An Ensemble Machine Learning Model based on Multiple Filtering and Supervised Attribute Clustering Algorithm for Classifying Cancer Samples

Shilpi Bose [Corresp., 1] , **Chandra Das** [1] , **Abhik Banerjee** [1] , **Kuntal Ghosh** [2] , **Matangini Chattopadhyay** [3] , **Samiran Chattopadhyay** [4] , **Aishwarya Barik** [1]

[1] Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, West Bengal, India

[2] Machine Intelligence Unit & Center for Soft Computing Research, Indian Statistical Institute, Kolkata, West Bengal, India

[3] School of Education Technology, Jadavpur University, Kolkata, West Bengal, India

[4] Department of Information Technology, Jadavpur University, Kolkata, West Bengal, India

Corresponding Author: Shilpi Bose
Email address: shilpi.bose@nsec.ac.in

Background: Machine learning is one kind of machine intelligence technique that learns from data and detects inherent patterns from large, complex datasets. Due to this capability, machine learning techniques are widely used in medical applications, especially where large-scale genomic and proteomic data are used. Cancer classification based on bio-molecular profiling data is a very important topic for medical applications since it improves the diagnostic accuracy of cancer and enables a successful culmination of cancer treatments. Hence, machine learning techniques are widely used in cancer detection and prognosis.

Methods: In this article, a new ensemble machine learning classification model named Multiple Filtering and Supervised Attribute Clustering algorithm based Ensemble Classification model (MFSAC-EC) is proposed which can handle class imbalance problem and high dimensionality of microarray datasets. This model first generates a number of bootstrapped datasets from the original training data where the oversampling procedure is applied to handle the class imbalance problem. The proposed MFSAC method is then applied to each of these bootstrapped datasets to generate sub-datasets, each of which contains a subset of the most relevant/informative attributes of the original dataset. The MFSAC method is a feature selection technique combining multiple filters with a new supervised attribute clustering algorithm. Then for every sub dataset, a base classifier is constructed separately, and finally, the predictive accuracy of these base classifiers is combined using the majority voting technique forming the MFSAC-based ensemble classifier. Also, a number of most informative attributes are selected as important features based on their frequency of occurrence in these sub-datasets.

Results: To assess the performance of the proposed MFSAC-EC model, it is applied on different high-dimensional microarray gene expression datasets for cancer sample classification. The proposed model is compared with well-known existing models to establish its effectiveness with respect to other models. From the experimental results, it has been found that the generalization performance/testing accuracy of the proposed classifier is significantly better compared to other well-known existing models. Apart from that, it has been also found that the proposed model can identify many important attributes/biomarker genes.

1 # An Ensemble Machine Learning Model based on
2 # Multiple Filtering and Supervised Attribute Clustering
3 # Algorithm for Classifying Cancer Samples

5 Shilpi Bose[1], Chandra Das[1], Abhik Banerjee[1], Kuntal Ghosh[2], Matangini Chattopadhyay[3],
6 Samiran Chattopadhyay[4], Aisarya Barik[1]

8 [1]Department of CSE, Netaji Subhash Engineering College, Kolkata, West Bengal, India
9 [2]Machine Intelligence Unit & Center for Soft Computing Research, Indian Statistical Institute,
10 Kolkata, West Bengal, India
11 [3]School of Education Technology, Jadavpur University, Kolkata, West Bengal, India
12 [4]Department of Information Technology, Jadavpur University, Kolkata, West Bengal, India

14 Corresponding Author:
15 Shilpi Bose
16 Department of CSE, Netaji Subhash Engineering College, Panchpota, Garia, Kolkata, 700152,
17 West Bengal, India
18 Email address: shilpi.bose@nsec.ac.in

19 ## Abstract

20 **Background:** Machine learning is one kind of machine intelligence technique that learns from
21 data and detects inherent patterns from large, complex datasets. Due to this capability, machine
22 learning techniques are widely used in medical applications, especially where large-scale
23 genomic and proteomic data are used. Cancer classification based on bio-molecular profiling
24 data is a very important topic for medical applications since it improves the diagnostic accuracy
25 of cancer and enables a successful culmination of cancer treatments. Hence, machine learning
26 techniques are widely used in cancer detection and prognosis.
27 **Methods:** In this article, a new ensemble machine learning classification model named Multiple
28 Filtering and Supervised Attribute Clustering algorithm based Ensemble Classification model
29 (MFSAC-EC) is proposed which can handle class imbalance problem and high dimensionality of
30 microarray datasets. This model first generates a number of bootstrapped datasets from the
31 original training data where the oversampling procedure is applied to handle the class imbalance
32 problem. The proposed MFSAC method is then applied to each of these bootstrapped datasets to
33 generate sub-datasets, each of which contains a subset of the most relevant/informative attributes
34 of the original dataset. The MFSAC method is a feature selection technique combining multiple
35 filters with a new supervised attribute clustering algorithm. Then for every sub dataset, a base
36 classifier is constructed separately, and finally, the predictive accuracy of these base classifiers is
37 combined using the majority voting technique forming the MFSAC-based ensemble classifier.
38 Also, a number of most informative attributes are selected as important features based on their
39 frequency of occurrence in these sub-datasets.

40 **Results:** To assess the performance of the proposed MFSAC-EC model, it is applied on different
41 high-dimensional microarray gene expression datasets for cancer sample classification. The
42 proposed model is compared with well-known existing models to establish its effectiveness with
43 respect to other models. From the experimental results, it has been found that the generalization
44 performance/testing accuracy of the proposed classifier is significantly better compared to other
45 well-known existing models. Apart from that, it has been also found that the proposed model can
46 identify many important attributes/biomarker genes.

Introduction

49 Cancer is one of the most fatal diseases around the globe (Tabares-Soto 2020, Hambali 2020).
50 According to the World Health Organization report, Cancer is marked as the second most deadly
51 disease and an estimated 9.7 million deaths around the world in 2018 have occurred due to this
52 signature disease (Hambali 2020). Generally, one in every 6 deaths all over the world, occurs due
53 to cancer. So, within 2030, the number of new cancer patients per year will increase
54 approximately by 25 million (Hambali 2020, NIH 2019). Although several advanced techniques
55 are already developed for the detection of cancer, the proper prognosis of cancer patients, till
56 date, is very poor and the survival rate is also very low (Tabares-Soto 2020, Hambali 2020,
57 Konstantina 2015). It has been already found that for very accurate cancer sample classification
58 or prediction, adequate information is not available from the clinical, environmental, and
59 behavioral characteristics of patients (Kourou 2015, Hambali 2020, Tabares-Soto Reinel 2020).
60 Recently, due to different types of bio-molecular data analysis, several genetic disorders with
61 different biological characteristics have been revealed which are very helpful for early
62 identification and prognosis of cancer and also to discern the responses for different types of
63 treatment (Colozza2005, Greller 1999, Li 2018, Liu2011, Pilling 2017, Su 2001, Swan 2013).
64 With the rapid advancements in genomic, proteomic, and imaging high-throughput technologies
65 (Colozza2005, Greller 1999, Li 2018, Liu2011, Pilling2017, Su 2001, Swan 2013), now it is
66 possible to accumulate huge amount (in the order of thousands) of different bio-molecular
67 information of patients. Using this huge amount of information, researchers have been trying to
68 develop more advanced techniques for early detection and proper prognosis of cancer, and also
69 to improve cancer therapy for improvement of patients' survival rate. To analyze this huge
70 amount of information, lab-based approaches are not adequate as these methods are costly and
71 time-consuming. So, computational or in-silico methods like statistical methods, machine
72 learning, deep learning, etc. have been being used extensively in this field.
73 It is well-known fact that in cancer-causing cells, gene expression is either overexpressed or
74 under expressed (Tabares-Soto 2020). So, measurement of gene expression in cancer cells can
75 give adequate information to improve cancer diagnostic procedures. Nowadays, different
76 developing countries have been using this procedure for cancer sample detection. It is already
77 known that using DNA microarray technology it is possible to measure the expression level of a
78 numerous number of genes for a single experiment/sample simultaneously. The outcome of
79 DNA microarray technology is a gene expression data matrix. This matrix carries information

80 about the expression level of a huge number of genes for a limited number of samples (such as
81 diseased patient samples and normal samples). The presence of the limited number of samples in
82 this data matrix is due to the lack of availability of samples. So, based on information of gene
83 expression data matrix, cancer sample classification is one of the essential tasks in the field of
84 cancer research (Chin 2016, Dashtban 2017, Ding 2005, Elyasigomari 2017, Furey2000, Golub
85 1999, Nada 2019, Tabares-Soto 2020).

86 Using computational or in-silico approaches, gene expression-based cancer sample classification
87 task has been reviewed extensively in different papers (Chin 2016, Dashtban2017, Ding 2005,
88 Elyasigomari 2017, Furey2000, Golub 1999, Nada 2019, Tabares-Soto 2020). However, the
89 main difficulties in the sample classification task arise due to several factors. Firstly, in these
90 data sets, a substantially small number of samples is available (generally in the order of
91 hundreds) compared to the availability of a huge number of genes (generally in the order of
92 thousands) (Chin 2016, Hambali 2020, Nada 2019). For sample classification, genes are treated
93 as features/attributes. So, the high-dimensional gene space is an overhead for most classification
94 algorithms. Secondly, only a very few genes are informative (differentially expressed) and the
95 rest of the section is non-informative (noisy) (Chin 2016, Hambali 2020, Nada 2019) for sample
96 classification and responsible for degrading the classifier's performance. Gene dimension
97 reduction by identification of informative genes as biomarkers can improve the classification
98 accuracy of classifiers. Apart from the improvement of classification accuracy, the identification
99 of informative biomarkers (here, informative genes) has great prospects from a biomedical point
100 of view. These are beneficial for finding the biological reason for a disorder, assessing disease
101 risk, and developing therapeutic targets. The third problem arises due to the small sample size
102 which creates an overfitting problem in classifier construction. Another problem that degrades
103 classifier performance is the sample class imbalance problem. This problem occurs due to the
104 presence of more instances/samples of one class (majority class) with respect to other class(es)
105 (minority class) in a dataset.

106 A fairly large number of works have been already developed for sample classification. These
107 works are divided into two categories. In the first category (Chin 2016, Hambali 2020, Nada
108 2019), the major emphasis is given to the selection of relevant genes for the reduction of feature
109 space. Then based on this reduced feature space, predictive/classification accuracy of the
110 samples is measured using different existing single classification models like naïve Bayes,
111 support vector machine, relevance vector machine, K-nearest neighbor, decision tree, logistic
112 regression, etc. As gene selection is a feature selection task, so based on feature selection
113 techniques, these methods are divided into different categories. These are (1) filter methods (2)
114 wrapper methods (3) embedded methods and (4) hybrid methods. Before we mention the second
115 category of classification methods, let us first elaborate on the first category methods one by one.
116 Filter methods (Chin 2016, Hambali 2020, Nada 2019) select a subset of features without taking
117 any information from any classification model. These methods select features that are
118 differentially expressed with respect to sample class labels. The filter methods rank individual
119 features according to their class discrimination power based on some statistical score function

120   and then select a number of high-ranked features to form a reduced and relevant feature subset.
121   The popular statistical score functions used in filter methods are Fisher's score, Signal to Noise
122   ratio (SNR), correlation coefficient, mutual information, Relief (Das 2019), etc. Filter methods
123   are computationally simple, fast, and unbiased in favor of any specific classifier as these methods
124   do not consider any knowledge from any classifier at the feature selection phase. The drawback
125   of filter methods is that the number of selected features is based solely on the trial-error method.
126   Wrapper methods (Chin 2016, Hambali 2020, Nada 2019), on the other hand, judge
127   discrimination capability of a feature subset using classification error rate or prediction accuracy
128   of a classifier as the feature evaluation function. It selects the most discriminative feature subset
129   via minimizing the classification error rate or maximizing the classification accuracy of a
130   classifier. The wrapper methods generally achieve better classification accuracy than the filter
131   methods because the selection of feature subset is classifier-dependent. One drawback of these
132   methods is that these are biased to used classifiers and another drawback is that these are
133   computationally more expensive than the filter methods as generation of the best feature subset
134   for the high-dimensional dataset is an NP-complete problem. Due to these reasons, these
135   methods are not applicable for high-dimensional datasets.
136   In Embedded methods (Chin 2016, Hambali 2020, Nada 2019), the optimal feature subset is
137   selected through the unique learning procedure of a specific classifier at the time of classifier
138   construction. Actually, in these methods, the optimal feature subset selection part is embedded as
139   part of classifier construction. These methods are faster than wrapper methods but are biased to
140   the specific classifier. In embedded approaches, the feature selection process is specific for a
141   particular classifier and is not applicable to other classifiers. These are also computationally
142   expensive. Due to these reasons for high-dimensional datasets, these methods are not applicable.
143   On the other hand, recently hybrid feature selection methods (Chin 2016, Hambali 2020, Nada
144   2019) are also developed. In hybrid methods, different category-based methods are combined to
145   take advantage of all of these methods for improving classification accuracy.
146   Apart from these methods, clustering techniques (Chin 2016, Hambali 2020) are also used for
147   feature selection purposes. Clustering techniques divide the data space in such a manner that
148   objects in the same cluster are similar while in different clusters they are dissimilar. For the
149   feature selection task, clustering methods (famous as attribute clustering in feature selection
150   domain) (Au 2005) divide the features into several distinct clusters and then reduce the feature
151   dimension by selecting a small number of significant features from each cluster. A lot of
152   unsupervised gene (attribute) clustering algorithms (Au2005, Chin 2016, Hambali 2020) are
153   already developed for this task. However, these methods are unsuccessful to find informative
154   functional groups of genes for sample classification as in clustering genes, no supervised
155   information from sample classes is considered (Au2005, Chin 2016, Hambali 2020). So,
156   scientists have developed a number of supervised gene (attribute) clustering algorithms (Dettling
157   2002, Hastie 2000, Hastie 2001, Maji 2012) in which genes are grouped using supervised
158   information from sample classes and a reduced gene set is formed via selecting the most
159   informative genes from each cluster.

160 All the above-mentioned variants deliver comparable feature selection and classification
161 accuracy. Quite often this type of classification models with only a few genes and with a limited
162 number of training samples can classify the majority of training samples correctly, but the
163 generalization capability of such classification models cannot be guaranteed (Bolo2012, Ghorai
164 2011, Nagi 2013, Wang 2006, Wang 2012, Yang 2010). So, the most important task for a
165 medical diagnosis system is to improve the classification accuracy of unknown samples
166 (generalization performance) which cannot be solved by this type of classification model.
167 Apart from this problem, the microarray data is related to several uncertainties due to fabrication,
168 hybridization, and image processing procedure in microarray technology. These uncertainties
169 introduce various types of noise in microarray data. Due to the presence of these uncertainties
170 with a limited number of training samples, the conventional machine learning approaches face
171 challenges to develop reliable classification models.
172 To overcome the above-mentioned problems, it is therefore essential to develop general
173 approaches and robust methods. In this regard, researchers are motivated to develop the second
174 category-based model. These are the different robust ensemble classification models (Bolo2012,
175 Ghorai 2011, Nagi 2013, Osareh 2013, Wang 2006, Wang 2012, Yang 2010) which can
176 overcome small sample size problems and are capable of removing uncertainties of gene
177 expression data.
178 Ensemble methods (Dietterich 2000) are a class of machine learning technique which combines
179 multiple base learning algorithms to produce one optimal predictive model. Ensemble
180 classification model refers to a group of individual/base classifiers that are trained individually
181 on the trained dataset in a supervised classification system and finally, an aggregation method is
182 used to combine the decisions produced by the base classifiers. These ensemble classification
183 models have the potential to alleviate the small sample size problem by applying multiple
184 classification models on the same training data or on bootstrapped samples (sampling with
185 replacement) of the training data to decrease the chance of overfitting in the training data. In this
186 way, the training dataset is utilized more efficiently, and as a consequence, the generalization
187 ability is improved.
188 Although different category-based ensemble classification models exist in the literature but these
189 ensemble models are not capable of addressing all the above-mentioned problems (small sample
190 size, high dimensional feature space, and sample class imbalance problem) related to microarray
191 data.
192 In this regard, here a new Multiple Filtering and Supervised Attribute Clustering algorithm-based
193 ensemble classification model named MFSAC-EC is proposed. In this model, first, a number of
194 bootstrapped versions of the original training dataset are created. At the time of the creation of
195 bootstrapped versions, an oversampling technique (Blaszczynski 2013) is adopted to solve the
196 class imbalance problem. For every bootstrapped dataset a number of sub-datasets (each with a
197 subset of genes) are generated using the proposed MFSAC method. The MFSAC is a hybrid
198 method combining multiple filters with a new supervised attribute clustering method. Then for
199 every sub dataset, a base classifier is constructed. Finally, based on the prediction accuracy of all

200   these base classifiers of all sub-datasets for all bootstrapped datasets an ensemble classifier (EC)
201   is formed using the majority voting technique.
202   The novelty of the proposed MFSAC-EC model is that here the emphasis is given
203   simultaneously on the high dimensionality problem of gene expression data, small sample size
204   problem as well as the class imbalance problem. All of these problems at the same time are not
205   considered in any existing ensemble classification model. First of all, due to the use of
206   bootstrapping method with a class balancing strategy, the proposed model can handle a small
207   sample size and overfitting problem. Secondly, in MFSAC, different filter methods are used with
208   their unique characteristics. So, different characteristics-based relevant gene subsets are selected
209   via different filters to form different sub-datasets from every bootstrapped dataset. Finally, every
210   gene subset is modified using a supervised attribute clustering algorithm. In this way, the high-
211   dimensionality problem of gene expression data is handled here. Apart from this, from the
212   MFSAC generated sub-datasets, the frequency of occurrence is counted for every gene and
213   informative genes are ranked accordingly. The prediction capability of the proposed model is
214   experimented with over different microarray datasets and compared with the existing well-
215   known models. Experimental outputs demonstrate the superiority of the proposed model over
216   existing models.

