Multi-filter Wrapper Enhanced Machine Learning Model for Cancer Diagnosis

Bibhuprasad Sahu[1](✉) and Sujata Dash[2]

[1] Department of Computer Science and Information Technology, Maharaja Sriram Chandra Bhanja Deo University, Baripada, Odisha, India
prasadnikhil176@gmail.com

[2] Department of Computer Application, Maharaja Sriram Chandra Bhanja Deo University, Baripada, Odisha, India

Abstract. The classification accuracy of the high dimensional dataset degrades due to the redundant and irrelevant features. Feature selection (FS) is used to reduce the dimensionality of the dataset by removing the noisy features. Each filter has its statistical approach. So the feature selected by a single filter may ignore the important one. We have presented a multifilter (MF) wrapper hybrid model. The advantage of using the MF method is to select the important feature by one filter which one may ignore by the other. Here, we have used an aggregator approach to combine the most efficacious features among the four individual filter methods (information gain (IG), chi-square (Chi-sq), minimum redundancy maximum relevance (mRMR), and relief). The accuracy assessment is carried out in a multiple filter wrapper (Jaya-SVM, GA-SVM, PSO-SVM, and FA-SVM). The evaluation and prediction of the subset of features are carried out with four classifiers with excellent performance, such as the support vector machine (SVM), Naive Bayes (NB), decision tree (DT), and linear discriminant analysis (LDA) were tested respectively. Four (breast cancer, leukemia, ovarian, and central nervous system (CNS)) cancer datasets are used to implement the model. The performance of the MF wrapper is excellent in comparison to a single filter. According to the findings of this study, the proposed hybrid approach is a more efficient and trustworthy feature selection technique for selecting highly discriminative features.

Keywords: Filter · Multifilter (MF) · Wrapper · SVM Classifier

1 Introduction

Rapid technological improvements in various domains of life generate data volumes at an unprecedented rate. This may appear to be beneficial to the decision-making process, but it is not when it comes to data dimensions. In microarray data analysis, for example, each sample contains the measurement of tens of thousands of variables, but issues arise in terms of the dimension of the dataset. In the microarray dataset, each sample contains a thousand featured genes. Nevertheless, existing machine learning methods are not designed to cope with huge data sets because a rising ratio of variables

S. Nandan Mohanty et al. (Eds.): ICISML 2022, LNICST 470, pp. 64–78, 2023.
https://doi.org/10.1007/978-3-031-35078-8_7

to sample size has a severe influence on the capacity to develop models with scientific validity. This is called the "curse of dimensionality". The microarray datasets contain a high number of features that enhance the noise and directly impact the accuracy of the machine learning algorithm. Such issues can be addressed by feature selection [18]. It decreases data dimensionality by eliminating features that are irrelevant or redundant [20, 21]. Wrappers, filters, and embedded methods are the basic supervised learning models used for feature selection. The filter identifies the features in the preprocessing step and works independently without giving importance to the classification algorithm [19]. But the wrapper and embedded approaches use machine learning algorithms for feature evaluation. Because of their easy ranking procedures, filter approaches such as IG, GR, Chi-square, and Relief-F (RF) have been recommended as the most eminent and convenient filter algorithms for addressing high-dimensional data in recent years. To remove irrelevant features, the filter method uses a statistical ranking score evaluation and a set of threshold values. One ranked feature higher than the threshold value is selected as the significant one and the rest are excluded for classification. In some cases, the performance of the SVM degrades drastically when the number of selected features is either too big or too small. The main reason behind this imbalance is that each filter algorithm always focuses on the dependencies among the features, so it represents poor performance to generate classifier performance. To enhance the performance of the machine learning model, we have used the ensemble MF combined with a robust feature aggregation technique with the wrapper for the selection of optimal features from the microarray datasets.

2 Literature Survey

Chyh-Ming Lai et al. [1] presented a VIKOR method that adopted MF (MF) feature subset selection followed by simplified swarm optimization as a wrapper to identify the optimal feature subset. The proposed model was evaluated using the most preferred SVM classifier and achieved 100% accuracy in 15 datasets out of 17 taken into consideration for the experimental study. A multi-stage MF with a wrapper Harris Hawks optimization (HHO) model is proposed by Ali Dabba et al. [2]. After preprocessing with the Min-Max method, followed by five filter approaches (F-score, improved F-score, mutual information (MI), mutual information maximization (MIM), and random forest (RF)), to recognize the top-ranked feature subset from the high-dimensional datasets, followed by wrapper HHO. The performance of the model is evaluated using a support vector machine (SVM) -LOOCV and KNN classifier. Similarly, the author [3] proposed an MF wrapper model to identify optimal feature subsets. In the first stage, three filter methods named CS, IG, and ReliefF are adopted to identify the most informative feature subset by removing redundant features. A nature-inspired algorithm is used to enhance the performance in the second phase of the model. Different classifiers are implemented in parallel with PSO for the evaluation of the proposed model. Two stages of MF supervised gene clustering algorithm is implemented to select the most significant clustered features [4]. In SA-EFS, an aggregation-based ensemble feature selection was proposed. The author used different filter algorithms (chi-square, maximum information coefficient, and XGBoost) [5]. The main focus of the study is the classification of the

