Performance Analysis of a new Filter and Wrapper Sequence for the Survivability Prediction of Breast Cancer Patients

Original Scientific Paper

E. Jenifer Sweetlin
Research Scholar,
Centre for Information Technology and Engineering,
Manonmaniam Sundaranar University, Tirunelveli, India
jsweetlin@gmail.com

S. Saudia
Assistant Professor,
Centre for Information Technology and Engineering,
Manonmaniam Sundaranar University, Tirunelveli, India
saudiasubash@msuniv.ac.in

Abstract *– Feature selection is an essential preprocessing step for removing redundant or irrelevant features from multidimensional data to improve predictive performance. Currently, medical clinical datasets are increasingly large and multidimensional and not every feature helps in the necessary predictions. So, feature selection techniques are used to determine relevant feature set that can improve the performance of a learning algorithm. This study presents a performance analysis of a new filter and wrapper sequence involving the intersection of filter methods, Mutual Information and Chi-Square followed by one of the wrapper methods: Sequential Forward Selection and Sequential Backward Selection to obtain a more informative feature set for improved prediction of the survivability of breast cancer patients from the clinical breast cancer dataset, SEER. The improvement in performance due to this filter and wrapper sequence in terms of Accuracy, False Positive Rate, False Negative Rate and Area under the Receiver Operating Characteristics curve is tested using the Machine learning algorithms: Logistic Regression, K-Nearest Neighbour, Decision Tree, Random Forest, Support Vector Machine and Multilayer Perceptron. The performance analysis supports the Sequential Backward Selection of the new filter and wrapper sequence over Sequential Forward Selection for the SEER dataset.*

Keywords*: accuracy, filter-wrapper, Sequential forward selection, Sequential backward selection*

1. INTRODUCTION

Breast cancer is one of the most serious medical reasons associated with death of women on earth. The disease is caused by many factors such as age, obesity, alcoholism, lack of physical activity, menopausal status and family history of breast cancer [1]. Data Mining and Machine Learning (ML) techniques have been used to develop various breast cancer prediction models [2]. The digital form of clinical breast cancer datasets is available in huge volumes and are multidimensional. This multidimensional dataset contains many independent features with more missing values, and some of the features are irrelevant for analysis. Applying these datasets to ML based classifiers drastically reduces accuracy. So, finding the relevant and optimal feature set combination is more important for enhancing and improving the accuracy of ML based classifiers. The training phases of the ML model design includes data preprocessing, feature selection and feature extraction stages [3-5]. Medical datasets are mostly imbalanced, so to reduce the effect of skewed class distribution in the model, studies have focused on data balancing methods [6]. To build a more effective ML classifier, feature selection techniques are used to filter out and find more optimal features from multidimensional datasets with irrelevant independent features [7]. Feature selection focuses on selecting significant independent features to improve the ability of classifiers to discriminate between classes. Furthermore, feature selection reduces feature dimensionality and computational complexity during the training phase of an ML algorithm [8].

The feature selection algorithms are categorized as filter, wrapper and embedded methods depending on how they combine feature selection sequences while determining informative independent features. Ranking technique is the principal criterion of filter methods which use rank ordering for feature selection based on a statistical score. The method filters irrelevant features which degrade the relationship between independent and dependent features, thereby selecting the highly ranked independent features to be applied to the training phase of an ML algorithm. These filter approaches are independent to any ML algorithm, computationally quick and scalable. Some feature ranking based filter techniques are the Mutual Information (MI), Pearson Correlation Coefficient (PCC) and Chi-Square (CS) methods [9]. Compared to filter methods, the wrapper methods show better performance because the independent feature selection mainly depends on a classification algorithm. Wrapper methods determine the quality of different subsets of independent features which are more suited for the classifier. But if the dimensionality of the dataset is large, the wrapper methods are very expensive in terms of time and computational speed since each feature set considered must be evaluated with the ML classifiers used [10]. In embedded methods [11], both filter and wrapper methods are used for feature selection and a classifier is used to evaluate the quality of the selected subset of independent features.

Liou et al. [12] used a Genetic Algorithm (GA) based approach along with an Artificial Neural Network (ANN), Decision Tree (DT) and Logistic Regression (LR) on the Wisconsin Breast Cancer Database-Original (WBCD) dataset for predicting breast cancer. The feature selection step was carried out based on Information Gain (IG). It is indicated in the paper that the GA model yielded better results while classifying the breast cancer data with an accuracy of 98.78%. Saygili [13] studied and compared the diagnosis of cancer using six ML methods such as Support Vector Machine (SVM), K-Nearest Neighbour (K-NN), Naive Bayes (NB), DT, Random Forest (RF) and Multilayer Perceptron (MLP) on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Gain Ratio (GR) was used as a feature selection technique. The results showed that the RF performed better with an accuracy of 98.77%. Omondiagbe et al. [14] proposed an automated method for diagnosing breast cancer on the WDBC dataset using SVM, ANN and NB with Correlation Based Filters, Recursive Feature Elimination (RFE), Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) in the preprocessing stage. The proposed SVM-LDA and ANN-LDA approaches achieved an accuracy of 98.82%. Islam et al. [15] compared five supervised ML techniques, namely SVM, K-NN, RF, ANN and LR on the WBCD dataset. PCC was used to identify the relationship between the attributes. The results showed that ANNs performed well with the highest accuracy of 98.57%. Alickovic et al. [16] used two datasets, WDBC and WBCD to construct an automated breast cancer diagnosis system to select significant features from the datasets; GA was used as a feature selection technique. The authors compared the performance of classifiers with GA feature selection and without GA feature selection on data mining algorithms such as DT, LR, Bayesian Network (BN), RF, Radial Basis Function Networks (RBFN), MLP, SVM and Rotation Forest (RoF). GA with RoF achieved the highest classification accuracy of 99.48%.