217   ## Materials & Methods
218   The proposed MFSAC-EC model is composed of different filter score functions, a new
219   supervised attribute clustering method, and an ensemble classification method. In the following
220   subsections, first, a brief overview is given on different filter score functions and then the
221   proposed MFSAC-EC model is described.

222   **Preliminaries**
223   In this paper, a data set (here, a microarray gene expression data set) is represented by a data
224   matrix,$K_{U \times V}$, with U data objects (samples) and V features (genes). The set of objects or
225   samples is represented as E = $\{E_1, E_2,.....,E_s,...E_U\}$ while the set of genes is represented as $G = \{$
226    $G_1,G_2,......,G_t,....G_V\}$. Here, each sample is a V-dimensional feature vector containing V number of
227   gene expression values. Similar way, every gene is a U-dimensional vector containing U number
228   of sample values. Here, $C_{U \times 1}$is a class vector representing the associated class label for every
229   sample. The class label is taken from a set $DC = \{d_1,d_2,.....,d_j,...d_N\}$ with N distinct class labels.

230   **Brief overview of Filter score functions used in MFSAC**
231   The filter score functions used in the proposed MFSAC-EC model are modified Fisher score (Gu
232   2011), modified T-test (Zhou 2007), Chi-square (Das 2019), Mutual information (Das 2019),
233   Pearson correlation coefficient (Leung 2010), SNR(Leung 2010) and Relief-F (Das 2019). A
234   summary of these 7 filters used in the MFSAC-EC model is given in the supplemental Table S1.

235   **Proposed MFSAC-EC Model**
236   In the proposed MFSAC-EC model, initially, bootstrapping (sampling with replacements) with a
237   class balancing procedure of samples is applied on training dataset K to create D number of
238   different bootstrapped versions from the training dataset. Here, every bootstrapped dataset with

239   U samples is formed by random sampling with replacements U times from the original dataset K.

240   After that oversampling procedure is applied to each minority class to achieve data balance.

241   Oversampling consists of increasing the minority class instances by their random replication to

242   exactly balance the cardinality of the minority and majority classes in each bootstrapped dataset.

243   Due to oversampling each bootstrapped dataset will contain more instances than the original

244   dataset.

245   The MFSAC method of the MFSAC-EC model, which is an integration of multiple filters and a

246   new supervised attribute (gene) clustering method, is applied on every newly created

247   bootstrapped (BK_l) training dataset. The proposed MFSAC method first calculates the class

248   relevance score of every gene present in the bootstrapped training dataset using each filter score

249   function (FT_x), $x = 1\ to\ 7$ mentioned above. Then for each filter score function, a sub dataset S

250   D_{lx} with a gene subset (GS_{lx}) is created by selecting a predefined number (let P) of the most

251   relevant genes from the full gene set G. So, $|GS_{lx}| = P$. After that on every gene subset (GS_{lx}) of

252   every sub dataset SD_{lx}, the SAC (Supervised Attribute Clustering) method is applied and a set of

253   clusters CGS_{lx} and corresponding cluster representatives (considered as modified features) are

254   formed. Finally, Q numbers of most relevant cluster representatives are selected as modified

255   features and a reduced sub dataset RSD_{lx} of the sub dataset SD_{lx} is formed. How the SAC

256   method works on GS_{lx} of every sub dataset SD_{lx} is discussed below.

257   For any sub dataset SD_{lx} , the SAC method starts by selecting the gene from the subset (GS_{lx})

258   with the highest FT_x value. Let gene $G_{li} \in GS_{lx}$ with the highest FT_x value be selected as the first

259   member (let $FT_x(G_{li}, C) = A$) and it also becomes the initial cluster representative $R(R = G_{li})$)

260   of the first cluster C_1GS_{lx} and G_{li} is deleted from GS_{lx}. In effect, $G_{li} \in C_1GS_{lx}$, and $GS_{lx} = GS_{lx}$

261    $- \{G_{li}\}$ and so $FT_x(R, C) = A$. This cluster is then grown up in parallel with the cluster

262   representative refinement process which is described next. In this process, the gene (let G_{lm}) with

263   next highest FT_x value is taken from GS_{lx} subset and is merged with the current cluster

264   representative R. The merging is done in two ways. Firstly the expression profile of $G_{lm}$is

265   directly added with R and a temporary augmented representative TR^+ is formed and its FT_x

266   value (let B_1) is calculated. The second one is that the sign-flipped value of the expression

267   profile of G_{lm} is added with R and another temporary augmented representative TR^- is formed

268   and its FT_xvalue (let B_2) is calculated. If $FT_x(TR^+, C) \geq FT_x(TR^-, C)$ that is $B_1 \geq B_2$ then

269   TR^+ is chosen else TR^- is chosen. Let TR^+ is chosen. Now if $FT_x(TR^+, C) > FT_x(R, C)$

270   then $R = TR^+$ otherwise, R is unaltered. Similar way if TR^- is chosen and if $FT_x(TR^-, C) >$

271    $FT_x(R, C)$ then $R = TR^-$ otherwise, R remains unchanged. If R is modified then the gene $G_{lm}$is

272   included in the cluster and G_{lm} is deleted from GS_{lx}. In effect, $G_{lm} \in C_1GS_{lx}$, and $GS_{lx} = GS_{lx} -$

273    $\{ G_{lm} \}$. So, the next chosen gene is included in the current cluster if it improves the class

274 relevance value of the current cluster representative. The merging process is described in Figure
275 1.
276 Here g0 represents the current cluster representative (R) and its class relevance score ((FT_x, R),
277 here Pearson score), is shown. Now among all the genes g1, g2, g3, g4, and g5, the Pearson score
278 of g1 is the highest. So, g1 is chosen for the merging process. Then g1 is added with R to create
279 the temporary augmented representative ($TR^+ = R + g1$) and also its sign-flipped value is
280 added with the R to form the temporary augmented representative ($TR^- = R - g1$). The Pearson
281 score of TR^+ is greater than the Pearson score of TR^-, so TR^+ is chosen. Now the Pearson
282 score of TR^+ is greater than the Pearson score of R, so TR^+ is considered as the current cluster
283 representative and $R = TR^+$. This process is continued for all other genes. Now, g3 is chosen as
284 it is the gene with the next highest Pearson value. g3 and its sign-flipped value are added
285 individually with current cluster representative R to form $TR^+ = R + g3$ and $TR^- = R - g3$
286 respectively. In this case, Pearson score of TR^- is greater than the Pearson score of TR^+. So,
287 TR^- is chosen. Then Pearson score of TR^- is Checked with the Pearson score of R and here
288 Pearson score of TR^- is greater than the Pearson score of R. So, TR^- is considered as current
289 cluster representative and $R = TR^-$. In this way, cluster representative is refined. This process is
290 repeated for every member of GS_{lx} subset.
291 After the formation of the first cluster and its corresponding augmented representative, R is
292 assigned to AR_{lx1} that means $AR_{lx1} = R$, and the supervised clustering process is repeated to
293 form the second cluster with the gene (let G_{lz}) with next highest FT_x value from GS_{lx} subset. In
294 this way a set of clusters $CGS_{lx} = \{C_1GS_{lx}, C_2GS_{lx}, ..., C_kGS_{lx}, ...\}$ and their corresponding
295 augmented cluster representatives $AR_{lx} = \{AR_{lx1}, ..., AR_{lxk}, ...\}$ are formed. After that Q
296 number of most powerful augmented cluster representatives are chosen (as modified features)
297 according to their FT_x value from the generated clusters and with these Q number of modified
298 features, a reduced sub dataset RSD_{lx} of sub dataset SD_{lx} is formed.
299 In this way, for every bootstrapped version (BK_l) of the training dataset, 7 number of RSD_{lx} sub-
300 datasets are created and for every RSD_{lx} an individual classifier is constructed using any existing
301 classifier and finally, an ensemble classifier (EC) is formed by combining all these classifiers of
302 all bootstrapped versions using the majority voting technique. To classify every sample using
303 this ensemble classifier, each classifier votes or classifies the sample for a particular class, and
304 the class for which the highest number of votes is obtained is considered as the output class.
305 **MFSAC method based Informative Attribute Ranking**
306 For every gene (feature/attribute), the frequency of occurrence that means the total number of
307 times it appears in all sub-datasets generated by the MFSAC method for all bootstrapped
308 versions is calculated. Then according to their frequency of occurrence, those genes are ordered
309 or ranked. The top-ranked genes with the highest occurrence frequency are considered the most
310 informative cancer-related genes.

311 The block diagram of the proposed MFSAC-EC model is shown in Figure 2, while the block
312 diagram of the MFSAC method is shown in Figure 3. The algorithm of the proposed model is
313 described below.
314

315 **Algorithm:** MFSAC-EC
316

317 **Input:** A $K_{U \times V}$ data matrix (here, gene expression data matrix) containing U number of data
318 objects (here, cancer samples) and V number of attributes (here, genes).
319 **Output:** An ensemble classifier MFSAC-EC is formed to classify test samples. From MFSAC
320 generated sub-datasets, informative genes are selected according to their rank. Every gene is
321 ranked according to its frequency of occurrence.
322 **Definitions:**
323 $E = \{E_1, E_2,....,E_s,...E_U\}$ is the set of objects or samples of $K_{U \times V}$ data matrix. Every sample E_s
324 is a V dimensional vector.
325 $G = \{G_1,G_2,......,G_t,....G_V\}$ is the set of features or genes of $K_{U \times V}$ data matrix. Every gene G_t is a
326 U dimensional vector.
327 $BK = \{BK_1, BK_2,...,BK_l,.....,BK_D\}$ is a set of the bootstrapped version of the original training
328 dataset. In every bootstrapped dataset the number of samples varies from the original dataset but
329 the number of features is the same as the original dataset.
330 $C_{U \times 1}$ is a class vector representing the associated class label for every sample. For a data matrix
331 N distinct class labels exist and class labels are taken from a set $DC = \{d_1,d_2,.....,d_k,...d_N\}$.
332 $FT_x(G_t, C)$ is x^{th} filter score function which returns the class relevance value of G_t gene with
333 respect to class vector C using FT_x score function, for $x = 1\ to\ 7$ as 7 represents the total
334 number of filtering score functions used here.
335 $GS_{lx} (GS_{lx} = P)$ is a set of top-ranked genes of G selected using FT_x score function and SD_{lx} is
336 corresponding sub dataset of BK_l. Here SD_{lx} is a data matrix containing P number of genes.
337 $CGS_{lx} = \{C_1GS_{lx}, C_2GS_{lx}, ..., C_kGS_{lx}, ...\}$ and $AR_{lx} = \{AR_{lx1}, ..., AR_{lxk},......\}$ are the set of
338 clusters and corresponding cluster representatives respectively generated from the corresponding
339 subset GS_{lx} of SD_{lx} . Here every AR_{lxk} is a vector.
340 TR^+, TR^-, R are vectors similar to a gene vector.
341 $RSD_l = \{RSD_{l1}, RSD_{l2},........,RSD_{lx},......RSD_{l7}\}$ is a set of sub-datasets each containing Q number
342 of most relevant cluster representatives formed for every bootstrapped dataset BK_l.
343 $CF_l = \{IC_{l1}, IC_{l2},........,IC_{lx},......IC_{l7}\}$ is a set of classifiers formed for every bootstrapped dataset.
344   1. Create D number bootstrapped version of training dataset K.
345   2. For Every bootstrapped dataset BK_l repeat step 3
346   3. Repeat for $x = 1\ to\ 7$
347   A. Repeat for $t = 1\ to\ V$

348   a) Calculate class relevance score $FT_x(G_t, C)$ of G_t gene, where $G_t \in G$, with

349   respect to class vector C

350   B. Select P number of top-ranked genes from G based on FT_x score function and

351   form GS_{lx} gene subset with corresponding SD_{lx} sub dataset

352   C. Set $k = 0$

353   D. Repeat until $GS_{lx} = \emptyset$

354   a) Set $k = k + 1$

355   b) Set $AR_{lxk} = 0$, $R = 0$, and $i = 0$

356   c) Select the gene (let G_{li}) whose FT_x score value is maximum among all

357   genes of GS_{lx} and set $R = G_{li}$

358   d) Add G_{li} to C_kGS_{lx}, and delete G_{li} from GS_{lx}

359   e) Set count =1

360   f) Repeat for $j = 1\ to\ |GS_{lx}|$

361   I.   Compute first augmented representatives TR^+ by adding

362        $G_{lj} \in GS_{lx}$ with R that means $TR^+ = R + G_{lj}$

363   II.   Compute second augmented representatives TR^- by

364   adding sign-flipped version of $G_{lj} \in GS_{lx}$ with R that

365   means $TR^- = R - G_{lj}$

366   III.   Compute class relevance value $FT_x(TR^+, C)$ and FT_x

367   (TR^-, C) using FT_x score function

368   IV.   If $FT_x(TR^+, C) \geq FT_x(TR^-, C)$then

369   If $FT_x(TR^+, C) > FT_x(R, C)$ then

370   • Set $R = R + G_{lj}$ and add G_{lj} to C_kGS_{lx} and

371   delete G_{lj} from GS_{lx}

372   • count = count +1

373   V.   If $FT_x(TR^-, C) > FT_x(TR^+, C)$ then

374   If $FT_x(TR^-, C) > FT_x(R, C)$ then

375   • Set $R = R - G_{lj}$ and add G_{lj} to C_kGS_{lx} and

376   delete G_{lj} from GS_{lx}

377   • count = count + 1

378   g) Set $R = R/count$

379   h) Set $AR_{lxk} = R$

380   E. Select Q number of most relevant cluster representatives according to FT_x score

381   from AR_{lx} set and form RSD_{lx} sub data set.

