A Hybrid Gene Selection Strategy Based on Fisher and Ant Colony Optimization Algorithm for Breast Cancer Classification

https://doi.org/10.3991/ijoe.v17i02.19889

Mohammed Hamim (✉)
ENSAM, Casablanca, Morocco
mohamed.hamim@gmail.com

Ismail El Moudden
EVMS-Sentara Healthcare Analytics and Delivery Science Institute,
Eastern Virginia Medical School, Va, USA

Mohan D Pant
Eastern Virginia Medical School, Va, USA

Hicham Moutachaouik, Mustapha Hain
ENSAM, Casablanca, Morocco

Abstract—Breast cancer poses the greatest threat to human life and especially to women's life. Despite the progress made in data mining technology in recent years, the ability to predict and diagnose such fatal diseases based on gene expression data still reveals a limited prediction performance, which may not be surprising since most of the genes in expression data are believed to be irrelevant or redundant. The dimensionality reduction process may be considered as a crucial step to analyze gene expression data, as it can reduce the high dimensionality of the breast cancer datasets, which may result into a better prediction performance of such diseases. The paper suggests a new hybrid approach-based gene selection that combines the filter method and the Ant Colony Optimization algorithm to find the smallest subset of informative genes (genes markers) among 24,481 genes. The proposed approach combines four machine learning algorithms - C5.0 Decision Tree, Support Vector Machines, K-Nearest Neighbors algorithm, and Random Forest Classifier - to classify each of the selected samples (patients) into two classes which have cancer or not. Compared with existing methods in the literature, experimental results indicate that our proposed gene selection approach achieved globally higher classification accuracies with a relatively smaller number of genes.

Keywords—Breast cancer, Gene Selection, Ant Colony Optimization, Microarray technology

1 Introduction

According to WHO, breast cancer is ranked second on the list of cancer-related deaths in women after lung cancer, affecting around two million women each year [1]. Diagnosis of breast cancer at an early stage may allow for adequate and effective treatment to be adopted, which may increase the survival rate for this disease. This fact puts in evidence a strong need to develop a prediction system that can detect breast cancer at an early stage, so that prompt treatment is started. With the development of microarray technology, gene expression analysis has become an effective tool in biomedical research since it enables to evaluate the expression levels of thousands of genes simultaneously, which has attracted a number of researchers' interest in prediction and diagnosis of different kinds of cancers [2]. However, using microarray technology to predict breast cancer is not without challenges, because the existence of a large number of genes against a small number of specimens may negatively influence the credibility of any prediction system. For this reason, and in order to improve breast cancer risk prediction performance, we proposed a new approach-based gene selection that combines two feature selection (FS) methods in a two-step hybrid system. The first step extracts most informative genes by using Fisher-score based filter method in order to reduce the search space, and then we use Ant Colony Optimization (ACO) based wrapper method to select the smallest subset of genes that allows the highest prediction performance. Our proposed hybrid approach is evaluated using four classifiers: C5.0 Decision Tree, Support Vector Machines (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN) algorithm.

In the next section of this work, we briefly review the existing literature. Techniques and tools are described in the third section. The penultimate section is devoted to discussing experimental results. The final section summarizes the contribution of this study.

2 Existing Literature

With the development of Machine Learning (ML) techniques, several breast cancer prediction approaches are available. In this section, we will present the most recent common techniques carried out in this research area.

Aldryan et al. have developed a breast cancer risk prediction model that combines MBP (Modified Backpropagation with Conjugate Gradient Polak-Ribiere) classifier with Ant Colony Optimization based gene selection. They tested their approach on five public microarray datasets (Breast cancer, Colon Tumor, Leukemia, Ovarian Cancer, and Lung Cancer). For Breast cancer, the best accuracy of 64.12% was achieved by involving 2448 genes (10% of genes). However, the proposed work differs from ours as its result is based on a large number of genes with a low accuracy compared to our proposed approach [3].

Al-Quraishi et al. proposed a new breast cancer prediction approach based on an ensemble of Deep Neural Network (DNN) and Support Vector Machine (SVM). By combining the ensemble classifier DNN+SVM with the Correlation-based filter method

(FCBF). Based on the holdout evaluation technique (80% training set and 20% test set), experiments achieved an accuracy of 96.11% using 112 genes. This proposed study differs from ours as we used a stratified k-fold cross-validation technique to evaluate the performance of our proposed models [4].

Kumari and Singh have designed a system that can predict breast cancer at an early stage based on a Wisconsin breast cancer database. The proposed system is a combination of FS using Correlation-Based Measures with classification using linear Regression (LR), SVM, and KNN algorithm. Experimental results show that the best results in terms of accuracy were achieved with KNN classifier. This research is different from our work as we predict breast cancer risk based on gene expression data [5].

Shen et al. have introduced a deep learning system detecting breast cancer on screening mammograms by using end-to-end training approach. This research focused on sensitivity and specificity as evaluation metrics, while in our study, we used accuracy to evaluate the quality of our models. Moreover, The proposed approach differs from ours as our prediction system is based on gene expression data [6].

Hajiabadi et al. have integrated a new objective function (a combination of three loss functions: Correntropy, Hinge, and Cross-Entropy) to a simple ANN architecture. They used precision, recall, f1-Score, and accuracy to evaluate the performance of the proposed objective function. However, the new method was evaluated by doing experiments on Wisconsin Breast Cancer Diagnosis (WBCD) dataset, which is not the case for our study [7].

In order to improve breast cancer risk prediction using gene expression data, Hamim et al. have proposed a new two-phase gene selection approach. First, they used Fisher score-based filter method to reduce the research space complexity. Then in the second phase, they used C5.0 Decision Tree algorithm to find the smallest subset of genes to predict breast cancer with high performance. The experiment results have shown that their prediction framework achieved a performance of 93.28% in term of accuracy by involving only five genes predictors [8].