microarray cancer dataset by recognizing the informative features. Ab Hamid et al. [6] calculate the correlation between different features by removing the noisy features with the adoption of MF (ensemble) algorithms such as IG, CS, RF, and gain ratio (GR). In the later stage, the evaluation of the proposed one is done using harmonize particle swarm optimization (PSO) and SVM. An ensemble MF feature selection approach named EnRank is proposed by combining the ranked feature subset using T-test, CS, Ridge regression, and LASSO. Later, five classification algorithms are used for the performance evaluation of the model. This model is used for the detection of pulmonary hypertension biomarkers [7]. B. Seijo-Pardo et al. [8] presented a homogeneous and heterogeneous ensemble feature selection approach using a combination method called aggregators, and both methodologies were evaluated using SVM. Sumant, A.S. et al. [9] proposed a multi-stage filter for best feature subset identification. In the first stage of the model, CS and SU filters are hybridized with RF, and in the second stage, the top (5–10)% feature subset is identified. Then the overall model is evaluated using random forest (RF), K-nearest neighbor (KNN), and support vector machine (SVM) classifiers. In [10], the MF concept is introduced in approaches such as univariate (CS, IG, GR, OneR) and multivariate (RF, SVM-AW, SVM-RFE). The high-ranked feature subset is combined using a sampling technique called bootstrap aggregation, followed by the threshold value to identify the final ensemble feature subset. Manosij Ghosh et al. [11] presented a two-phase MF-wrapper model. In phase 1, the top n features are identified using RF, CS, and SU with a union and intersection approach. To find the optimal feature subset wrapper, GA is included in the second stage, followed by NLP, SVM, and KNN. The novel hybrid MF-wrapper Grasshopper optimization algorithm (GOA) is incorporated with SA to handle the slow convergence and exploitation of search done by the GOA. Sharifai, A.G. et al. [12] presented an integrated MF feature selection with wrapper sequential forward selection (SFS) to enhance the performance of the model. In this study, six filter methods are used. Feature rank greater than the threshold value is considered an input for SFS. Moumita Mandal et al. [13] proposed an Ensemble MIR-FCS Method, in which three filters called MI, RF, and CS are used to select the best feature from each filter. After a combination of all the features, the efficiency of the model was evaluated using various classifiers such as RNF, KNN, and NB. Uzma et al. [14] proposed a two-stage MF-local search-based Feature Selection (LSFS) wrapper for optimal feature selection followed by SVM, KNN, and NB for evaluation of the proposed model. Like [13], Namrata Singh et al. proposed a multi-stage MF-wrapper model for feature selection. In this study, two variants of filter methods are used, such as subset-based filters (CFS and CONS) and rank-based filters (CS, IG, and RF). Ensemble feature aggregation is carried out after removing duplicate features and then fed to the wrapper-based sequential forward selection algorithm for optimal feature subset selection. SVM with RBF, Random Forest, and KNN classifiers are used for the evaluation of the model [15]. Decorate: a Meta-learning ensemble technique was proposed to achieve better accuracy from the high-dimensional dataset by minimizing the misclassification error [17].

3 MF Aggregator Model

Saeys et al. first presented an ensemble feature selector-based feature selection. The key point of adopting an ensemble MF aggregator mode is to prepare a diverse set of feature selections [16]. It is classified into two types, such as homogeneous and heterogeneous ensembles. In the homogeneous, the same base learner is used for a different sample of data, whereas in the heterogeneous, different base learners are used. An aggregator model aggregates the feature weights or ranks. Various aggregator models are used by various researchers, such as weighted mean aggregation (the feature selection is done based on the features having the highest weight mean), and complete linear aggregation. This method employs the complete ranking of all features, followed by the ranks from all ranking lists that are summed for each feature. The best features have the lowest sum of ranks. The robust rank aggregate method finds features that consistently rank higher than expected under the null hypothesis of uncorrelated inputs and gives each feature a significance score. Feature occurrence frequency (feature selection is performed by determining the number of occurrences of every feature across all lists and ranking them based on their frequency of occurrence). And in classification accuracy-based aggregation (feature selection is accomplished by counting the number of occurrences of each feature across all lists and ranking them according to their frequency of occurrence). We used heterogeneous ensembles in this study because the goal of heterogeneous ensembles is to use the best parts of different algorithms to get strong subsets of features.

