Received: 7 December 2022 | Revised: 4 January 2023 | Accepted: 10 January 2023
DOI: 10.37917/ijeee.19.2.6 Vol. 19 | Issue 2 | December 2023

Open Access

Iraqi Journal for Electrical and Electronic Engineering
Original Article

IJEEE University of Basrah College of Engineering

Using Pearson Correlation and Mutual Information (PC-MI) to Select Features for Accurate Breast Cancer Diagnosis Based on a Soft Voting Classifier

Mohammed S. Hashim*, Ali A. Yassin
Department of Computer science - Education College for Pure Sciences, University of Basrah, Basrah, 61004, Iraq

Correspondance
*Mohammed S. Hashim
Department of Computer science,
Education College for Pure Sciences,
University of Basrah, Basrah, Iraq
Email: moh.salah@uobasrah.edu.iq

Abstract

Breast cancer is one of the most critical diseases suffered by many people around the world, making it the most common medical risk they will face. This disease is considered the leading cause of death around the world, and early detection is difficult. In the field of healthcare, where early diagnosis based on machine learning (ML) helps save patients' lives from the risks of diseases, better-performing diagnostic procedures are crucial. ML models have been used to improve the effectiveness of early diagnosis. In this paper, we proposed a new feature selection method that combines two filter methods, Pearson correlation and mutual information (PC-MI), to analyse the correlation amongst features and then select important features before passing them to a classification model. Our method is capable of early breast cancer prediction and depends on a soft voting classifier that combines a certain set of ML models (decision tree, logistic regression and support vector machine) to produce one model that carries the strengths of the models that have been combined, yielding the best prediction accuracy. Our work is evaluated by using the Wisconsin Diagnostic Breast Cancer datasets. The proposed methodology outperforms previous work, achieving 99.3% accuracy, an F1 score of 0.9922, a recall of 0.9846, a precision of 1 and an AUC of 0.9923. Furthermore, the accuracy of 10-fold cross-validation is 98.2%.

Keywords
Breast Cancer, Feature Selection, Soft Voting Classifier, Cross-Validation.

I. Introduction

Breast cancer is one of the most well-known and common diseases in the world, and its prevalence has been steadily rising in recent years. Women are the most likely to suffer from breast cancer, as 685,000 deaths and 2.3 million infections have been discovered, according to the World Health Organization reports. This cancer manifests as a lump in the breast that can either be benign or malignant and, in the latter case, spread to other parts of the body. Breast cancer risk is significantly influenced by genetic mutations [1].

The adoption of early detection methods, which aid in the treatment of this tumour and raise the likelihood of survival by 90%, helps increase survival rates [1]. Given that computers and other technologies are used to be able to learn, identify and diagnose the disease effectively and to provide treatment recommendations based on the data gathered from the patient, artificial intelligence (AI) and machine learning (ML) assist clinicians in the early identification of breast cancer [2]. In the medical field, ML algorithms for classification and prediction are frequently utilised[3], particularly on datasets related to breast cancer, to determine if a tumour is benign or malignant [4].

Many studies (details in the related work section) have been conducted in the field of early diagnosis of breast cancer using AI models. Some of these studies employ ML algorithms (models) to determine if the tumour is benign or malignant. Some studies use artificial neural networks, and some use an ensemble classifier as a classifier that combines several models.

As for the methods used in selecting features, many methods have been used in these studies, some of which are dependent on the filter method, and some are dependent on the wrapper method. The wrapper method removes redundant features that affect the model learning process and lead to a huge error in classification.

These studies are limited in terms of the accuracy of diagnosis and prediction, and the reason for this is due to several reasons, including the imbalance of the dataset, which leads to the bias of the ML model to the majority side [5]. Moreover, previous work were limited to the available feature selection methods, and they did not use a method that combines two simple methods to yield the best results and lowest cost, as well as models each one separately to find the best classifier among them.

The following is a summary of the study's main contributions:

- To attain the best diagnosis accuracy, we adopted a new feature selection method called PC-MI that combines two methods, namely, correlation analysis based on Pearson correlation and feature selection based on mutual information.
- The dataset balancing process is performed using SMOTE to avoid bias of the ML model to a specific party.
- The soft voting classifier is used, which integrates three models into one model that carries the strength of these models.
- The impact of employing a soft voting classifier on prediction accuracy was analysed.
- To aid physicians in the accuracy of the diagnosis, a web page was designed that diagnoses the type of breast cancer tumour.

The remaining portions of the paper are organised as follows. In section II we present previous studies related to our work. Section III provides a detailed explanation of the proposed methodology for diagnosing the type of breast cancer tumour. Section IV presents the results reached using the proposed methodology and discusses these results. Section V introduces the conclusion.

II. Related work

Given the pressing need for accurate diagnosis, healthcare is one of the most significant fields in which AI has been employed. ML and deep learning algorithms are used in several tests on datasets related to breast cancer, and they produce classification results with a high degree of accuracy. In this part, we present some previous studies related to diagnosing breast cancer using AI algorithms.

Hazra et al. [6] used the Wisconsin Diagnostic Breast Cancer (WDBC) dataset where they performed the feature selection process using the PC coefficient to obtain the least number of features. These features were passed on to three models, namely, support vector machine, naïve Bayes and ensemble classifiers to compare the results and achieve the best model classifying the disease, where the results showed that support vector machine with 19 features had the best accuracy of 98.51%.

Khuriwal and Mishra [7] used the WDBC dataset in their study and applied chi-square as a feature selection method to filter the dataset and keep the best features that diagnose the type of tumour present. Only 16 features were selected, and these features were passed to a voting classifier that included logistic regression (LR) and artificial neural network, which gave this classifier an accuracy score of 98.50%. Allam and Nandhini [8] used binary teaching learning-based optimisation, one of the wrapper methods for selecting the best features that represent a dataset. In their study, they used the WDBC dataset to diagnose tumour type. Five classification models have been applied: support vector machine (SVM), discriminant analysis, decision tree (DT), k-nearest neighbours (KNN) and Naive Bayes; amongst them, SVM gave the highest accuracy of 98.43% with nine features.