To diagnosis breast cancer at its early stages, Rajamohana et al. have employed several ML algorithms such as random forests, decision trees, KNN, and SVM. The experiments were conducted on WBCD dataset, and results show that random forest gives a good result in predicting breast cancer with an accuracy of 93.34% [9].

3 The Proposed Framework

Figure 1 summarizes the main steps of our proposed framework to improve breast cancer risk prediction. Using Fisher-score based filter method and ACOC5 based wrapper method; ACO is used to implement the gene selection, and C5.0 algorithm serves as a fitness function. C5.0, SVM, KNN and RF are used to classify the selected genes.

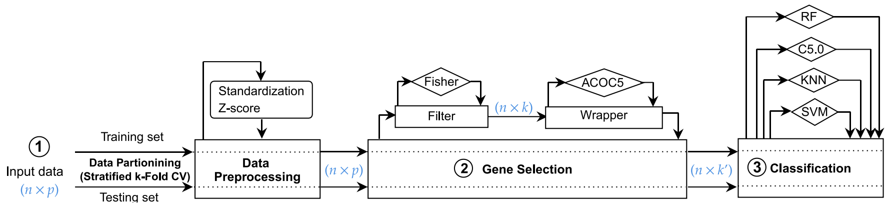

Fig. 1. Our proposed Framework

3.1 Gene selection strategy: Hybrid fisher ACOC5 (HFACOC5)

Frequently, gene-expression data analysis refers to a large number of features (genes), p, versus a small number, k, of samples (patients) ($k << p$). In contrast, in traditional classification research, the number of genes is smaller than the number of samples. Given this fact, technically, performing classification using microarray data may appear to be time and resource consuming. Feature selection (Gene selection in the context of microarray data analysis) is a powerful tool because it allows us to significantly reduce our research space by selecting only relevant and informative features and removing irrelevant and redundant ones [10], which may improve computational speed and prediction accuracy. Various approaches for FS exist in the literature; the present paper proposes a combination of two FS approaches, the Fisher-score based filter method, and the ACOC5 based wrapper method. The proposed gene selection approach is called Hybrid Fisher ACOC5 (HFACOC5), which is illustrated in the flowchart shown in Figure 2. The overall pseudo-code of our framework is illustrated in Algorithm 1 - discussed in the following subsections:

Algorithm 1: Pseudo code of our whole framework

```
Function Classification(Training_set, Test_set):
    $List\_Classifiers = [SVM, C5.0, KNN, RF]$ /* iterate over all proposed machine learning algorithms */
    for each $Classifiers$ in $List\_Classifiers$ do
        $Model$ ←Train $classifier$ on the $Training\_set$
        Test the $Model$ on the $Testg\_set$
        Calculate average performance (Accuracy , F1-score, and the AUC) using Equation (7).
    Return All obtained $Models$ with their average performances

/* Main program */
Input : -A p-diemsional DNA microarray dataset $D = [y, x_1, x_2, ..., x_p]_{n \times 1}$ , with $n$ is the number of samples, $x$ is the gene vector , $y$ is the target vector and $p$ is
        the number of features
        $k$ : the number of selected features using Fisher score based Filter method
        $k\prime$ : the number of selected features using our ACOC5 based wrapper approach
Output: List of generated prediction models with their average performance and running time for each one.
Split data $D$ using the stratified K-fold cross-validation technique.
for each $fold$ in $D$ do
    $D_{test} \leftarrow fold$
    $D_{train} \leftarrow$ remaining $(10-1) folds$
    $SD_{train}, SD_{test} \leftarrow Standardization(D_{train}, D_{test})$/* Standardization of $D_{train}$ and $D_{test}$ using Equation (8) */
    $F_{train}, F_{test} \leftarrow Filter(k, SD_{train}, SD_{test})$/* Filtering data using Fisher score Equation (1) */
    for each $k\prime$ in $[5, 10, 15, 20]$ do
        $Sub_{train}, Sub_{test} \leftarrow ACOC5(k\prime, F_{train}, F_{test})$/* Selection of the optimal gene subset using our ACOC5 wrapper approach */
        $Prediction\_Models \leftarrow Classification(Sub_{train}, Sub_{test})$/* Classification process using the optimal gene subset */
Return all generated prediction models with their evaluation performances
```

Filter method using Fisher score: As they act independently of the machine learning process, filters-based feature selection methods are faster than the wrapper methods; wherefore, these methods are more commonly used when it comes to dealing with a high dimensional data set [11]. Many filter techniques exist [12]–[14]; in the present work, the Fisher score denoted by F was applied to select the most relevant features (genes). As a supervised strategy used in binary prediction problem, Fisher score focuses on a subset of features (genes) for which the distances between data points from different labels should be the largest possible, whereas distances between the data points from the same labels should be reduced as much as possible [15]. Thus, the gene subset is determined in two steps:

- At the first step, the score of each gene G^i is computed using Equation (1)

$$F(G^i) = \frac{\sum_{k=1}^{c} \eta_k(\mu_k^i - \mu^i)^2}{\sum_{k=1}^{c} \eta_k(\sigma_k^i)^2} \tag{1}$$

  μ_k^i, σ_k^i are the mean and standard deviation of *k-th* class, corresponding to the *i-th* gene. μ^i denotes the mean of the whole *i-th* feature in the X matrix.
- At the second step, all genes in breast cancer dataset are ranked by their importance, and the top 100 ranked genes with high scores are selected as the most informative.

ACOC5 based wrapper approach: If filter methods evaluate the goodness of features independently of any classification process, the wrapper approaches, on the contrary, use the learning algorithm to evaluate the importance of feature subsets, the reason why they are very slow and computationally more intensive [16]. However, wrappers generally guarantee better FS results than filters in most of cases. In the proposed work, we explore our research space by using ACO search engines. Our motivation beneath this choice resides in the ability of ACO to efficiently scan the search space to find the optimal gene subset.