In this study, we have combined four filters named IG, Chi-square, mRMR, and ReliefF to obtain an initial gene subset that can be improved in a subsequent stage by a wrapper approach along with a classifier.

3.1 Algorithm: Aggregator Method

Data: A – No. of filter method
Data: B – Threshold (No. of feature to be selected)
Result: Combined feature subset of all filters.
1. For each (a) from 1 to A do
2. Calculate the rank (R_a) using filter method (a)
3. End
4. C= Select the ranking (R_a) with the ranking combination method.
5. Ct= Identify (T) attributes or features.
6. Input the feature subset to the wrapper.

3.2 Study of Filters

In this study, we have used three different filter techniques. The brief study is presented below.

3.2.1 Chi-Square

Independence between two features is calculated using the CS ($\chi 2$) test. It is one of the statistical tests most preferred if the occurrence of the feature is independent of the class value of the dataset. The ($\chi 2$) value of the feature (f) is mentioned in Eq. (1).

$$\chi 2 = \sum_{a=1}^{c} \sum_{b=1}^{k} \frac{(Y_{ab} - V_a)^2}{V_a} \tag{1}$$

Let Y presents the number of examples. Y_{ab} presents the sample numbers in B_a class with b^{th} interval. The expected frequency is presented as Y_{ab}. Here, k and c present the interval count and class count of the dataset, respectively. If $prob_a$ is the probability occurs for the event j, then the expected value can be evaluated using the formula, $V_a = T^* prob_a$. Here (T) is the total number of events. So the lower chi-square value denotes more dependence between the features.

3.2.2 ReliefF

Compared to Relief, ReliefF is the most robust and suitable filter model to deal with incomplete and noisy data. This filtering approach randomly selects an instance R_i and starts searching for k of its neighbors from the same class (called nearest hits H_j) and a different class called (nearest misses $M_j(C)$). The weight W [A] for all of its attributes A is then updated depending on R_i value with corresponding H_j and M_j. A Diff function is used to deal with incomplete data and missing value attributes are handled probabilistically. The final equation of ReliefF is presented in Eq. (2).

$$W[A] = W[A] - \sum_{j=1}^{k} \text{diff}(A, R_i, H_j)/(m.k) + \sum_{C \neq \text{class}(R_i)} \left[\frac{P(C)}{1 - P(\text{class}(R_i))} \sum_{j=1}^{k} \text{diff}(A, R_i, M_j(C)) \right] / m.k \tag{2}$$

3.2.3 mRMR

The MRMR identifies an optimal gene subset with maximal relevance to the target class and minimal redundancy in a gene set. Relevance and redundancy were evaluated using mutual information (I). Let G be the gene subset of a given dataset with class level (c). Let X and Y be the genes in the subset. Then the relevance (V_x) and redundancy (W_x) can be evaluated using Eqs. (3) and (4). The mutual information quotient (MIQ) is calculated using Eq. (5). mRMR is used to rank all the features according to their ranks.

$$V_x = I(x, c) \tag{3}$$

$$W_x = \frac{1}{|S|} \sum_{y \in S} I(x, y) \tag{4}$$

$$\max_{x \in \varphi^c} MIQ_x = \max_{x \in \varphi^c} I(x,c) / \frac{1}{|S|} \sum_{y \in S} I(x,y) \tag{5}$$

3.2.4 Information Gain (IG)

The evaluation of the features is done based on the information gained by considering one feature each time. Information entropy can be used as a metric for feature ranking. Equation (6) represents the entropy of a class feature.

$$H(X) = -\sum^{P}(X)\log_2 P(X) \tag{6}$$

Here, P(X) presents the marginal probability distribution for the random variable X. The value of IG for the attribute feature X is then given by Eq. (7).

$$IG(X/Y) = H(X) - H(X/Y) \tag{7}$$

where H(X) presents the entropy of dataset X, which quantifies the degree of uncertainty in predicting the value of a random variable. And H(X/Y) presents conditional entropy (i.e., uncertainty based on the known variable Y). Each feature's order is determined by the IG value, and high-ranking genes are chosen as input features.

4 MF Wrapper: The Proposed Model

4.1 Algorithm: Pseudo Code of Proposed Rank Aggregation Enhanced MF Wrapper Model

Input: Training dataset D = {y_1, $y_2 \ldots y_{n-1}$, xn}, α, filter algorithm set = {IG, mRMR, Chisq, ReliefF}, Classifier = {SVM, DT, NB, LDA}

Output: Classify the result of D.