Liu et al. [17] proposed predictive models for breast cancer survivability using the DT algorithm on an imbalanced Surveillance, Epidemiology, and End Results (SEER) dataset. In the feature selection stage, LR backward selection was implemented for data reduction and an undersampling method for data balancing. To increase the predictive performance of the classification, the bagging algorithm was implemented and the Area under curve (AUC) results obtained are 76.78%. Miri et al. [18] used the SEER dataset to predict the survivability of patients with breast cancer. To solve the class imbalance problem in the dataset, two oversampling methods such as Borderline-Synthetic Minority Oversampling Technique (SMOTE) and Density-based Synthetic Oversampling (DSO) were used. For feature selection a combination of Correlation-based Feature Selection (CFS) and Particle Swarm Optimisation (PSO) was used. DT, BN and LR were used as classifiers for prediction and an accuracy of 94.33% was achieved when DSO with CFS-PSO was used with DT. Aavula et al. [19] used MLP, DT, LR and SVM classifiers on SEER datasets to predict the survivability of patients with breast cancer. For feature selection, entropy and gain were used to select relevant features using Representative Feature Subset Selection (RFSS) algorithm on the classifiers. SVM-RFSS produced a higher accuracy of 96.78%. Zand et al. [20] predicted breast cancer survivability on the SEER breast cancer dataset by using three classification techniques such as NB, MLP and DT. IG was used to rank features. The DT produced a prediction accuracy of 86.7%. Manikandan et al. [21] used supervised classifiers such as DT, NB and ensemble learning techniques such as AdaBoost, XGBoost and Gradient Boosting classifier for the classification of breast cancer. To select the features Variance Threshold (VT) and PCA was used. The results showed that DT performed better with an accuracy of 98%. Simsek et al. [22] constructed a hybrid DM based methodology to differentiate the importance of variables for survival change over time for three different time periods: 1 year, 5 years and 10 years on the SEER dataset. Least Absolute Shrinkage and Selection Operator (LASSO) technique and GA were used for the independent feature selection. Since the data sets are unbalanced, to balance the number of living and deceased labels in the dataset, two resampling methods such as Random Under Sampling (RUS) and SMOTE were applied. ANN and LR, were applied along with ten-fold cross-validation technique to determine and evaluate the performance of the classification model. A performance analysis was conducted for each model to identify the importance of each variable in the model for time periods of 1 year, 5 years and 10 years and the accuracy obtained has a maximum value of 84%.

Based on the above literature review, it was observed that in the majority of the papers, various feature selection approaches have been explored to improve the accuracy of predicting breast cancer or the survivability of breast cancer patients using the WDBC, WBCD and SEER datasets. This paper proposes a new filter and wrapper sequence using filter methods such as MI, CS and wrapper methods: Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) for the clinical breast cancer dataset, SEER for the classification of the survivability of breast cancer patients as living and deceased with improved accuracy. A breast cancer patient at a later stage of the disease has to pass through mental and physical trauma when subjected to heavy dose chemotherapy or radiotherapy [23]. Upon determining the criticality of breast cancer for survivability, patients can be relieved from such trauma by deciding between second-line treatment or ending the treatment. A second-line palliative or hospice treatment can improve quality of life [24]. So, optimal independent features are identified from the SEER dataset in this paper for predicting the survivability of breast cancer patients as the intersection of independent features obtained from filter methods, MI, CS which are then applied as input to the wrapper method, SFS or SBS. The informative features obtained as output from the wrapper method are subjected to ML algorithms: Multiple Logistic Regression (MLR), K-NN, DT, RF, SVM and MLP. The results are compared using different ML classifiers based on Accuracy (ACC), False Positive Rate (FPR), False Negative Rate (FNR) and Area under the Receiver Operating Characteristics curve (AUC-ROC). The remainder of this paper is organized into four sections. In Section 2, the methodology for the new filter and wrapper sequence is discussed. The experimental results and discussion are presented in Section 3. Finally, Section 4 concludes the paper.

2. PROPOSED METHODOLOGY

The paper proposes a new filter and wrapper sequence to obtain an optimal informative set of independent features for the classification of survivability of breast cancer patients using the clinical breast cancer dataset, SEER and a comparative performance analysis of different ML models in terms of ACC, FPR, FNR and AUC-ROC. The workflow of the proposed filter and wrapper sequence is shown in Fig.1 and discussed in the following subsections.