Inspired by the food searching system of real-life ants, Ant Colony Optimization is a popular metaheuristic algorithm introduced by Marco Dorigo in the early 1990s [17]. In nature, when a source of nourishment is found, the ants communicate between them to find the shortest path between the nourishment source and their nest. The communication process is done via a special chemical known as pheromone. So, when ants travel down to get the source of food, they deposit an amount of pheromone on the chosen path to cross. As the pheromone is a volatile chemical, the more ants deposit pheromone on a path, the more that path becomes more attractive for being followed, and the other paths become less attractive, and in this way, the optimal path is chosen [18].

As a powerful optimization technique used in many research areas [19], [20], ACO is a promising approach that has been widely employed in FS [19], [21]. In the context of FS, the main idea of ACO is to model the problem of selection as a problem of finding the optimal path in a graph, where nodes in the graph are features (genes), and edges between them represent the choice of the next features. Thus, searching for the optimal path in the graph is the synonym of finding the optimal feature subset in a features space. In the context of gene selection using ACO, we can reformulate our approach as follows:

- **Step 1:** Initialize the parameters of ACO, such as the number of ants m, the maximum number of iterations t_{max}, the amount of pheromone in the search space, pheromone evaporation coefficient $0 \leq \rho \leq 1$, the heuristic desirability η, and tunable parameters ($\alpha \geq 0$ decides the relative influence of pheromone, $\beta \geq 0$ controls the influence of η , and Q a constant multiplier defines the amount of pheromones that should put each ant)

- **Step 2:** For each iteration t, each ant k starts in a randomly selected feature, and to construct a candidate feature subset from it, ants are supposed to follow the probabilistic transition rule of Equation (2).

$$P_{ij}^{k}(t) = \begin{cases} \frac{\tau_{ij}^{\alpha}(t)\cdot\eta_{ij}^{\beta}(t)}{\sum_{l\in S_i^k}\tau_{il}^{\alpha}(t)\cdot\eta_{il}^{\beta}(t)} & \forall j \in S_i^k \\ 0 & otherwise \end{cases} \tag{2}$$

where, S_i^k denotes features set (nodes) that have not been selected yet, $\tau_{ij}(t)$ the amount of pheromone trail on the edge ij (between nodes (features) i and j), $\eta_{ij}(t)$ the heuristic desirability to visit feature j (to select feature j) when the ant k is in the feature i.

- **Step 3:** Evaluate each candidate feature subset S_k (constructed by each ant k) using the classifier C5.0 (described in the fourth section). The evaluation step is carried using Equation (3).

$$Accuracy = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{2}(Accuracy_i^{Train} + Accuracy_i^{Test}) \tag{3}$$

where K denotes the number of folds as we used the stratified K-fold cross-validation technique to evaluate the candidate feature subset.

- **Step 4:** At the end of each iteration, find the ant with the best feature subset using the third step and update the local pheromone trail in the search space according to Equation (4). If t_{max} is reached, go to the fifth step, otherwise go to the second step.

$$\tau_{ij}(t+1) = \rho\cdot\tau_{ij}(t) + \sum_{k=1}^{m}\Delta\tau_{ij}^{k} \tag{4}$$

With: $\Delta\tau_{ij}^{k} = \begin{cases} \frac{Q}{Accuracy_k} & if\ edge(i,j)\ \ is\ part\ of\ S_k \\ 0 & otherwise \end{cases}$

where, $Accuracy_k$ denotes the accuracy corresponding to the constructed subset by ant k.

- **Step 5:** Find the best feature subset with the highest average classification accuracy among all best solutions.

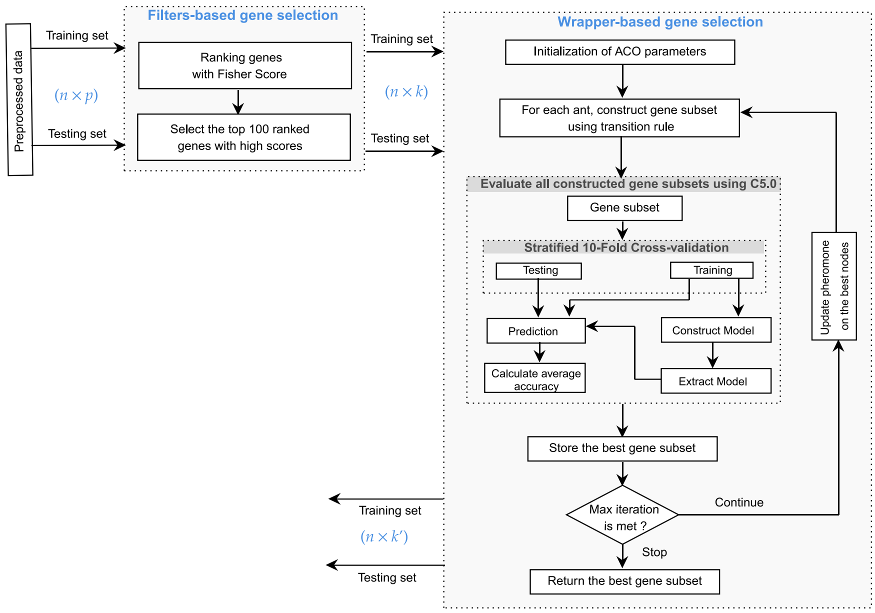
Preprocessed data
Training set
$(n \times p)$
Testing set
Filters-based gene selection
Ranking genes with Fisher Score
Select the top 100 ranked genes with high scores
Training set
$(n \times k)$
Testing set
Wrapper-based gene selection
Initialization of ACO parameters
For each ant, construct gene subset using transition rule
Evaluate all constructed gene subsets using C5.0
Gene subset
Stratified 10-Fold Cross-validation
Testing
Training
Prediction
Construct Model
Calculate average accuracy
Extract Model
Update pheromone on the best nodes
Store the best gene subset
Max iteration is met ?
Continue
Stop
Return the best gene subset
Training set
$(n \times k')$
Testing set

Fig. 2. Flowchart explaining the HFACOC5 gene selection approach

3.2 Gene classification

After selecting the subset of the most informative genes, we classify data using the following algorithms: KNN, SVM, C5.0 decision tree, and Random Forest (RF). The prediction results of these classifiers are then used to evaluate the effectiveness of our proposed gene selection approach.