// Stage1: Aggregator Approach

1. Procedure aggregation of MFs
2. Aggregation ← { }
3. F-score ← { }
4. a ← Number of filters used for aggregation
5. Ct ←Number of genes selected by each filter.
6. For $filter_i$ ← 1, a do
7. [score] ← calculate ($filter_i$, a)
8. [Rank] ← Rank(score, Ct)
9. Aggregation = Aggregation ∪ {Aggregation Score (Rank, a)
10. End for
11. F-score ← sort[Aggregation]
12. Return F-score
13. End procedure

// Stage2: Wrapper Model

14. Input the feature subset to Wrapper= {Jaya -SVM, PSO-SVM, GA-SVM, PSO - SVM }
15. Obtain the best optimal features.

// Stage3: Classification

16. Do classify by classifiers, Classifiers = {SVM, DT, NB, LDA}
17. For each classifier, classifier $_i$ in the classifier do
18. Select the $\alpha\%$ optimal features from the $F_{optimal\ best}$
19. Learn the classifier$_i$ based on dataset D.
20. Return classification result.

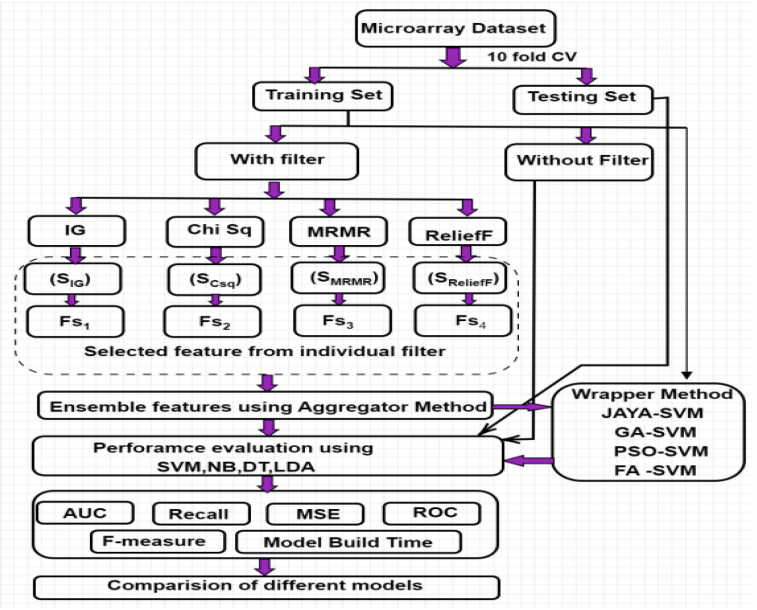

Fig. 1. Conceptual framework of the MF wrapper model

In this study, the MF wrapper model is proposed to handle high-dimensional datasets. The idea behind this study is to minimize the search time and the time complexity to extract the feature subset to enhance the accuracy with a lower number of features. In other words, preprocessing of the dataset is carried out for fitness valuation using

statistical measures and feature selection using filter algorithms. The proposed model combines the filter and the wrapper into a hybrid model. It consists of two stages (Fig. 1).

- In the first stage, i.e., the filter stage, we used four distinguished filters named IG, Chi-square, mRMR, and ReliefF. We have used an aggregation method to prepare a feature subset that can be considered as input for the wrapper. The main idea behind the implementation of a multi-filter instead of a single filter is that the implementation of a single filter may generate an imbalanced feature subset as it just ignores the impact of the feature subset (selected). And each filter has its statistical approach, so the feature selected in a single filter may ignore the important one. So to avoid this gap, we used MF in the first stage to recognize the important features from the high-dimensional dataset.
- In stage two, i.e., the wrapper phase, the final selected feature subset from stage 1 is considered as an input for the wrapper. In this study, we used four different meta-heuristic algorithms. For unbiased comparisons of the performance of the proposed model, three-parameter-based (GA, PSO, and FA) and one non-parameter-based (Jaya) metaheuristic algorithm were used.

4.2 Outline of Classifiers

In this study, four efficient classifiers were used, namely SVM, NB, DT, and LDA. Different classifiers were carefully selected because their performance varies depending on the type of dataset. And the result is not similar when the dataset varies. To involve the entire dataset, we have preferred 10CV on training and test datasets.