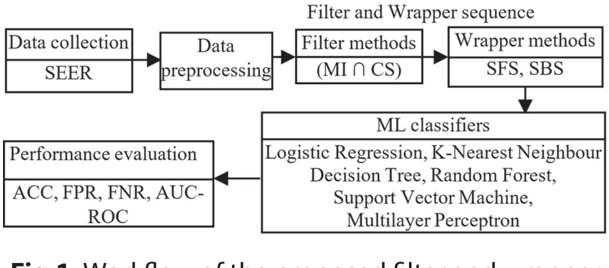

Fig.1. Workflow of the proposed filter and wrapper sequence

2.1. DATA COLLECTION

The SEER dataset-Surveillance, Epidemiology, and End Results is taken from the website, www.seer.cancer.gov. The dataset is an authentic source for cancer statistics updated every year by the Surveillance Research Program (SRP) of Cancer Control and Population Sciences (DCCPS), a division of the National Cancer Institute (NCI), USA. There are 7,755,157 records and 258 features in the dataset collected between the years, 1975 and 2018. The records which have breast cancer information are taken for analysis and so after removing all other types of cancer records only 1,073,477 breast cancer records with 20 features are considered for the later stages of data preprocessing, feature selection, training and testing. The independent features selected from SEER for the classification of survivability of breast cancer patients are age, sex, grade, primary tumour laterality, summary stage, surgery, tumour size, nodes examined, nodes positive, oestrogen receptor (ER) status, progesterone (PR) status, human epidermal growth factor receptor 2 (HER2) status, and the dependent feature is the survival status with labels, living and deceased.

2.2. DATA PREPROCESSING

The datasets collected are preprocessed through steps such as data transformation, data cleaning and data balancing [25]. In the data transformation, the downloaded SEER dataset is transformed into a .csv format as required for implementation in Python. In the data cleaning step, missing records are removed to improve the information content of the data. After cleaning the data, 10,838 data records with 12 features are obtained. The dataset has both categorical and numerical features. The nominal categorical feature values are converted into numerical values by a one-hot encoding technique [26] to work in the Python scikit-learn library. After applying one-hot encoding, the SEER dataset produces 10,838 records with 22 features. The SEER dataset is highly imbalanced. The dataset must be balanced before it is applied to classifiers. In the data balancing stage, the SMOTE-Edited Nearest Neighbour (SMOTE-ENN) [27] technique is applied to handle the imbalance in the training set. SMOTE-ENN is a combination of SMOTE and ENN where SMOTE is an oversampling technique that generates synthetic data of minority samples according to their nearest neighbours. ENN performs data cleaning. After data balancing, 8,769 records corresponding to the majority class, living and 10,233 records corresponding to the minority class, deceased are obtained for the SEER dataset. These records are then subjected to the new filter and wrapper sequence.

2.3. FILTER AND WRAPPER SEQUENCE

The data records obtained after data balancing are subjected to the new filter and wrapper sequence where the optimal independent features are identified with the filter methods: MI, CS and the wrapper methods: SFS or SBS for the classification of the survivability of breast cancer patients as living and deceased. The intersection of independent features obtained from MI and CS is then applied individually to the SFS and SBS wrapper methods. The filter and wrapper methods are briefed in the following subsections.

2.3.1 Filter Methods

The filter methods used in the new filter and wrapper sequence are Mutual Information and Chi-Square.

Mutual Information

Mutual Information is a filtering method which determines the correlation between independent and dependent features. If each of the n independent features is defined as S_i and the dependent feature as T, then the formula to calculate the Mutual Information, MI (S_i, T) between S_i and T is defined in equation 1.

$$MI(S_i,T) = H(S_i) + H(T) - H(S_i|T) \qquad (1)$$

where $1 \le i \le n$, $H(S_i)$ is the entropy of independent feature S_i, $H(T)$ is the entropy of dependent class T, and $H(S_i|T)$ is the conditional entropy of S_i and T. The higher values of MI specifies that the independent feature, S_i contains more information for classification [28]. Therefore, k_1 number of independent features whose MI values are greater than 0 are selected to take part in the successive stages of the filter wrapper sequence. The MI values of the independent features obtained by applying the MI filter in this work are listed in Table 1.

Table 1. Independent features selected after applying the MI filter

S.No	Independent features	MI values
1.	Summary stage-localized	0.1046
2.	Age	0.1004
3.	Tumour size	0.0924
4.	ER status-positive	0.0891
5.	PR status-positive	0.0773
6.	Grade-2	0.0722
7.	Laterality-right	0.0634
8.	Nodes positive	0.0554
9.	Grade-1	0.0548
10.	Nodes examined	0.0363
11.	HER2 status-positive	0.0250
12.	HER2 status-negative	0.0165
13.	Laterality-left	0.0135
14.	Surgery-surgery performed	0.0104
15.	Grade-3	0.0096
16.	Sex-female	0.0089
17.	Summary stage-regional	0.0076
18.	ER status-negative	0.0034
19.	Summary stage-distant	0.0008

Chi-Square

Chi-Square is a statistical method which evaluates the independence of the features in a dataset. In this technique, the independence of two events namely the occurrence of independent features and the occurrence of dependent features are evaluated [29]. The equation to calculate the chi-square value, χ^2 for each independent feature is defined in equation 2.

$$CS = \sum_{f=1}^{p} \frac{(O_f - E_f)^2}{E_f} \qquad (2)$$

where O_f is the frequency of the different possible p combinations of S_i independent feature values and T dependent feature values and $1 \le f \le p$. E_f is the expected frequency of association between the f^{th} combination of independent feature values and dependent feature values as defined in equation 3. Here Tr_{S_i} is the sum of the records corresponding to each value of the ith independent feature, S_i under consideration, Tc_T is the sum of the records corresponding to each value of the dependent feature, T and m is the total number of records in the training set. When the CS score is higher than the chi-square value, χ^2, determined from the chi-square distribution table, corresponding to the degrees of freedom, dof, the features are highly related. B_{S_i} in the independent feature, S_i and the number of values in the dependent feature B_T. The dof is calculated as the product of B_{S_i} and B_T as in equation 4.