382   F. Construct a classifier C_{lx} for RSD_{lx} sub data set

383   4. Apply a test sample over all the classifiers of all bootstrapped dataset and calculate the
384   prediction accuracy of each classifier
385   5. Apply simple voting over all predictions to form an ensemble classifier EC and get final
386   prediction.
387   6. Calculate number of occurrences for every gene for all RSD_{lx} sub datasets across all
388   bootstrapped versions and rank them according to their count.
389   7. Select a number of top-ranked genes as informative genes.
390   8. End
391
392

393 **Description and Preprocessing of the Datasets**
394 The experimentation has been carried out over ten publicly available different gene expression
395 binary class and multi-class datasets. Among these datasets, eight datasets are cancer datasets
396 and two arthritis datasets. The eight cancer datasets are Leukemia (Golub1999), Colon (Alon
397 1999), Prostate (Singh 2002), Lung (Gordon 2002), RBreast (Veer 2002), Breast (West 2001),
398 MLL (Armstrong 2001), and SRBCT (Khan 2001). To show the accuracy of the proposed model
399 with respect to other than cancer datasets here two arthritis datasets RAHC (Pouw Kraan 2003)
400 and RAOA (Pouw Kraan 2007) are also considered. The summary of the datasets is represented
401 in Table 1.
402 In the Leukemia dataset (Golub1999), the gene expression data matrix is prepared using
403 Affymetrix oligonucleotide arrays. The original dataset consists of two datasets: the training
404 dataset and the testing dataset. The training dataset consists of 38 samples (27 Acute
405 Lymphoblastic Leukemia (ALL) and 11 Acute Myeloid Leukemia (AML)) while the test dataset
406 consists of 34 samples (20 Acute Lymphoblastic Leukemia (ALL) and 14 Acute Myeloid
407 Leukemia (AML)), each with 7129 probes from 6817 genes. For the Leukemia dataset, training
408 and test datasets are merged here and genes with missing values are removed and finally, the
409 dataset with 7070 genes and 72 samples is prepared.
410 In the Colon cancer dataset (Alon 1999), gene expression of 6500 genes for 62 samples is
411 measured using Affymetrix oligonucleotide arrays. Among these 62 samples, 40 are Colon
412 cancer samples and 22 are normal samples. Among these 6500 genes, 2000 genes are selected
413 based on the confidence of measured expression levels.
414 Prostate cancer dataset (Singh 2002) also consists of training and testing datasets. In the training
415 dataset, among 102 samples, 50 are normal samples and 52 are prostate cancer samples. In the
416 test dataset among 34 samples, 25 are prostate cancer samples and 9 are normal prostate samples.
417 Gene expression of every sample is measured with respect to 12600 genes using Affymetrix
418 chips. Here, training and test datasets are merged, and a dataset with 12600 genes and 136
419 samples is formed.
420 The Lung cancer dataset (Gordon 2002) consists of 181 samples. Among these samples, 31 are
421 malignant pleural mesothelioma and rest150 adenocarcinoma of lung cancer. Each sample is

422   represented by 12533 genes and the gene expression of every sample is measured using
423   Affymetrix human U95A oligonucleotide probe arrays.
424   In Rbreast data set (Veer 2002), the patients, who are considered as breast cancer patients after 5
425   years intervals of initial diagnosis, fall under the category of relapse and rest as no relapse of
426   metastases. 97 samples have been provided in which 46 patients developed distance metastases
427   within 5 years and they are considered as relapse while the remaining remained healthy and are
428   labeled as non-relapse. This dataset comprises 24481 genes and among them, 293 are removed.
429   In the Breast cancer dataset (West 2001), the gene expression of 49 samples is measured using
430   HuGeneFL Affymetrix microarray arrays. Breast tumors are positive or negative in the presence
431   or absence of estrogen receptors (ER). In this dataset, 25 samples are ER+ tumors and 24
432   samples are ER- tumors.
433   MLL (Armstrong 2001) is a type of dataset which comprises of training data set of 57 leukemia
434   samples including 20 ALL, 17 MLL, and 20 AML and the test dataset including 4 ALL, 3 MLL,
435   and 8 AML samples. For MLL cancer dataset training and test, datasets are merged here and
436   finally, the dataset with 12582 genes and 72 samples are prepared.
437   SRBCT dataset (Khan 2001) is introduced as a dataset comprising of gene-expression for
438   identifying small round blue-cell tumors of childhood SRBCT and samples of this dataset are
439   further divided into four class which are neuroblastoma, rhabdomyosarcoma, non-Hodgkin
440   lymphoma, and Ewing family of tumors and they are obtained from cDNA microarrays. A
441   training set consisting of 63 SRBCT tissues, a test set consisting of 20 SRBCT and 5 non-
442   SRBCT samples are available. Here we have considered only the training dataset. Each tissue
443   sample is already standardized to zero mean value and has a unit variance across the genes.
444   RAHC commonly known as Rheumatoid Arthritis versus Healthy Controls is a data set (Pouw
445   Kraan 2003) which comprises of gene expression characterizing as peripheral blood cells of 32
446   patients with RA, 3 patients with probable RA, and 15 age with sex-matched healthy controls
447   performed under microarrays with a complexity of 26000 unique genes of 46000 elements.
448   RAOA commonly known as Rheumatoid Arthritis versus Osteoarthritis is a dataset (Pouw Kraan
449   2007) that includes the gene expression of thirty patients in which 21 of them are with RA and
450   the remaining 9 of them are with OA. The Cy5 labeled experimental cDNA and Cy3 labeled
451   common reference sample were pooled and hybridized to the lymphochips (consisting of 18000
452   cDNA spots which symbolize immunology in the genes of relevance).

453   Results

454   To assess the performance of the proposed MFSAC-EC model, four well-known existing
455   classifiers named K-Nearest Neighbor (Duda 1999), Naive Bayes (Duda 1999), Support vector
456   machine (Vapnik 1995), and Decision tree(c4.5) (Duda 1999) are applied independently in this
457   model and four different ensemble classification models are formed. To prove the superiority of
458   the proposed model, it is compared with existing well-known filter methods (used here) and
459   existing recognized gene selection methods (Ding 2005, Au2005, Maji2005) and also with
460   different existing ensemble classifiers (Bolo2012, Nagi 2013, Osareh 2013, Wang 2006, Wang
461   2012). To analyze the performance, the methods are applied to different publicly available

462 cancer and other disease-related gene expression datasets. The major metrics used here for
463 evaluations of the performance of the proposed classifier are the Cross-validation method
464 (LOOCV, 5-fold, and 10-fold), ROC Curve, and Heat map.

465 **Tools Used**

466 The algorithms are implemented using Python programming language and Scikit-learn libraries
467 (Pedregosa 2011) which are explained in (Komer 2014) for ML algorithms. The programs are
468 executed on an online Colab platform with 12 GB RAM and Intel(R) Xeon(R) processor
469 available in the "CPU" Runtime Type at the time of writing. Figures and tables are generated in
470 the Matplotlib library (Hunter 2007) and also in Microsoft Excel. The python codes used here are
471 available at https://github.com/NSECResearchCD-SLB/PEERJ_MFSAC_EC.
472 In the following subsections, first, the different types of metrics used here are discussed, and then
473 the performance of the proposed MFSAC-EC model is verified with respect to these metrics.
474 This is followed by comparing the classification performance of the proposed model with
475 different existing methods in terms of 10-fold cross-validation. The proposed model does not
476 only perform the task of classification but also ranks every attribute or gene in descending order
477 based on its information present in the dataset. To show the effectiveness of this ranking
478 procedure topmost eight genes from Colon cancer and Leukemia cancer datasets are represented
479 with their corresponding names, symbols, and references in significant cancer-related journals to
480 demonstrate their significant roles in these cancers.

481 **Evaluation Metrics**

482 The performance of the proposed MFSAC-EC classifier is established with respect to the
483 following measures.

484 **Cross-Validation method**

485 The first well-known metric used here to evaluate the classification model performance is the k-
486 fold cross-validation method (Wang 2012). In the k-fold cross-validation method, the dataset is
487 randomly divided into k number of folds and k-1 folds are used for training and one fold is used
488 for testing. The process is repeated for k number of times and average classification accuracy is
489 taken. When k is set at 1 that means the fold size is equal to the size of the dataset (training
490 dataset size is equal to one less than the number of samples in the dataset and validation is done
491 using the remaining sample) then it is considered as Leave one out cross-validation method
492 (LOOCV). For k is equal to 2, the cross-validation method is named the household method. It
493 has been found that when k is set at a very small value that means the fold size is large then the
494 accuracy of the classification model is affected by low bias and high variance problems. On the
495 other hand, if k is set at a high value that means the fold size is not so large then the classification
496 accuracy of the classification model has a high bias but low variance. It has been found that 10-
497 fold cross-validation method outperforms the LOOCV method (Breiman 1992, Ambroise 2002,
498 Asyali 2006) and it has been also endorsed that the 10-fold cross-validation method as a better
499 measure for classification.
500 In training-testing random splitting the dataset is initially randomly partitioned into training set
501 ($2/3^{rd}$ of the dataset) and testing set ($1/3^{rd}$ of the dataset) with 50 runs.

502 **ROC curve analysis**

503 The performance of the proposed classifier for two-class datasets is also judged using Receiver
504 Operator Characteristic (ROC) analysis (Wang 2012). It is a visual method for evaluating binary
505 classification models. Under this analysis, the following measures are considered to judge the
506 binary classification model.

507 Classification accuracy (Acc) is defined as,

508 $$Acc = \frac{TP + TN}{TP + FP + TN + FN} \qquad\qquad 0 \leq Acc \leq 1$$

509 The sensitivity (SN) or True Positive Rate (TPR) can be defined as,

510 $$SN = TPR = \frac{TP}{TP + FN}$$

511 The specificity (SP) or True Negative Rate (TNR) can be defined as,

512 $$SP = TNR = \frac{TN}{TN + FP}$$

513 The False Positive Rate (FPR) can be defined as:

514 $$FPR = (1 - specificity) = \frac{FP}{FP + TN}$$

515 The Positive Predicted Value (PPV) can be defined as:

516 $$PPV = \frac{TP}{TP + FP}$$

517 The Negative Predicted Value (NPV) can be defined as:

518 $$NPV = \frac{TN}{TN + FN}$$

519 Where TP, TN, FP, FN are true positive, true negative, false positive, and false negative
520 respectively.

521 The ROC curve is plotted considering TPR along the y-axis and FPR along the x-axis. The area
522 under the ROC curve (AUC) is used to represent the performance of the binary classification
523 model. The higher AUC value of a ROC curve for a particular classification model signifies the
524 better performance of the classification model in differentiating positive and negative examples.
525 The range of AUC value is 0<=AUC<=1.

526 **Heat map analysis**

527 A heatmap is a data representation diagram in which the values for a variable of interest are
528 portrayed using a data matrix. In this data matrix, the values of the variable are represented
529 across two-axis variables as a grid of colored squares. The axis variables are divided into ranges
530 and each cell's color represents the intensity of that variable for the particular ranges of values of
531 axis variables.

532 Here, the performance of the proposed classifier for multi-class datasets is judged using Heat
533 map representation of confusion matrix (Liu 2014), where a confusion matrix is a tabular
534 representation to visualize the performance of a classification model in terms of true positive,
535 true negative, false positive and false negative.

536 **Parameter Estimation**

537 Before running the MFSAC-EC, the parameters are settled down. In MFSAC-EC the input
538 training dataset is bootstrapped. The proposed MFSAC-EC model is run here varying the number
539 of bootstrapped datasets (D) from 5 to 30 and the classification accuracy of this model is more or
540 less the same from 10 to the rest of the range. So, the number of bootstrapped datasets for every
541 training dataset for this model is set at 10.
542 In MFSAC method, initially P number of genes is selected by each filter method. Here in Table
543 2, the classification accuracy of the proposed model is shown with respect to different values of
544 P. From Table 2, it has been found that the proposed model gives the best result for P=100 for
545 RAOA and RAHC datasets. In case of Breast cancer, Lung cancer, MLL and SRBCT datasets it
546 gives the best result at P = 200. For Leukemia datasets it gives the best result at P= 500. So, it
547 can be said that MFSAC-EC gives best result for P value within 200 to 500 in all cases for all
548 datasets except Colon and Prostate. In Colon and Prostate, it shows the best result for $P = 1500$.
549 Here we have used SVM, DT (C4.5), NB, and KNN classifiers individually for forming different
550 ensemble classification models. All the classifiers are implemented using Scilit-learn libraries of
551 Python. For all classifiers, we have set parameters with default parameter values. For DT as
552 default setting we have used splitting function = Gini, Splitting criterion = best, height = none (
553 that means for every sample it reaches a leaf/class node). For SVM, we have used the RBF
554 kernel function. For KNN we have chosen K (number of nearest neighbor) value from 3 to 7.
555 The overall execution time of a single run of the MFSAC-EC model (considering bootstrapped
556 dataset creation, feature selection using MFSAC, and then generating classification accuracy of
557 test samples using LOOCV, 5-fold, 10-fold, and random splitting) and testing time using only
558 10-fold are shown for different datasets in Table 3.
559
560 **Classification Performance of the Proposed MFSAC-EC Classifier**
561 In Table 4, using the LOOCV method, the classification accuracy of our proposed MFSAC-EC
562 model is 100% for different datasets (Leukemia, Breast, RBreast, Lung, RAOA, and RAHC) for
563 all cases. In the Prostate dataset, we did not get 100% accuracy using our model with respect to
564 any type of existing classifier. In MLL, Colon, and SRBCT it also gives 100% accuracy using all
565 types of ensemble classifiers.
566 In Table 5 and Table 6, it has been shown that using 5 -fold and 10-fold cross-validation,
567 MFSAC-EC does not provide 100% accuracy only for Colon and Prostate cancer datasets. For
568 other datasets, it provides 100% accuracy with respect to all types of ensemble classifiers.
569 To show the generalization property of the proposed ensemble classifiers, the classification
570 accuracy of these classifiers is also measured repeatedly with respect to the random splitting of
571 the dataset into a training set (2/3 data of original dataset) and test set (1/3 data of original
572 dataset). Random splitting is done with care such that class proportion is alike in the training set
573 and test set. In Table 7, the classification accuracy of the above mentioned four different types of
574 ensemble classifiers for the different number of cluster representatives is shown in different
575 datasets which are based on the best result of 50 random splitting of the dataset into a training set
576 (2/3 data of original dataset) and test set (1/3 data of original dataset).

577   From the results of Table 4 to Table 7, it has been observed that classification accuracy in the
578   LOOCV method, 5-fold cross-validation, and 10-fold cross-validation methods is higher than the
579   random splitting of the dataset, and the overall generalization performance of the proposed
580   classification model is also good.
581   The performance of the proposed model for different two-class datasets with respect to different
582   parameters like SN, SP, PPV, NPV, FPR is shown in Table 8. From this table, it is found that the
583   performance of the proposed model is very good with respect to all these parameters for all two-
584   class datasets.
585   In Figures 4, the ROC curve is shown for different two-class datasets. In Figures 4a, 4b, and 4c,
586   the ROC curves are shown for Breast cancer using LOOCV, for Colon cancer using 5-fold cross
587   validation, and for RAHC dataset using 10-fold cross-validation respectively. The ROC curves
588   for Leukemia Cancer, and Lung cancer datasets using LOOCV are given in Supplemental
589   Figures F1a and F1b respectively. For Breast cancer, Leukemia cancer, and Lung cancer, the
590   AUC value is equal to 1.0 in every case. The ROC curves are shown for RAOA, and RBreast
591   cancer datasets using 5-fold cross-validation in Supplemental Figures F2a, and F2b respectively.
592   For these datasets also the prediction accuracy using 5-fold cross validation is very high
593   according to the AUC value. In Supplemental Figures F2c, the ROC curves are shown for
594   Prostate cancer using 10-fold cross-validation. From these curves of 10-fold cross validation, it
595   may be seen that except for Prostate cancer, for all other datasets the AUC value is 1 and for
596   Prostate cancer, the AUC value is close to 1.
597   In Figures 5a and 5b, heatmap representation of the confusion matrix are shown for multi-class
598   datasets: SRBCT and MLL with respect to 5-fold cross-validation, and 10-fold cross-validation
599   respectively. From these figures, it is clear that for the proposed model prediction accuracy is
600   accurate in most cases.
601
602   **Comparison of MFSAC-EC Model with Well-Known Existing Filter Methods used**
603   **in this model**
604   In Supplemental Figure F3, the proposed MFSAC-EC model in combination with different
605   existing classifiers is compared with different filter methods used in this model with respect to
606   SRBCT, RAHC, Prostate, and Colon datasets in terms of 10-fold cross-validation. In all cases,
607   the performance of the proposed model is significantly better with respect to all filters.
608
609   **Comparison of MFSAC-EC Model with Well-Known Existing Gene Selection**
610   **Methods**
611   In Figure 6, the MFSAC-EC model with different existing classifiers as base classifiers are
612   compared with existing well-known supervised gene selection methods named mRMR
613   (minimum redundancy maximum relevance framework) (Ding 2005), MSG (mutual information
614   based supervised gene clustering algorithm) (Maji 2012), CFS (Correlation-based Feature
615   Selection) (Ruiz 2006), and FCBF (Fast Correlation-Based Filter) (Ruiz 2006) with respect to

616   different classifiers using 10-fold cross-validation method. From these results, it has been found
617   that the proposed model outperforms in most of the cases.
618   In Figure 7, the MFSAC-EC model is compared with well-known existing unsupervised gene
619   selection methods named MGSACO (Tabakhi 2015), UFSACO (Tabakhi 2014), RSM(Lai
620   2006), MC (Haindl 2006), RRFS (Ferreira 2012), TV (Theodoridis 2008), and LS (Liao 2014)
621   with respect to DT, SVM, NB classifiers using random splitting method. From these results, it
622   can be said that the MFSAC-EC model outperforms in all cases.
623
624   **Comparison of MFSAC-EC Model with Well-Known Existing Ensemble**
625   **Classification and DEEP learning Models**
626   In Table 9, the proposed MFSAC-EC model using the DT classifier is compared with well-
627   known existing ensemble classification models with respect to 10-fold cross-validation. These
628   models are PCA-basedRotBoost (Osareh 2013), ICA-based RotBoost (Osareh 2013), AdaBoost
629   (Osareh 2013), Bagging (Osareh 2013), Arcing (Osareh 2013), Rotation Forest (Osareh 2013),
630   EN-NEW1 (Wang 2006), and EN-NEW2 (Wang 2006). From Table 9, it is clear that the
631   proposed model using DT classifier outperforms in all cases.
632   In Table 10, the proposed MFSAC-EC model using DT, NB, KNN as base classifiers are
633   compared with different existing ensemble classifiers with respect to 10-fold cross-validation.
634   These classifiers are Bagging based ensemble classifier (Nagi 2013), Boosting based ensemble
635   classifier (Nagi 2013), Stacking based ensemble classifier (Nagi 2013), Heuristic breadth-first
636   search-based ensemble classifier (HBSA) (Wang 2012), Sd_Ens (Nagi 2013), and Meta_Ens
637   (Nagi 2013). In Table 11 our model using SVM and KNN as base classifiers is compared with
638   auto-encoder-based deep learning models (Nabendu 2020) in terms of random splitting. Here,
639   results are shown only for the datasets for which results are available in the literature, and all
640   other fields are marked as "Not Found". In all cases, the MFSAC-EC model outperforms all the
641   well-known existing ensemble models (except for the Colon cancer dataset) and deep learning
642   models which in turn validates the usefulness of the proposed model.
643
644   **Biological Significance Analysis**
645   The top 8 genes selected by the MFSAC-EC model for Colon cancer and Leukemia are listed in
646   Table 12. For every gene, the name and symbol of the gene as well as the Accession number of
647   the Affymetrix chip are listed. Apart from this information, to validate those genes, biomedical
648   literature of the genes is searched and for every gene, the corresponding reference about its role
649   and significance for a particular disease is provided.
650