C5.0 decision tree: Based on decision trees, C5.0 is a new popular classification algorithm developed from C4.5 by [22]. Compared to its ancestor, C5.0 takes its reputation from many advantages: its ability to handle different kinds of data, dealing with missing values and outliers, its high speed and high classification performance, especially with the high-dimensional datasets, supporting boosting and cross-validation process, and automatically allowing removal of unhelpful features. To have better results in terms of performance, the maximum number of boosting trials is set to 100.

Support vector machine: As a binary classifier algorithm, the SVM aims at finding the linear separation (hyperplane) between two classes of observations with the idea that the more the border between them is maximum, the more robust the classification [23]. However, in most real classification cases, datasets are often linearly non-separable, which may necessitate transforming the original space into a new space, and then a linear separation is constructed using the new space [24]. To handle the problem of non-linearity in the present work, the transformation using the Radial Basis Function (RBF) is used, in which the gamma value is set using the formula $(1/ Number\_features)$.

K-Nearest Neighbors algorithm: KNN is one of the simplest supervised ML algorithms used for pattern classification and regression. The KNN is recognized as being a non-parametric algorithm because it does not use any mathematical functions to predict labels for new observations; instead, the prediction process is based on the majority (for classification) or average (for regression) of the k nearest neighbors of the new observations (with $k > 0$) in training set by using "feature similarity" [25]. The similarity process is defined using the distance metric between two observations. In the present work, as we have continuous features, we used Euclidean distance as the distance metric, and the number of neighbors, k is set to 4.

Random forest: Random Forest (RF) is a supervised ML algorithm used for pattern classification and regression. In the context of classification, RF is an ensemble of independent tree classifiers. Each tree classifier is constructed using randomly selected subset features. Thus, new observations are classified by taking the most popular class (using a majority voting function) among all predicted classes by all the tree predictors in the RF (we calculate the average in regression case)[26]. To construct a decision tree classifier, many techniques are used, the most frequent ones are the Information Gain (IG) and the Gini Index (GI) [27]. In the present paper, we used the GI for the randomly feature selection measure.

Performance evaluation: To evaluate the efficiency of our HFACOC5 gene selection approach on breast cancer risk prediction, the well-known metrics accuracy and F-score are considered. Based on the confusion matrix (Table 1), the metrics are calculated using Equation (5) and (6) presented below:

Table 1. Confusion matrix representation

		Predicted classes	
		Positives	***Negatives***
Actual classes	**Positives**	**True Positive (TP)** Patients diagnosed with cancer, and also the system predicted them with cancer.	**False Negative (FN)** Patients diagnosed with cancer, but the system predicted them as healthy.
	Negative	**False Positive (FP)** Patients diagnosed as healthy, but the system predicted them with cancer.	**True Negative (TN)** Patients diagnosed as healthy, and also the system predicted them as healthy.

$$Accuracy(\%) = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (5)$$

$$F - score = \frac{2TP}{2TP+FP+FN} \quad (6)$$

To further validate the classification performance of our prediction models, another popular metric used for performance comparison, the well-known AUC (area under the ROC curve) is used [28], which is the sum of successive trapezoid areas under the ROC (Receiver Operating Characteristic) curve [29]. The model that gives 100% of correct predictions (TP + TN = 100% of samples) has an AUC of 1, while the model that gives 100% of wrong predictions (FN + FP = 100% of samples) has an AUC of 0.

To have a well understanding of how our proposed approach behaves well, each metric of performance described above is computed using the formula in Equation (7).

$$Metric = \frac{1}{2}(Metric_{Train} + Metric_{Test}) \tag{7}$$

4 Experimental Results and Discussion

4.1 Dataset source

The dataset was obtained from ELVIRA Biomedical Data Set Repository [30]. The dataset contains 24,481 scanned gene expressions with 97 instances, 51 of which are healthy, and the rest are diagnosed with cancer. The "NaN" symbol in the original data was replaced with the mean. Table 2 summarizes the dataset description.

Table 2. Breast cancer dataset characteristics

Dataset	Features	Samples	Classes	Description	Ref
Breast Cancer	24481	97	2	51 malignant samples and 46 benign samples	[31]

4.2 Data preprocessing

To improve the prediction performance of our models, prior to feeding our gene expression data to any process of selection or classification, the gene expression levels of each gene were standardized using z-score formula as follow:

$$z = \frac{(x-\mu)}{\sigma} \tag{8}$$

Where x denotes the gene, μ is the mean and σ the standard deviation of that gene.

4.3 Stratified k-fold cross-validation

To avoid the statistical problem of over-estimating due to data partitioning, we used the stratified k-Fold Cross-Validation technique to split our data. Thus, samples were randomly divided into k equal-sized folds, with the same proportions of instances in terms of classes in all folds. The k-1 partitions (folds) are used to fit the model and the remaining partition is used to test de trained model; thus, we ensure that each class in the dataset has the chance to appear in the training folds and testing folds. Moreover, All the process of gene selection was run on the training set to obtain gene subsets. Then the test set was used to testify the classification accuracy of the obtained gene subsets. The Max, Min, average, and standard deviation results of performance metrics of classification were calculated to correctly evaluate the performance of our gene selection strategy. As the data contains fewer samples than the number of genes, we used the stratified 10-fold cross-validation on the whole breast cancer dataset as it is the most common practice in cross-validation.