4.3 Experimental Environment Setup

Python with a 2.2 Hz Core i7 CPU with 8 GB RAM was used to carry out the overall execution of the proposed hybrid machine-learning model. It is an open-source programming language and contains powerful libraries. To design the experimental study, we used four microarray datasets containing diversity in the number of samples, gene numbers, and classes. After preprocessing (by converting real and categorical values to numeric data, the datasets are prepared for the task), each dataset is randomly divided into two batches (80% as a training set and 20% as a testing set) to estimate the performance based on the confusion matrix. The multi-filter algorithms are parameterized in terms of counts of features in the early stages of our configured model to draw out those amounts of significant features from various features in the datasets. The number of features for each dataset is user-specified, and we have personalized the number for the convenience of our model. The datasets were obtained from the UCI repository. The details of the dataset descriptions are presented in Table 1.

5 Results and Discussions

To evaluate the efficiency of the proposed model, different performance metrics such as precision, recall, F-score, confusion matrix, and AUC/ROC graph are used. The metrics [8, 9] are defined as follows

$$\text{Accuracy (Acc)} = \frac{\text{True}_{+\text{ve}} + \text{True}_{-\text{ve}}}{\text{True}_{+\text{ve}} + \text{True}_{-\text{ve}} + \text{False}_{+\text{ve}} + \text{False}_{-\text{ve}}} \tag{8}$$

$$\text{F - score} = \frac{\text{True}_{+\text{ve}}}{\text{True}_{+\text{ve}} + \frac{1}{2}(\text{False}_{+\text{ve}} + \text{False}_{-\text{ve}})} \tag{9}$$

where, $\text{True}_{+\text{ve}}$ specifies % of correctly classified (positive examples), $\text{False}_{+\text{ve}}$ means % of incorrectly classified as positive (but incorrect examples), $\text{True}_{-\text{ve}}$ means % of correctly classified (negative examples) and $\text{False}_{-\text{ve}}$ % of means incorrectly classified (positive examples).

Table 1. Dataset Study

Datasets	Genes #	Classes #	Samples #
Breast	24481	02 (relapse-34, Non relapse-44)	78
Leukaemia3	7129	03 (ALL-47, AML-25)	72
CNS	7129	02 (Survivors-21, Failures-39)	60
Ovarian	24482	02 (Cancer-162, normal -91)	253

In the experiment, we evaluated the performance of the proposed model with four different datasets. From the resulting study in Table 2, we found the average accuracy for breast cancer and ovarian dataset are 95% and 96% respectively, but in the case of leukemia it is 97%. The proposed model performed best in the case of the CNS dataset with 98%. While focusing on the performance of the individual dataset, we found the wrapper FA performs with 100% accuracy, whereas wrapper Jaya and GA achieved an average accuracy of 93% and PSO with 95% respectively. Similarly, in the case of the leukemia dataset wrapper, FA performs with high accuracy with 99%. Other methods such as wrapper Jaya, GA, and PSO perform with an accuracy of 96%, 95%, and 97% respectively. In the case of the CNS dataset wrapper, Jaya and PSO achieve an accuracy of 99%, but wrapper GA and FA perform with 96% and 98% respectively. Wrapper PSO and FA perform with an accuracy of 98% in the case of the ovarian dataset and Wrapper Jaya and GA achieve accuracy of 92% and 97% respectively. In the case of the ovarian dataset, the MF-Jaya with NB classifier performs worst with an accuracy of 76%.

Table 2. Performance of MF Wrapper

Approach	Dataset	Classifier	Precision	Recall	F-score	Accuracy
MF-Jaya	BC	SVM	0.98	0.98	0.98	0.91
		DT	0.98	0.94	0.96	0.93
		NB	0.95	0.93	0.94	0.94
		LDA	0.97	0.92	0.94	0.95

(*continued*)

Table 2. (*continued*)

MF-GA		SVM	0.97	0.97	0.98	0.94
		DT	0.94	0.94	0.95	0.91
		NB	0.97	0.97	0.96	0.94
		LDA	0.98	0.97	1	0.96
MF-PSO		SVM	0.96	0.97	0.9	0.92
		DT	0.96	0.91	0.9	0.95
		NB	0.94	0.93	0.98	0.97
		LDA	0.94	0.95	0.94	0.96
MF-FA		SVM	1	0.94	0.96	1
		DT	1	0.97	0.96	1
		NB	1	0.94	0.98	1
		LDA	1	0.97	0.98	1
Approach	Dataset	Classifier	Precision	Recall	F-score	Accuracy
MF-Jaya	Leukemia	SVM	0.9	1	0.94	0.98
		DT	1	1	1	0.94
		NB	0.9	1	0.94	0.97
		LDA	0.9	1	0.94	0.98
MF-GA		SVM	0.97	0.94	0.96	0.97
		DT	0.96	0.91	0.93	0.96
		NB	0.95	0.89	0.96	0.92
		LDA	0.96	0.97	0.97	0.98
MF-PSO		SVM	0.97	0.98	0.97	0.98
		DT	0.97	0.98	0.97	0.99
		NB	0.98	0.97	1	0.96
		LDA	0.94	0.92	0.93	0.96
MF-FA		SVM	0.98	0.99	0.98	0.99
		DT	1	0.97	0.96	1
		NB	1	0.98	0.95	1
		LDA	0.96	0.98	0.96	1
Approach	Dataset	Classifier	Precision	Recall	F-score	Accuracy
MF-Jaya	CNS	SVM	0.9	0.81	0.85	1
		DT	0.9	0.81	0.85	1
		NB	1	0.9	0.95	1
		LDA	0.9	0.9	0.9	0.96