Discussion

652   In this paper, a new Multiple Filtering and Supervised Attribute Clustering algorithm-based
653   ensemble classification model named MFSAC-EC is proposed. The main motivation behind this
654   work is to develop a machine learning-based ensemble classification model to overcome the
655   over-fitting problem which arises due to the presence of sample class imbalance problem, small

656 sample size problem, and also high dimensional feature set problem in the microarray gene
657 expression dataset, to enhance the prediction capability of the proposed model.
658 Nowadays, in designing machine learning models, the use of ensemble methodology has been
659 increasing day by day as it incorporates multiple learning algorithms and also training datasets in
660 different efficient manners to improve the overall prediction accuracy of the model. Due to the
661 inclusion of prediction accuracy of multiple learning models and also the use of different
662 bootstrapping datasets, the chances of potential overfitting in training data is greatly reduced in
663 the ensemble models, and as a consequence the prediction accuracy increases. One necessary
664 condition of the superior performance of an ensemble classifier with respect to its individual
665 member/base classifier is that every base classifier should be very accurate and diverse (Osareh
666 2013). A classifier is considered accurate if its generalization capability is high and two
667 classifiers satisfy diverse property if their prediction in classifying the same unknown samples
668 vary from each other. The general principle of ensemble methods is to rearrange training datasets
669 in different ways (either by resampling or reweighting) and build an ensemble of base classifiers
670 by applying a base classifier on every rearranged training dataset (Osareh 2013).
671 In our proposed ensemble model, at first, a number of bootstrapped datasets of the original
672 training dataset is created. In every bootstrapped dataset, the class imbalance problem is solved
673 using the oversampling method. Then for every bootstrapped dataset, a number of sub-datasets
674 are created using the MFSAC method (which is a hybrid method combining multi-filters and a
675 new supervised attribute/gene clustering method), and then for every generated sub dataset, a
676 base classifier is constructed using any existing classification model. After that, a new ensemble
677 classifier (EC) is formed using the majority voting scheme by combining the prediction accuracy
678 of all those base classifiers.
679 The prediction accuracy of the proposed model is verified by applying it to high-dimensional
680 microarray gene expression data From Figure 6, and Figure 7 it has been found that the
681 classification accuracy of the MFSAC-EC model is much better than the well-known existing
682 gene selection methods. From Table 9, Table 10, and Table 11, it has been also found that the
683 proposed MFSAC-EC classification model is superior to the existing ensemble classification
684 models in almost every case. The superior performance of the proposed model is due to the
685 following reasons:
686   • The generation of the different bootstrapped versions of training data and also the use of
687   the oversampling procedure to balance the cardinality of majority class and minority class
688   in every bootstrapped dataset reduces the chances of the overfitting problem of a
689   classifier.
690   • Different types of filter methods are used in the MFSAC method. It has been already
691   observed that one filter gives better performance for one dataset while the same gives
692   poor results for other datasets. This is because every filter uses separate metrics and so
693   the choice for a filter for a specific dataset is a very complex task. As different filter
694   methods are used in the MFSAC method, so different sub-datasets with different
695   characteristics-based attributes/genes are formed from each dataset. This is shown using

696	Venn diagram in Supplemental Figures F4a and F4b. Here for Leukemia and Prostate
697	cancer datasets, the first twenty genes, selected by each filter are shown. In case of
698	Leukemia dataset, Relief measure generates non-overlapping gene subset while using
699	other filter metrics presence of a small number of overlapping genes in different gene
700	subsets are observed. In Prostate cancer dataset, Relief generates non-overlapping gene
701	subset and also maximum number of genes are non-overlapping in gene subsets formed
702	by Fisher score, MI (mutual information). From these figures, it is clear that using
703	different filter methods different subsets of genes are selected and different sub datasets
704	are formed. It shows diversity of those filter methods. As a consequence, the base
705	classifiers prepared on these diverse datasets are become diverse. This diversity increases
706	the power of ensemble classifier.
707	• Moreover, the genes selected by different filter methos are good biomarker also. In Table
708	12, the top ranked 8 genes selected by MFSAC-EC model are shown for Leukemia and
709	Colon cancer datasets. Among these genes, gene MPO (with column number 1720),
710	CST3 (with column number 1823), ZYX (with column number 4788), CTSD (with
711	column number 2062), CD79A/MB-1(with column number 2583), LYZ (with column
712	number 6738) in Leukemia dataset are important biomarkers as these are selected by
713	different filter methods mentioned in Supplemental Figure F4.
714	• In MFSAC, at first, a sub dataset of the most relevant genes is selected by each filter
715	method. Then on each sub dataset, the proposed supervised gene clustering algorithm is
716	applied and a reduced sub dataset of modified attributes/features in the form of
717	augmented cluster representatives is generated. In this method, at the time of cluster
718	formation, genes are augmented based on their supervised information. In other words,
719	such augmentation is considered where it increases the class discrimination power. Thus
720	effectively, the class relevance of any augmented cluster representative is greater than
721	that of any single gene involved in that process. So, this modified sub dataset containing
722	a reduced feature set in the form of augmented cluster representatives is more powerful
723	according to class discrimination power than the sub dataset containing a subset of the
724	most relevant genes. Apart from this, it is well known fact in gene expression data that
725	two genes are functionally similar if they are pattern-based similar (either positively co-
726	expressed or negatively co-expressed) (Das 2016). So, at the time of the augmentation
727	procedure, two types of augmentations are considered here. One is that a gene is added
728	with its original value with the current cluster representative and another one is that the
729	gene is added with its sign-flipped value with the current cluster representative. This is
730	because if the current cluster representative and a gene are positively co-expressed then
731	normal addition is considered but if they are negatively co-expressed then normal
732	addition will hamper the addition process and in that case, sign-flipping of that gene will
733	give proper result. The effect of augmentation with respect to every filter method is
734	shown in Figure 8. In Figure 8, for the Breast cancer dataset, at the time of supervised
735	cluster formation from each filter generated subset, the original gene, and its

736 corresponding class relevance value, and also augmented gene and its corresponding
737 class relevance are shown. From Figure 8, it is clear that for every filter method the class
738 relevance score of every original gene is increased with respect to that filter after
739 augmentation. In Figure 8, different class labels are distinguished by different colors.
740 • Finally, for each sub dataset with modified attributes in the form of augmented cluster
741 representatives, a classifier is constructed using any existing classifier, and these
742 classifiers are combined using the majority voting technique to form an ensemble
743 classifier (EC). The use of different sub-datasets with optimal gene subsets in the form of
744 augmented cluster representatives and the formation of a classifier for every sub dataset
745 can solve the overfitting problem of any single classifier. This is due to the reason that
746 not all sub-datasets can consistently perform well on all types of cancer datasets (due to
747 inherent characteristics of the datasets), but due to the use of majority voting in ensemble
748 classifiers, this problem can be solved or reduced.
749 Another outcome of our proposed model is to rank informative genes for every cancer dataset.
750 For this task, the frequency of occurrence of each gene present in the form of augmented cluster
751 representatives in every sub dataset is counted and these genes are ranked according to the
752 counted value to measure the importance of those genes for any specific disease, here cancer. To
753 establish the biological significance of those selected genes for every cancer dataset, their
754 contribution has been confirmed by other existing studies where they are referred already. From
755 these existing studies, it is clear that the selected genes are important for cancer class
756 discrimination and also are important as cancer biomarkers for molecular treatment targets.
757

Conclusions

759 Many machine learning and statistical learning-based classifiers for sample classification already
760 exist in the literature, but these methods are prone to suffer from overfitting due to small sample
761 size problems, class imbalance problems, and the curse of the high dimensionality of microarray
762 data. Although some of the existing methods can mitigate these issues to quite an extent, the
763 problems have still not been satisfactorily overcome. Due to this reason, here a novel feature
764 selection-based ensemble classification model named MFSAC-EC is proposed. It has been
765 shown that the proposed model can handle the above-mentioned issues present in existing
766 models. To check the performance of the proposed MFSAC-EC model, this classifier is applied
767 to test sample classification accuracy in high dimensional microarray gene expression data, a
768 domain that will be beneficial in the field of cancer research. From the experimental results, it
769 has been found that the proposed model outperforms all other well-known existing classification
770 models combined with the different recognized feature selection methods and also the newly
771 developed ensemble classifiers for all types of cancer datasets mentioned here. Apart from this
772 classification task, the proposed model can also rank informative attributes according to their
773 importance. The efficiency of the proposed model in this task is vindicated by finding the most
774 informative genes for Colon cancer and Leukemia cancer datasets using this model. These genes
775 are biologically validated based on other well-known existing studies. Consequently, it is clear

776 that the selected genes are vital for sample class discrimination and are also important
777 biomarkers for molecular treatment targets of deadly diseases.
778

779 **Acknowledgments**

780
781

782 **References**

783 Alon U, Barkai N, Notterman D. A., Gish K., Ybarra S., Mack D., and Levine A. J. 1999. Broad patterns of gene
784 expression revealed by clustering analysis of tumor and normal colon tissues probed by Oligonucleotide arrays.
785 Proceedings of National Academy of Sciences, USA. 96(12): 6745–6750.
786 Ambroise C, McLachlan GJ. 2002. Selection bias in gene extraction on the basis of microarray gene-expression
787 data. Proceedings of National Academy of Sciences, USA. 99(10):6562–6566.
788 Armstrong S, Staunton J, Silverman L, Pieters R, den Boer M, Minden M, Sallan S, Lander E, Golub T and
789 Korsmeyer S. 2001. MLL translocations specify a distinct gene expression profile that distinguishes a unique
790 leukemia. Nature Genetics. 30:41–47.
791 Asyali MH, Colak D, Demirkaya O, Inan MS. 2006. Gene expression profile classification: A review. Current
792 Bioinformatics. 1(1):55–73.
793 Austin H Chen, Yin-Wu Tsau and Ching-Heng Lin. 2010. Novel methods to identify biologically relevant genes for
794 leukemia and prostate cancer from gene expression profiles. BMC Genomics.
795 Au W. H., Chan K. C. C., Wong A.K.C., and Wang Y. 2005. Attribute clustering for grouping, selection,
796 classification of gene expression data. IEEE/ACM Transactions of Computational Biology and Bioinformatics.
797 2(2):83–101.
798 Bai YX, Yi Ji-Lin, Li Jian-Feng, Hong Sui. 2007 Clinicopathologic significance of BAG1 and TIMP3 expression in
799 colon carcinoma. World Journal of Gastroenterology. 13(28): 3883-3885.
800 Błaszczyński Jerzy, StefanowskiŁ Jerzy, Idkowiak ukasz.2013. Extending Bagging for Imbalanced Data.
801 Proceedings of the 8th International Conference on Computer Recognition Systems (CORES 2013).
802 Bolo´ n-Canedo V, Sánchez-Maroño Noelia, Alonso-Betanzos Amparo. 2012. An ensemble of filters and classifiers
803 for microarray data classification. Pattern Recognition. 45(1):531-539.
804 Botchkina Inna L, Roweh Rebecca A, Rivadeneira David E, Karpeh Martin S Jr, Crawford Howard, Dufour
805 Antoine, Ju Jingfang, Wang Yuan, Leyfman Yan, Botchkina Galina I. 2009. Phenotypic Subpopulations of
806 Metastatic Colon Cancer Stem Cells: Genomic Analysis. Cancer Genomics & Proteomics. 6(1): 19-30.
807 Breiman L, Spector P. 1992. Submodel selection and evaluation regression – the X-random case. International
808 Statistical Review. 60(3):291–319.
809 Brown Adam R, Simmen C.M. Rosalia, Raj Vinay R., Van Trang T., MacLeod Stewart L, Simmen Frank A. 2015.
810 Krüppel-like factor 9 (KLF9) prevents colorectal cancer through inhibition of interferon-related signaling.
811 Carcinogenesis. 36(9): 946–955.
812 Chen Xiao, Gao Bing, Ponnusamy Murugavel, Lin Zhijuan, Liu Jia. 2017. MEF2 signaling and human diseases.
813 Oncotarget. 8(67):112152–112165.
814 Chin A J, Mirzal A, Haron H, Hamed H N A. 2016. Supervised, Unsupervised and Semi-supervised   Feature
815 Selection: A Review on Gene Selection. IEEE Transactions on Computational Biology and Bioinformatics.
816 13(5):971-989.
817 Colozza M, Cardoso F, Sotiriou C, Larismont D, Piccart M J. 2005. Bringing molecular prognosis and prediction to
818 the clinic. Clin Breast Cancer. 6(1):61-76.
819 Das Chandra, Bose Shilpi, Chattopadhyay Matangini, Chattopadhyay Samiran. 2016. A novel distance-based
820 iterative sequential KNN algorithm for estimation of missing values in microarray gene expression data.
821 IJBRA.12(4):312.
822 Das Chandra, Bose Shilpi, Banerjee Abhik, Dutta Sourav, Ghosh Kuntal, Chattopadhyay Matangini. 2019.
823 Comparative Performance Analysis of Different Measures to Select Disease Related Informative Genes from

824   Microarray Gene Expression Data. International Conference on Innovation in Modern Science and Technology
825   (ICIMSAT-2019), Springer.
826   Dashtban M. and Balafar M. 2017. Gene selection for microarray cancer classification using a new evolutionary
827   method employing artificial intelligence concepts. Genomics.109(2): 91-107.
828   Dettling M and Buhlmann P. 2002. Supervised Clustering of Genes. Genome Biology. 3(12): 0069.1-0069.15.
829   Dieterich T. G.. 2000. An experimental comparison of three methods for constructing ensembles of decision trees:
830   bagging, boosting, and randomization. Machine Learning. 40(2): 139–157.
831   Ding C. and Peng H. 2005. Minimum redundancy feature selection from microarray gene expression data. Journal of
832   Bioinformatics and Computational Biology. 3(2): 185–205.
833   Duda R. O., Hart P. E., and Stork D. G.1999. Pattern Classification and Scene Analysis. NewYork:Wiley.
834   Durai Rajaraman, Yang Y Shi, Seifalian M Alexander, Goldspink Geoffrey, Winslet C Marc. 2007. Role of
835   insulin-like growth factor binding protein-4 in prevention of colon cancer. World Journal of Surgical Oncology.
836   5:128.
837   Elyasigomari V, Lee D. A., Screen H. R.C, Shaheed M.H. 2017. Development of a two-stage gene selection method
838   that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer
839   classification. Journal of Biomedical Informatics. 67:11-20.
840   Feng Hailiang, Liu Yanyan, Bian Xiaocui, Zhou Fangying, Liu Yuqin. 2018. ALDH1A3 affects colon cancer in
841   vitro proliferation and invasion depending on CXCR4 status. British Journal of Cancer. 118: 224–232.
842   Ferreira A.J., Figueiredo M.A.T. 2012. An unsupervised approach to feature discretization and selection, Pattern
843   Recognition. 45: 3048-3060.
844   Furey T.S., Cristianini N, Duffy N, Bednarski D.W, Schummer M, and Haussler D. 2000. Support Vector Machine
845   Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. Bioinformatics. 16(10):
846   906-914.
847   Gardina Paul J, Clark Tyson A, Shimada Brian, Staples Michelle K, Yang Qing, Veitch James, Schweitzer Anthony,
848   Awad Tarif, Sugnet Charles, Dee Suzanne, Davies Christopher, Williams Alan, Turpaz Yaron. 2006. Alternative
849   splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC
850   GENOMICS. 7: 325
851   Ghorai Santanu, Mukherjee Anirban, Sengupta Sanghamitra, Dutta Pranab K.. 2011. Cancer Classification from
852   Gene Expression Data by NPPC Ensemble. IEEE/ACM Transactions on Computational Biology and
853   Bioinformatics. 8(3).
854   Giorgio Eros Di, Hancock Wayne W, Brancolini Claudio. 2018. MEF2 and the tumorigenic process, hic sunt leones.
855   BBA - Reviews on Cancer. 1870(2):261-273.
856   Golub T. R., Slonim D. K., Tamayo P, Huard C, Gaasenbeek M, Mesirov J.P., Coller H, Loh H.L., Downing J. R.,
857   Caligiuri M. A., Bloomfield C. D., and Lander E. S.. 1999. Molecular Classification of Cancer: Class Discovery and
858   Class Prediction by Gene Expression Monitoring. Science. 286(5439):531-537.
859   Gordon G. J., Jensen R. V., Hsiao L.-L., Gullans S. R., Blumenstock J. E., Ramaswamy S., Richards W. G.,
860   Sugarbaker D. J., and Bueno R. 2002. Translation of microarray data into clinically relevant cancer diagnostic tests
861   using gene expression ratios in lung cancer and mesothelioma. Cancer Research. 62: 4963–4967.
862   Greller LD, Tobin FL. 1999. Detecting selective expression of genes and proteins. Genome Research. 9:282–296.
863   Gu Q., Li Z., Han J.. Generalized Fisher Score for Feature Selection. 2011. UAI'11: Proceedings of the Twenty-
864   Seventh Conference on Uncertainty in Artificial Intelligence.
865   Hai Wang, Hu Haiyan, Zhang Qian, Yang Yadong, Li Yanming, Hu Yang, Ruan Xiuyan, Yang Yaran, Zhang
866   Zhaojun, Shu Chang, Yan Jiangwei, Wakeland Edward K, Li Quanzhen, Hu Songnian, Fang Xiangdong. 2013.
867   Dynamic transcriptomes of human myeloid leukemia cells. Genomics. 102:250–256.
868   Haindl M, Somol P, Ververidis D, Kotropoulos C.2006. Feature Selection Based on Mutual Correlation. Pattern
869   Recognition, Image Analysis and Applications. Springer Berlin Heidelberg. 569-577.
870   Hambali Moshood A., Oladele Tinuke O., Adewole Kayode S. 2020. Microarray cancer feature selection: Review,
871   challenges and research directions. International Journal of Cognitive Computing in Engineering.1: 78-97.

