Journal of Basrah Researches (Sciences) **49**(1), 44 (2023)
DOI: https://doi.org/10.56714/bjrs.49.1.5

Feature Selection Using a Hybrid Approach Depends on Filter and Wrapper Methods for Accurate Breast Cancer Diagnosis

Mohammed S. Hashim, Ali A.Yassin*

Department of Computer science, Education College for Pure Sciences, University of Basrah, Basrah, Iraq.

ARTICLE INFO

Received 13 December 2022
Accepted 08 January 2023
Published 30 June 2023

Keywords:

Breast cancer, machine learning, feature selection, voting classifier, cross-validation.

Citation: M.S. Hashim, A. A.Yassin, J. Basrah Res. (Sci.) **49**(1), 44 (2023).
DOI:https://doi.org/10.56714/bjrs.49.1.5

ABSTRACT

Breast cancer is the biggest cause of mortality in women, outscoring all other malignancies. Diagnosing breast cancer is hard because the disease is complicated, treatment methods change, and there are many different kinds of patients. Information technology and artificial intelligence contribute to improve diagnostic procedures, which are critical for care and treatment as well as reducing and controlling cancer recurrence. The primary part of this research is to develop a new feature selection strategy based on a hybrid approach that combines two methods for selecting features: the filter and the wrapper. In two stages, this method reduces the number of features from 30 to 15 to increase and improve classification accuracy. The suggested method was tested using the Wisconsin Breast Cancer Dataset dataset (WDBC). To enhance the classification of breast cancer tumors, a soft voting classifier was used in this study. The proposed methodology outperforms previous research, achieving 1 for the F1 score, 1 for AUC, 1 for recall, 1 for precision, and 100% for accuracy. Furthermore, 10-fold cross-validation has a 98.2% accuracy rate.

1. Introduction

Breast cancer is a neoplastic illness that presents a significant threat to women's health. It is regarded as the most common cause of cancer death in women. According to estimates from the World Health Organization, women are more prone than males to get breast cancer, as shown by the 685,000 deaths and 2.3 million infections that have been identified [1]. This cancer is in the form of a tumor in the breast and is classified as either benign, in the form of a mass in a specific position that can be removed, or malignant, which spreads to neighboring parts of the body.Artificial intelligence (AI) plays an important role in healthcare where cognitive technology is used in medical contexts. The most basic way to describe AI is as the ability of computers and other systems to mimic human cognition and build the ability to learn, reason, and make decisions. Therefore, AI has important implications in diagnostics and prediction, where it may assist doctors and other medical professionals in effectively detecting and diagnosing diseases and developing treatment plans based on the patient's information[2].Early diagnosis of the type of breast cancer tumor is crucial in preserving human life by 90% [1]. For this, many AI algorithms have been used in this field for individualized care and treatment as well as to lessen and manage cancer recurrence. In order to

***Corresponding author email :** ali.yassin@uobasrah.edu.iq

ISSN: 1817-2695 (Print); 2411-524X (Online)
Online at: https://jou.jobrs.edu.iq

specify whether a tumor is benign or malignant, machine learning (ML) algorithms for classification and prediction are commonly used in breast cancer research, especially on datasets relevant to breast cancer [3].Early breast cancer detection has been the subject of several studies using machine and deep learning algorithms. Several feature selection methods have been applied, which are filtering, covering, and hybrids that combine more than one method, as well as several classification algorithms have been put into practice to accurately diagnose the kind of breast cancer tumor.The imbalance of the dataset, which causes the ML model to be biased to the majority side [4], is one of the factors that contribute to the limitations of these researchers in terms of the accuracy of diagnosis and prediction. A new approach that combines more than two ways to get the greatest results in terms of accuracy in diagnosis was not used in prior work, which was restricted to using the existing feature selection methods. Details are provided in the related work section.
The following are the main contributions made by this study:

- SMOTE is used in the dataset balancing procedure to prevent bias in the ML model toward a certain party.
- To achieve the highest classification accuracy, we develop a new hybrid feature selection strategy that combines the filter and wrapper feature selection methods.
- We use a soft voting classifier, which combines the strengths of three models into a single model to improve classification accuracy.

The following sections make up the remaining text of the paper: In the second section, we present previous studies related to our work. Section 3 presents a background that explains the feature selection methods as well as the ML algorithms used in this study. Section 4 outlines the proposed methodology for identifying breast cancer tumor types. Section 5 presents and discusses the results acquired utilizing the proposed methodology. The conclusion is introduced in Section 6.

