel Hybrid K-Means and GMM Machine Learning Model for Breast Cancer Detection," *IEEE Access,* vol. 9, 2021, doi: 10.1109/ACCESS.2021.3123425.

[10] M. R. Karim, G. Wicaksono, I. G. Costa, S. Decker, and O. Beyan, "Prognostically relevant subtypes and survival prediction for breast cancer based on multimodal genomics data," *IEEE Access,* vol. 7, pp. 133850-133864, 2019.

[11] 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, pp. 1-14, 2020.

[12] S. R. Gupta, "Prediction time of breast cancer tumor recurrence using Machine learning," *Cancer Treatment and Research Communications,* vol. 32, p. 100602, 2022.

[13] S. SJ, P. K. SC, and T. A. Assegie, "A cost-sensitive logistic regression model for breast cancer detection," *The Imaging Science Journal,* pp. 1-9, 2023.

[14] D. W. H. Wolberg, W. N. Stree, and O. L. Mangasarian. "Breast Cancer Wisconsin (Diagnostic) Data Set." (2023). UCI machine learning repository. https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

[15] W. H. Wolberg, W. N. Street, D. M. Heisey, and O. L. Mangasarian, "Computerized breast cancer diagnosis and prognosis from fine-needle aspirates," *Archives of Surgery,* vol. 130, no. 5, pp. 511-516, 1995.

[16] G. Alfian *et al.*, "Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method," *Computers,* vol. 11, no. 9, p. 136, 2022.

[17] A. R. Kharwar and D. V. Thakor, "An ensemble approach for feature selection and classification in intrusion detection using extra-tree algorithm," *International Journal of Information Security and Privacy (IJISP),* vol. 16, no. 1, pp. 1-21, 2022.

[18] " Decision Trees." Scikit learn. (2023) https://scikit-learn.org/stable/modules/tree.html#tree-mathematical-formulation (accessed 20-3-2023, 2023).

[19] Z. Huang and D. Chen, "A breast cancer diagnosis method based on VIM feature selection and hierarchical clustering random forest algorithm," *IEEE Access,* vol. 10, pp. 3284-3293, 2021.

[20] A. Ramchandani, C. Fan, and A. Mostafavi, "Deepcovidnet: An interpretable deep learning model for predictive surveillance of covid-19 using

heterogeneous features and their interactions," *Ieee Access,* vol. 8, pp. 159915-159930, 2020.

[21] M. A. Tahir, C.-H. Chan, J. Kittler, and A. Bouridane, "Face recognition using multi-scale local phase quantisation and linear regression classifier," in *2011 18th IEEE International Conference on Image Processing*, 2011: IEEE, pp. 765-768.

[22] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert systems with applications,* vol. 36, no. 2, pp. 3240-3247, 2009.

[23] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," *arXiv preprint arXiv:2008.05756,* 2020.

[24] s.-k. l. developers. "Compare the effect of different scalers on data with outliers." (2023). Sci-kit learn. https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html#:~:text=make_plot(0)-,StandardScaler,empirical%20mean%20and%20standard%20deviation. (accessed 2-2-2023, 2023).

[25] A. A. Tokuç. "Why Feature Scaling in SVM?" (2023). Baeldung. https://www.baeldung.com/cs/svm-feature-scaling#:~:text=Feature%20scaling%20is%20crucial%20for,data%20points%20from%20different%20classes.