International Journal of Cognitive Informatics and Natural Intelligence
Volume 15 • Issue 4 • October-December 2021

Classification of Gene Expression Data Using Feature Selection Based on Type Combination Approach Model With Advanced Feature Selection Technology

Siddesh G. M., Ramaiah Institute of Technology, Bengaluru, India
Gururaj T., Ramaiah Institute of Technology, Bengaluru, India

ABSTRACT

A key step in addressing the classification issue was the selection of genes for removing redundant and irrelevant genes. The proposed type combination approach-feature selection (TCA-FS) model uses the efficient feature selection methods, and the classification accuracy can be enhanced. The three classifiers, K nearest neighbour (KNN), support vector machine (SVM), and random forest (RF), are selected for evaluating the opted feature selection methods and prediction accuracy. The effects of three new approaches for feature selection are improved recursive feature elimination (IRFE), revised maximum information co-efficient (RMIC), as well as upgraded masked painter (UMP). These three proposed techniques are compared with existing techniques and are validated with (1) stability determination test, (2) classification accuracy, (3) error rates of three proposed techniques. Due to the selection of proper threshold on classification, the proposed TCA-FS method provides a higher accuracy compared to the existing system.

KEYWORDS

Classification, Embedded, Ensemble Feature Selection, Feature Extraction, Filter, Gene Expression, Hybrid, Machine Learning, Wrapper

1. INTRODUCTION

There are some issues in the gene expression data. For example by selecting the best extraction method and by reducing the dimensionality of the data. The efficient dimension reduction technique needs to be chosen to reduce the number of non-relevant features present in the dataset. Gene selection is also an important factor in removing essential elements which improve precision (Lamba et al., 2018).

Due to very high dimensionality of gene expression data, biologists would find it difficult to handle the data on gene expression(Bennet et al., 2015). Hence it is tedious to identify such microarray results. In addition, the irrelevant characteristics and noisy data of the gene expression dataset are also present. The statistical approaches are the optimal solution to such a problem. Automatic statistical computation is required to avoid the errors caused during manual calculations. Such problems can be addressed using the learning methods of the machine.

Additionally, irrelevant features may also be available along with noisy data in the gene expression data set. Therefore essential pre-processing methods are needed. The dominant elements that facilitate

DOI: 10.4018/IJCINI.20211001.oa46

the prediction must be extracted from the enormous dataset. This reduction in technology has the advantage of enhancing accuracy, avoiding overfitting, decreasing model complexity and reducing training time. The selection of features (FS) allows the models to efficiently predict by using the remaining functions in the Machine Learning method(Aouf et al., 2019). The test results show that, if and only if the FS is included in the classification stage, the prediction precision can be increased. The accuracy will decrease if the FS is not included in the classification phase(Vanjimalar, Ramyachitra, & Manikandan, 2018).

The methods of feature selection can be broadly divided into three categories due to feature analysis being combined with the nature of the classification model. The division is based on how feature searching is combined with the design of a classification model - filter methods, wrapping methods, and embedded methods. Filter methods measure the relevance of features only by examining the data's intrinsic characteristics. (Haar, Anding, Trambitckii, & Notni, 2019) The quest for an ideal property sub-set, is included in the field of selection techniques known as embedded techniques, which can be found in the integrated areas of feature subsets and expectations. Embedded approaches are much more computational than wrapper models because they need to be interacted. The selection of features can also increase learning accuracy, reduce learning time, and improve learning performance utilising features(Zhao et al., 2010). The selection and extraction of functions(Sun et al., 2005) are two ways of reducing dimensions.

1.1. Big Data for Machine Learning

Extreme, broad or wide samples display very high dimensional data and imbalanced class naming. In many fields of data mining, big data with large data sets and high size have emerged such as text extraction and information recovery methods for the selection of extremes which can sometimes disable conventional methods. However, with various forms of data sets, the value of analysis is significant. The scalability of the machine output of the procedure for data selection is defined as the sensitivity. This includes accuracy, the complexity of time and space efficiency. The selection process should be suitable for data sets of various sizes and is highly time-scalable. The reliability of the selection outcomes is defined as being immune to such variations. If the effects of feature selection vary when samples are introduced or decreased, the reliability of the feature selection process is not deemed to be enough. The stability indicators are measured with weighted consistency, Hamming distance, Tanimoto average index, and so on. Increased estimation and exposure to differential size within function subsets of the characteristics of similarity indexes can measure the cross-sectional size of two subcomponents which would also help estimate the stability loss.

The random forest and random subspace system (RSM) can be named the ensemble feature selection approach since the base classifier is going to be learned by utilizing the various sample sets of space. This enhances features consistency and gives specific training details for the output of ensemble classifiers(Lazar et al., 2012; Bock et al., 2010) has introduced the RSM and Bagging binary ensemble classification approach and this is the GAM. The GAMbag, GAMrsm, GAMens and other methods use the Basic Classificator GAMs and these methods.