2. Related Work

Healthcare is one of the most crucial sectors to use artificial intelligence because of the critical necessity for correct diagnosis. Therefore, to improve the accuracy and speed of classification, numerous researchers have applied artificial intelligence approaches to the early detection of breast cancer. In this section, we summarized several pertinent research that used machine learning and deep learning techniques to diagnose the type of breast cancer tumor, whether it is benign or malignant.Dhahri et al. [5] conducted a comparative study between several AI algorithms, namely random forest, gradient boosting, SVM, gaussian naïve Bayes, k-nearest neighbor, extra trees classifier, AdaBoost, linear discriminant analysis, quadratic discriminant analysis, LR, and DT. In this study, they used a Wisconsin Diagnostic Breast Cancer (WDBC) dataset and the genetic algorithm was applied to it to choose the most advantageous subset of features, as the number of features was reduced from 30 to 12 features. With a rate of 98.24%, the AdaBoost classifier ultimately achieved the best accuracy in comparison to the others. Memon et al. [6] used the linear SVM as an ML model to solve the breast cancer classification problem. They used the WDBC dataset and implemented Recursive Feature Elimination to choose the best subset of features to train and test the model on. The number of features was reduced from 30 to 18. In the result, high specificity (99%), accuracy (99%), and sensitivity (98%) were attained by the SVM model.HAQ et al. [7] have used SVM as an ML classifier and three different ways to choose the best subset of features that have been applied to the WDBC dataset, namely autoencoder algorithms, principal component analysis, and relief, as each of these methods gives a different number of features. They compared the performance results of SVM when applied with all previous feature selection methods where only 18 features were used in SVM with Principal Component Analysis to reach the best accuracy of 97.45%. Sharma and Mishra [8] used a WDBC dataset and applied three different methods to select the best subset, namely correlation-based feature selection, information gain, and sequential feature selection. They then run ten classification algorithms on the original dataset and the resulting subsets of the three methods of feature selection to find the best features to use. When compared to other methods, correlation-based feature selection produces features with higher accuracy. After that, they selected the three best

algorithms that gave the highest classification accuracy from this method for building the voting classifier, namely the artificial neural network (ANN), LR, and SVM, where the results showed that this classifier gave a classification accuracy of 99.41%. HUANG and CHEN [9] utilized two datasets in their research, which are the WBC and WDBC. They applied the Variable Importance Measure (VIM) method to the two datasets used to extract the best subset of the original dataset, where they reduced the number of features from 30 to 24 for the WDBC dataset and for the WBC dataset the features shrank from 9 to 8. Then, they compared the performance of their proposed model Hierarchical Clustering Random Forest (HCRF) with three models, AdaBoost, random forest, and DT, using both datasets. The comparison's outcome revealed that the HCRF model had the best accuracy in the WBCD dataset (97.05%) and the WBC dataset (97.76%). Ibrahim et al. [10] conducted a comparative study between seven classification algorithms, in addition to the soft and hard voting classifier, which work to diagnose the type of breast cancer tumor, whether it is benign or malignant. In their study, they used the WDBC dataset and applied a set of feature selection methods to it for choosing the best subset, namely correlation analysis and principal component analysis, and the wrapper method. The soft vote classifier ultimately reached the greatest accuracy of 99% by utilizing 21 features selected using the correlation analysis and principal component analysis methodologies.Jumanto et al. [11] utilized a backpropagation artificial neural network as a classifier model to determine the type of breast cancer tumor. A breast cancer WDBC dataset was used, and the forward feature selection method based on the random forest was implemented to decide which subset of features was the best on which the model was trained and test its performance. The results demonstrated the model's 98.3% accuracy.

3. Background

In this part, we will explain the feature selection methods as well as the machine learning algorithms used in this study.

3.1. Pearson Correlation Coefficient

The degree and direction of the association between two variables are statistically quantified using the Pearson correlation, which is one of the filter methods for choosing the best features from the dataset. It details how strongly two variables are linearly correlated where their value is confined between (-1 and 1). When the value (0) that means there is no correlation between the two features. There is a positive correlation between the two features when the value is between zero and one and a negative correlation when the value is between zero and a negative one [12].

3.2. Mutual Information

Mutual information (MI) is one of the filter methods for choosing the best features from the dataset. it measures the dependency between two variables and is mainly used to measure if there is a strong correlation between any feature of the data set with the target class. The value of MI ranges between (0,1) where [13]:

- value(1): Strong dependency
- value(0): No dependency

Through the value of MI, we conclude whether a particular feature is of great importance through its strong association with the target class, as well as a particular feature of weak importance through its weak association with the target class.