Memon et al. [9] applied recursive feature elimination (RFE) as a method for selecting the feature on the WDBC dataset for diagnosing whether the breast cancer tumour is benign or malignant. This method produced 18 features out of 30 features that were passed to the SVM model, which achieved high specificity (99%), accuracy (99%) and sensitivity (98%). Dhahri et al. [10] used genetic programming as a method to select the best features from the WDBC dataset. This method resulted in extracting 12 features out of 30 features, where several models were used to compare their performance on this dataset. These models were AdaBoost, LR, Gaussian Naïve Bayes, quadratic discriminant analysis, random forest, gradient boosting, SVM, linear discriminant analysis, KNN, DT and extra trees classifier. By contrast, the AdaBoost classifier obtained the highest accuracy relative to the others with a rate of 98.24%.

Ibrahim et al. [11] used two methods of feature selection in their study, which are correlation analysis and principal component analysis, and wrapper methods to select the best features where these methods were applied to a WDBC dataset. Seven classification algorithms were applied, and a soft and hard voting classifier was used from these algorithms to achieve the best accuracy. The results of all these classifiers were compared. The soft voting classifier obtained the highest accuracy of 99% with the use of 21 features selected using correlation analysis and principal component analysis. HAQ et al. [12] used the WDBC dataset in their research, and three methods of feature selection were applied, namely, principal component analysis, relief, and autoencoder algorithms. SVM was used as a classification model and applied to all results of feature selection methods to compare results. SVM with principal component analysis using only 18 features achieved the highest accuracy of 99%. HUANG and CHEN [13] used the variable Importance Measure (VIM) as a method for selecting a feature from the WDBC and WBC datasets. They developed a new model known as hierarchical clustering random Forest (HCRF), which is based on a DT and random forest. Three models were applied, namely, AdaBoost, DT and random forest. We then compared the results of these models on both datasets. As a result, the HCRF model obtained the highest accuracy in the WDBC dataset by 97.05% and 97.76% in the WBC dataset.

Jumanto et al. [14] used forward feature Selection and random forest for selecting features from the WDBC dataset. As a classifier, backpropagation ANN was used to predict whether breast cancer tumour is malignant or benign. The results showed that the classifier used had an accuracy of 98.3%.

Furthermore, we notice the previous works have suffered from limitations in the accuracy of the diagnosis, and this could be due to the bias that occurred during the training of the AI models as a result of the imbalance of the dataset, or it could be that the feature selection method and the AI model are not significantly proportional to this dataset. Therefore, these works need to improve the accuracy of early diagnosis, which in turn helps preserve the patient's life.

III. Methodology

In this study, we propose a methodology that uses the voting classifier method, which combines multiple models to produce the best prediction accuracy for the diagnosis of breast cancer. Sub section (C) explains the voting classifier in further detail. The proposed methodology consists of three main phases: pre-processing phase, feature selection phase and prediction phase (Fig. 1). Before we explain the main phases, we will describe the dataset used in this study.

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset obtained from the UCI ML repository is used in this paper. The University of Wisconsin originally provided a dataset containing two classes: malignant (M) and benign (B). It comprised 569 samples (B=357 and M=212) and 32 features. These features display the fundamental properties of the breast cell. Two of these features are not used on the practical side (id, Unnamed:32). The diagnosis field uses the remaining 30 features that contain a real value [15]. In the proposed methodology, we will develop a method that combines two filter methods (feature selection based on Pearson correlation and feature selection based on mutual information), which in turn reduces the number of features from 30 to 18 for increasing classification accuracy, which will be explained later.

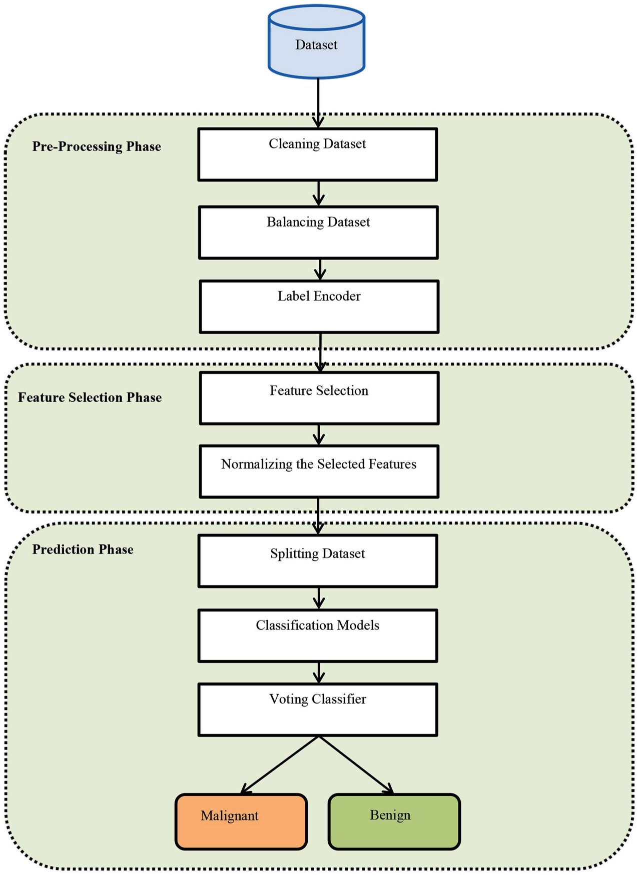


Fig. 1. Proposed methodology.

A. *Pre-processing Phase*

At this phase, we perform a set of initial operations on the dataset to improve the quality of the data and ensure that the classification model works well. The main operations in this phase are cleaning, balancing and label encoding for the dataset.

1) Cleaning Dataset: The first process focuses on cleaning the dataset, which involves identifying data errors and then editing, updating or removing data for overcoming errors, where we filter the data for the next stage. The cleaning

dataset performs two main processes as follows. Firstly, the number of features that are actually used is only 30. As the dataset consists of 32 features, it involves two unimportant features: ‘id’, and ‘Unnamed:32’, where ‘id’ is simply an identifier and ‘Unnamed:32’ is a column whose rows are all empty values, so we will drop this feature. Secondly, when most of the values for each column or row are missing, we drop that row or column to ensure the quality and correctness of the data. In another case, if some column or row values are missing, the mean will be calculated to restore data.