4.4 Experimental settings

Using parallel processing, all experiments of our strategy were implemented in python 3.7 and tested on a machine with Intel E5-2637 v2 3.5 GHz and 64 GB of RAM using the operating system MS Windows 10.

To achieve better convergence, the parameter setting of our whole prediction system was empirically determined. However, we don't claim that these parameter values are optimal. Parameter optimization may be the subject of future research. For example, the ACOC5 algorithm was implemented with 10 ants and with a maximum number of iterations of 100. The initial pheromone intensity $\tau_{ij}(t = 0)$ of each edge was set to 1, and pheromone evaporation value ρ was set to 0.5. For the parameters that determine the relative importance (α) of the pheromone and the heuristic information (β) were set to $\alpha = 1$ and $\beta = 5$, respectively. In the experiments conducted for performance evaluation of our proposed gene selection strategy, we varied the size of the gene subset between 5 and 20 with an increment of 5 (i.e., 5, 10, 15, and 20).

4.5 Results and discussions

In this section, we explain the experimental results using the proposed gene selection framework. To measure the performance of the proposed strategy, 10 experiments were conducted using the stratified 10-fold cross-validation evaluation technique (section 4.3). The overall experiment results in terms of classification performance (accuracy, F1-score, and AUC) are reported in Table 3 and Figure 3, including mean, max, min, and the standard deviation (SD) of the four classifiers (SVM, KNN, C5.0, and RF) for each of the selected gene subset. According to these results, the research space was reduced two times using our proposed gene selection approach. First, it passed from *p* = 24481 (the original number of genes in the input dataset) to k = 100 genes using Fisher-score based filter method, and then the new space passed in its turn from k genes to k′ = (5, 10, 15, and 20) using ACOC5 based wrapper approach. Each obtained subset of k′ genes was used to construct four classifier-based models (HFACOC5-SVM, HFACOC5-KNN, HFACOC5-C5.0, and HFACOC5-RF). As shown in Table 3 and Figure 3, the most relevant performance results were achieved by models based on decision tree classifiers (HFACOC5-C5.0 and HFACOC5-RF), because they achieved a higher performance rate (> 91%) in terms of all evaluation measures, irrespective of the size of the selected gene subset, while, the lowest performance rate was achieved by models based on KNN algorithm (HFACOC5-KNN). Also, it can be noticed that the classification performance slightly decreased with the increase of gene subset size, especially for models based on decision tree classifiers. For example, for the classification model HFACOC5-C5.0, the performance accuracy decreased from 95.44% (with F1-Score = 0.95 and AUC = 0.96) for five genes to 91.33% (with F1-Score = 0.91 and AUC = 0.91) for 20 genes, which may explain the positive impact of dimensionality reduction process on the prediction performance. The gene accession numbers for each selected gene subset are listed in Table 4.

Because our main aim is to predict the risk of breast cancer with high performance, based on the results shown in Table 3 and Figure 3, the shrinkage model HFACOC5-

C5.0 with five genes was deemed to be the best model because it achieved the best performance prediction (Accuracy of 95.44%, F1-Score = 0.95, and AUC = 0.96) with the smallest number of involved genes (5 genes). Table 5 and Figure 4 gives more details about the experiments performances of our voted prediction model (HFACOC5-C5.0). As it can be noticed from Table 5 and also Figure 4, our favorite model achieved a maximum classification accuracy of 99-100% in 50% of all experiments (10 folds), and classification accuracy of 90-95% in 40% of the remaining experiments. Figure 5 also gives a better overview of the performance of our generated models that involve only k'=5 genes predicators. As we can notice from this figure, for our favorite shrinkage model HFACOC5-C5.0, the roc curves of five out of 10 experiments (folds) are almost superimposed on the "perfect performance" shown in dotted lines, which can confirm the choice of this model as the best generated one using our new gene selection strategy.

5 Conclusion and Future Work

The main purpose beneath this study was to develop and evaluate a classification prediction model for predicting the risk of breast cancer using gene expression data. A new hybrid approach-based gene selection (HFACOC5) was proposed to identify small gene subsets able to achieve high prediction performance. The idea of the proposed approach was to take advantage of both filters (Fisher-score) and wrappers (ACOC5). The Fisher-score selects the most informative genes by first filtering out irrelevant genes and then running ACOC5 over the resulting subset (to achieve maximum accuracy and minimum redundancy). After conducting experiments using the stratified 10-fold cross-validation evaluation technique, using far fewer genes, our proposed strategy achieves high prediction performance in terms of all evaluation measures when it is coupled with Decision tree-based classifiers (a maximum accuracy performance of 99-100% in 50% of all experiments involving five genes). Moreover, as far as we know in the context of our research objective, this is the first time that the data partitioning process using the cross-validation technique was applied before the gene selection approach, which makes our results in terms of selected gene subset and prediction performance more credible than any previous work.

As future work, our proposed approach can be further improved on different aspects, such as considering other bio-inspired algorithms. Also, including experimentation on new microarray data can enable us to test the effectiveness of our strategy far more.