3.3. Sequential Feature Selection (SFS)

It is one of the wrapper methods to select the best features from the dataset by removing unimportant features, it relies on the artificial intelligence algorithm used to determine the best features that can be kept. The first phase of the sequential forward feature selection adds one feature to an empty set in order to produce the highest value for the objective function. Following the first

phase, each of the remaining features is separately added to the existing subset, and the new subset is then assessed. The subset is permanently updated with the individual attributes that provide the highest level of classification accuracy. Until we have the required amount of features, the process is repeated. Because the dependency between the features is not taken into account, this algorithm is referred to as a naive SFS algorithm [14].

3.4. Support Vector Machine

The term "Support Vector Machines (SVM)" refers to a discriminant technique that is often utilized in machine learning for classification. The SVM algorithm separates two classes with the greatest possible margin when a task can be split up linearly. Soft SVM cannot locate a robust separating hyperplane when a problem is not linearly separable in input space, which would minimize the amount of wrongly classified data points and be well generalizable. Therefore, SVM in the kernel approach employs the kernel trick to convert the data into a higher-dimensional space before tackling the machine-learning task [15].

3.5. Logistic Regression

Logistic regression (LR) is a useful statistical ML algorithm that could aid with classification and regression issues. This statistical model is utilized to predict binary values. LR was used to evaluate the relationship among the dichotomous dependent variable, often referred to as the response variable, and the independent variable sometimes referred to as the predictor variable. To normalize the prediction to be between 0 and 1, LR uses a sigmoid function. It is commonly utilized in the medical field because it gives specific output values [16].

3.6. Decision Tree

A Decision Tree (DT) is considered one of the commonly used algorithms because its construction process is simple and does not require any parameters where it is used for classification and regression functions. The decision tree contains three types of nodes, namely the root node, internal nodes, and leaf nodes, and each of these nodes represents a decision source. The tree was built from the top to the bottom, and at each level the best feature is chosen based on a specific criterion, on the basis of which the decision is made to continue to the lower level [17].

3.7. Voting Classifier

The voting classifier (VC) is one of the ensemble strategies used to create a potent classifier with greater classification accuracy than conventional ML classifiers. On the majority of the dataset, ensemble-based algorithms often outperform others [18]. The VC takes more than one artificial intelligence model and votes among the prediction results of these models to produce a powerful model that carries the power of the input models, where There are two types of voting hard and soft [19].

4. Methodology

In this section, we present the proposed methodology, which consists of two stages, namely the Dataset Setup Phase and Prediction Phase, where it works on first preparing the dataset before using it by the machine learning model. Then we show the proposed model that we use, which in turn gives the best classification accuracy for the type of breast cancer tumor and as shown in Fig. 1.

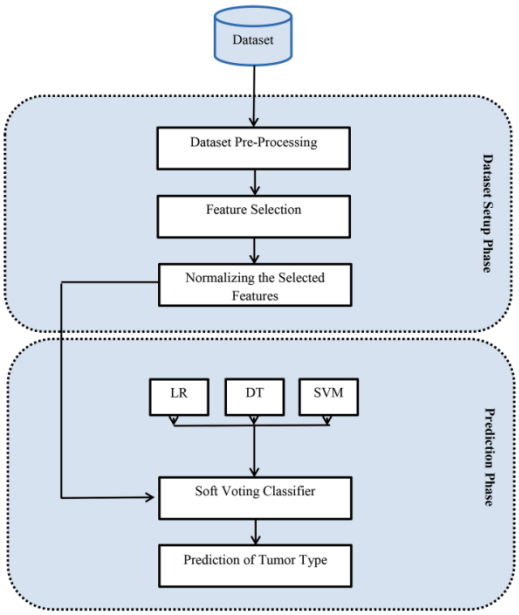

Fig. 1. Our proposed methodology.

In the beginning, we carry out preliminary processing of the data set to ensure the quality of the data and then choose the best features that represent this data set by developing a new method that combines filter and wrapper methods to reduce the number of features from 30 to 15 in order to ensure accurate and correct learning, which in turn gives accurate prediction results when these 15 features are used by the soft voting classifier which also acts as a powerful model that carries the power of the models built into it.

In this study, we use a Wisconsin Diagnostic Breast Cancer dataset that outlines the features of a breast mass' cell nuclei. The UCI machine learning repository provided this dataset [20]. This dataset consists of 569 samples and 32 features whereas only 30 features are used in practice. The feature "diagnosis" is the target class as it contains two types of tumor, benign(B)= 357 samples and malignant(M)=212 samples.