2) Balancing Dataset: The importance of a balanced dataset for a model is to generate higher accuracy models devoid of bias. Thus, a balanced dataset is important for a classification model. An uneven class distribution of the dataset may cause trouble in later phases of training and classification as classifiers will have very fewer data to learn features of a particular class. SMOTE is one of the best techniques used to balance the dataset. Unlike normal upsampling, SMOTE makes use of the nearest neighbour algorithm to generate new and synthetic data that can be used to train the models. It will generate new data points for the minority class (in this case, for class M) to balance the dataset where SMOTE gives the minority class an increased likelihood of being successfully learned. Fig. 2 shows how to create new data by SMOTE[16].

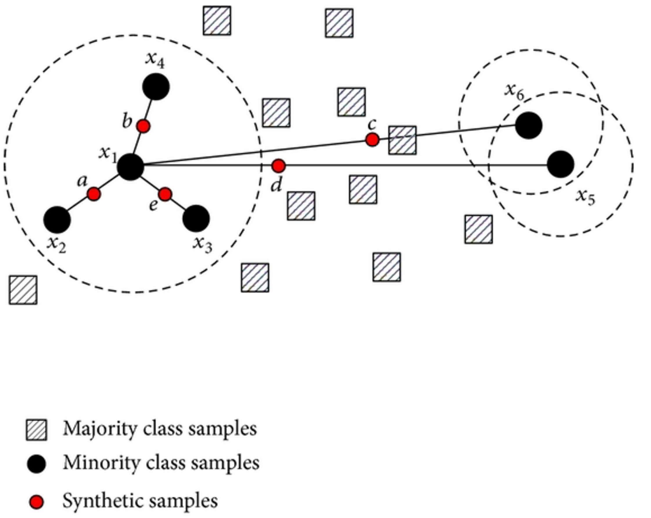
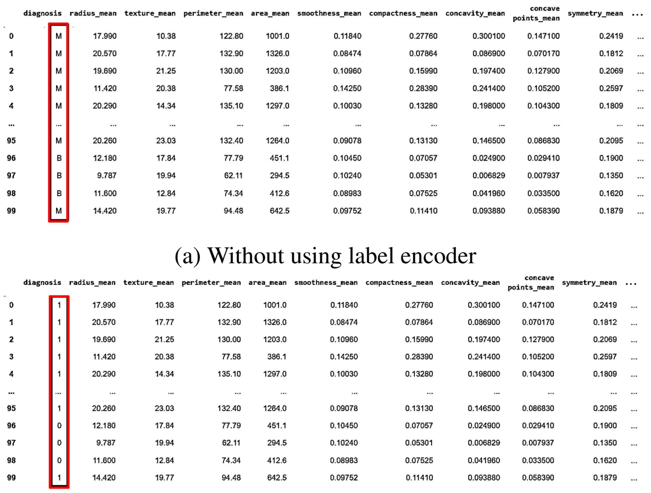


Fig. 2. Smote technique [16].

Fig. 2 shows two classes in the dataset: minority and majority. The SMOTE technique works by using the nearest neighbour algorithm to create new data points for the minority class located on the line connecting two data points of the same class represented by (a, b, c, d, e). The main benefit of this process is the elimination of innate inclinations to favour and overfit toward the majority classes due to the disparity in samples’ proportions of minority and majority classes. Finally, SMOTE balances the dataset between majority and minority classes.

3) Label Encoder: In this stage, after performing the balancing of the dataset, we will encode the target class ‘diagnosis’ via transformation (Malignant to 1 and Benign to 0). In classification analysis, the dependent variable is usually affected by qualitative factors and ratio scale variables. Hence, these category variables must be encoded into numerical values using encoding techniques because ML algorithms only accept numerical inputs [17]. Fig. 3 shows the result of the label encoder on the diagnosis field in the dataset.



	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	...
0	M	17.990	10.38	122.80	1001.0	0.11840	0.27760	0.300100	0.147100	0.2419	...
1	M	20.570	17.77	132.90	1326.0	0.08474	0.07864	0.086900	0.070170	0.1812	...
2	M	19.690	21.25	130.00	1203.0	0.10960	0.15990	0.197400	0.127900	0.2069	...
3	M	11.420	20.38	77.58	386.1	0.14250	0.28390	0.241400	0.105200	0.2597	...
4	M	20.290	14.34	135.10	1297.0	0.10030	0.13280	0.198000	0.104300	0.1809	...
...	...	...	...	...	...	...	...	...	...	...	...
95	M	20.260	23.03	132.40	1264.0	0.09078	0.13130	0.146500	0.086830	0.2095	...
96	B	12.180	17.84	77.79	451.1	0.10450	0.07057	0.024900	0.029410	0.1900	...
97	B	9.787	19.94	62.11	294.5	0.10240	0.05301	0.006829	0.007937	0.1350	...
98	B	11.600	12.84	74.34	412.6	0.08983	0.07525	0.041960	0.033500	0.1620	...
99	M	14.420	19.77	94.48	642.5	0.09752	0.11410	0.093880	0.058390	0.1879	...

(a) Without using label encoder

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	...
0	1	17.990	10.38	122.80	1001.0	0.11840	0.27760	0.300100	0.147100	0.2419	...
1	1	20.570	17.77	132.90	1326.0	0.08474	0.07864	0.086900	0.070170	0.1812	...
2	1	19.690	21.25	130.00	1203.0	0.10960	0.15990	0.197400	0.127900	0.2069	...
3	1	11.420	20.38	77.58	386.1	0.14250	0.28390	0.241400	0.105200	0.2597	...
4	1	20.290	14.34	135.10	1297.0	0.10030	0.13280	0.198000	0.104300	0.1809	...
...	...	...	...	...	...	...	...	...	...	...	...
95	1	20.260	23.03	132.40	1264.0	0.09078	0.13130	0.146500	0.086830	0.2095	...
96	0	12.180	17.84	77.79	451.1	0.10450	0.07057	0.024900	0.029410	0.1900	...
97	0	9.787	19.94	62.11	294.5	0.10240	0.05301	0.006829	0.007937	0.1350	...
98	0	11.600	12.84	74.34	412.6	0.08983	0.07525	0.041960	0.033500	0.1620	...
99	1	14.420	19.77	94.48	642.5	0.09752	0.11410	0.093880	0.058390	0.1879	...