872   Handschuh Luiza. 2019. Not Only Mutations Matter: Molecular Picture of Acute Myeloid Leukemia Emerging from
873   Transcriptome Studies. Journal of Oncology.
874   Hastie T, Tibshirani R, Eisen M.B., Alizadeh A, Levy R, Staudt L, Chan W. C., Botstein D, and Brown P. 2000.
875   'Gene Shaving' as a Method for Identifying Distinct Sets of Genes with Similar Expression Patterns. Genome
876   Biology. 1(2):1-21.
877   Hastie T, Tibshirani R, Botstein D, and Brown P. 2001. Supervised Harvesting of Expression Trees. Genome
878   Biology. 1: 1-12.
879   Hunter JD. 2007. Matplotlib: a 2D graphics environment. Computing In Science & Engineering 9(3):90–95 DOI
880   10.1109/MCSE.2007.
881   Kamal Amany M, El-Hefny Nadia H, Hegab Hany M, El-Mesallamy Hala O. 2016. Expression of thioredoxin-1
882   (TXN) and its relation with oxidative DNA damage and treatment outcome in adult AML and ALL: A comparative
883   study. Hematology. 21(10):567-575.
884   Karlenius Therese Christina, Tonissen Kathryn Fay. 2010. Thioredoxin and Cancer: A Role for Thioredoxin in all
885   States of Tumor Oxygenation. Cancers (Basel). 2(2): 209–232.
886   Kim Yundeok , Yoon Sulhee, Kim Soo Jeong, Kim Jin Seok, Cheong Jun-Won, Hong Yoo Min. 2012.
887   Myeloperoxidase Expression in Acute Myeloid Leukemia Helps Identifying Patients to Benefit from Transplant.
888   Yonsei Med J. 53(3): 530–536.
889   Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson
890   C, Meltzer S Paul. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and
891   artificial neural networks. Nat Medicine. 7(6):673–679.
892   Klimiankou Maksim, Uenalan Murat, Kandabarau Siarhei, Nustede Rainer, Steiert Ingeborg, Mellor-Heineke
893   Sabine, Zeidler Cornelia, Skokowa Julia, Welte Karl. 2019. Ultra-Sensitive CSF3R Deep Sequencing in Patients
894   With Severe Congenital Neutropenia. Front. Immunol.
895   Komer B, Bergstra J, Eliasmith C. 2014. Hyperopt-sklearn: automatic hyperparameter configuration for scikit-learn.
896   In: Proceedings of the 13th Python in Science Conference (SCIPY 2014). 33–39.
897   Kourou Konstantina, Exarcos T P, Exarcos K P, Karmouzis M V, Fotaidis D. 2015. Machine learning applications
898   in cancer prognosis and prediction. Computational and Structural Biotechnology Journal. 13: 8–17.
899   Kozlov Igor, Beason Kevin, Yu Cheng, Hughson Michael. 2005, CD79a expression in acute myeloid leukemia
900   t(8;21) and the importance of cytogenetics in the diagnosis of leukemias with immunophenotypic ambiguity. Cancer
901   Genetics and Cytogenetics 163(1):62-7.
902   Kozovska. Z, Patsalias A., Bajzik V, Durinikova E, Demkova L, Jargasova S, Smolkova B, Plava J, Kucerova L,
903   Matuskova M. 2018. ALDH1A inhibition sensitizes colon cancer cells to chemotherapy. BMC Cancer. 18:656.
904   Lagunas-Rangel Francisco Alejandro, Chávez-Valencia Venice, Gómez-Guijosa Miguel Ángel, Cortes-Penagos
905   Carlos. 2017. Acute Myeloid Leukemia—Genetic Alterations and Their Clinical Prognosis. Int J Hematol Oncol
906   Stem Cell Res. 11(4): 328–339.
907   Lai C, Reinders MJT, Wessels L. 2006. Random subspace method for multivariate feature selection. Pattern
908   Recognition Letters., 27: 1067-1076.
909   Lance Amanda, Druhan Lawrence J, Vestal C Greer, Steuerwald Nury M, Hamilton Alicia, Smith Mathew, Price
910   Andrea, Tjaden Elise, Fox Andee N, Avalos Belinda R. 2020. Altered expression of CSF3R splice variants impacts
911   signal response and is associated with SRSF2 mutations. Leukemia. 34(2):369-379.
912   Leung Yukyee and Hung Yeungsam. 2010. A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and
913   Microarray Data Classification", IEEE/ACM Transactions on Computational Biology and Bioinformatics. 7(1).
914   Léveillard Thierry, Aït-Ali Najatel. 2017. Cell Signaling with Extracellular Thioredoxin and Thioredoxin-Like
915   Proteins: Insight into Their Mechanisms of Action. Oxidative Medicine and Cellular Longivity.
916   Li Z, Xie W, Liu T. 2018. Efficient feature selection and classification for microarray data. Plos one.
917   Liao B, Jiang Y, Liang W, Zhu W, Cai L, Cao Z. 2014. Gene selection using locality sensitive Laplacian score.
918   IEEE/ACM Transactions on Computational Biology and Bioinformatics. DOI: 10.1109/TCBB.2014.2328334.
919   Liu Jian, Cheng Yuhu, Wang Xuesong, Zhang Lin, Wang Z. Jane. 2018. Cancer Characteristic Gene Selection via
920   Sample Learning Based on Deep Sparse Filtering. Scientific Reports, Nature. 8: 8270.

921 Liu Q, Sung A H, Chen Z, Liu J, Chen L, Qiao M, Wang Z, Huang X, Deng Y. 2011. Gene selection and
922 classification for cancer microarray data based on machine learning and similarity measures. BMC Genomics.
923 12(5):S1.
924 Liu Sheng, Dissanayake Shamitha, Patel Sanjay V, Dang Xin. 2014. Learning accurate and interpretable models
925 based on regularized random forests regression. BMC Systems Biology. 8 Suppl 3(Suppl 3):S5.
926 M Lizet. Waals van der, M. H. Inne. Rinkes Borel, Kranenburg Onno. 2018. ALDH1A1 expression is associated
927 with poor differentiation, 'right-sidedness' and poor survival in human colorectal cancer. PLOS ONE.
928 Maji P, Das C. 2012. Relevant and Significant Supervised Gene Clusters for Microarray Cancer Classification. IEEE
929 Transactions on Nanobioscience. 11(2).
930 Nabendu Bhui, Pintu Kumar Ram, Pratyay Kuila. 2020. Feature Selection from Microarray Data based on Deep
931 Learning Approach. ICCCNT2020.
932 Nada A, Alshamlan H. 2019. A survey on hybrid feature selection methods in microarray gene expression data for
933 cancer classification", IEEE Access.
934 Nagi Sajid, Bhattacharyya Kr. Dhruba. 2013. Classification of microarray cancer data using ensemble approach.
935 Netw Model Anal Health Inform Bioinformatics. 2:159–173.
936 Osareh Alireza, Bita Shadgar. 2013. An Efficient Ensemble Learning Method for Gene Microarray Classification.
937 BioMed Research International, Hindawi Publishing Corporation.
938 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R,
939 Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. 2011. Scikit-learn:
940 machine learning in python. Journal of MachineLearning Research 12:2825–2830.
941 Pilling M J, Henderson A, Gardner P. 2017. Quantum Cascade Laser Spectral Histopathology: Breast Cancer
942 Diagnostics Using High Throughput Chemical Imaging. Analytical Chemistry. 89 (14), 7348-7355.
943 Pouw Kraan T.C.T.M. Van der, Gaalen F.A. van, Kasperkovitz P.V., Verbeet N.L., Smeets T.J.M., Kraan M.C.,
944 Fero M., Tak P.-P., Huizinga T.W.J. , Pieterman E., Breedveld F.C., Alizadeh A.A. , and Verweij C.L. 2003.
945 Rheumatoid Arthritis is a Heterogeneous Disease:Evidence for Differences in the Activation of the STAT-1
946 Pathway between Rheumatoid Tissues. Arthritis and Rheumatism. 48(8):2132-2145.
947 Pouw Kraan T.C.T.M. van der, Wijbrandts C.A., Baarsen L.G.M. van, Voskuyl A.E., Rustenburg F., Baggen J.M.,
948 Ibrahim S.M., Fero M., Dijkmans B.A.C., Tak P.P., and Verweij C.L. 2007. Rheumatoid Arthritis Subtypes
949 Identified by Genomic Profiling of Peripheral Blood Cells: Assignment of a Type I Interferon Signature in a
950 Subpopulation of Pateints. Annals of the Rheumatic Diseases. 66: 1008-1014.
951 Ritter Malte, Klimiankou Maksim, Klimenkova Olga, Schambach Axel, Hoffmann Dirk, Schmidt Amy, Kanz
952 Lothar, Link Daniel C., Welte Karl, Skokowa Julia. 2020. Cooperating, congenital neutropenia–associated Csf3r and
953 Runx1 mutations activate pro-inflammatory signaling and inhibit myeloid differentiation of mouse HSPCs. Annals
954 of Hematology. 99: 2329–2338.
955 Ruiz R., Riquelme J.C., Aguilar-Ruiz J.S. 2006. Incremental wrapper-based gene selection from microarray data for
956 cancer classification. Journal of Pattern Recognition. 39(12): 2383–2392.
957 Simmen Frank A, Su Ying, Xiao Rijin, Zeng Zhaoyang, Simmen CM Rosalia. 2008. The Krüppel-like factor 9
958 (KLF9) network in HEC-1-A endometrial carcinoma cells suggests the carcinogenic potential of dys-regulated
959 KLF9 expression. Reproductive Biology and Endocrinology.
960 Singh D., Febbo P.G., Ross K., Jackson D. G., Manola J., Ladd C., Tamayo P., Renshaw A. A., Amico A. V. D,
961 Richie J. P. , Lander E. S., Loda M., Kantoff P.W., Golub T.R., and Sellers W.R..2002. Gene expression correlates
962 of clinical prostate cancer behavior. Cancer Research..1: 203–209.
963 Singh P, Dai B, Dhruva B, Widen S G 1994. Episomal Expression of Sense and Antisense Insulin-like Growth
964 Factor (IGF) binding Protein-4 Complementary DNA Alters the Mitogenic Response of a Human Colon Cancer Cell
965 Line (HT-29) by Mechanisms That Are Independent of and Dependent upon IGF-11. Cancer Research. 54: 6563-
966 6570.
967 Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, Schultz PG, Powell SM, Moskaluk CA, Frierson
968 HF, Hampton GM. 2001. Molecular classification of human carcinomas by use of gene expression signatures.
969 Cancer Research. 61:7388–7393.

970 Su Chun-Wen, Lin Chiao-Wen, Yang Wei-En, Yang Shun-Fa. 2019. TIMP-3 as a therapeutic target for cancer.
971 Therapeutic Advances in Medical Oncology.
972 Su Li, Luo Yongli, Yang Zhi, Yang Jing, Yao Chao, Cheng Feifei, Shan Juanjuan, Chen Jun, Li Fangfang, Liu
973 Limei, Liu Chungang, Xu Yanmin, Jiang Lupin, Deyu Guo, Prieto Jesus, Ávila A Matías, Shen Junjie, Qian Cheng.
974 2016. MEF2D Transduces Microenvironment Stimuli to ZEB1 to Promote Epithelial–Mesenchymal Transition and
975 Metastasis in Colorectal Cancer. Molecular and Cellular Pathobiology. 76(17):5054-5067.
976 Swan A L, Mobasheri Ali, Allaway David, Liddell Susan, Bacardit Jaume. 2013. Application of Machine Learning
977 to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology. OMICS. 17(12): 595–
978 610.
979 Szuber Natasha, Tefferi Ayalew. 2018. Chronic neutrophilic leukemia: new science and new diagnostic criteria.
980 Blood Cancer Journal. 19.
981 Tabares-Soto Reinel, Orozco-Arias Simon, Romero-Cano Victor, Segovia Bucheli Vanesa, Rodríguez-Sotelo José
982 Luis , Jiménez-Varón Cristian Felipe. 2020. A comparative study of machine learning and deep learning algorithms
983 to classify cancer types based on microarray gene expression data. PeerJ Computer Science.
984 Tabakhi S, Moradi P, Akhlaghian F. 2014. An unsupervised feature selection algorithm based on ant colony
985 optimization. Engineering Applications of Artificial Intelligence. 32: 112-123.
986 Tabakhi S, Najafi, A., Ranjbar, R., &Moradi, P. 2015. Gene selection for microarray data classification using a
987 novel ant colony optimization. Neurocomputing, 168: 1024-1036.
988 Tang Zhenjie, Yuan Shuqiang, Hu Yumin, Zhang Hui, Wu Wenjing, Zeng Zhaolei, Yang Jing, Yun Jingping, Xu
989 Ruihua, Huang Peng. 2012. Over-expression of GAPDH in human colorectal carcinoma as a preferred target of 3-
990 Bromopyruvate Propyl Ester" Published in final edited form as: J Bioenerg Biomembr. 44(1): 117–125.
991 Theodoridis S, Koutroumbas K. 2008. Pattern Recognition, fourth ed., Elsevier Science.
992 Thorsen Kasper, Sørensen D Karina, Brems-Eskildsen Sofie Anne, Modin Charlotte, Gaustadnes Mette, Hein K
993 Anne-Mette, Kruhøffer Mogens, Laurberg Soren, Borre Michael, Wang Kai, Brunak Søren, Krainer R Adrian,
994 Tørring Niels , Dyrskjøt Lars, Andersen Claus L, Orntoftet Torben F.2008. Alternative Splicing in Colon, Bladder,
995 and Prostate Cancer Identified by Exon Array Analysis. Molecular & Cellular Proteomics. 7: 1214-1224.
996 Tong Dong Ling, Ball Graham R. 2014. Exploration of Leukemia Gene Regulatory Networks Using A Systems
997 Biology Approach. 2014 IEEE International Conference on Bioinformatics and Biomedicine.
998 Vapnik V. 1995. The Nature of Statistical Learning Theory. New York: Springer-Verlag. Veer L. J. V., Dai H.,
999 Vijver M. J. V. D., He Y. D., Hart A. A.M., Mao M., Peterse H. L., Kooy K. v. d., Marton M. J., Witteveen A. T.,
1000 Schreiber G.J. , Kerkhoven R.M., Roberts C. , Linsley P. S., Bernards R., and Friend S. H..2002. Gene expression
1001 profiling predicts clinical outcome of breast cancer. Nature. 415(6871): 530–536.
1002 Wang Ching Wei. 2006. New Ensemble Machine Learning Method for Classification and Prediction on Gene
1003 Expression Data. Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA.
1004 Wang Shu-Lin, Li Xue-Ling, Fanget Jianwen. 2012. Finding minimum gene subsets with heuristic breadth-first
1005 search algorithm for robust tumor classification. BMC Bioinformatics.13:178.
1006 West M., Blanchette C., Dressman H., Huang E., Ishida S., Spang R., Zuzan H. , Olson J. A., Marks J. R., and
1007 Nevins J. R.. 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc.
1008 Natl. Acad. Sci. USA. 98(20):11 462–11 467.
1009 Yang P, Zhou B Bing, Zhang Zili, Zomaya Y, Albert. 2010. A multi-filter enhanced genetic ensemble system for
1010 gene selection and sample classification of microarray data. BMC Bioinformatics. 11: S5.
1011 Yan Zhi, Li Jiangeng, Xiong Yimin, Xu Weitian, Zheng Guorong. 2012. Identification of candidate colon cancer
1012 biomarkers by applying a random forest approach on microarray data. Oncology Reports, SPANDIDOS
1013 Publications. 28(3):1036-42.
1014 Yunsong Q, Yang Xibei. 2013. Interval-valued analysis for discriminative gene selection and tissue sample
1015 classification using microarray data. Genomics. 101(1):38-48.
1016 Ying Mingyao, Tilghman Jessica, Wei Yingying, Guerrero-Cazares Hugo, Alfredo Quinones-Hinojosa, Ji Hongkai,
1017 Laterra John. 2014. KLF9 Inhibits Glioblastoma Stemness through Global Transcription Repression and Integrin-α6
1018 Inhibition. Journal for Biochemistry and Molecular Biology. 289(47): 32742–32756.