4.1. Dataset Setup Phase

At this phase of our study, we perform a set of operations that prepare and configure the data set used to be ready for use by machine learning models to predict the type of breast cancer tumor, which is Dataset pre-processing, Feature selection, and Normalizing the selected features.

4.1.1. Dataset Pre-processing

In this part, we will explain a set of preliminary operations that we perform on the dataset used to maintain the quality of this data as well as prepare it for the later stages, which in turn helps to conduct better training for machine learning models.

Firstly, the dataset used consists of 32 features, the actual number that we use is 30 features only because the first feature is “ id ” and the last feature is “unnamed:32” all rows have empty values, so we perform the drop function for both features.

Secondly, the dataset consists of two classes, the minority class (M=212) and the majority class (B=357). so, we perform is the process of balancing the dataset by using SMOTE method which works to balance the number of samples of the minority and majority categories, as it works to raise the number of minority samples from 212 to 357 by creating new samples located on the lines that connect every two samples of this class, as it relies on K-nearest neighbors in selecting samples. The main goal of balancing the dataset is to remove the bias that occurs in machine learning models during the training process towards the majority class [21].

Finally, we do the Labelencoder process for the Class Label on the feature " diagnosis " where we convert the label (M) into 1 and label (B) into 0. The main goal of this process is to encode the data into numeric values such that a number is used to represent each categorical feature because only numeric values can be entered into machine learning models [22].

4.1.2. Feature Selection

The process of selecting the best features from the data set is an important step in the classification and prediction process. Choosing a subset of the original data set helps improve classification accuracy and remove unimportant features that help increase learning errors as well as make learning algorithms' computational complexity less complicated [23].

In our study, we developed a new approach to feature selection, as it is considered a hybrid approach that combines two methods for selecting features the first is the filter and the second is the wrapper, where this method is in two stages which reduce the number of features from 30 features to 15 features to increase and improve classification accuracy.

In the first stage, we use the filter method, where we combine two methods of the filter method, namely: Pearson correlation coefficient and mutual information. Pearson correlation measures the degree of relationship between one feature and another in the used dataset, where a group of pairs is formed based on the degree of relationship between them. We take the relationships with which the degree of correlation is greater than or equal to 0.89, and then we collect the relationships that have common features in totals. Next, we extract the common features in each group. After extracting the common features from each group, it is the turn of mutual information, by selecting the feature that has the highest value from mutual information to represent this group and dropping the rest of the features of this group, this combined method is called PC-MI. Table 1 shows the mutual information values for each feature.

Table 1. Mutual information values for features.

Feature	Score	Feature	Score
texture_se	0.002271	compactness_mean	0.276540
smoothness_se	0.023746	radius_se	0.277863
fractal_dimension_mean	0.023849	compactness_worst	0.283761
symmetry_se	0.027308	perimeter_se	0.284167
fractal_dimension_se	0.048942	concavity_worst	0.358751
symmetry_mean	0.070514	area_se	0.366311
fractal_dimension_worst	0.097076	area_mean	0.405551
symmetry_worst	0.101290	radius_mean	0.407724
smoothness_mean	0.108997	concavity_mean	0.421748
compactness_se	0.119047	perimeter_mean	0.427724
smoothness_worst	0.120389	concave points_worst	0.471331
texture_worst	0.139945	concave points_mean	0.474548
texture_mean	0.145510	radius_worst	0.483495
concave points_se	0.177446	area_worst	0.490916
concavity_se	0.178546	perimeter_worst	0.499442

Table 2 shows the first stage of the proposed approach for feature selection.

Table 2. The first stage of the proposed feature selection approach.

Groups	Correlated features	Pearson Score	Common features	Chosen feature of a high mutual information value
1	[radius_mean, perimeter_mean]	0.998	radius_mean	perimeter_worst
	[radius_worst, perimeter_worst]	0.994		
	[radius_mean, area_mean]	0.989	perimeter_mean	
	[perimeter_mean, area_mean]	0.988		
	[radius_worst, area_worst]	0.986	area_mean	
	[perimeter_worst, area_worst]	0.980		
	[perimeter_mean, perimeter_worst]	0.972	radius_worst	
	[radius_mean, radius_worst]	0.971		
	[perimeter_mean, radius_worst]	0.970	perimeter_worst	
	[radius_mean, perimeter_worst]	0.967		
	[area_mean, radius_worst]	0.965	area_worst	
	[area_mean, area_worst]	0.962		
	[area_mean, perimeter_worst]	0.961		
	[perimeter_mean, area_worst]	0.947		
	[radius_mean, area_worst]	0.947		
2	[radius_se, perimeter_se]	0.97	radius_se	area_se
	[radius_se, area_se]	0.96	perimeter_se	
	[perimeter_se, area_se]	0.94	area_se	
3	[concavity_mean, concave points_mean]	0.93	concavity_mean	concave points_mean
	[concave points_mean, concave points_wors	0.91	concave points_mean	
	[compactness_mean, concavity_mean]	0.89	concave points_wors	
			compactness_mean	
4	[compactness_worst, concavity_worst]	0.90	compactness_worst	concavity_worst
			concavity_worst	
5	[texture_mean, texture_worst]	0.91	texture_mean	texture_worst
			texture_worst	