(b) Using label encoder

Fig. 3. Label encoder on the dataset

B. Feature Selection Phase

In the beginning, and before choosing the model that fits with our dataset, we should choose the appropriate features that our model will train on to yield the best results. Less redundant data means greater modeling accuracy, less misleading data means fewer opportunities for decisions based on noise and less data equals faster algorithms. As a result, the main objective of feature selection is to improve accuracy, reduce training time and decrease over-fitting [18]. In this phase, we present a proposed method that combines two methods from the filter method, which is correlation analysis using Pearson correlation and mutual information. In the first stage, we analyse the relationships in the dataset by finding the correlation matrix that uses Pearson correlation as a measure and then we collect the highly correlated features that contain common elements in one set. Our processing keeps the common feature with the highest value mutual information and drops the rest of the features in each group.

1) Correlation Analysis Based on Pearson Correlation: Pearson Correlation (PC) is a measure of the degree of rela-

tionship (correlation) between features. This scale measures the degree of correlation between all features, where the value of the relationship ranges between [-1,1]:

Score (1): This value indicates that the correlation between the two features is completely directly proportional.

Score (0): This value denotes the absence of correlation between the two features.

Score (-1): This value indicates that the correlation between the two features is inversely proportional.

The PC coefficient between two features can be measured through (1) [19].

$$r=\frac{[N\sum_{i=1}^{N}x_i.y_i-\sum_{i=1}^{N}x_i.\sum_{i=1}^{N}y_i]}{\sqrt{[N\sum_{i=1}^{N}x_i^2-(\sum_{i=1}^{N}x_i)^2].[N\sum_{i=1}^{N}y_i^2-(\sum_{i=1}^{N}y_i)^2]}} \quad (1)$$

Fig. 4 displays the heat map of Pearson correlation scores between WDBC features.

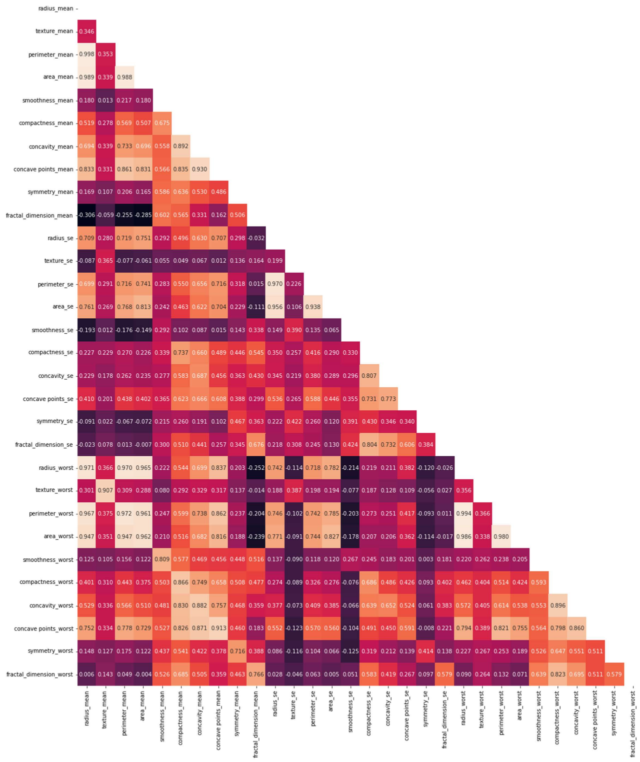


Fig. 4. Heat map of correlations between WDBC features.

2) Feature Selection Based on Mutual Information: Mutual Information (MI) is a measure of the dependency between each feature and the target class. The importance of the current measure is to find the best features that are closely related to the goal. The resulting value ranges between [0,1], where value (0) represents independent features, and value (1) refers to dependent features. The MI value is computed by (2) [20]:

$$I(X,Y)=\sum_{i=1}^{n}\sum_{j=1}^{m}P(X_i,Y_i).\log[\frac{P(X_i|Y_i)}{P(X_i)}] \quad (2)$$

After applying (2) to the dataset used in this study, Table I shows each feature and its value in descending order.

TABLE I.
MUTUAL INFORMATION VALUE FOR EACH FEATURE

Feature	value
perimeter_ worst	0.499442
area_ worst	0.490916
radius_ worst	0.483495
concave points_ mean	0.474548
concave points_ worst	0.471331
perimeter_ mean	0.427724
concavity_ mean	0.421748
radius_ mean	0.407724
area_ mean	0.405551
area_ se	0.366311
concavity_ worst	0.358751
perimeter_ se	0.284167
compactness_ worst	0.283761
radius_ se	0.277863
compactness_ mean	0.276540
concavity_ se	0.178546
concave points_ se	0.177446
texture_ mean	0.145510
texture_ worst	0.139945
smoothness_ worst	0.120389
compactness_ se	0.119047
smoothness_ mean	0.108997
symmetry_ worst	0.101290
fractal_dimension_worst	0.097076
symmetry_ mean	0.070514
fractal_dimension_ se	0.048942
symmetry_ se	0.027308
fractal_dimension_mean	0.023849
smoothness_ se	0.023746
texture_ se	0.002271

3) Feature Selection Based on Pearson Correlation and Mutual Information (PC-MI): The filtering method is considered the best and least complicated and costly way to select the feature, because this method selects the feature based on correlation analysis and is separate from the ML model used [21]. PC has been used to find the degree of relationship between one feature and another in the dataset, and using MI separately helps determine the degree of relationship of each feature to the target class. Therefore, these two methods will

combine first to obtain the highly interrelated features with each other and know which of these features has the highest degree of correlation with the target class. We create a new filter method that combines PC and MI. This method is called feature selection based on PC and MI (PC-MI).