1019   Yu Herbert, Rohan Thomas. 2000. Role of the Insulin-Like Growth Factor Family in Cancer Development and
1020   Progression. Journal of the National Cancer Institute. 92(18): 1472–1489.
1021   Zhang Jin-Ying , Zhang Fan, Hong Chao-Qun, Giuliano Armando E., Cui Xiao-Jiang, Zhou Guang-Ji , Zhang Guo-
1022   Jun, Cui Yu-Kun. 2015. Critical protein GAPDH and its regulatory mechanisms in cancer cells. Cancer Biol Med
1023   12(1):10-22.
1024   Zhang Yang, Wang Fang, Chen Xue, Zhang Yu, Wang Mingyu, Liu Hong, Cao Panxiang, Ma Xiaoli, Wang Tong,
1025   Zhang Jianping, Zhang Xian, Lu Peihua, Liu Hongxing 2018. CSF3R Mutations are frequently associated with
1026   abnormalities of RUNX1, CBFB, CEBPA, and NPM1 genes in acute myeloid leukemia. Cancer. 124(16):3329-
1027   3338.
1028   Zhou Nina and Wang Lipo. 2007. A Modified T-test Feature Selection Method and Its Application on the HapMap
1029   Genotype Data. Genomics Proteomics Bioinformatics. 5(3-4): 242-9.
1030   Zhu Kongxi, Wang Yunxia, Liu Lan, Li Shuai, Yu Weihua. 2019. Long non-coding RNA MBNL1-AS1 regulates
1031   proliferation, migration, and invasion of cancer stem cells in colon cancer by interacting with MYL9 via sponging
1032   microRNA-412-3p. Clinics and Research in Hepatology and Gastroenterology. 44(1):101-114.

Description of Cancer Gene Expression Datasets

Table 1. Description of Cancer Gene Expression Datasets

Dataset	Data Dimension Gene × Sample (Original)	Data Dimension Gene × Sample (Used)	Sample Class Labels	Dataset	Data Dimension Gene × Sample (Original)	Data Dimension Gene × Sample (Used)	Sample Class Labels
Leukemia	7129 × 72	7070 × 72	2	Breast	7129 × 49	7129 × 49	2
Colon	2000 × 62	2000 × 62	2	MLL	12582 × 72	12582 × 72	3
Prostate	12600 × 136	12600 × 136	2	SRBCT	2308 × 63	2308 × 63	4
Lung	12533 × 181	12533 × 181	2	RAHC	41057 × 50	41057 × 50	2
Rbreast	24481 × 97	24188 × 97	2	RAOA	18433 × 30	18433 × 30	2

Classification Accuracy of MFSAC-EC depending on varying number of genes selected by each Filter

This table shows the impact of parameter P with respect to sample classification accuracy(%) in terms of both LOOCV and 10-Fold Cross Validation approach. P defines the number of top ranked genes selected by each filter method.

Table 2: Classification Accuracy of MFSAC-EC depending on varying number of genes selected by each Filter

Dataset	Evaluation Metric	MFSAC-EC																							
		P=100				P=200				P=500				P=1000				P=1200				P=1500			
		NB	KNN	DT	SVM	NB	KNN	DT	SVM	NB	KNN	DT	SVM	NB	KNN	DT	SVM	NB	KNN	DT	SVM	NB	KNN	DT	SVM
Leukemia	LOOCV	98.6	98.6	98.6	98.6	98.6	98.6	98.6	98.6	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	10 Fold	98.6	98.6	98.6	98.6	98.6	98.6	98.6	98.6	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
RAHC	LOOCV	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	10 Fold	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
MLL	LOOCV	98.6	100	100	97.2	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	10 Fold	97.2	100	100	97.2	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
RAOA	LOOCV	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	10 Fold	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
SRBCT	LOOCV	98.4	98.4	100	98.4	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	10 Fold	100	98.4	98.4	98.4	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Breast	LOOCV	98	95.9	93.9	95.9	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	10 Fold	100	95.9	95.9	98	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Lung	LOOCV	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	10 Fold	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Rbreast	LOOCV	92.6	93.7	93.7	95.8	99	97.9	100	99	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	10 Fold	91.6	96.8	95.8	97.9	97.9	99	99	99	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
COLON	LOOCV	91.9	91.9	93.6	91.9	91.9	91.9	96.8	91.9	98.4	98.4	96.8	98.4	98.4	98.4	98.4	98.4	98.4	98.4	98.4	100	100	100	98.4	100
	10 Fold	91.9	91.9	93.6	91.9	91.9	91.9	95.2	91.9	98.4	96.8	96.8	96.8	98.4	100	98.4	98.4	100	98.4	98.4	100	100	100	98.4	100
Prostrate	LOOCV	83.8	90.4	92.7	86.8	88.2	92.7	92.7	88.2	91.2	95.6	97.1	91.9	94.9	97.1	97.8	94.1	98.5	98.5	97.8	98.5	98.5	99.3	98.5	99.3
	10 Fold	85.3	88.2	91.9	86.8	88.2	91.9	93.4	89.7	91.9	95.6	97.8	91.2	95.6	97.8	97.8	94.1	99.3	98.5	97.8	98.5	99.3	99.3	97.8	99.3

Total Execution Time in a single run of MFSAC-EC on Different Datasets

Total execution time in a single run of MFSAC-EC including Bootstrapped dataset creation, Feature Selection by filter methods and supervised attribute clustering approach, Training, Testing using LOOCV, 5-Fold, 10-Fold, and Random Splitting is given in the first row. While execution time using only 10-Fold Cross Validation is given in the 2nd row. Here the time for the best P value is shown here.

1   **Table 3: Total Execution Time (Bootstrapping, Feature Selection, Training, Testing for LOOCV, 5-Fold, 10-Fold, Random Splitting) in a single run of MFSAC-EC on Different Datasets**

	Leukemia	RAHC	MLL	RAOA	SRBCT	Breast	Lung	Rbreast	COLON	Prostrate
No. of Feature selected for best result	500	100	200	100	200	200	100	500	1200	3000
Total Time Taken	8mins 23secs	7mins 32secs	7mins 54secs	4mins 43secs	5mins 17secs	4mins 2secs	11mins 14secs	10mins 22secs	17mins 40secs	1hr 18mins 41secs
Time Taken for only 10 fold	35secs	30secs	41secs	36secs	30secs	32secs	36secs	33secs	30secs	36secs

Classification Accuracy of the proposed MFSAC-EC model with respect to LOOCV

Classification accuracy (%) of MFSAC-EC model has been shown in terms of LOOCV with respect to four ensemble classifiers MFSAC-EC + NB, MFSAC-EC+KNN, MFSAC-EC+DT, and MFSAC-EC+SVM. Every ensemble classifier is run 50 times using LOOCV for every dataset and the accuracy is shown which is obtained maximum number of times.

Table 4. Classification Accuracy of the proposed MFSAC-EC model with respect to LOOCV

Dataset	Proposed Model		Cluster Representatives			Dataset	Proposed Model		Cluster Representatives		
			1	2	3				1	2	3
COLON	MFSAC-EC	NB	100	98.39	98.39	MLL	MFSAC-EC	NB	100	100	100
		KNN	98.39	100	100			KNN	100	100	100
		DT	98.39	98.39	98.39			DT	100	100	100
		SVM	100	98.4	98.4			SVM	100	100	100
Prostate		NB	97.06	97.79	98.53	SRBCT		NB	96.83	100	100
		KNN	97.79	97.79	98.53			KNN	96.83	100	100
		DT	97.79	98.53	97.79			DT	96.83	98.41	100
		SVM	98.53	99.26	99.26			SVM	82.54	98.41	100
Leukemia		NB	100	100	100	Lung		NB	100	100	100
		KNN	100	100	100			KNN	100	100	100
		DT	100	100	100			DT	100	100	100
		SVM	100	100	100			SVM	100	100	100
RAOA		NB	100	100	100	RAHC		NB	100	100	100
		KNN	100	100	100			KNN	100	100	100
		DT	100	100	100			DT	100	100	100
		SVM	100	100	100			SVM	100	100	100
Breast		NB	100	100	100	RBreast		NB	100	100	100
		KNN	100	100	100			KNN	100	100	100
		DT	100	100	100			DT	100	100	100
		SVM	100	100	100			SVM	100	100	100

Classification Accuracy of the proposed MFSAC-EC model with respect to 5-Fold Cross Validation

Classification accuracy (%) of MFSAC-EC model has been shown in terms of 5-Fold Cross Validation with respect to four ensemble classifiers MFSAC-EC + NB, MFSAC-EC+KNN, MFSAC-EC+DT, and MFSAC-EC+SVM. Every ensemble classifier is run 50 times using 5-Fold Cross Validation for every dataset and the accuracy is shown which is obtained maximum number of times.

Table 5. Classification Accuracy of the proposed MFSAC-EC model with respect to 5-Fold Cross Validation

Dataset	Proposed Model		Cluster Representatives			Dataset	Proposed Model		Cluster Representatives		
			1	2	3				1	2	3
COLON	MFSAC-EC	NB	96.77	96.77	96.77	MLL	MFSAC-EC	NB	100	100	100
		KNN	98.39	96.77	96.77			KNN	98.61	100	100
		DT	98.39	96.77	98.39			DT	98.61	100	100
		SVM	98.39	96.77	96.77			SVM	100	100	100
Prostate		NB	97.06	97.79	98.53	SRBCT		NB	98.41	100	100
		KNN	97.79	97.79	99.26			KNN	96.83	100	100
		DT	97.06	97.79	94.85			DT	96.83	98.41	100
		SVM	97.79	98.53	99.26			SVM	96.83	100	100
Leukemia		NB	100	100	100	Lung		NB	100	100	100
		KNN	100	100	100			KNN	100	100	100
		DT	100	100	100			DT	100	99.44	100
		SVM	100	100	100			SVM	100	100	100
RAOA		NB	100	100	100	RAHC		NB	100	100	100
		KNN	100	100	100			KNN	100	100	100
		DT	100	100	100			DT	100	100	100
		SVM	100	100	100			SVM	100	100	100
Breast		NB	100	100	100	RBreast		NB	100	100	100
		KNN	100	100	100			KNN	100	100	100
		DT	100	100	100			DT	100	100	100
		SVM	100	100	100			SVM	100	100	100

Classification Accuracy of the proposed MFSAC-EC model with respect to 10-Fold Cross Validation

Classification accuracy (%) of MFSAC-EC model has been shown in terms of 10-Fold Cross Validation with respect to four ensemble classifiers MFSAC-EC + NB, MFSAC-EC+KNN, MFSAC-EC+DT, and MFSAC-EC+SVM. Every ensemble classifier is run 50 times using 10-Fold Cross Validation for every dataset and the accuracy is shown which is obtained maximum number of times.

Table 6. Classification Accuracy of the proposed MFSAC-EC model with respect to 10-Fold Cross Validation

Dataset	Proposed Model		Cluster Representatives			Dataset	Proposed Model		Cluster Representatives		
			1	2	3				1	2	3
COLON	MFSAC-EC	NB	98.39	98.39	98.39	MLL	MFSAC-EC	NB	100	100	100
		KNN	98.39	98.39	100			KNN	100	100	100
		DT	98.39	98.39	98.39			DT	100	100	100
		SVM	98.39	98.39	98.39			SVM	100	100	100
Prostate		NB	97.06	97.79	98.53	SRBCT		NB	96.83	96.83	100
		KNN	97.79	97.79	99.26			KNN	92.06	100	100
		DT	97.06	97.79	94.85			DT	95.24	96.83	100
		SVM	97.79	98.53	99.26			SVM	80.95	92.06	100
Leukemia		NB	100	100	100	Lung		NB	100	100	100
		KNN	100	100	100			KNN	100	100	100
		DT	100	100	100			DT	100	100	100
		SVM	100	100	100			SVM	100	100	100
Breast		NB	100	100	100	RBreast		NB	100	100	100
		KNN	100	100	100			KNN	100	100	100
		DT	100	100	100			DT	100	100	100
		SVM	100	100	100			SVM	100	100	100
RAOA		NB	100	100	100	RAHC		NB	100	100	100
		KNN	100	100	100			KNN	100	100	100
		DT	100	100	100			DT	100	100	100
		SVM	100	100	100			SVM	100	100	100

Classification Accuracy of the proposed MFSAC-EC model with respect to Random Splitting of the Datasets

Classification accuracy (%) of MFSAC-EC model has been shown in terms of random splitting with respect to four ensemble classifiers MFSAC-EC + NB, MFSAC-EC+KNN, MFSAC-EC+DT, and MFSAC-EC+SVM. Every ensemble classifier is run 50 times using random splitting for every dataset and the accuracy is shown which is obtained maximum number of times. For random splitting the dataset is divided into training (2/3) and testing (1/3) part 50 times randomly.

Table 7. Classification Accuracy of the proposed MFSAC-EC model with respect to Random Splitting of the Datasets

Dataset	Proposed Model		Cluster Representatives			Dataset	Proposed Model		Cluster Representatives		
			1	2	3				1	2	3
COLON	MFSAC-EC	NB	98.39	98.39	98.39	MLL	MFSAC-EC	NB	100	100	100
		KNN	98.39	98.39	98.39			KNN	100	100	100
		DT	98.39	98.39	98.39			DT	98.61	100	98.61
		SVM	98.39	100	98.39			SVM	100	100	100
Prostate		NB	94.68	95.74	93.62	SRBCT		NB	95	85	95
		KNN	97.87	96.81	92.55			KNN	95	100	90
		DT	94.68	94.68	94.68			DT	80	90	95
		SVM	94.68	96.81	94.68			SVM	65	75	95
Leukemia		NB	100	100	100	Lung		NB	100	100	100
		KNN	100	100	100			KNN	100	100	100
		DT	100	100	100			DT	100	100	100
		SVM	100	100	100			SVM	100	100	100
RAOA		NB	100	100	100	RAHC		NB	100	100	100
		KNN	100	100	100			KNN	100	100	100
		DT	100	100	100			DT	100	100	81.25
		SVM	100	100	100			SVM	100	100	81.25
Breast		NB	100	100	100	RBreast		NB	91.94	91.94	91.94
		KNN	100	100	100			KNN	85.48	87.10	83.87
		DT	100	100	100			DT	83.87	79.03	80.65
		SVM	100	100	100			SVM	93.55	91.94	91.94

Evaluation of MFSAC-EC classifier based on SN, SP, PPV, NPV, FPR for two class data sets with respect to LOOCV

The performance of the MFSAC-EC model for two class datasets is represented using Receiver Operator Characteristic (ROC) analysis. SN represents Sensitivity, SP represents Specificity, PPV represents Positive Predicted Value, NPV represents Negative Predicted Value, and FPR represents False Positive Rate.

Table 8. Evaluation of MFSAC-EC classifier based on SN, SP, PPV, NPV, FPR for two class data sets with respect to LOOCV

Dataset	Proposed Model		SN	SP	PPV	NPV	FPR	Dataset	Proposed Model		SN	SP	PPV	NPV	FPR
Leukemia	MFSGC-EC	NB	100	100	100	100	0	Breast	MFSGC-EC	NB	100	100	100	100	0
		KNN	100	100	100	100	0			KNN	100	100	100	100	0
		DT	100	100	100	100	0			DT	100	100	100	100	0
		SVM	100	100	100	100	0			SVM	100	100	100	100	0
Prostate		NB	98.7	98.3	98.7	98.3	1.7	Rbreast		NB	100	100	100	100	0
		KNN	98.7	98.3	98.7	98.3	1.7			KNN	100	100	100	100	0
		DT	100	96.61	97.46	100	3.4			DT	100	100	100	100	0
		SVM	100	98.3	98.7	100	1.7			SVM	100	100	100	100	0
Colon		NB	100	100	100	100	0	Lung		NB	100	100	100	100	0
		KNN	100	100	100	100	0			KNN	100	100	100	100	0
		DT	100	100	100	100	0			DT	100	100	100	100	0
		SVM	100	100	100	100	0			SVM	100	100	100	100	0
RAHC		NB	100	100	100	100	0	RAOA		NB	100	100	100	100	0
		KNN	100	100	100	100	0			KNN	100	100	100	100	0
		DT	100	100	100	100	0			DT	100	100	100	100	0
		SVM	100	100	100	100	0			SVM	100	100	100	100	0

Comparison of MFSAC-EC + DT with different existing Ensemble Classifiers using DT in terms of 10-Fold Cross Validation

Here MFSAC-EC + DT model is compared with existing ensemble classifiers where DT is used as base classifier. C4.5 algorithm is used as DT.