We conclude from the first stage of the proposed approach to feature selection, that the number of features has been reduced from 30 to 18, as 12 features have been dropped, which are (area_worst, radius_worst, perimeter_mean, radius_mean, area_mean, texture_mean, compactness_mean, concavity_mean, concave points_worst, radius_se, perimeter_se, compactness_worst).In the second stage, we use one of the wrapper methods which is sequential forward feature selection. After selecting the best 18 features from the first stage of the proposed approach, we include these 18 features on sequential forward feature selection using the soft voting classifier as a machine learning model because all the wrapper methods depend on the feature selection on the model used. After entering 18 features on this method, we get 15 features that give the best average score this method, which is (smoothness_mean, fractal_dimension_mean, texture_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, fractal_dimension_se, texture_worst, perimeter_worst, smoothness_worst, concavity_worst, symmetry_worst, fractal_dimension_worst).

We conclude by passing the used data set through the two phases of the proposed hybrid approach that the number of features has been reduced from 30 features to 15 features only, and thus the undesirable features that affected the classification process have been removed.
Figure 2 shows the proposed hybrid approach for selecting the top 15 features from the used dataset.

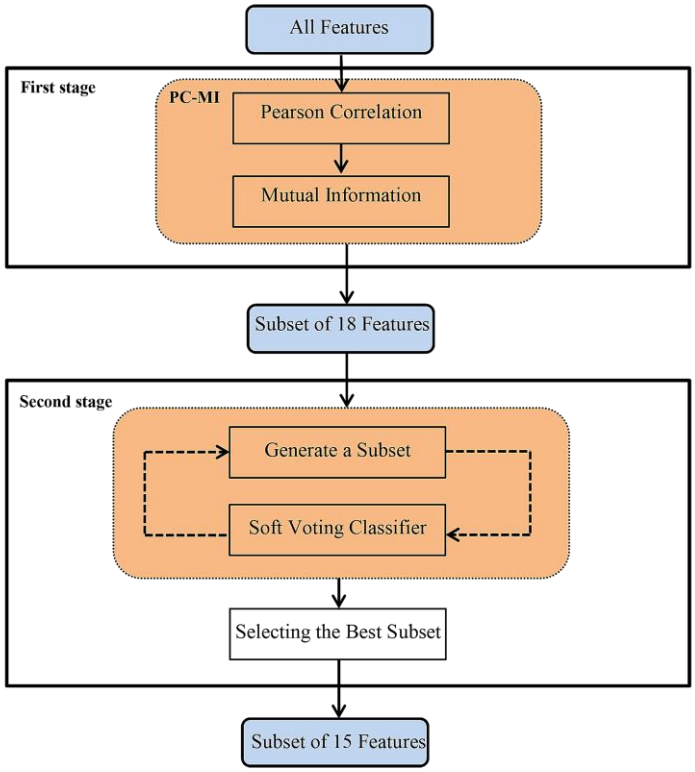

Fig. 2. The proposed feature selection approach.

4.1.3. Normalizing the Selected Features

After applying the proposed approach to choosing the best subset of the features of the original data set, we normalize these features using Standard Scaler. It works to turn values into standard units free from the arithmetic mean's effect. The main benefit of feature normalization is to give good results in terms of learning as well as reduce the time required for the training process [24]. It can be calculated by Eq.(1) [25].

$$\bar{x}_i = [x_i - \bar{x}] / s \tag{1}$$

Where :

- $\bar{x}_i$: The normalizing value
- x_i : The sample value
- $\bar{x}$: The arithmetic mean
- S : The standard deviation

4.2 Prediction Phase

Upon completing the initial processing of the data set, selecting the best subset, and normalizing these features, the dataset, which eventually contains 15 features, becomes ready for the machine learning model to be trained on.