Firstly, this method finds the features that have a PC value greater than or equal to 0.89 by analysing the correlation heat map of features shown in Fig. 4. Secondly, it works to merge the set of interrelated features that contain common features into one group. Thirdly, one common feature is chosen from each group, which obtains the highest MI value. Finally, the remaining features in each group are dropped from the dataset (Table II). Fig. 5 shows the proposed method of our work.

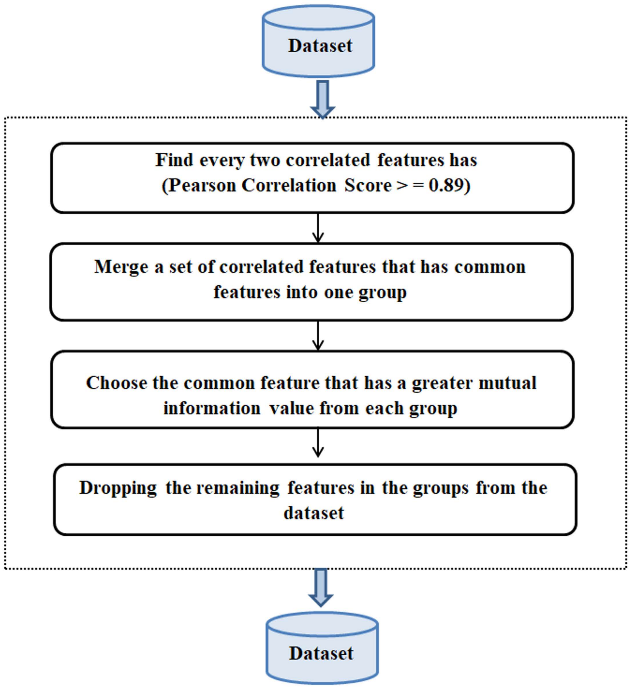

Dataset
Find every two correlated features has (Pearson Correlation Score > = 0.89)
Merge a set of correlated features that has common features into one group
Choose the common feature that has a greater mutual information value from each group
Dropping the remaining features in the groups from the dataset
Dataset

Fig. 5. Proposed feature selection method.

The result of the proposed method for selecting the feature is to drop 12 features from the dataset; therefore, the remaining features that will be used are only 18. We remove unimportant features that hinder the work of the ML model and keep the features that help the model learn correctly and give the best classification accuracy.

TABLE II.
FEATURES DROPPED FROM THE DATASET

No	Features	No	Features
1	*area_worst*	7	*compactness_mean*
2	*radius_worst*	8	*concavity_mean*
3	*perimeter_mean*	9	*concave points_worst*
4	*radius_mean*	10	*radius_se*
5	*area_mean*	11	*perimeter_se*
6	*texture_mean*	12	*compactness_worst*

We note from previous table that 12 features are dropped from the dataset, where five features are dropped from the first group, one feature is dropped from the second group, three features are dropped from the third group, two features are dropped from the fourth group and one feature is dropped from the fifth group. As a result, we have the 18 best features.

Table III shows that five groups are found. Each group contains pairs of features that are strongly connected to each other, and all these pairs that belong to one group have common features. One feature from each group with the highest MI value is selected as the highest feature with a strong correlation with the target class.

4) Normalizing the Selected Features: After selecting the best features from the dataset through the proposed method for feature selection, we normalise the remaining features using StandardScaler. The main objective of StandardScaler is to convert feature values into standard units free from the influence of the arithmetic mean and dispersion, where the resulting values are free from the units of measurement. It can be computed from eq. 3 [22]:

$$Z = [X - \bar{X}]/S \tag{3}$$

where: • Z: StandardScaler Score
- X: Sample
- $\bar{X}$: Arithmetic mean
- S: Standard deviation

C. Prediction Phase

After the pre-processing of the dataset and the selection of the appropriate features, the dataset is ready to work with the ML model for making predictions. Therefore, in this section, we explain the mechanism for dividing the dataset, the ML models used and the proposed model that will be used in the prediction process.

1) Splitting Dataset: The dataset will be split into two parts. The first part is the training, which is a set of data used in training and building the model. The second part is the testing, which is a set of data in which the performance of the model is tested using a specific scale. In this paper, two methods of splitting are used as follows:

TABLE III.
PC-MI METHOD FOR FEATURE SELECTION

groups	Correlated features	Pearson Correlation Score	chosen feature of a high mutual information value
1	[radius_mean,area_worst]	0.947	perimeter_ worst
	[radius_mean,perimeter_worst]	0.967	
	[radius_mean,radius_worst]	0.971	
	[radius_mean,area_mean]	0.989	
	[radius_mean,perimeter_mean]	0.998	
	[area_mean,area_worst]	0.962	
	[area_mean,perimeter_worst]	0.961	
	[area_mean,radius_worst]	0.965	
	[radius_worst,area_worst]	0.986	
	[radius_worst,perimeter_worst]	0.994	
	[perimeter_worst,area_worst]	0.980	
	[perimeter_mean,area_worst]	0.947	
	[perimeter_mean,perimeter_worst]	0.972	
	[perimeter_mean,radius_worst]	0.970	
	[perimeter_mean,area_mean]	0.988	
2	[texture_mean,texture_worst]	0.907	texture_ worst
3	[compactness_mean,concavity_mean]	0.892	concave points_ mean
	[concavity_mean,concave points_mean]	0.930	
	[concave points_mean,concave points_worst]	0.913	
4	[radius_se,area_se]	0.956	area_ se
	[radius_se,perimeter_se]	0.970	
	[perimeter_se,area_se]	0.938	
5	[compactness_worst,concavity_worst]	0.896	concavity _ worst

•train–test–split (training=0.8,testing=0.2) k–fold cross validation (k=10)

2) Classification Models: In this part, the ML models that are used in this study will be explained and clarified.

Logistic Regression: A statistical model known as LR uses a qualitative dependent variable that can only use discrete values to represent the connection between two independent variables. It is used to investigate the influence of predictor variables on categorical outcomes. In an epidemiologic study, logistic models are frequently used to analyse the connections between risk factors and the development of the disease. In medical publications that do not specialize in epidemiology and public health, these models are often utilised [23].