Table 9. Comparison of MFSAC-EC using DT with different existing Ensemble Classifiers using DT in terms of 10-Fold Cross Validation

	MFSAC-EC	PCA-based RotBoost	ICA-based RotBoost	AdaBoost	Bagging	Arcing	Rotation Forest	EN-NEW1	EN-NEW2
Colon	**98.39**	95.48	96.1	94.97	94.92	69.35	95.21	79.03	83.87
Leukemia	**100**	98.75	98.77	98.22	97.47	Not Found	97.97	Not Found	Not Found
Breast	**100**	94.39	97.88	98.89	92.74	80.41	98.6	94.85	95.88
Lung	**100**	98.11	99.54	96.3	97.08	97.24	97.56	98.34	99.45
Prostate	**97.79**	Not Found	Not Found	90.44	94.12	87.5	Not Found	94.85	97.06
MLL	**100**	98.86	99.31	97.63	97.11	91.67	97.61	93.06	98.61
SRBCT	**100**	99.5	99.59	98.16	96.46	Not Found	97.44	Not Found	Not Found

Comparison of MFSAC-EC using DT, KNN, NB, SVM with different existing Ensemble Classifiers using DT, KNN, NB, SVM in terms of 10-Fold Cross Validation

Here classification accuracy (%) of four ensemble classifiers MFSAC-EC + NB, MFSAC-EC + KNN, MFSAC-EC+DT, and MFSAC-EC+SVM are shown with respect to results of other existing ensemble classifiers with the same base learners. The best accuracy (%) for every dataset is shown in bold.

Table 10. Comparison of MFSAC-EC using DT, KNN, NB, SVM with different existing Ensemble Classifiers using DT, KNN, NB, SVM in terms of 10-Fold Cross Validation

Dataset	MFSAC-EC				Bagging			Boosting			Stacking			HBSA		SD_Ens	Meta_Ens
	DT	NB	KNN	SVM	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	KNN	SVM		
Leukemia	**100**	**100**	**100**	**100**	94.12	88.23	73.53	91.18	88.24	75.53	91.18	91.18	91.18	88.46	88.46	92.45	94.12
Colon	98.39	98.39	**100**	98.39	95.16	66.13	90.32	98.39	87.1	91.94	98.39	93.59	93.59	75	85	94.4	99.21
Prostate	97.79	**99.26**	**99.26**	**99.26**	26.47	26.47	38.24	26.47	26.47	52.94	26.47	26.47	52.94	85.29	97.06	52.94	52.94
Lung	**100**	**100**	**100**	**100**	91.28	96.64	97.32	81.88	95.3	97.99	97.99	97.99	96.64	Not Found	Not Found	81.88	97.99
Breast	**100**	**100**	**100**	**100**	78.95	36.84	68.42	68.42	36.84	68.42	68.42	68.42	68.42	Not Found	Not Found	73.49	79.87

Comparison of MFSAC-EC using SVM and KNN with respect to different existing deep learning Classifiers using random splitting

Here classification accuracy (%) of two ensemble classifiers MFSAC-EC + KNN, and MFSAC-EC+SVM are shown with respect to results of other existing ensemble classifiers with the same base learners. The best accuracy (%) for every dataset is shown in bold.

Table 11. Comparison of MFSAC-EC using SVM and KNN with respect to different existing deep learning Classifiers using random splitting

Dataset	SVM			KNN		
	MFSAC-EC	Folded Autoencoder	Autoencoder	MFSAC-EC	Folded Autoencoder	Autoencoder
Colon	**100**	90.15	73.11	**98.39**	81.09	56.97
Prostate	**96.81**	84.16	64.3	**97.87**	76.48	52.1
Leukemia	**100**	93.62	84.12	**100**	85.24	77.13

List of genes selected by MFSAC-EC model for Colon and Leukemia cancer Datasets

Here second column represents the gene names while third column indicate the gene accession number. The fourth column indicates the description of the gene while the fifth column indicates the literature where it has been referred as cancer biomarker.

Table 12. List of genes selected by MFSAC-EC model for Colon and Leukemia cancer Datasets

Dataset	Gene Name	Accession Number	Description	Validation of Genes
Colon	TPM1	Hsa.1130	Human tropomyosin isoform mRNA, complete cds.	Gardina 2006, Thorsen 2008, Botchkina 2009
	IGFBP4	Hsa.1532	Human insulin-like growth factor binding protein-4 (IGFBP4) gene, promoter and complete cds.	Durai 2007, Singh 1994, Yu 2000
	MYL9	Hsa.1832	Myosin Regulatory Light Chain 2, Smooth Muscle Isoform (Human); contains element TAR1 repetitive element	Yan 2012, Zhu 2019
	ALDH1L1	Hsa.10224	Aldehyde Dehydrogenase, Mitochodrial X Precursor (Homo sapiens)	Feng 2018, Waals 2018, Kozovska 2018
	KLF9	Hsa.41338	Human mRNA for GC box binding protein/ Kruppel Like Factor 9, complete cds	Brown 2015, Ying 2014, Simmen 2008
	MEF2C	Hsa.5226	Myocyte-Specific Enhancer Factor 2, Isoform MEF2 (Homosapiens)	Chen 2017, Giorgio 2018, Su2016
	GADPH	Hsa.1447	Glyceraldehyde 3-Phosphate Dehydrogenase	Zhang 2015, Tang 2012
	TIMP3	Hsa.11582	Metalloproteinase Inhibitor 3 Precursor	Su 2019, Bai 2007
Leukemia	TXN	X77584_at	TXN Thioredoxin	Kamal 2016, Léveillard 2017, Karlenius 2010
	CSF3R	M59820_at	CSF3R Colony stimulating factor 3 receptor (granulocyte)	Zhang 2018, , Ritter 2020, Klimiankou 2019, Lance 2020
	MPO	M19508_xpt3_s_at	MPO from Human myeloperoxidase gene	Szuber 2018, Kim 2012, Lagunas-Rangel 2017, Handschuh 2019
	LYZ	M21119_s_at	LYZ Lysozyme	Wang 2013, Liu 2018, Tong 2014
	CST3	M27891_at	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	Austin 2010
	ZYX	X95735_at	Zyxin	Austin 2010, Yunsong 2013
	CTSD	M63138_at	CTSD Cathepsin D (lysosomal aspartyl protease)	Wang2013
	CD79A/ MB-1 gene	U05259_rna1_at	MB-1 membrane glycoprotein	Wang 2013, Kozlov 2005

Figure 1

Cluster Representative Refinement Procedure

Each row of the table represents the gene with its class relevance value in terms of Pearson correlation coefficient with respect to sample class row. TR+ and TR- represent the augmented gene with their class relevance score in terms of Pearson correlation coefficient with respect to sample class row.

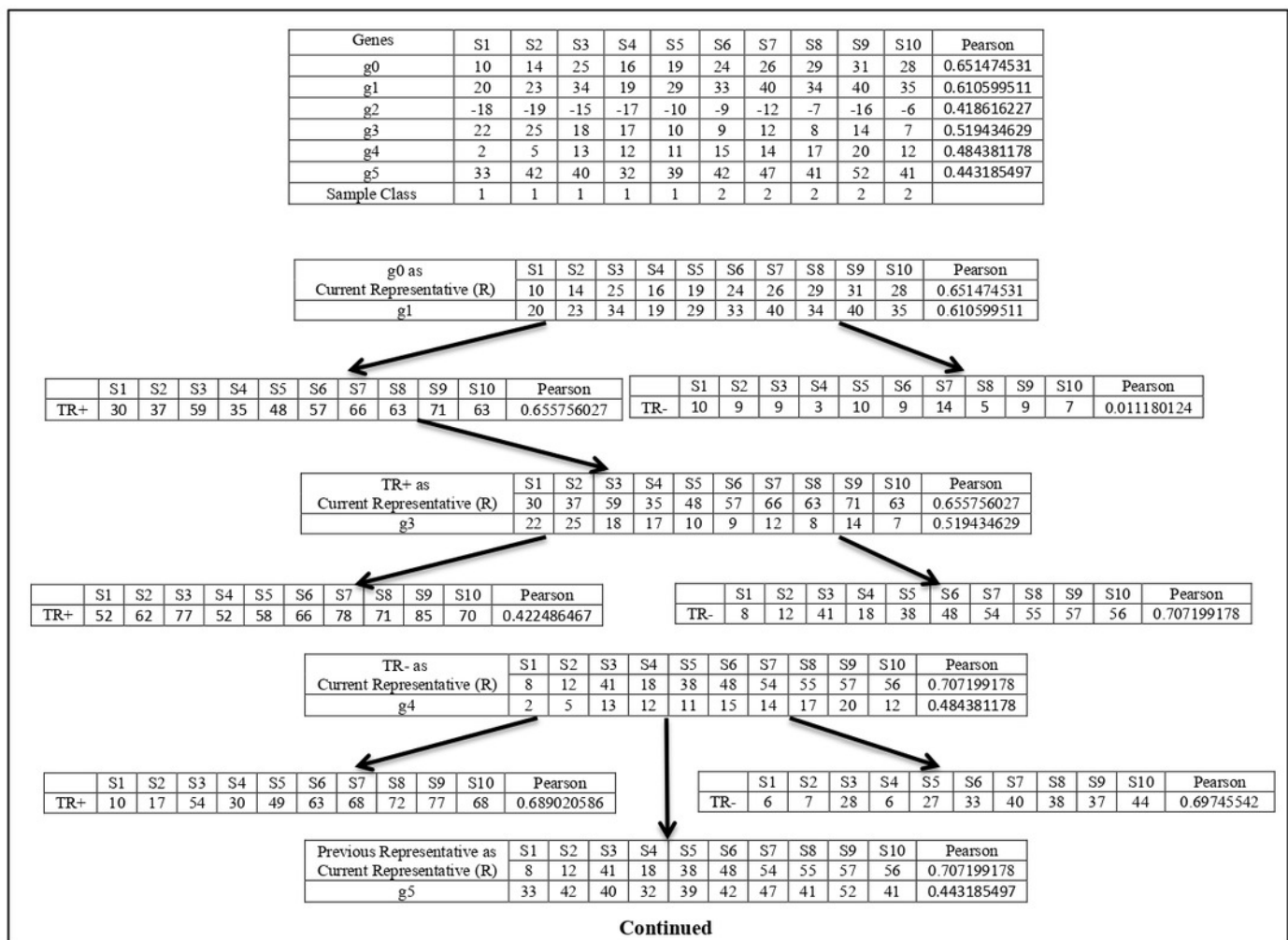



Figure 2

Block Diagram of the Proposed MFSAC-EC Model

Here BK_1, BK_2 ..... BK_D are D number of bootstrapped datasets. RSD_{11} ....RSD_{17} represent different reduced sub datasets of BK_1 bootstrapped datasets after applying MFSAC method. IC_{11} to IC_{17} represent individual classifiers applied on RSD_{11} ... RSD_{17} respectively.

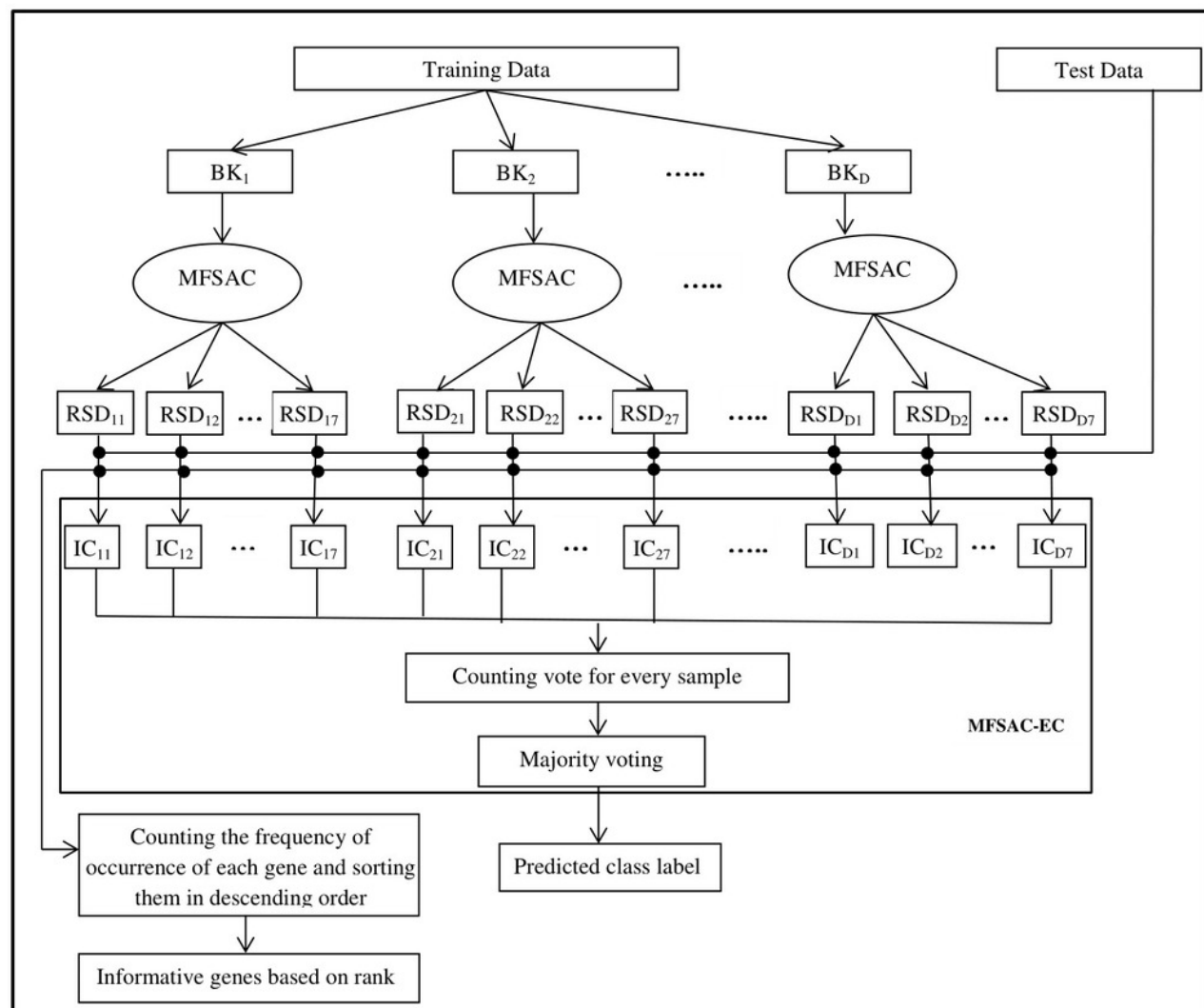



Figure 3

Block Diagram of MFSAC Method

BK_l is the lth bootstrapped dataset. FT_1 ...... FT_7 are the seven filter score functions as

Supplemental Table S1. SD_{11} .... SD_{17} are sub datasets created after applying filter score

functions. SAC is the Supervised attribute clustering method applied to generate RSD_{11}.....

RSD_{17} reduced sub datasets.

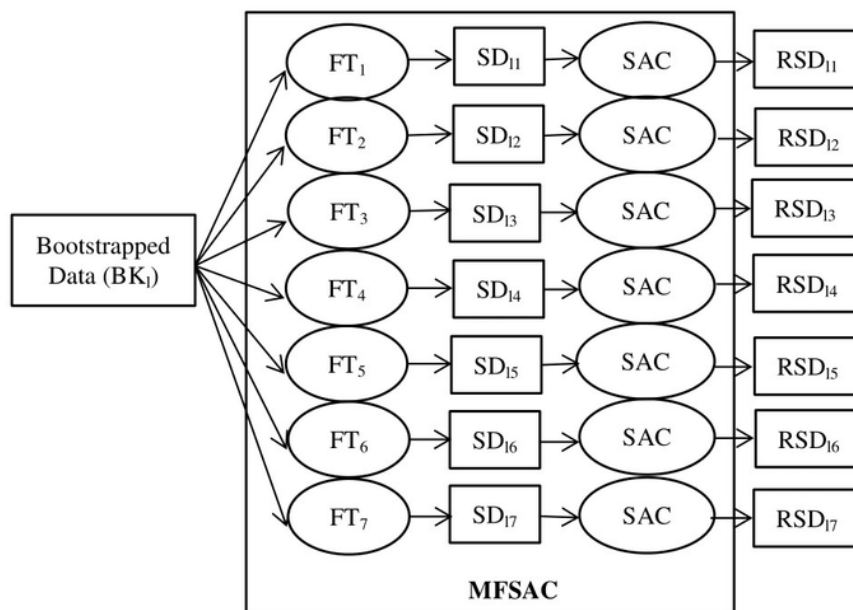



Figure 4

AUC for for three datasets using MFSAC-EC+ KNN, MFSAC-EC+NB, MFSAC-EC+ DT and MFSAC-EC+ SVM Classifiers.