$$E_f = \frac{Tr_{S_i} \times Tc_T}{m} \qquad (3)$$

$$dof = \left(B_{S_i} - 1\right) \times \left(B_T - 1\right) \qquad (4)$$

If the CS score is lower than the χ^2 score, the features are less correlated. The independent features with low CS scores are not included when modelling the classifier. The CS values of the independent features obtained after applying the CS filter in the proposed work are listed in Table 2.

Table 2. Independent features selected after applyingthe CS filter

S.No	Independent features	CS values
1.	Tumour size	28640.032
2.	Nodes positive	8185.748
3.	Age	5188.570
4.	Nodes examined	3048.862
5.	Grade-2	1407.698
6.	Summary stage-localized	1329.466
7.	Grade-1	1255.024
8.	Laterality-right	1194.877
9.	PR status-positive	1030.023
10.	ER status-positive	717.781
11.	HER2 status-positive	638.850
12.	Laterality-left	231.628
13.	Summary stage-regional	204.313

S.No	Independent features	CS values
14.	HER2 status-negative	82.197
15.	Grade-3	70.376
16.	Sex-male	39.709
17.	ER status-negative	25.409
18.	PR status-negative	17.142
19.	Summary stage-distant	4.046
20.	Sex-female	2.078
21.	Surgery-not performed	1.384
22.	Surgery-surgery performed	0.922

The intersection of independent features obtained from both filter methods, *MI* and *CS* is used to find the optimal set of features. The independent features, Summary stage-localized, ER status-positive, Nodes examined, Laterality-right, HER2 status-negative, HER2 status-positive, Tumour size, Surgery-surgery performed, Grade-1, Laterality-left, Age, Summary stage-regional, ER status-negative, PR status-positive, Grade-2, Summary stage-distant, Sex-female, Grade-3 and Nodes positive obtained are subjected as inputs to the wrapper methods: SFS or SBS to find a more optimal set of independent features from the SEER training set for the classification of the survivability of breast cancer patients.

2.3.2 Wrapper Methods

The wrapper methods used in this new filter and wrapper sequence are SFS and SBS. SFS is a greedy search algorithm that starts with an empty feature set and adds one independent feature at a time to determine the performance of the classifier until a desired number of independent features are obtained in the independent feature subset. It stops adding independent features when no improvement in classification performance is observed, or all features are added to the model [30]. In this work, the DT classifier is used as an estimator to select 15 independent features and the average accuracy score achieved after 15 iterations is 97.92% as shown in Fig. 2 (a). The selected independent features are listed in Table 3. SBS which is also known as Sequential Backward Elimination, works just the opposite to SFS. It starts with a full set of independent features in the training set and eliminates the least significant independent feature in each iteration until the classification performance does not change further. This method works best with a large number of independent features in the training set [30]. The estimator used in this technique is the DT classifier and the average accuracy score obtained is 98.23% as shown in Fig. 2(b). The selected independent features are listed in Table 3.

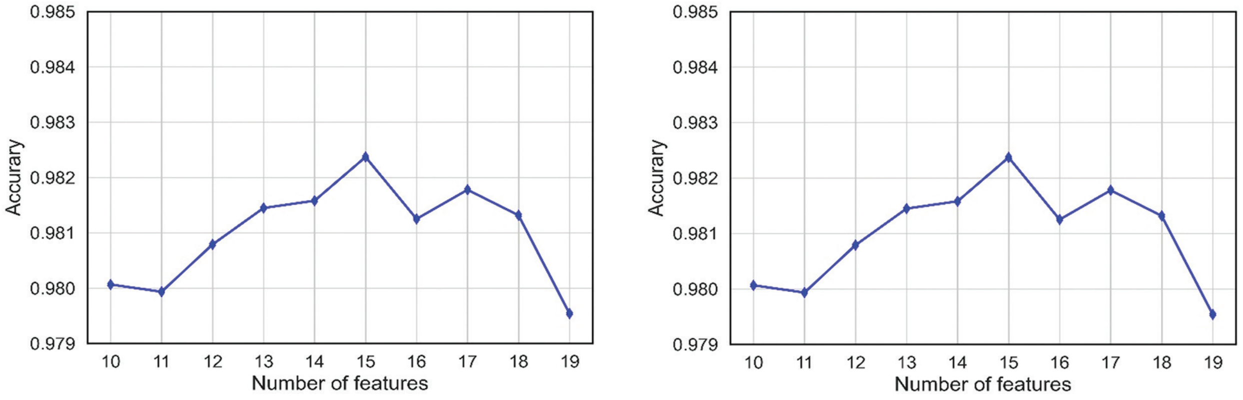

Fig. 2. Line plot between independent features and the accuracy scores a) SFS b) SBS

Table. 3. Independent Features obtained after applying wrapper methods, SFS and SBS