Support Vector Machine: When learning the parameters of the SVM model during the training phase, SVM, one of the most significant and potent ML models, needs access to all of the training data. Support vectors, a subset of these training examples, are the only ones on which SVM relies to make predictions in the future. The hyperplanes' margins are determined by support vectors. Finding the greatest number of hyperplanes that may be used to divide two classes is the major goal of the training phase. When an issue is not linearly separable in the input space, a kernel can transfer the data into a higher-dimensional space called kernel space, where the data will be linearly separable. Linear hyperplane can be obtained in the kernel space to divide the several classes involved in the classification job. This approach is appealing because, compared with learning a nonlinear surface, the cost of moving to kernel space is minimal [24].

Decision Tree: A DT is one of the most important models in decision-making processes, as it is widely used in the field of ML. The trees are built from top to bottom, and nodes of these trees representing features are selected based on a certain scale (information gain in this study). In each node of the tree, a specific decision is made, and this decision directs you to another level of the tree until the root node, which is the source of the decision, is reached[25].

Voting Classifier: It is a type of ensemble classifier that depends on AI models, where it works to combine a certain set of models to produce one model that carries the strength of the models that have been combined, which gives the best prediction accuracy [26]. Here, we use a soft voting classifier and input three ML models (LR, SVM and DT), which are considered the best models that work with a voting classifier on this dataset based on a set of experiments. This classifier works on a probabilistic basis, as each of the input models of the classifier produces a probability value for class 0 and class 1. In the final result, the soft voting classifier uses the highest probability rate of all the input models, as shown in Fig.6. Finally, we can summarize the proposed methodology as the following. Firstly, we carry out some preliminary treatments for improving the dataset. Secondly, the best features are selected through the use of the proposed method (PC-MI) to be used by the proposed model (soft voting classifier) to give the best classification fit of the tumour type, whether it is benign or malignant.

IV. Results and discussion

In this section, the performance results based on the proposed methodology are shown and discussed in terms of the F1 score, precision, recall, accuracy, AUC, and ROC curves. We conduct three experiments, where the first experiment includes comparing the performance of the soft voting classifier with the models included (LR, SVM and DT) separately. The second experiment includes displaying the results of the soft voting classifier and comparing them with the previous work, as both experiments use train–test–split as a way to split the dataset. In the third experiment, the dataset is split into 10-fold, and the results of the performance of the soft voting classifier are presented. In addition, we explain the place of our proposed methodology and the contribution made to the applied side of early diagnosis of breast cancer.

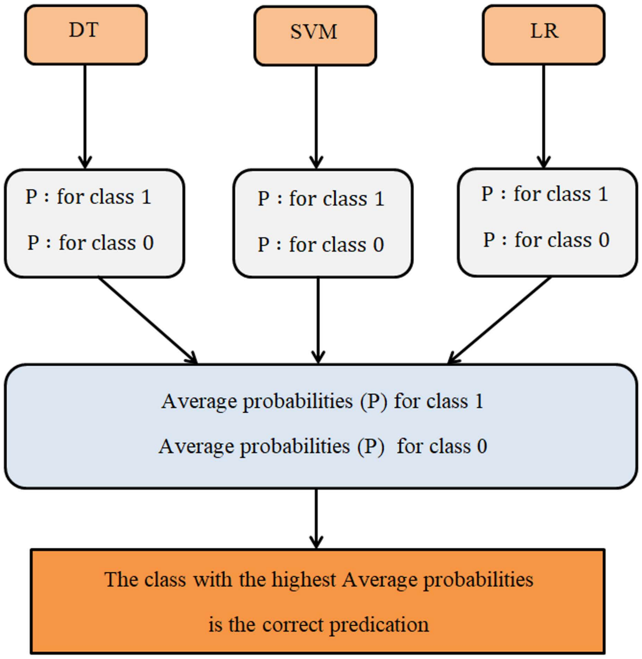

DT
SVM
LR
P : for class 1
P : for class 0
P : for class 1
P : for class 0
P : for class 1
P : for class 0
Average probabilities (P) for class 1
Average probabilities (P) for class 0
The class with the highest Average probabilities
is the correct predication

Fig. 6. The proposed soft voting classifier.

Experiment (1): In this experiment, we compare the performance results of the models used (LR, DT and SVM) with the proposed model (soft voting classifier) using the presented methodology, where train–test–split is adopted as a method for splitting the dataset. Table IV refers to the comparison results of these models.

TABLE IV.
Comparison between the performance of our work and machine learning models used

Model	Accuracy (%)	Precision	Recall	F1 Score
Logistic Regression (LR)	98.6%	0.9846	0.9846	0.9846
Decision Tree (DT)	96.5%	0.9412	0.9846	0.9624
Support Vector Machine (SVM)	96.5%	0.9839	0.9385	0.9606
Soft VotingClassifier [LR, DT, SVM]	**99.3%**	**1**	**0.9846**	**0.9922**

Table IV shows that the soft voting classifier obtains the highest degree of accuracy (99.3%), F1 score (0.9922), recall (0.9846), and precision (1) because the voting classifier depends on integrating the three models into one model that carries the strength of these combined models, which leads to the best prediction accuracy.

Figure 7 shows the ROC curves for the soft voting classifier with the models included in it (LR, SVM and DT).

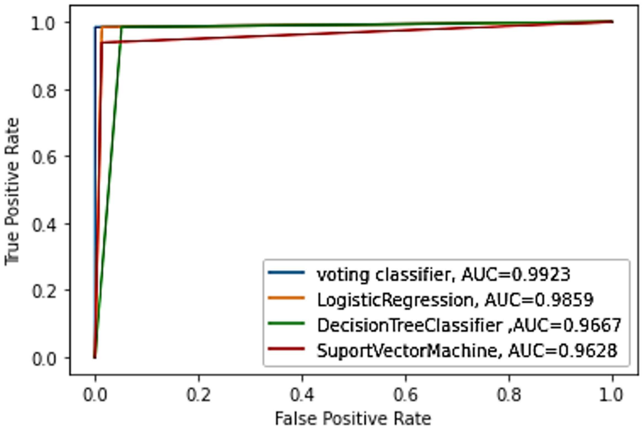


Fig. 7. ROC curves for models that used.