Table 3. Performance measurement for our proposed hybrid strategy (HFACOC5) for each k'

Dataset	Number of genes (P)	HFACOC5		Classification Model	$(\boldsymbol{Mean}{}_{Min}^{Max} \mp \boldsymbol{Standard\ Deviation})$		
		Fisher (k)	*ACOC5 (k')*		*Accuracy (%)*	*F1-score*	*AUC*
Breast Cancer	24481	100	5	HFACOC5-C5.0	$(95.44_{85.00}^{100} \mp 5.45)$	$(0.95_{0.83}^{1.00} \mp 0.06)$	$(0.96_{0.84}^{1.00} \mp 0.06)$
				HFACOC5-RF	$(92.83_{85.00}^{100} \mp 5.81)$	$(0.92_{0.83}^{1.00} \mp 0.06)$	$(0.93_{0.85}^{1.00} \mp 0.06)$
				HFACOC5-KNN	$(85.64_{75.98}^{95.45} \mp 7.64)$	$(0.85_{0.72}^{0.95} \mp 0.08)$	$(0.86_{0.76}^{0.96} \mp 0.08)$
				HFACOC5-SVM	$(87.98_{71.55}^{95.40} \mp 7.93)$	$(0.88_{0.69}^{0.96} \mp 0.09)$	$(0.88_{0.71}^{0.95} \mp 0.08)$
			10	HFACOC5-C5.0	$(93.78_{85.00}^{100} \mp 6.75)$	$(0.93_{0.79}^{1.00} \mp 0.08)$	$(0.95_{0.85}^{1.00} \mp 0.06)$
				HFACOC5-RF	$(92.22_{83.33}^{100} \mp 5.11)$	$(0.93_{0.86}^{1.00} \mp 0.05)$	$(0.92_{0.82}^{1.00} \mp 0.05)$
				HFACOC5-KNN	$(86.58_{73.80}^{94.89} \mp 6.94)$	$(0.87_{0.76}^{0.95} \mp 0.06)$	$(0.87_{0.73}^{0.95} \mp 0.07)$
				HFACOC5-SVM	$(88.29_{68.81}^{97.13} \mp 8.14)$	$(0.89_{0.69}^{0.97} \mp 0.08)$	$(0.88_{0.69}^{0.97} \mp 0.08)$
			15	HFACOC5-C5.0	$(92.44_{85.00}^{100} \mp 7.15)$	$(0.92_{0.79}^{1.00} \mp 0.08)$	$(0.95_{0.88}^{1.00} \mp 0.05)$
				HFACOC5-RF	$(92.00_{85.00}^{100} \mp 6.75)$	$(0.92_{0.83}^{1.00} \mp 0.07)$	$(0.92_{0.85}^{1.00} \mp 0.07)$
				HFACOC5-KNN	$(77.51_{66.95}^{89.66} \mp 6.06)$	$(0.74_{0.56}^{0.88} \mp 0.08)$	$(0.78_{0.67}^{0.90} \mp 0.06)$
				HFACOC5-SVM	$(82.71_{70.98}^{92.61} \mp 6.56)$	$(0.83_{0.68}^{0.93} \mp 0.07)$	$(0.83_{0.71}^{0.93} \mp 0.07)$
			20	HFACOC5-C5.0	$(92.39_{88.89}^{100} \mp 4.34)$	$(0.93_{0.88}^{1.00} \mp 0.04)$	$(0.93_{0.80}^{1.00} \mp 0.07)$
				HFACOC5-RF	$(91.33_{85.00}^{100} \mp 5.80)$	$(0.91_{0.83}^{1.00} \mp 0.07)$	$(0.91_{0.85}^{1.00} \mp 0.06)$
				HFACOC5-KNN	$(81.49_{75.95}^{90.91} \mp 5.43)$	$(0.81_{0.76}^{0.90} \mp 0.05)$	$(0.82_{0.76}^{0.91} \mp 0.05)$
				HFACOC5-SVM	$(87.90_{81.55}^{96.59} \mp 4.14)$	$(0.88_{0.80}^{0.97} \mp 0.05)$	$(0.88_{0.82}^{0.97} \mp 0.04)$

Table 4. Best gene subsets obtained using our gene selection strategy

HFACOC5		Selected Genes
Fisher (k)	*ACOC5 (k')*	
100	5	Contig47544_RC, AJ011306, NM_001168, AF055033, NM_013262
	10	NM_001787, NM_003882, NM_001961, NM_004994, AL049689, NM_006115, NM_004368, AL137615, NM_018074, Contig43454_RC
	15	NM_003662, NM_007292, NM_003600, NM_002811, Contig51882_RC, NM_004219, Contig38901_RC, AL050227, NM_001168, Contig36744_RC, NM_005744, Contig6238_RC, NM_020132, AL080059, Contig33814_RC
	20	NM_016577, Contig47405_RC, NM_012261, NM_001207, Contig32185_RC, Contig47544_RC, Contig30047_RC, Contig46218_RC, Contig38726_RC, NM_002449, AJ011306, NM_013262, Contig46421_RC, NM_002808, Contig55725_RC, Contig51800, AB002324, NM_020120, NM_001168, NM_004994

Table 5. Performance measures for HFACOC5-C5.0 model (with k′ = 5 genes) over the 10-fold (10 experiments)

	experiments (over stratified 10-fold cross validation)										Mean	Max	Min	Standard deviation
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*				
Accuracy (%)	85	95	90	95	90	100	100	100	99.43	100	95.44	100	85	5.45
AUC	0.84	0.98	0.88	0.9	0.96	1	1	1	1	1	0.96	1.00	0.84	0.06
F1-score	0.83	0.95	0.9	0.95	0.92	1	1	1	0.99	1	0.95	1.00	0.83	0.06

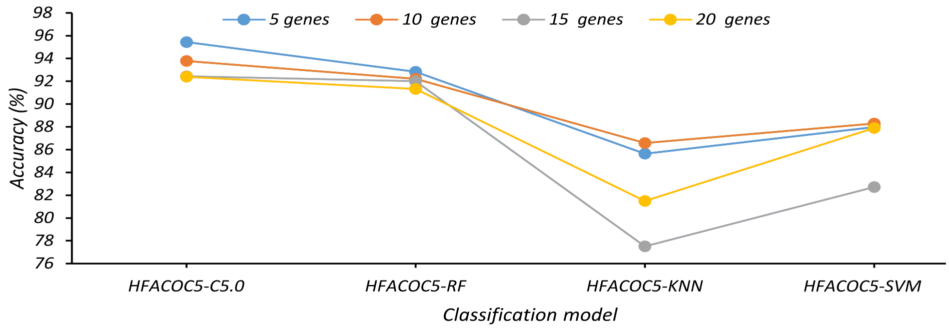

Fig. 3. Classification accuracy for the proposed hybrid strategy (HFACOC5)