SFS	SBS
Summary stage-localized	Summary stage-localized
Nodes examined	ER status-positive
Laterality-right	Nodes examined
HER2 status-negative	Laterality-right
HER2 status-positive	Tumour size
Tumour size	Grade-1
Surgery-surgery performed	Laterality-left
Grade-1	Age
Laterality-left	Summary stage-regional
Age	ER status-negative
Summary stage-regional	Grade-2
ER status-negative	Summary stage-distant
PR status-positive	Sex-female
Summary stage-distant	Grade-3
Sex-female	Nodes positive

The main objective of the new filter and wrapper sequence is to remove irrelevant independent features from the training set and reduce the dimension of the training set. From the independent feature subset obtained from the wrapper methods, SFS and SBS, the first 10 and 15 independent features are selected and subjected to ML algorithms to analyse the performance of classifiers. The independent features that are distinctly identified by SFS are HER2 status-positive, HER2 status-negative, Surgery-surgery performed and PR status-positive and by SBS are ER status-positive, Grade-2, Grade-3 and Nodes positive.

2.4. CLASSIFIERS

Classification is a supervised learning algorithm that identifies the value of dependent feature of a given independent test data record based on the classifier model produced from a labelled training set. In binary classification, the classifier obtained after the training phase predicts one of the two values of the dependent feature. Six ML based binary classifiers such as MLR, K-NN, DT, RF, SVM and MLP are used to analyse the performance of the new filter and wrapper sequence for predicting the survivability of breast cancer patients as living and deceased. Multiple Logistic Regression (MLR) [31] predicts the probability of a dependent feature using a logistic function. To identify the value of a dependent feature, the threshold value set corresponding to the different independent features is determined during the training phase of the algorithm. The test record whose values lie above the threshold value set falls into one class and the test record whose values which lie below falls into another class. K-Nearest Neighbour (K-NN) [32] identifies the class of a test record by using the dependent feature values of 'K' neighbours nearest to the test record under consideration. The K-nearest neighbours are identified using the distance measures such as Euclidean, Manhattan, Minkowski or Hamming distance. Decision Tree (DT) [33] determines the relationship between independent and dependent features in the form of a tree like structure based on measurements of information content in the independent features. The branches of the DT represent a decision rule set for identifying the class of the test record. Random Forest (RF) [34] is a top-down approach in which a number of decision trees are obtained using various subsets of the training set and the ensemble of their results is predicted as the dependent feature value of the incoming test records. The more the number of decision trees in the forest, the greater the accuracy. Support Vector Machine (SVM) [35] segregates the n-dimensional space of the independent features of the training set by an optimal hyperplane which lays the maximum distance between the support vectors on either side. The parameters of this optimal hyperplane help classify the incoming test records. Multilayer Perceptron (MLP) [36] is a feed forward neural network where the weights of the links connecting the input and the hidden layer, hidden and output layer are optimised during back propagation-based training to obtain an optimised weight vector which can predict the dependent feature value of the test record. The performance of these ML classifiers is evaluated using the objective metrics namely ACC, FPR, FNR and AUC-ROC. The results are compared and discussed in the next section.

3. EXPERIMENTAL RESULTS AND DISCUSSION

In the proposed new filter and wrapper sequence, the SEER clinical breast cancer dataset is used in the experimental analysis. The training set is subjected to the data preprocessing stage and the new filter and wrapper sequence. In the data preprocessing step, data transformation, data cleaning and data balancing steps are carried out. The filter methods, MI and CS are applied to the preprocessed training sets. The intersection of more relevant independent features identified from the filter stage are subjected to the SFS and SBS wrapper methods to obtain a more optimal set of independent features. The optimal set of 10 and 15 independent features are applied to the ML algorithms and the performance of the analysis are evaluated in this section in terms of the evaluation metrics: ACC, FPR, FNR and AUC-ROC. These evaluation metrics are defined in the following subsection.

3.1. EVALUATION METRICS

The performance evaluation of the classifiers is mainly based on the correct and incorrect predictions made by the model. The confusion matrix provides more insight into the performance of a prediction model and also identifies the classes which are correctly and incorrectly predicted by the model. Accuracy provides the ratio of the number of correct predictions to the total number of predictions made by the model. False Positive Rate refers to the number of predictions where the classifier incorrectly predicted the deceased class as living [18]. False Negative Rate refers to the number of predictions where the classifier incorrectly predicted the living class as deceased [15]. AUC-ROC curve is drawn by plotting the FPR and TPR values. The curve is plotted for different probability thresholds of the models while predicting the probability of different classes [18]. The ROC curve corresponding to the largest area has a better ability to classify between living and deceased classes.

3.2. RESULTS

The values of different evaluation metrics obtained from the testing stage are tabulated in Table 4. The results obtained show that, when the number of independent features identified from the new filter wrapper sequence is 15, SVM produced an accuracy of 99% and DT, an accuracy of 98.1% from SFS. Similarly, when the selected independent features are 10, DT has an accuracy of 85.9%. In the case of the SBS wrapper technique, SVM yields the highest accuracy of 99.5% and K-NN produced an accuracy of 98.7% when 15 selected features are used. When 10 independent features are selected, DT obtained an accuracy of 86.3%. SBS performed better than SFS across all ML algorithms when 15 independent features are selected. When SBS was used, K-NN produced an FPR of zero, and when SFS was used with 15 independent features, it produced an FPR of 0.001. When 15 independent features were used, SVM produced an FNR value of 0.005 with SBS and 0.006 with SFS. Based on the results, K-NN and SVM perform better in terms of FPR and FNR. The accuracy values obtained from different classifiers for 10 and 15 selected independent features from the wrapper technique, SFS and SBS are shown in the Fig. 3 and Fig. 4. Comparatively, the SBS wrapper technique produced higher accuracy than the SFS.