Experiment (2): In this experiment, we use the balanced dataset after selecting the best features through the proposed method PC-MI, where only 18 features are used. Train–test–split is used as a method for splitting the dataset, as the data are entered into a soft voting classifier to predict the type of tumour that may appear in some persons, whether it is benign or malignant. Fig. 8 displays the result of the performance of the soft voting classifier based on the important performance scaling factors.

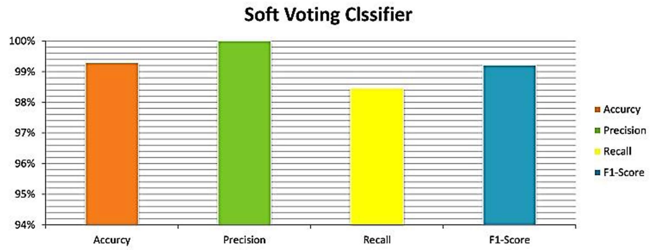


Fig. 8. Performance results of soft voting classifier.

Experiment (3): In this experiment, we use 18 features obtained from the proposed feature selection method (PC-MI) where the dataset is split into 10-fold for training and testing. The training data are passed to the voting classifier, and our proposed model is evaluated by cross-validation. Our proposed model's soft voting classifier (LR, DT and SVM) exhibits 98.2% test accuracy.

Results on Applied Side: We have proposed a methodology that can create an applied health system (web page) that helps many health institutions in the speed and accuracy of diagnosing the type of breast cancer tumour based on ML models. This method helps preserve the patient's life through early treatment and disposal of the tumour. This page is implemented using Spyder, which is a development environment that uses the Python language to create software applications.

Firstly, a sample of the mass in the breast is obtained, and this sample is analysed by a specialist called a pathologist. After that the values of the required features are extracted. The values of these features are entered into the Breast Cancer Tumor Diagnostic website, which is built based on our proposed model (soft voting classifier) that predicts whether the tumour is benign or malignant as shown in Fig. 9.

Table V presents a comparison between our proposed method and the related studies that use the feature selection process on a WDBC dataset where the table shows that our proposed method gives the highest degree of accuracy by 99.3, making it superior to all previous studies that we compared in the last years. This superiority in accuracy is because of the methods that we have used such as dataset balancing and selecting the feature proposed, as well as the proposed model (soft voting classifier).

V. Conclusion

Breast cancer should be detected early for effective treatment. Being one of the top causes of mortality in women, early diagnosis is crucial. The developed ML models enhance early breast cancer tumour prediction. However, false positive and false negative instances are important in medical research. Therefore, we focus not just on accuracy in our work but also on F1 score, precision, recall, AUC and ROC curve. In this work, feature selection is performed by developing a method that combines two filtering techniques, PC and mutual information (PC-MI), to select the best features before passing them to a classification model. The proposed model (soft voting classifier) is used to enhance the performance where it includes three models (LR, SVM and DT). A comparison is made between the performance of this models and the proposed model to prove the efficiency and strength of our proposed model in the prediction process. The proposed methodology outperforms previous work, achieving 99.3% accuracy, an F1 score of 0.9922, a recall of 0.9846, a precision of 1 and an AUC of 0.9923. Furthermore, the accuracy of 10-fold cross-validation is 98.2%. Finally, a web page is created using spyder and streamlit to make the proposed methodology workable from the practical side, thereby helping many health institutions in the speed and accuracy of diagnosing the type of breast cancer tumour. This study's future goals include using more feature selection techniques in conjunction with the WDBC dataset to improve breast cancer diagnosis. In addition, deep learning models will also be used for breast cancer detection.

Conflict of Interest

The authors have no conflict of relevant interest to this article.

References

[1] W. H. O. . WHO, “http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/,” *World Breast Cancer Rep.*, 2020.

[2] A. B. Nassif, M. A. Talib, Q. Nasir, Y. Afadar, and O. Elgendy, “Breast cancer detection using artificial intelligence techniques: A systematic literature review,” *Artificial Intelligence in Medicine*, vol. 127, p. 102276, 2022.

[3] A. Haleem, M. Javaid, and I. H. Khan, “Current status and applications of artificial intelligence (ai) in medical field: An overview,” *Current Medicine Research and Practice*, vol. 9, no. 6, pp. 231–237, 2019.

[4] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016.

[5] S. Guo, Y. Liu, R. Chen, X. Sun, and X. Wang, “Improved smote algorithm to deal with imbalanced activity classes in smart homes,” *Neural Processing Letters*, vol. 50, pp. 1503–1526, 2019.

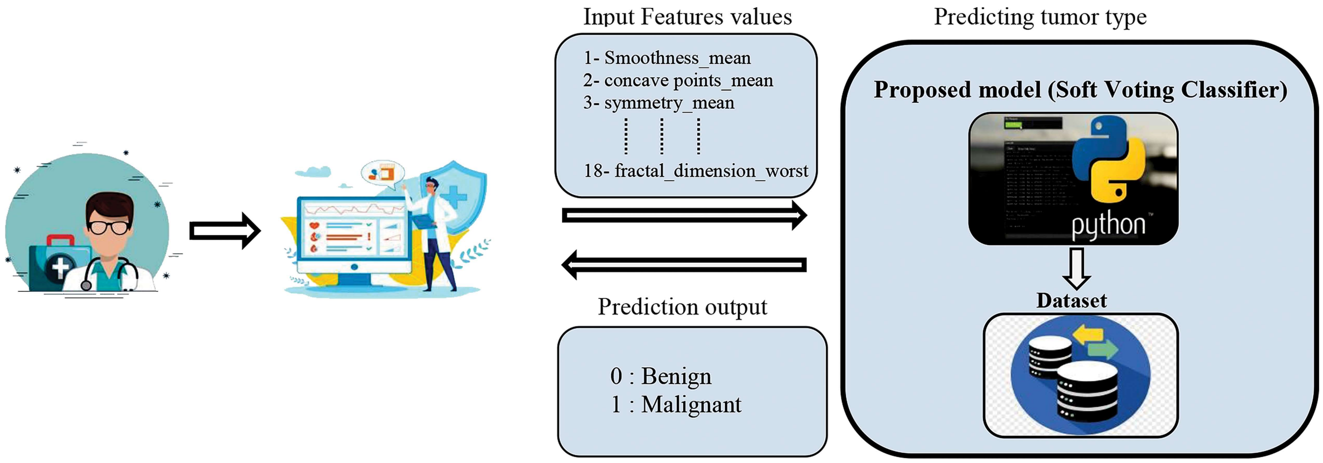


Fig. 9. Performance results of soft voting classifier.