In this study, we used the soft voting classifier as a machine learning model to predict the type of breast cancer tumor if it is benign or malignant. Three models were used to include it in this classifier, which is (DT, SVM, and LR). These models are considered the best models that work with the soft voting classifier based on the experiments that have been performed. This classifier builds one powerful model that holds the power of the models embedded in it. The soft voting classifier

decides whether a tumor is benign or malignant based on the highest probability average of all three models used, where the basis of its work is on the probabilistic (P) principle, as shown in Fig. 3.

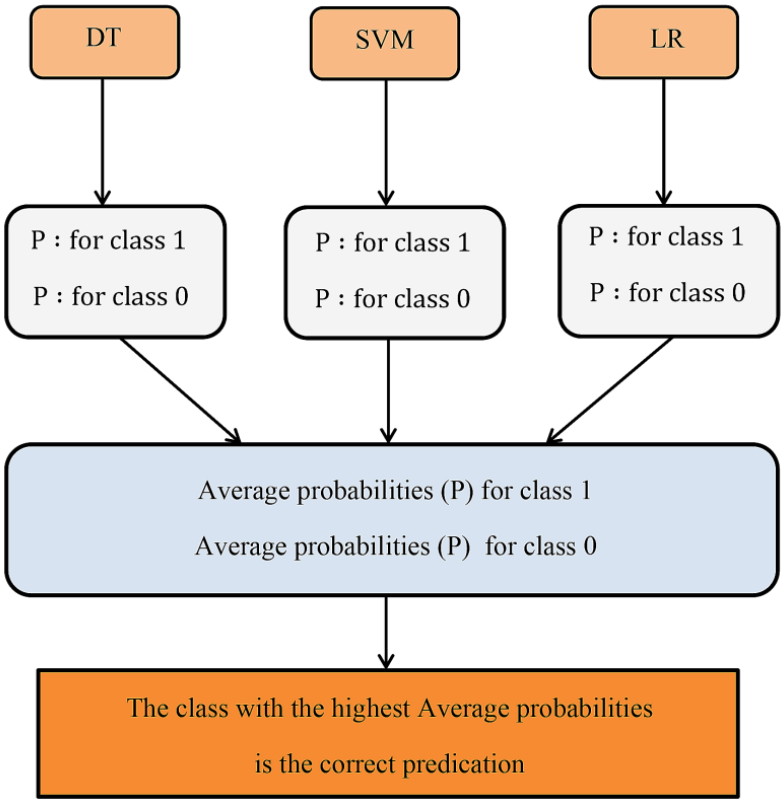

Fig. 3. Our soft voting classifier.

5. Result and Discussion

In this section, we present two experiments for diagnosing breast cancer tumor types using the proposed feature selection approach as well as using the proposed soft voting classifier as a machine learning model for the classification and prediction process. In addition, we compared the results obtained with previous works that work on the same dataset used. These experiments using the proposed methodology were applied to the WDBC dataset. In addition, we employed a number of metrics in order to assess and gauge how well our proposed methodology performed which are precision, accuracy, AUC, recall, F1 score, and confusion matrix.

5.1. Experiment (1)

In this experiment, we divided the used dataset (WDBC) after applying the proposed approach to feature selection and obtaining only 15 features into two parts, the first part is the training set (80%), and the second is the test set (20%). We used the first set to train the soft voting classifier (DT, SVM, LR). We used the second group to test and measure the performance of our proposed model, as this model got 1 for precision, 100% of accuracy, 1 for AUC, 1 for recall, and 1 for F1 score. The confusion matrix for the findings is shown in Fig. 4.

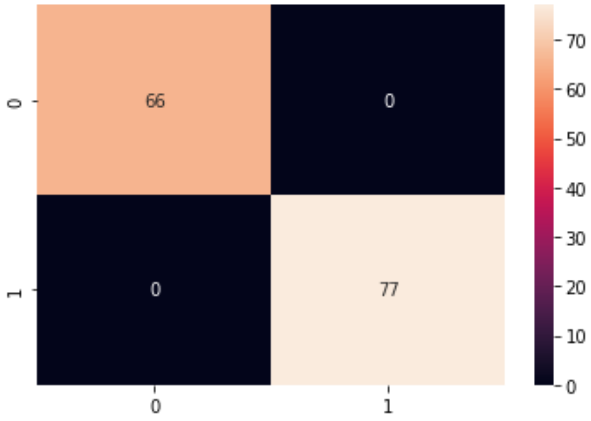

Fig. 4. Confusion matrix based on our findings.