Table. 4. Comparative analysis on different evaluation metrics using SEER dataset

Classifiers	Features=10							
	SFS				SBS			
	ACC	FPR	FNR	AUC-ROC	ACC	FPR	FNR	AUC-ROC
MLR	83.4	0.290	0.032	89.9	83.7	0.168	0.158	90.9
K-NN	84.5	0.151	0.159	92.3	84.2	0.132	0.186	92.7
DT	85.9	0.220	0.055	93.4	86.3	0.143	0.131	93.4
RF	85.8	0.219	0.059	93.4	85.7	0.133	0.153	93.7
SVM	85.7	0.234	0.043	90.9	86.1	0.127	0.153	92.4
MLP	85.7	0.223	0.057	93.1	85.3	0.144	0.150	93.4
Features=15								
MLR	89.3	0.156	0.053	95.8	90.7	0.110	0.074	97.4
K-NN	97.9	0.001	0.044	99.3	98.7	0	0.027	99.5
DT	98.1	0.013	0.026	98.1	97.9	0.012	0.030	97.9
RF	96	0.035	0.045	99.4	96.9	0.028	0.033	99.6
SVM	99	0.006	0.014	100	99.5	0.005	0.005	100
MLP	90.3	0.111	0.082	96.6	91.5	0.090	0.079	97.6

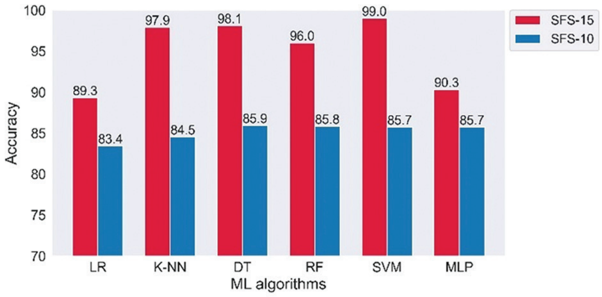

Fig. 3. Accuracy values of different ML algorithms using SFS

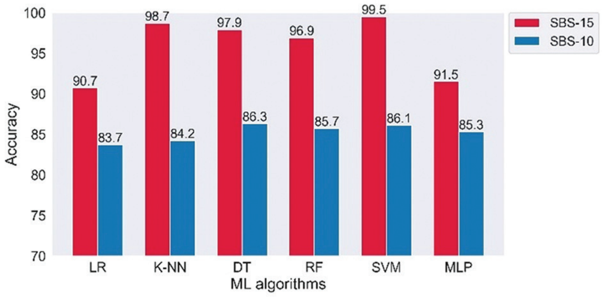

Fig. 4. Accuracy values of different ML algorithms using SBS

The ROC curves are drawn between the FPR and TPR. When the model predicts the probability of belonging to different classes, curves are plotted for different thresholds of the ML models under comparison. The ROC curves are plotted between FPR and TPR for the classifiers, MLR, K-NN, DT, RF, SVM and MLP which correspond to the proposed feature selection sequence using SFS and SBS for 15 independent features from the SEER dataset, as shown in Fig.5 and Fig.6. The AUC-ROC curves are higher for the ML classifiers when 15 independent features identified from the proposed feature selection sequence are used. According to Fig. 5 and Fig. 6, the AUC-ROC curve of the SVM classifier is larger for both SFS and SBS.

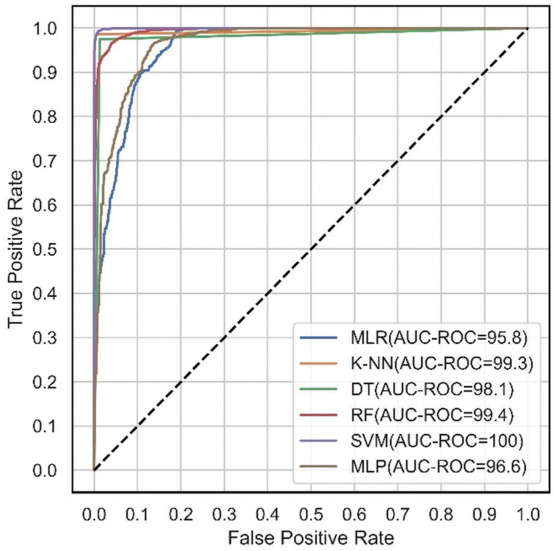

Fig. 5. ROC curves for the proposed feature selection sequence-SFS with 15 independent features