TABLE V.
SIGNIFICANCE OF THE PROPOSED METHOD COMPARED WITH THE RELATED STUDIES

Authors	Year	Balanced the dataset	Practical Application	Feature selection method	Model	Accuracy (%)
Hazra [6]	2016	NO	NO	Pearson correlation coefficient	SVM	98.51
Khuriwal [7]	2018	NO	NO	Univariate feature selection (chi2)	VotingClassifier(ANN,LR)	98.50
Allam [8]	2018	NO	NO	Binary teaching learning-based optimisation (FS-BTLBO)	SVM	98.43
Memon [9]	2019	NO	NO	Recursive feature elimination (RFE)	SVM	99
Dhahri [10]	2019	NO	NO	Genetic algorithm	Adaboost classifier	98.24
Ibrahim [11]	2021	NO	NO	Correlation analysis and principal component analysis	Soft Voting	99.00
HAQ [12]	2021	NO	NO	Principal component analysis	SVM	97.45
HUANG [13]	2021	NO	NO	Variable importance measure method (VIM)	Hierarchical clustering random forest (HCRF)	97.05
Jumanto [14]	2022	NO	NO	Forward feature selection	ANN	98.3
Proposed Method	**2022**	**YES**	**YES**	**Pearson correlation and mutual information (PC-MI)**	**Soft voting classifier** **(LR, DT, SVM)**	**99.3**

[6] A. Hazra, "Study and analysis of breast cancer cell detection using naive bayes , svm study and analysis of breast cancer cell detection using naive bayes , svm and ensemble algorithms," *Int. J. Comput. Appl.*, vol. 145, no. January 2017, pp. 39–45, 2016.

[7] N. Khuriwal, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," *2018 IEEMA Eng. Infin. Conf.*, pp. 1–5, 2018.

[8] M. Allam and M. Nandhini, "Optimal feature selection using binary teaching learning based optimization algorithm," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 2, pp. 329–341, 2022.

[9] M. H. Memon, J. P. Li, A. U. Haq, M. H. Memon, and W. Zhou, "Breast cancer detection in the iot health environment using modified recursive feature selection," *wireless communications and mobile computing*, vol. 2019, pp. 1–19, 2019.

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

[11] S. Ibrahim and S. Nazir, "Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis," *J. Imaging*, vol. 225, pp. 1–7, 2021.

[12] A. U. Haq, J. P. Li, A. Saboor, J. Khan, S. Wali, S. Ahmad, A. Ali, G. A. Khan, and W. Zhou, "Detection of breast cancer through clinical data using supervised and unsupervised feature selection techniques," *IEEE Access*, vol. 9, pp. 22090–22105, 2021.

[13] 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, 2022.

[14] J. Jumanto, M. F. Mardiansyah, R. N. Pratama, M. F. Al Hakim, and B. Rawat, “Optimization of breast cancer classification using feature selection on neural network,” *Journal of Soft Computing Exploration*, vol. 3, no. 2, pp. 105–110, 2022.

[15] D. W. H. Wolberg, “https://archive.ics.uci.edu/ml/datasets/breast cancer wisconsin (diagnostic),” *M.L Repos.*, 1995.

[16] K. Teh, P. Armitage, S. Tesfaye, D. Selvarajah, and I. D. Wilkinson, “Imbalanced learning: Improving classification of diabetic neuropathy from magnetic resonance imaging,” *PloS one*, vol. 15, no. 12, p. e0243907, 2020.

[17] K. Potdar, “A comparative study of categorical variable encoding techniques for neural network classifiers,” *Int. J. Comput. Appl*, vol. 175, no. 4, pp. 7–9, 2017.

[18] Q. Al-Tashi, S. J. Abdulkadir, H. M. Rais, S. Mirjalili, and H. Alhussian, “Approaches to multi-objective feature selection: A systematic literature review,” *IEEE Access*, vol. 8, pp. 125076–125096, 2020.

[19] R. Saidi, W. Bouaguel, and N. Essoussi, “Hybrid feature selection method based on the genetic algorithm and pearson correlation coefficient,” *Machine learning paradigms: theory and application*, pp. 3–24, 2019.

[20] B. Gierlichs and E. Prouff, “Mutual information analysis: a comprehensive study mutual information analysis: a comprehensive study,” *J. Cryptol*, vol. 24, no. 2, pp. 269–291, 2011.

[21] A. Alonso-betanzos, “Filter methods for feature selection – a comparative study filter methods for feature selection . a comparative study,” *Int. Conf. Intell. Data Eng. Autom. Learn. Springer, Berlin, Heidelb.*, vol. 4881, no. December, pp. 178–187, 2007.

[22] P. Ferreira, D. C. Le, and N. Zincir-Heywood, “Exploring feature normalization and temporal information for machine learning based insider threat detection,” in *2019 15th International Conference on Network and Service Management (CNSM)*, pp. 1–7, IEEE, 2019.

[23] W. T. Ambrosius, *Topics in biostatistics*. Springer, 2007.

[24] G. H. Lewes, “Support vector machines for classification,” *Effic. Learn. Mach. Apress, Berkeley, CA*, no. January, pp. 39–66, 2015.

[25] L. Rokach and O. Maimon, “Decision trees,” *Data Min. Knowl. Discov. handbook. Springer, Boston, MA*, no. January, pp. 165–192, 2005.

[26] M. A. Khan, M. A. Khan Khattk, S. Latif, A. A. Shah, M. Ur Rehman, W. Boulila, M. Driss, and J. Ahmad, *Voting classifier-based intrusion detection for iot networks*. 2022.