Table 3 presents a comparison between the result of our proposed methodology and previous works working on the same dataset used.

Table 3. Comparison among the results of our study with previous works that used the WDBC dataset.

Authors	Year	Balanced Dataset	Number of Features	Model	Accuracy (%)
Dhahri [5]	2019	NO	12	Adaboost classifier	98.24
Memon [6]	2019	NO	18	SVM	99
HAQ [7]	2021	NO	18	SVM	97.45
Sharma [8]	2021	NO	11	Voting Classifier (ANN - SVM -LR)	99.41
HUANG [9]	2021	NO	24	Hierarchical Clustering Random Forest	97.05
Ibrahim [10]	2021	NO	21	Soft Voting	99
Jumanto [11]	2022	NO	—	ANN	98.3
Proposed Methodology	**2023**	**YES**	**15**	**Soft Voting Classifier (DT-SVM-LR)**	**100**

In this experiment, we found that the soft voting classifier, which was based on our proposed method, was the most accurate compared to other methods. Its performance was tested on 143 cases, and in all of them, the correct prediction was made. Thus, we conclude that the model was trained correctly based on the training set.

5.2. Experiment (2)

In this experiment, we used 15 features according to the result of the proposed approach to feature selection. The performance of the proposed model (soft voting classifier) is measured using k-fold as k = 10, where the largest part is used for training and the remaining part is used for performance testing in each round. The accuracy of the model's performance is calculated by finding the final average of the accuracy in all rounds. During this experiment, the soft voting classifier obtained an accuracy of 98.2%.

6. Conclusion

Breast cancer is an extremely hazardous illness that affects women all over the globe. The accurate and efficient identification of breast cancer is a major medical concern, and many researchers have suggested many diagnostic techniques for its detection. However, these current techniques still need to be improved if breast cancer is to be accurately and efficiently detected. In this study, we proposed a new methodology that improves the accuracy and efficiency of breast cancer tumor type detection, whether it is benign or malignant. At the beginning of this methodology, we balanced the data set to eliminate bias during training. After that, we proposed a new approach to feature selection, which combines two methods of feature selection, filter, and wrapper, which in turn reduces the number of features from 30 to 15, and thus removes the features that negatively affect the training process. In addition, we used a soft voting classifier that includes DT, SVM, and LR which combines these three models to produce a single model that carries the power of these models, which in turn improves the accuracy of diagnosis. The performance of our proposed methodology was evaluated through several measures, namely: F1 score, AUC, recall, precision, and accuracy. The proposed methodology achieved 1 for the F1 score, 1 for the AUC, 1 for recall, 1 for precision, and 100 % accuracy when compared to previous works that employed the same dataset. Furthermore, 10-fold cross-validation has a 98.2% accuracy rate.

References

[1] World Health Orgnaization | WHO, World Breast Cancer Rep (2020).
[2] A. Haleem, M. Javaid, I.H. Khan, Current Medicine Research and Practice **9**(6), 231 (2019).
[3] H. Asri, H. Mousannif, H. Al Moatassime, T. Noel, Procedia Comput. Sci. **83**, 1064 (2016).
[4] S. Guo, Y. Liu, R. Chen, X. Sun, X. Wang, Neural Process. Lett. **50**(2), 1503 (2019).
[5] H. Dhahri, E. Al Maghayreh, A. Mahmood, W. Elkilani, M. F. Nagi, J. Healthc. Eng. **2019**, 11 (2019).
[6] M. H. Memon, J. P. Li, A. U. Haq, M. H. Memon, W. Zhou, Wirel. Commun. Mob. Comput. **2019**, 19 (2019).
[7] A.U. HAQ, J.P. LI, A. SABOOR, J. KHAN, S. WALI, S. AHMAD, A. ALI, G.A. KHAN, W. ZHOU, IEEE Access **9**, 99 (2021).
[8] A. Sharma, P.K. Mishra, International Journal of Information Technology, 1 (2021).
[9] Z. Huang, D. Chen, , IEEE Access **10**, 3284 (2022).
[10] S. Ibrahim, S. Nazir, J. Imaging **7**(11), 225 **(**2021).
[11] M. F. Mardiansyah, R. Pratama, M. F. Al Hakim, B. Rawat, J. Soft Comput. Explor **3**(2), 105 (2022).
[12] R. Saidi, W. Bouaguel, N. Essoussi, Mach. Learn. Paradig. theory Appl. Springer, Cham **7**, 3 (2019).
[13] L. Batina, B. Gierlichs, E. Prouff, M. Rivain, F. X. Standaert, N. Veyrat-Charvillon, J. Cryptol. **24(**2), 269 (2011).
[14] K. Pavya, D. B. Srinivasan, Int. J. Sci. Dev. Res. **2(**6), 594 (2017).
[15] M. Awad, R. Khanna, "Support Vector Machines for Classification,", 1st ed. , ch.3, 39 (2015).
[16] M. A. Naji, S. El Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, O. Debauche, Procedia Comput. Sci. **191**, 487 (2021).
[17] M. F. Ak, Healthcare, **8**(2), 111 (2020).