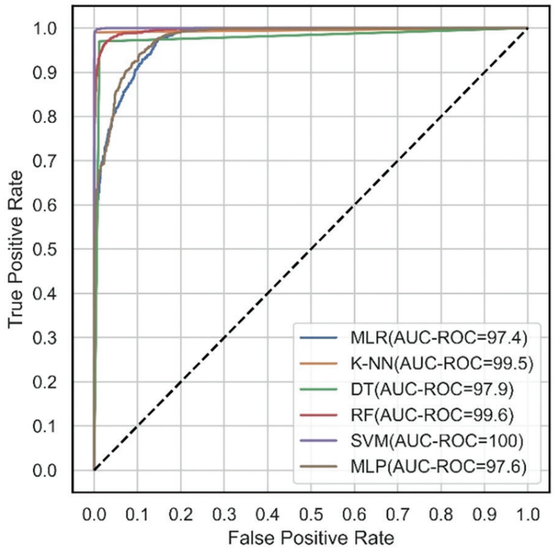

Fig. 6. ROC curves for the proposed feature selection sequence-SBS with 15 independent features

Table. 5. Comparison with results from other features selection techniques-SEER dataset

References	Number of features used	Feature selection techniques	ML which produces highest ACC & AUC (%)
Liu et al. [17]	11	LR backward selection	DT-76.78 (AUC)
Miri et al. [18]	10	CFS, PSO	DT-94.33
Zand et al. [20]	16	IG	DT-86.7
Manikandan et al. [21]	13	VT, PCA	DT-98
Proposed Method	10 and 15	Filter: MI, CS, Wrapper: SBS	SVM-99.5

In addition, the results produced by the proposed feature selection sequence is compared with the results obtained in previous studies [17-22] which use the respective feature selection sequences as mentioned in Table 5. The accuracy produced using the feature selection techniques in Table 5 based on SEER are less than the results produced by the proposed feature selection sequence.

4. CONCLUSION

In this paper, a new filter and wrapper feature selection sequence based on the filter methods, MI, CS and SFS, SBS is proposed. The intersection of independent features obtained from both the filter methods, MI and CS is used to find the optimal set of features. The independent features obtained are subjected as input to the wrapper methods, SFS or SBS to determine a more optimal set of independent features from the SEER training set for the classification of the survivability of breast cancer patients. The results show that SVM performed better than other algorithms, with 99.5% accuracy and higher AUC-ROC values. When SBS and 15 independent features were used, K-NN and SVM both produced lower FPR and FNR values. Compared to SFS, SBS produced better results when 15 independent features are selected. In addition, the results are compared with those obtained using other feature selection techniques in the SEER dataset. It is found that the proposed feature selection sequence with SBS produced higher values for all evaluation metrics when compared to other feature selection techniques in the comparative study while predicting the survivability of breast cancer patients.

5. REFERENCES:

[1] R. D. Kehm, A. A. Llanos, J. A. McDonald, P. Tehranifar, M. B. Terry, “Evidence-Based Interventions for Reducing Breast Cancer Disparities: What Works and Where the Gaps Are?”, Cancers, Vol. 14, No. 17, 2022, p. 4122.

[2] C.A.U. Hassan, M.S. Khan, M.A. Shah, “Comparison of machine learning algorithms in data classification”, Proceedings of the 24th International Conference on Automation and Computing, Newcastle Upon Tyne, UK, 6 September 2018, pp. 1-6.

[3] P. Misra, A. S. Yadav, “Impact of preprocessing methods on healthcare predictions”, Proceedings of the 2nd International Conference on Advanced Computing and Software Engineering, Sultanpur, India, 9 March 2019, pp. 144-150.

[4] N. Somu, M. G. Raman, K. Kirthivasan, V. S. Sriram, “Hypergraph based feature selection technique for medical diagnosis”, Journal of Medical Systems, Vol. 40, 2016, pp. 1-16.

[5] K. N. Neeraj, V. Maurya, “A review on machine learning (feature selection, classification and clustering) approaches of big data mining in different area of research”, Journal of Critical Reviews, Vol. 7, No. 19, 2020, pp. 2610-2626.

[6] V. Kumar, G. S. Lalotra, P. Sasikala, D. S. Rajput, R. Kaluri, K. Lakshmanna, M. Shorfuzzaman, A. Alsufyani, M. Uddin, “Addressing binary classification over class imbalanced clinical datasets using computationally intelligent techniques”, Healthcare, Vol. 10, No. 7, 2022, p. 1293.

[7] E. O. Omuya, G. O. Okeyo, M. W. Kimwele, “Feature selection for classification using principal component analysis and information gain”, Expert Systems with Applications, Vol. 174, 2021, p. 114765.

[8] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, J. Saeed, “A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction”, Journal of Applied Science and Technology Trends, Vol. 1, No. 2, 2020, pp. 56-70.

[9] D. H. Jeong, B. K. Jeong, N. Leslie, C. Kamhoua, S. Y. Ji, “Designing a supervised feature selection technique for mixed attribute data analysis”, Machine Learning with Applications, Vol. 10, 2022, p. 100431.

[10] Y. M. Sobhanzadeh, H. Motieghader, A. Masoudi-Nejad, “Feature Select: a software for feature selection based on machine learning approaches”, BMC Bioinformatics, Vol. 20, No. 1, 2019, pp. 1-7.

[11] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevinoo, J. Tang, H. Liu, “Feature selection: A data perspective”, ACM Computing Surveys, Vol. 50, No. 6, 2017, pp. 1-45.

[12] D. M. Liou, W. P. Chang, “Applying Data Mining for the Analysis of Breast Cancer Data”, Data Mining in Clinical Medicine, Vol. 1246, 2015, pp. 175-189.

[13] A. Saygili, “Classification and Diagnostic Prediction of Breast Cancers via Different Classifiers”, International Scientific and Vocational Studies Journal, Vol. 2, No. 2, 2018, pp. 48-56.