(*continued*)

Table 2. (*continued*)

MF-GA		SVM	0.98	0.97	0.97	1
		DT	0.89	0.91	0.92	0.94
		NB	0.96	0.89	0.91	0.92
		LDA	0.97	0.97	0.98	1
MF-PSO		SVM	0.97	0.98	0.98	1
		DT	0.98	0.98	0.98	1
		NB	0.94	0.97	0.98	1
		LDA	0.98	0.97	0.98	0.99
MF-FA		SVM	1	1	1	1
		DT	0.92	0.94	0.94	0.96
		NB	0.89	0.91	0.98	0.99
		LDA	0.98	0.98	0.97	1
Approach	Dataset	Classifier	Precision	Recall	F-score	Accuracy
MF-Jaya	Ovarian	SVM	1	1	1	0.99
		DT	1	1	1	0.98
		NB	1	1	1	0.76
		LDA	1	1	1	0.97
MF-GA		SVM	0.98	0.97	0.98	1
		DT	0.98	0.97	0.98	1
		NB	0.96	0.94	0.96	0.97
		LDA	0.91	0.89	0.91	0.94
MF-PSO		SVM	0.96	0.96	0.98	1
		DT	0.98	0.97	0.97	0.99
		NB	0.92	0.89	0.87	0.93
		LDA	0.99	0.98	0.98	1
MF-FA		SVM	0.97	0.98	0.98	1
		DT	0.98	0.99	1	1
		NB	0.98	1	1	1
		LDA	0.92	0.93	0.89	0.94

From the resulting study, it is noteworthy to explain that the MF wrapper performs better than its other counterparts. Average performance accuracy ranges from 92% to 100%. Similarly, the recall and F-score vary between 81% and 100%. The average accuracy achieved by the MF wrapper varies within a range of 95% and 98%. Due to the page limit, it is not possible to present the complete result analysis of each model. We presented the ROC graph of the CNS and breast cancer dataset with the MF wrapper

model. The performance of the hybrid model is compared with the existing models. The comparative study of our model with existing models is presented in Table 3. The performance of our model is mentioned in bold. From Table 3, it is clear that our proposed model performs better except for the ovarian dataset due to overfitting (Fig. 2).

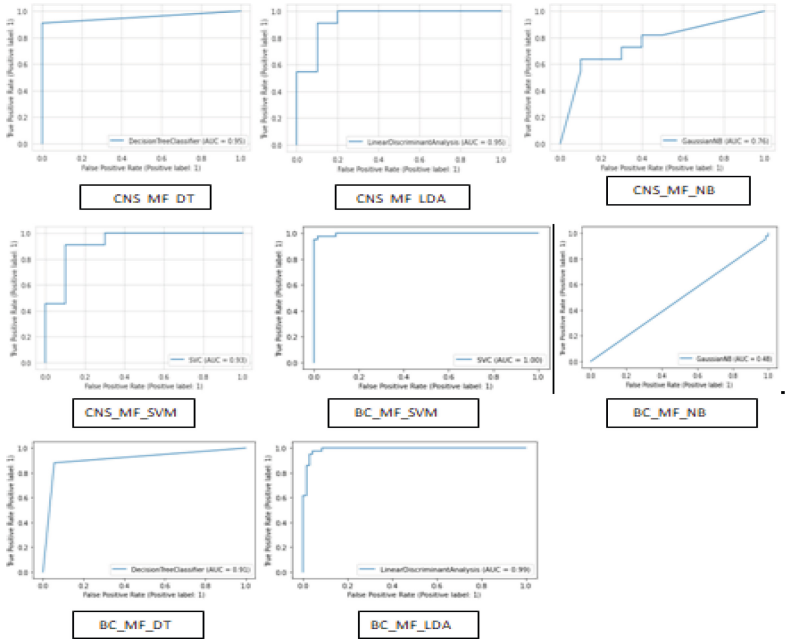

Fig. 2. ROC-AUC curve of MF-Wrapper with different datasets

Table 3. Performance comparisons proposed vs State-of-art Models

Dataset	Method	Filter	Classifier	Accuracy
Leukemia	MF-GE	MF	C4.5	84.51
Leukemia	EFHFS	MF	NB	69.3
Leukemia	AC-MOFOA	IG	KELM	96.78
Leukemia	AC-MOFOA	Chi-Square	KELM	96.4
Leukemia	AC-MOFOA	ReliefF	KELM	95.82