(a) For Breast Cancer dataset using LOOCV. (b) For Colon Cancer dataset using 5-Fold Cross Validation. (c) For RAHC dataset using 10-Fold Cross-Validation

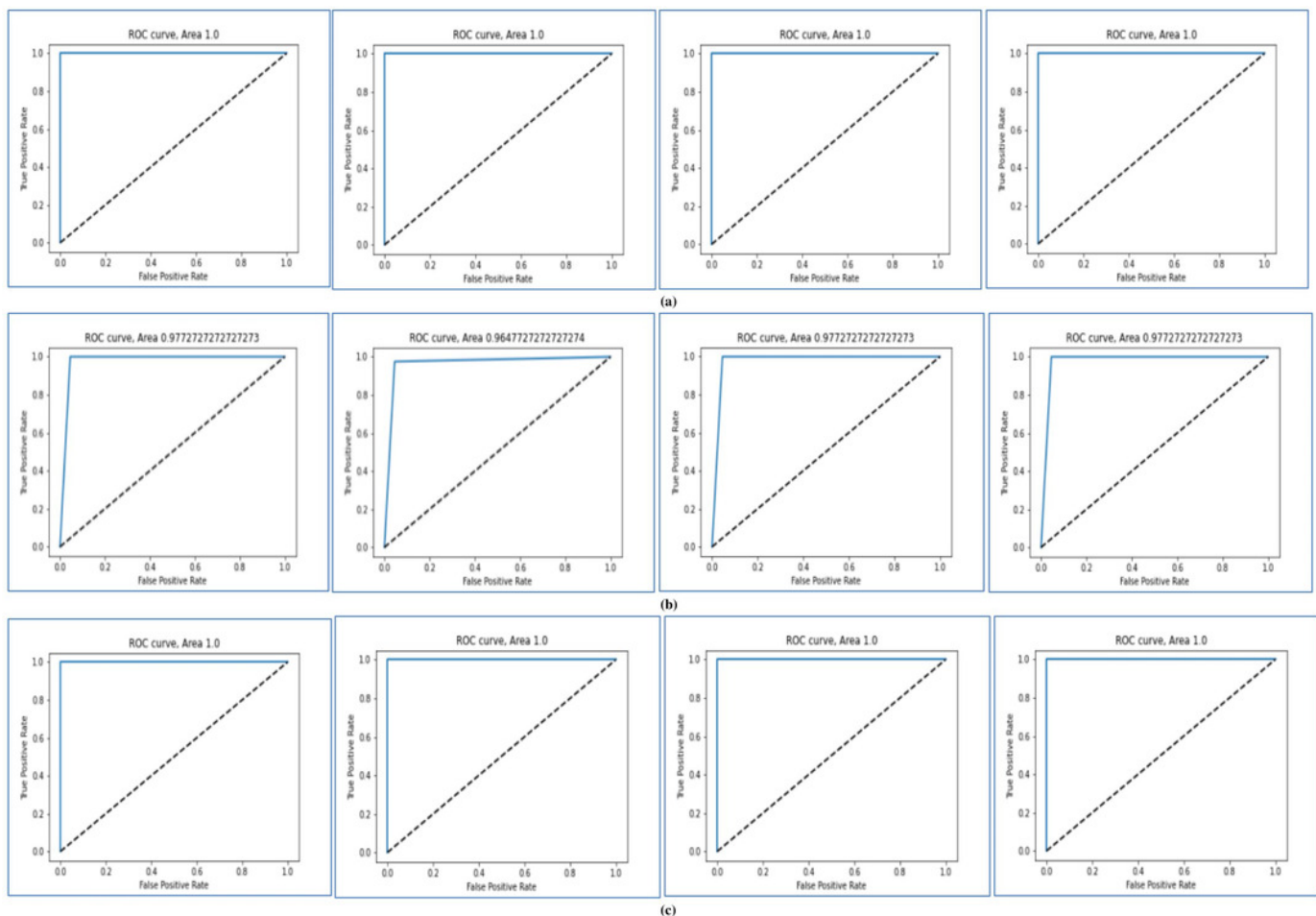



Figure 5

Heatmap of MFSAC-EC with base classifiers NB, KNN, DT and SVM respectively for multiclass datasets.

(a) For SRBCT dataset using 5-Fold Cross-validation. (b) For MLL dataset using 10-Fold Cross-Validation.

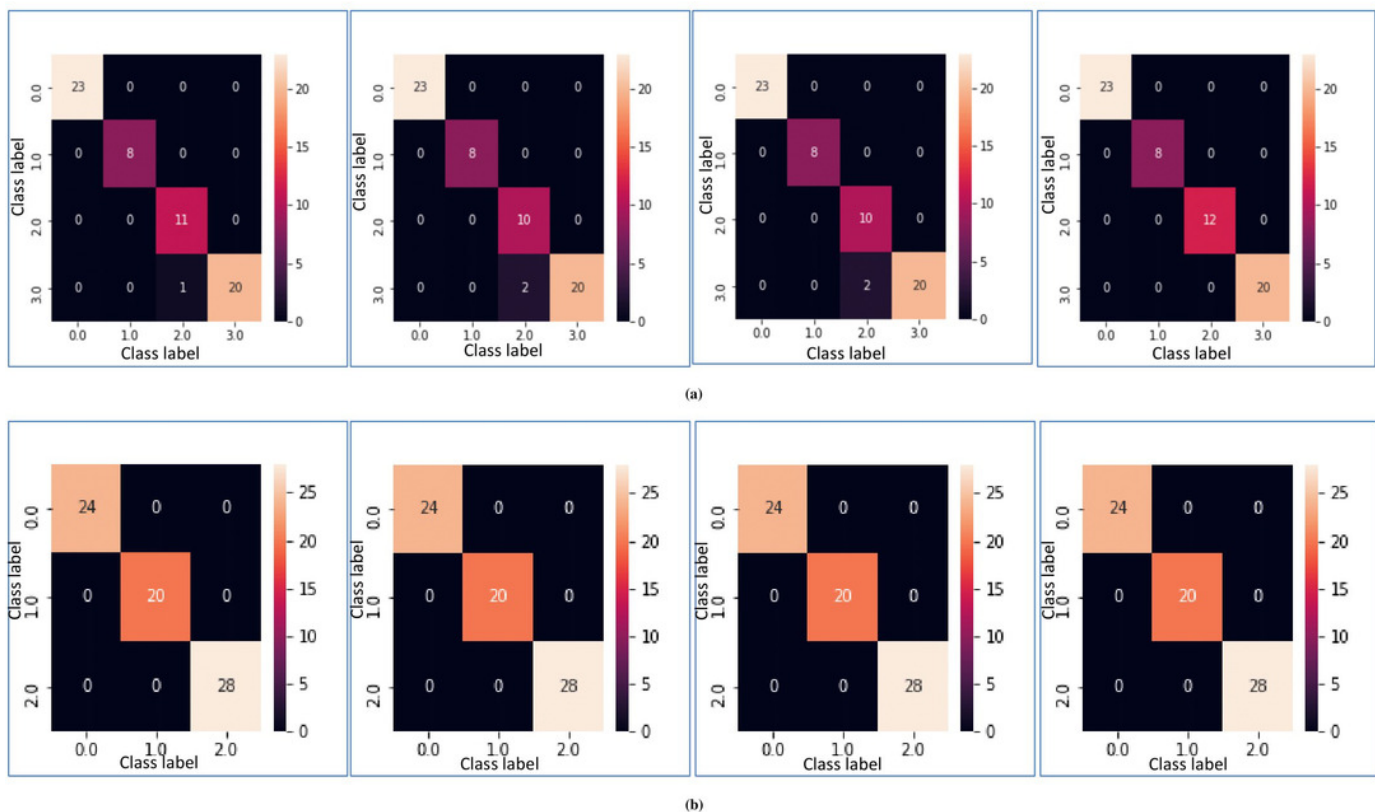



Figure 6

Comparison of MFSAC-EC with other well-known supervised gene selection methods and full gene set in terms of 10-Fold Cross-Validation for all Datasets.

In each figure classification accuracy(%) of MFSAC-EC model along with other supervised gene selection methods for all datasets are represented using different colored bars using (a) NB (b) KNN (c) DT and (d) SVM as base classifier.

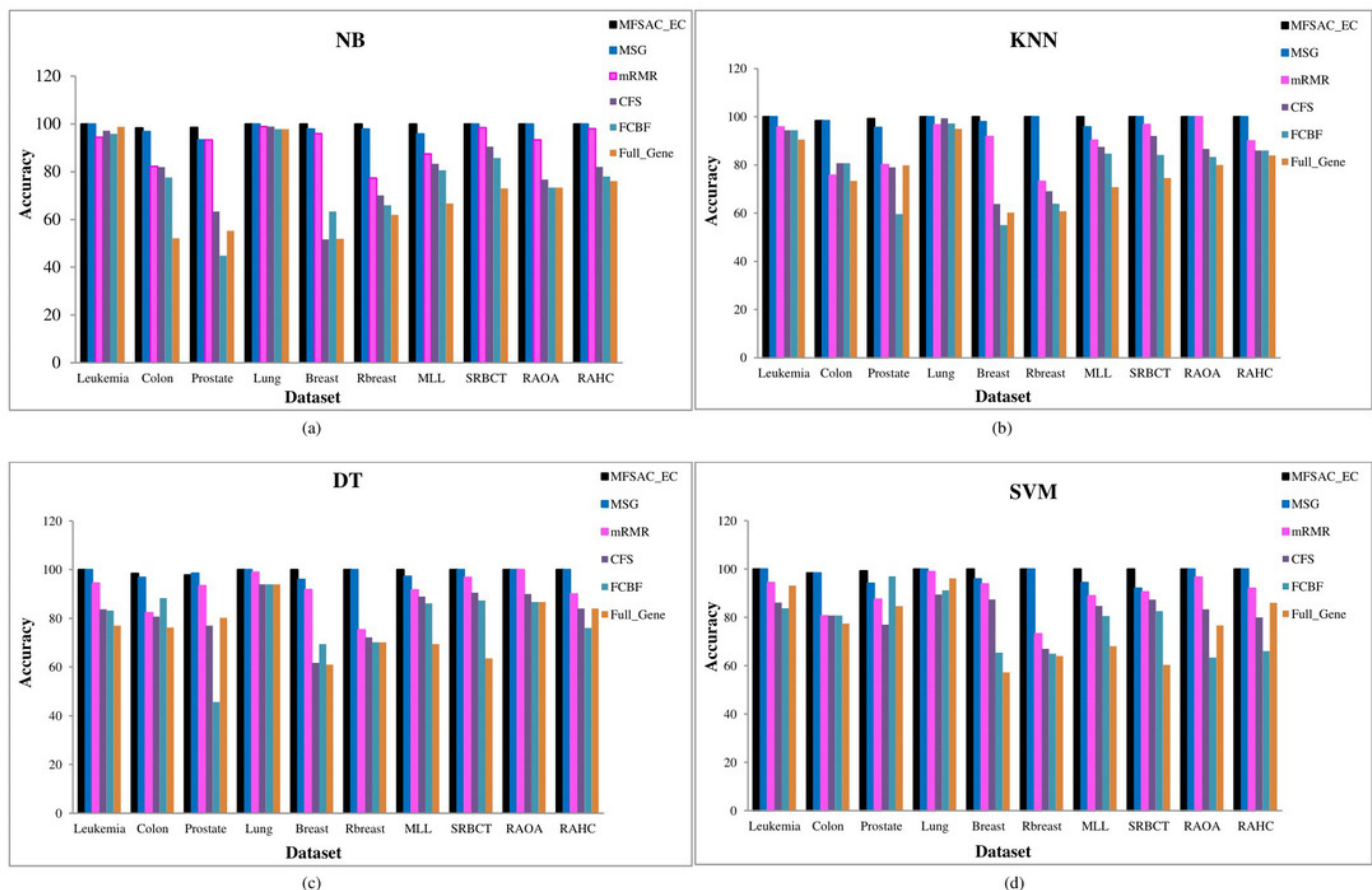



Figure 7

Comparison of MFSAC-EC with other well-known unsupervised gene selection methods in terms of random splitting for different datasets.

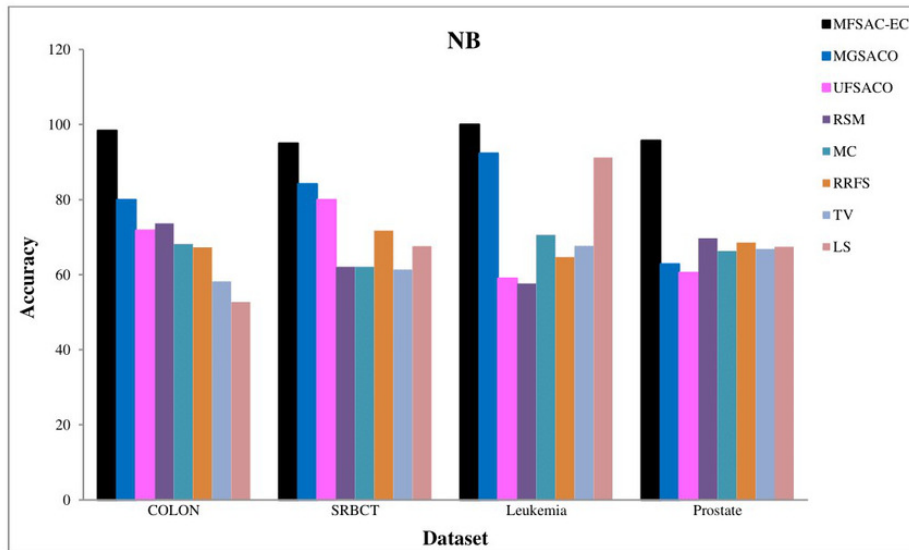
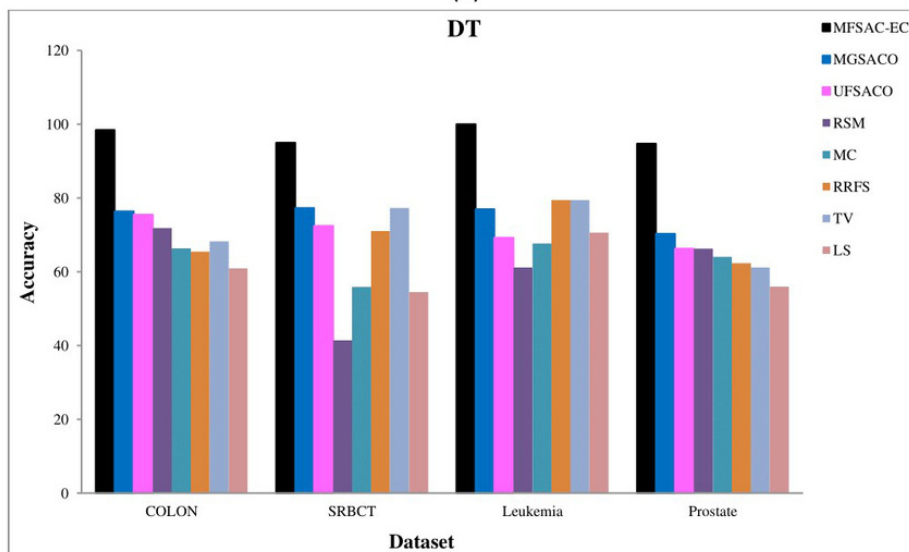
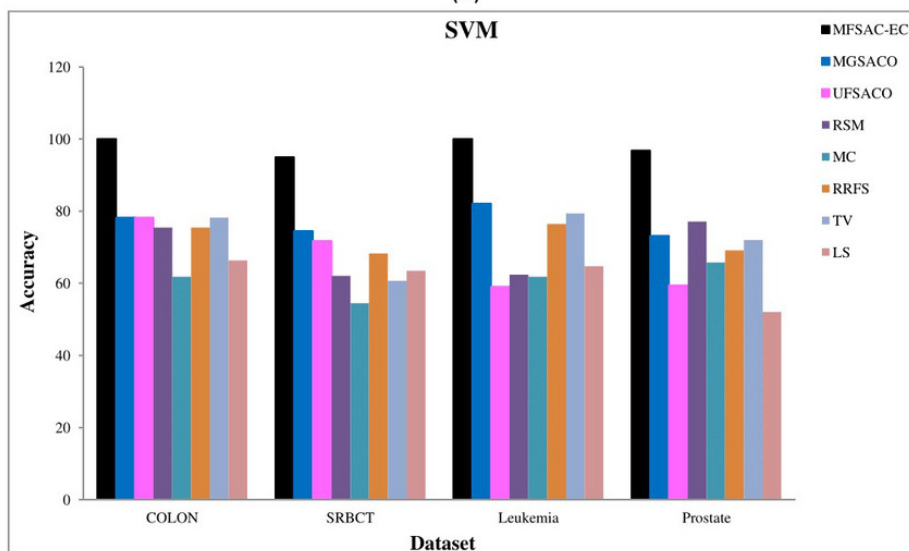
In each figure classification accuracy(%) of MFSAC-EC model along with other unsupervised gene selection methods for four datasets are represented with different colored bars using (a) NB (b) DT and (c) SVM as base classifier.

(a)

(b)

(c)

Figure 8

Original gene (different class label with different color) and corresponding Augmented gene with respect to different filter methods for Breast Cancer dataset

Seven figures for seven different filter score function are shown here. In each figure the original gene and augmented gene are plotted with respect to sample class label. X-axis represents class label while Y-axis represents expression value. Two different class labels are represented by Blue and Red color. The difference of expression values of two classes in the augmented gene shows class discrimination ability of that gene. Gene number is the column number in the original dataset.