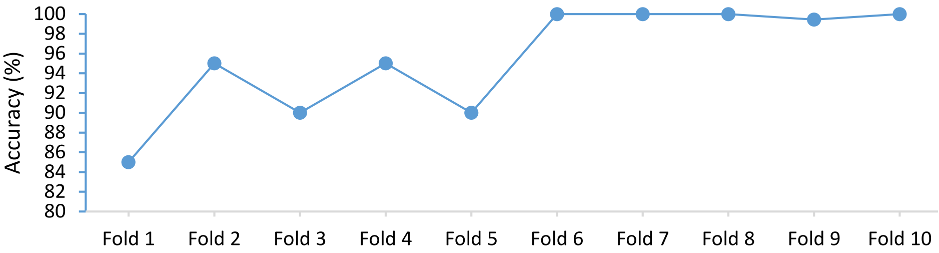

Fig. 4. Classification Accuracy for HFACOC5-C5.0 model (with k′ = 5 genes) for each fold over the 10-fold (10 experiments)

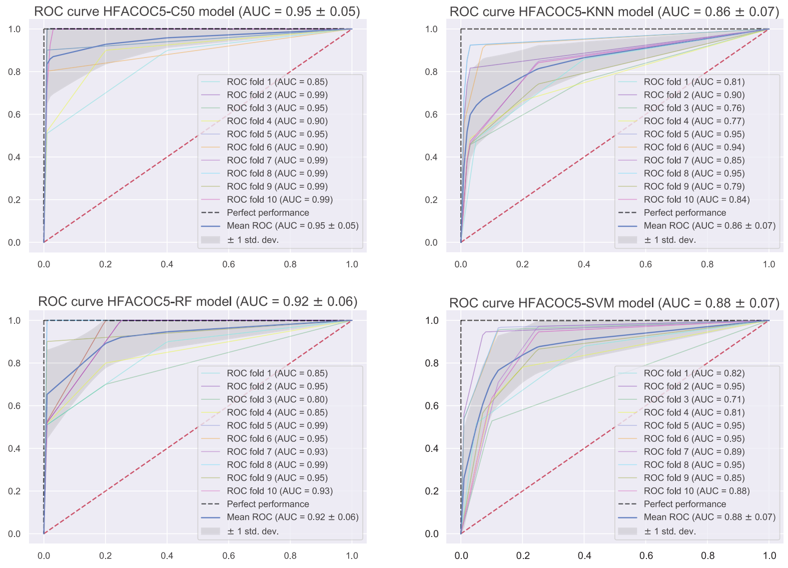

Fig. 5. Roc curves and AUC score over 10-fold (10 experiments) for all generated prediction model involving k′ = 5 genes predicators

6 References

[1] C. E. DeSantis *et al.*, "Breast cancer statistics, 2019," *CA A Cancer J Clin*, vol. 69, no. 6, pp. 438–451, Nov. 2019, doi: 10.3322/caac.21583.

[2] H. Moutachaouik and I. El Moudden, "Mining Prostate Cancer Behavior Using Parsimonious Factors and Shrinkage Methods," 2018. https://doi.org/10.2139/ssrn.3180967

[3] D. P. Aldryan, Adiwijaya, and A. Annisa, "Cancer Detection Based on Microarray Data Classification with Ant Colony Optimization and Modified Backpropagation Conjugate Gradient Polak-Ribiére," in *2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, Tangerang, Indonesia, Nov. 2018, pp. 13–16. https://doi.org/10.1109/ic3ina.2018.8629506

[4] T. Al-Quraishi, J. H. Abawajy, N. Al-Quraishi, A. Abdalrada, and L. Al-Omairi, "Predicting Breast Cancer Risk Using Subset of Genes," in *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, Paris, France, Apr. 2019, pp. 1379–1384. https://doi.org/10.1109/codit.2019.8820378

[5] M. Kumari and V. Singh, "Breast Cancer Prediction system," *Procedia Computer Science*, vol. 132, pp. 371–376, 2018. https://doi.org/10.1016/j.procs.2018.05.197

[6] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep Learning to Improve Breast Cancer Detection on Screening Mammography," *Sci Rep*, vol. 9, no. 1, p. 12495, Dec. 2019. https://doi.org/10.1038/s41598-019-48995-4

[7] H. Hajiabadi, V. Babaiyan, D. Zabihzadeh, and M. Hajiabadi, "Combination of loss functions for robust breast cancer prediction," *Computers & Electrical Engineering*, vol. 84, p. 106624, Jun. 2020. https://doi.org/10.1016/j.compeleceng.2020.106624

[8] M. Hamim, I. El Moudden, H. Moutachaouik, and M. Hain, "Decision Tree Model Based Gene Selection and Classification for Breast Cancer Risk Prediction," in *Smart Applications and Data Analysis*, vol. 1207, M. Hamlich, L. Bellatreche, A. Mondal, and C. Ordonez, Eds. Cham: Springer International Publishing, 2020, pp. 165–177. https://doi.org/10.1007/978-3-030-45183-7_12

[9] S. P. Rajamohana, K. Umamaheswari, K. Karunya, and R. Deepika, "Analysis of Classification Algorithms for Breast Cancer Prediction," in *Data Management, Analytics and Innovation*, Singapore, 2020, pp. 517–528. https://doi.org/10.1007/978-981-32-9949-8_36

[10] H. Liu and H. Motoda, *Computational methods of feature selection*. Chapman & Hall/CRC, 2008.