(*continued*)

Table 3. (*continued*)

Dataset	Method	Filter	Classifier	Accuracy
Leukemia	AC-MOFOA	mRMR	KELM	96.26
Leukemia	Ensemble PSO	MF	Ensemble	95.3
Leukemia	**Wrapper JAYA**	**MF**	**All**	**96(AVG)**
Leukemia	**Wrapper GA**	**MF**	**All**	**95(AVG)**
Leukemia	**Wrapper PSO**	**MF**	**All**	**97(AVG)**
Leukemia	**Wrapper FA**	**MF**	**All**	**99(AVG)**
Breast cancer	Ensemble	MF	SVM	72.91
Breast cancer	Ensemble PSO	MF	SVM	96.15
Breast cancer	**Wrapper JAYA**	**MF**	**All**	**93(AVG)**
Breast cancer	**Wrapper GA**	**MF**	**All**	**93(AVG)**
Breast cancer	**Wrapper PSO**	**MF**	**All**	**95(AVG)**
Breast cancer	**Wrapper FA**	**MF**	**Al**	**100(AVG)**
Breast cancer	AC-MOFOA	IG	KELM	75.11
Breast cancer	AC-MOFOA	Chi-Square	KELM	69.31
Breast cancer	AC-MOFOA	ReliefF	KELM	76.23
Breast cancer	AC-MOFOA	mRMR	KELM	77.29
Breast cancer	Ensemble PSO	MF	Ensemble	64.98
CNS	**Wrapper JAYA**	**MF**	**All**	**99(AVG)**
CNS	**Wrapper GA**	**MF**	**All**	**96(AVG)**
CNS	**Wrapper PSO**	**MF**	**All**	**99(AVG)**
CNS	**Wrapper FA**	**MF**	**All**	**98(AVG)**
CNS	Ensemble PSO	MF	Ensemble	66.6
Ovarian	**Wrapper JAYA**	**MF**	**SVM**	**92(AVG)**
Ovarian	**Wrapper GA**	**MF**	**DT**	**97(AVG)**
Ovarian	**Wrapper PSO**	**MF**	**NB**	**98(AVG)**
Ovarian	**Wrapper FA**	**MF**	**LDA**	**98(AVG)**
Ovarian	Ensemble PSO	MF	Ensemble	99.6

6 Conclusions

In this study, we have presented an MF wrapper machine learning model for gene selection and classification of high-dimensional microarray datasets. The proposed model begins with an aggregator model, with these high-ranked features selected and combined to prepare one dataset. The outcome of the MF aggregator model is considered as an input for the wrapper. We evaluated our model with four microarray datasets (breast,

leukemia, CNS, and ovarian cancer) using ten CVs. Then the performance of the proposed model is compared with other state-of-the-art machine learning models for the said dataset. The results show that the multiple filters outperform compared with individual ranking methods for initial gene selection. We can conclude that more filters used more possibilities to ensure that no relevant gene was left out or a potential biomarker in the early stages. Based on different experiments during our framework study, it suggests that choosing MF to improvise the performance of the model to select the best feature from the high-dimensional dataset. The experimental results carried out with four microarray datasets demonstrate that adopting this MF wrapper model is not only able to identify the best-featured genes, but also enhances performance. This proposed model is efficient and effective for high-dimensional datasets as per the result analysis in Table 3. In the future, we will use different metaheuristic approaches to deal with noisy, high-dimensional datasets to get the most accurate results by reducing the number of features.