[14] D. A. Omondiagbe, S. Veeramani, A. S. Sidhu, “Machine Learning Classification Techniques for Breast Cancer Diagnosis”, IOP Conference Series: Materials Science and Engineering, Vol. 495, No. 1, 2019, p. 012033.

[15] M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, M. N. Kabir, “Breast cancer prediction: a comparative study using machine learning techniques”, SN Computer Science, Vol. 1, No. 5, 2020, pp. 1-14.

[16] E. Alickovic, A. Subasi, "Breast cancer diagnosis using GA feature selection and Rotation Forest", Neural Computing and Applications, Vol. 28, No. 4, 2015, pp. 753-763.

[17] Y. Q. Liu, C. Wang, L. Zhang, "Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data", Proceedings of the 3rd International Conference on Bioinformatics and Biomedical Engineering, Beijing, China, 11-13 June 2009, pp. 1-4.

[18] S. M. Rostami, M. Ahmadzadeh, "Extracting predictor variables to construct breast cancer survivability model with class imbalance problem", Journal of AI and Data Mining, Vol. 6, No. 2, 2018, pp. 263-276.

[19] R. Aavula, R. Bhramaramba, "XBPF: an extensible breast cancer prognosis framework for predicting susceptibility, recurrence and survivability", International Journal of Engineering and Advanced Technology, Vol. 8, No. 5, 2019, pp. 2249-8958.

[20] H. Karim, K. Zand, "A comparative survey on data mining techniques for breast cancer diagnosis and prediction", Indian Journal of Fundamental and Applied Life Sciences, Vol. 5, No. 1, 2015, p. 4330.

[21] P. Manikandan, U. Durga, C. Ponnuraja, "An Integrative Machine Learning Framework for Classifying SEER Breast Cancer", https://ssrn.com/abstract=4308284 (accessed: 2022)

[22] S. Simsek, U. Kursuncu, E. Kibis, M. A. Abdellatif, A. Dag, "A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival", Expert Systems with Applications, Vol. 139, 2019, p. 112863.

[23] A. Lahousse, E. Roose, L. Leysen, S. T. Yilmaz, K. Mostaqim, F. Reis, J. Nijs, "Lifestyle and pain following cancer: State-of-the-art and future directions", Journal of Clinical Medicine, Vol. 11, No. 1, 2021, p. 195.

[24] M. Petrova, G. Wong, I. Kuhn, I. Wellwood, S. Barclay, "Timely community palliative and end-of-life care: a realist synthesis", BMJ Supportive and Palliative Care, 2021, pp. 1-15.

[25] A. Idri, H. Benhar, J. L. Fernandez-Aleman, I. Kadi, "A systematic map of medical data preprocessing in knowledge discovery", Computer Methods and Programs in Biomedicine, Vol. 162, 2018, pp. 69-85.

[26] R. Gupta, R. Bhargava, M. Jayabalan, "Diagnosis of Breast Cancer on Imbalanced Dataset Using Various Sampling Techniques and Machine Learning Models", Proceedings of the 14th International Conference on Developments in eSystems Engineering, Sharjah, United Arab Emirates, 7-10 December 2021, pp. 162-167.

[27] S. Fotouhi, S. Asadi, M. W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data", Journal of Biomedical Informatics, Vol. 90, 2019, p.103089.

[28] F. Souza, C. Premebida, R. Araujo, "High-order conditional mutual information maximization for dealing with high-order dependencies in feature selection", Pattern Recognition, Vol. 131, 2022, p. 108895.

[29] K. Juneja, C. Rana, "An improved weighted decision tree approach for breast cancer prediction", International Journal of Information Technology, Vol. 12, No. 3, 2020, pp. 797-804.

[30] Y. B. Wah, N. Ibrahim, H. A. Hamid, S. Abdul-Rahman, S. Fong, "Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximising Classification Accuracy", Pertanika Journal of Science and Technology, Vol. 26, No. 1, 2018, pp. 329-340.

[31] C. Shravya, K. Pravalika, S. Subhani, "Prediction of breast cancer using supervised machine learning techniques", International Journal of Innovative Technology and Exploring Engineering, Vol. 8, No. 6, 2019, pp. 1106-1110.

[32] R. Ehsani, F. Drablos, "Robust Distance Measures for k NN Classification of Cancer Data", Cancer Informatics, Vol.19, 2020, pp.1-9.

[33] R. Ahuja, A. Chug, S. Gupta, P. Ahuja, S. Kohli, "Classification and clustering algorithms of machine learning with their applications", Nature-inspired Computation in Data Mining and Machine Learning, Vol. 855, 2020, pp. 225-248.

[34] M. Schonlau, R. Y. Zou, "The random forest algorithm for statistical learning", The Stata Journal, Vol.20, No.1, 2020, pp. 3-29.

[35] S. Badillo, B. Banfai, F. Birzele, I. I. Davydov, L. Hutchinson, T. K. Thong, J. S. Polster, B. Steiert, J. D. Zhang, "An introduction to machine learning", Clinical Pharmacology and Therapeutics, Vol. 107, No. 4, 2020, pp. 871-885.

[36] M. Desai, M. Shah, "An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN)", Clinical eHealth, Vol. 4, 2021, pp. 1-11.