[11] N. Sánchez-Maroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter Methods for Feature Selection – A Comparative Study," in *Intelligent Data Engineering and Automated Learning - IDEAL 2007*, Berlin, Heidelberg, 2007, pp. 178–187. https://doi.org/10.1007/978-3-540-77226-2_19

[12] W. Duch, "Filter Methods," in *Feature Extraction: Foundations and Applications*, I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, Eds. Berlin, Heidelberg: Springer, 2006, pp. 89–117.

[13] M. S. Al-Batah, B. M. Zaqaibeh, S. A. Alomari, and M. S. Alzboon, "Gene Microarray Cancer Classification using Correlation Based Feature Selection Algorithm and Rules Classifiers," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 15, no. 08, Art. no. 08, May 2019. https://doi.org/10.3991/ijoe.v15i08.10617

[14] M. S. Al-batah, "Ranked Features Selection with MSBRG Algorithm and Rules Classifiers for Cervical Cancer," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 15, no. 12, Art. no. 12, Aug. 2019. https://doi.org/10.3991/ijoe.v15i12.10803

[15] Q. Gu, Z. Li, and J. Han, “Generalized Fisher Score for Feature Selection,” *arXiv:1202.3725 [cs, stat]*, Feb. 2012, Accessed: May 10, 2020. [Online]. Available: http://arxiv.org/abs/1202.3725

[16] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artificial Intelligence*, vol. 97, no. 1–2, pp. 273–324, Dec. 1997. https://doi.org/10.1016/s0004-3702(97)00043-x

[17] M. Dorigo and G. Di Caro, “Ant colony optimization: a new meta-heuristic,” in *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)*, Jul. 1999, vol. 2, pp. 1470-1477 Vol. 2,. https://doi.org/10.1109/cec.1999.782657

[18] S. Kashef and H. Nezamabadi-pour, “An advanced ACO algorithm for feature subset selection,” *Neurocomputing*, vol. 147, pp. 271–279, Jan. 2015. https://doi.org/10.1016/j.neucom.2014.06.067

[19] J. E. Bell and P. R. McMullen, “Ant colony optimization techniques for the vehicle routing problem,” *Advanced Engineering Informatics*, vol. 18, no. 1, pp. 41–48, Jan. 2004. https://doi.org/10.1016/j.aei.2004.07.001

[20] X. Huang, “Ant Colony Optimization Algorithm Model Based on the Continuous Space,” *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 12, no. 12, Art. no. 12, Dec. 2016. https://doi.org/10.3991/ijoe.v12i12.6451

[21] S. Nemati, M. E. Basiri, N. Ghasem-Aghaee, and M. H. Aghdam, “A novel ACO–GA hybrid algorithm for feature selection in protein function prediction,” *Expert Systems with Applications*, vol. 36, no. 10, pp. 12086–12094, Dec. 2009. https://doi.org/10.1016/j.eswa.2009.04.023

[22] Quinlan J R, “RuleQuest Research Data Mining Tools.” https://rulequest.com/ (accessed May 21, 2020).

[23] “Statistical Learning Theory | Wiley,” *Wiley.com*. https://www.wiley.com/en-us/Statistical+Learning+Theory-p-9780471030034 (accessed May 25, 2020).

[24] M. Marjanović, M. Kovačević, B. Bajat, and V. Voženílek, “Landslide susceptibility assessment using SVM machine learning algorithm,” *Engineering Geology - ENG GEOL*, vol. 123, pp. 225–234, 2011. https://doi.org/10.1016/j.enggeo.2011.09.006

[25] B. V. Dasarathy, Ed., *Nearest Neighbor: Pattern Classification Techniques*. Los Alamitos, Calif. : Washington: IEEE Computer Society, 1990.

[26] L. Breiman, “Random Forests--Random Features,” p. 29.

[27] M. Pal, “Random forest classifier for remote sensing classification,” *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, Jan. 2005. https://doi.org/10.1080/01431160412331269698

[28] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve.,” *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982. https://doi.org/10.1148/radiology.143.1.7063747

[29] F. Melo, “Area under the ROC Curve,” in *Encyclopedia of Systems Biology*, W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Eds. New York, NY: Springer, 2013, pp. 38–39.

[30] “Data Repository -- Breast Cancer.” http://leo.ugr.es/elvira/DBCRepository/BreastCancer/BreastCancer.html (accessed May 22, 2020).

[31] L. J. van ’t Veer *et al.*, “Gene expression profiling predicts clinical outcome of breast cancer,” *Nature*, vol. 415, no. 6871, pp. 530–536, Jan. 2002, doi: 10.1038/415530a.

7 Authors

Mohammed Hamim is a PhD student at ENSAM-Casablanca, University Hassan II. He received his M.S degree in “Big Data & Internet of Things” in 2018.

Ismail El Moudden is a faculty in the EVMS-Sentara Healthcare Analytics and Delivery Science Institute at Eastern Virginia Medical School, Norflok, Va, USA. He received his Ph.D. in statistics and data science from the Mohammed V University in Rabat.

Mohan D Pant is an Associate Professor in the School of Health Professions at Eastern Virginia Medical School (EVMS), Norflok, Va, USA. He received his PhD in Educational Measurement and Statistics (now, Quantitative Methods) and MS in Mathematics from Southern Illinois University at Carbondale in 2011 and 2006, respectively.

Hicham Moutachaouik is a Professor at ENSAM-Casablanca, University Hassan II, Casablanca, Morocco. He received his PhD in Information retrieval and habilitation degree in Artificial intelligence.

Mustapha Hain is a Professor at ENSAM-Casablanca, University Hassan II, Casablanca, Morocco. He received his PhD in Software Engineering and his habilitation degree in Artificial Intelligence.

Article submitted 2020-11-17. Resubmitted 2020-12-05. Final acceptance 2020-12-07. Final version published as submitted by the authors.