[18] M. A. Khan, M. A. K. Khattk, S. Latif, A.A. Shah, M.U. Rehman, W. Boulila, M. Driss, J. Ahmad, " Voting Classifier-Based Intrusion Detection for IoT Networks,", vol 1399. Springer , 313 (2021).
[19] A. Rasool, C. Bunterngchit, L. Tiejian, R. Islam, Q. Qu, Int. J. Environ. Res. Public Health **19**(6), 3211 (2022).
[20] D. W. H. Wolberg, M.L Repository., (1995).
[21] K. T. Id, P. Armitage, S. Tesfaye, D. Selvarajah, D. Wilkinson, PLoS One **15**(12), e0243907 (2020).
[22] K. Potdar, Int. J. Comput. Appl. **175**(4), 7 (2017).
[23] Q. Al-tashi, S. J. Abdulkadir, and S. Member, IEEE Access **8**, 125076 (2020).
[24] J. Sevilla, IEEE Trans. Nucl. Sci. **44**(3), 1464 (1997).
[25] P. Ferreira, D. C. Le, N. Zincir-heywood, 2019 15th Int. Conf. Netw. Serv. Manag. (CNSM). IEEE, Halifax, NS, Canada,1 (2019).

Journal of Basrah Researches (Sciences) **49**(1), 44 (2023)
DOI: https://doi.org/10.56714/bjrs.49.1.5

اختيار الميزات باستخدام نهج هجين يعتمد على طرق Filter و Wrapper لتشخيص دقيق لسرطان الثدي

محمد صلاح هاشم ، علي عادل ياسين*

قسم علوم الحاسوب ، كلية التربية للعلوم الصرفة ، جامعة البصرة ، البصرة ، العراق.

معلومات البحث

الاستلام 13 كانون الأول 2022
القبول 08 كانون الثاني 2023
النشر 30 حزيران 2023

الكلمات المفتاحية

سرطان الثدي ، التعلم الآلي ، اختيار الميزات ، مصنف التصويت ، التحقق المتبادل.

Citation: M.S. Hashim, A. A.Yassin, J. Basrah Res. (Sci.) **49**(1), 44 (2023).
DOI:https://doi.org/10.56714/bjrs.49.1.5

الملخص

سرطان الثدي هو أكبر سبب للوفيات عند النساء ، حيث يتفوق على جميع الأورام الخبيثة الأخرى. من الصعب تشخيص سرطان الثدي لأن المرض معقد ، وطرق العلاج تتغير ، وهناك أنواع عديدة من المرضى. تساهم تكنولوجيا المعلومات والذكاء الاصطناعي في تحسين إجراءات التشخيص ، والتي تعتبر ضرورية للرعاية والعلاج وكذلك الحد من تكرار الإصابة بالسرطان والسيطرة عليه. يتمثل الجزء الأساسي من هذا البحث في تطوير إستراتيجية جديدة لاختيار الميزات بناءً على نهج هجين يجمع بين طريقتين لاختيار الميزات: Filter and Wrapper. على مرحلتين ، تقلل هذه الطريقة عدد الميزات من 30 إلى 15 لزيادة دقة التصنيف وتحسينها. تم اختبار الطريقة المقترحة باستخدام مجموعة بيانات ويسكونسن لسرطان الثدي (WDBC). لتعزيز تصنيف أورام سرطان الثدي ، تم استخدام soft voting classifier في هذه الدراسة. تتفوق المنهجية المقترحة على الأبحاث السابقة ، حيث حققت 1 الى F1 score ، و 1 الى AUC ، و 1 الى recall ، و 1 الى precision ، و 100٪ الى accuracy . علاوة على ذلك ، يبلغ معدل الدقة في (10-fold cross-validation) 98.2٪.

***Corresponding author email :** ali.yassin@uobasrah.edu.iq

ISSN: 1817-2695 (Print); 2411-524X (Online)
Online at: https://jou.jobrs.edu.iq