References

1. Lai, C.-M., Huang, H.-P.: A gene selection algorithm using simplified swarm optimization with MF ensemble technique. Appl. Soft Comput. **100**, 106994 (2021)
2. Dabba, A., Tari, A., Meftali, S.: A new multi-objective binary Harris Hawks optimization for gene selection in microarray data. J. Ambient. Intell. Humaniz. Comput. **14**(4), 3157–3176 (2021). https://doi.org/10.1007/s12652-021-03441-0
3. Alrefai, N., Ibrahim, O.: Optimized feature selection method using particle swarm intelligence with ensemble learning for cancer classification based on microarray datasets. Neural Comput. Appl. **34**(16), 13513–13528 (2022). https://doi.org/10.1007/s00521-022-07147-y
4. Bose, S., Das, C., Banerjee, A., Chattopadhyay, M., Chattopadhyay, S.: An ensemble filtering and supervised clustering based informative gene selection algorithm in microarray gene expression data. In: 2020 4th International Conference on Computational Intelligence and Networks (CINE), pp. 1–7. IEEE (2020)
5. Wang, J., Jing, X., Zhao, C., Peng, Y., Wang, H.: An ensemble feature selection method for high-dimensional data based on sort aggregation. Syst. Sci. Control Eng. **7**(2), 32–39 (2019)
6. Ab Hamid, T.M.T., Sallehuddin, R., Yunos, Z.M., Ali, A.: Ensemble based filter feature selection with harmonize particle swarm optimization and support vector machine for optimal cancer classification. Mach. Learn. Appl. **5**, 100054 (2021). https://doi.org/10.1016/j.mlwa.2021.100054
7. Liu, X., Zhang, Y., Chunli, F., Zhang, R., Zhou, F.: EnRank: an ensemble method to detect pulmonary hypertension biomarkers based on feature selection and machine learning models. Front. Genet. **12**, 601 (2021)
8. Seijo-Pardo, B., Porto-Díaz, I., Bolón-Canedo, V., Alonso-Betanzos, A.: Ensemble feature selection: homogeneous and heterogeneous approaches. Knowl. Based Syst. **118**, 124–139 (2017)
9. Sumant, A.S., Patil, D.: Ensemble feature subset selection: integration of symmetric uncertainty and chi-square techniques with RReliefF. J. Inst. Eng. India: Ser. B **103**(13), 831–844 (2021). https://doi.org/10.1007/s40031-021-00684-5
10. Pes, B.: Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. Neural Comput. Appl. **32**(10), 5951–5973 (2019). https://doi.org/10.1007/s00521-019-04082-3

11. Ghosh, M., Adhikary, S., Ghosh, K.K., Sardar, A., Begum, S., Sarkar, R.: Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods. Med. Biol. Eng. Compu. **57**(1), 159–176 (2018). https://doi.org/10.1007/s11517-018-1874-4
12. Sharifai, A.G., Zainol, Z.B.: Multiple filter-based rankers to guide hybrid grasshopper optimization algorithm and simulated annealing for feature selection with high dimensional multiclass imbalanced datasets. IEEE Access **9**, 74127–74142 (2021). https://doi.org/10.1109/ACCESS.2021.3081366
13. Mandal, M., Ghosh, D., Acharya, S., Saha, N., Sarkar, R.: MIRFCS: an ensemble of filter methods for classification of disease data. In: Das, A.K., Nayak, J., Naik, B., Dutta, S., Pelusi, D. (eds.) Computational Intelligence in Pattern Recognition. AISC, vol. 1349, pp. 205–217. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-2543-5_18
14. Uzma, Z.H.: An ensemble filter-based heuristic approach for cancerous gene expression classification. Knowl. Based Syst. **234**, 107560 (2021). https://doi.org/10.1016/j.knosys.2021.107560
15. Singh, N., Singh, P.: A hybrid ensemble-filter wrapper feature selection approach for medical data classification. Chemom. Intell. Lab. Syst. **217**, 104396 (2021)
16. Saeys, Y., Abeel, T., Peer, Y.: Robust feature selection using ensemble feature selection techniques. In: Daelemans, W., Goethals, B., Morik, K. (eds.) Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-87481-2_21
17. Dash, S.: A diverse meta learning ensemble technique to handle imbalanced microarray dataset. In: Advances in Nature and Biologically Inspired Computing, pp. 1–13. Springer, Cham (2016)
18. Dash, S., Patra, B., Tripathy, B.K.: A hybrid data mining technique for improving the classification accuracy of microarray data set. Int. J. Inf. Eng. Electron. Bus. **4**(2), 43–50 (2012). https://doi.org/10.5815/ijieeb.2012.02.07
19. Sahu, B., Dash, S., Mohanty, S.N., Rout, S.K.: Ensemble comparative study for diagnosis of breast cancer datasets. Int. J. Eng. Technol. **7**(4), 281–285 (2018)
20. Dash, S., Patra, B.: Redundant gene selection based on genetic and quick-reduct algorithms. Int. J. Data Mining Intell. Inf. Technol. Appl. **3**(2), 1
21. Dash, S., Patra, B.: Feature selection algorithms for classification and clustering in bioinformatics. In: Tripathy, B.K., Acharjya, D.P. (eds.) Global Trends in Intelligent Computing Research and Development, pp. 111–130. IGI Global (2014). https://doi.org/10.4018/978-1-4666-4936-1.ch005