2. LITERATURE REVIEW AND SURVEY

Standard approaches to select the characters are divided into four groups: filter, wrapper, embedded and hybrid. Each function is assessed individually in filter approaches. These approaches can easily be expanded to broad data sets; they are low in sophistication and are separate for classification. Measurements such as t-testing and acquisition of information, minimum repetition, maximum relevance (MRR) and euclidean length are common for this purpose. The classifier's efficiency and interdependence of the features play a small part throughout filter selection techniques. Therefore, if the classifier performance is low or redundant, it is unpredictable.

Two deterministic and stochastic systems are separated in the wrapper strategy. Examples include SFS(sequential forward selection) & SBE (sequential reverse dispatch). Examples of these models include Genetic algorithms, and random scaling is the Ant colony as the stochastic paradigm. Graders perform well in the wrapping process, but space and time complexity are becoming more difficult.

Embedded methods are used to evaluate the problem and to select key features using model properties. Techniques like the Decision Tree and the neural network come from this strategy community, but often very complicated. (Guyon et al., 2002) has implemented the popular embedded technique for gene selection and cancer gradation based on the Vector and Recursive Support Function (SVM-RFE) and (Maldonado et al., 2011)suggested that an optimized approach to the dual wording of SVM be introduced as a punitive aspect. Any of the above approaches cannot solve all the problems. The literature then suggests approaches to groups(Ye et al., 2013). The program is hybridly chosen, with the tests included. These strategies have been used. The SVM-RFE and mRMR techniques are hybridised. (Saqib et al., 2020) proposed the MF-GARF: Hybridizing Multiple Filters and GA Wrapper for Feature Selection.

CFS-TGA a new approach, by the Taguchi-genetic algorithm (TGA) and hybrid feature selection(CFS) by (Chuang et al., 2011)have suggested the KNN as the classifier. (Liu et al., 2010) suggested generation by their occurrence frequency and assigning rank, the genetic dynamic algorithm (GADP), of all potential sub-set genes. (Yassi & Moattar, 2014) proposed a microarray collection method in tandem with wrapper strategies to address data shortages. In this proposed work, researchers suggested a feature selection approach to select the least selection of features that could be best classified.

(Apiletti et al., 2012) have implemented another way of selecting functions, called the painter decision-making technique, which selects functions. As a result of their core share of the gene joint estimates between various categories. A graphical machine algorithm paints the term "MaskedPainter." The painter's algorithm prefers the items that should be drawn on their overlap. In the same way, MaskedPainter prioritizes genes dependent on overlaps in their transmission. The masked name refers to the fact that the gene details is in a special format called the gene mask. Various microarray data sets are validated in our method. We have been mainly focusing on multi-class data sets, given that classification problems are often more difficult than binary classes and offer a more realistic assessment of the proposed methods(Jeffery et al., 2006).

Filter algorithms do not take device similarity into consideration. Nevertheless, they are also only cost-effective for massive data sets with their linear time computation complexity(Liu & Yu, 2005). A filter-based heuristic algorithm evaluation Correlation-based Feature Selection (CFS) (Gopika & M.e., 2018)believes it should be autonomous and closely connected with sample class labels for a rational form of functions.

The goal of a hybrid algorithm is to automate the generation of the optimally selected feature sub-set by combining the wrapping and filtering strategy heuristic(Liu & Yu, 2005)in connection with the filtering and wrapping, (Xing et al., 2001)suggested that a hybrid approach should be taken to select a high-size microarray feature subset and that the regulation strategy was overtaken by sufficient rates. A new RMIC algorithm is being provided in this analysis to pick the feature wrappers dependent on the Maximal Information Coefficient (MIC) measurement (Reshef et al., 2011)between two variables. The first phase of the RMIC system retains all the functionality for the MIC classifieds and each other and only those with severe classification are held for further screening. RMIC then implements the appropriate analysis technique in order to locate the function subset with the best classification results. The experimental findings indicate that the other algorithms are typically waged with considerably reduced functionality.

The proposed refined Recursive Feature Removal (RFE) (Ding & Wilkins, 2006)differs from each iteration deletion of the sum of the functions. Each iteration has removed 1/(j+1) left functions. (Yu et al., 2010) recommended that RFE be promoted as well. These strategies work best and somewhat reduce the use of time. The rest of the paper deals with the question of choosing functions in order

to ensure that the alternative is compatible. First of all, with a variable phase estimate, we propose a modern RFE method. The meaning of the scale in the iteration hold is a number. Precisely to claim that the estimation of the degree often decreases the sum of contains. When it comes to every point below, the first one stays unchanged.

3. PROBLEM STATEMENT

The work proposes the selection of features from the TCA-FS model, in order to obtain candidates sets of several optimal subsets. This model is first of all applied to different selection methods for features. The outcomes of multiple optimal function subset candidates can then be aggregated according to various laws in order to obtain the optimum sub-sets of functions. In conclusion, the proposed model is tested by three classification algorithms with strong results.

4. DESIGN AND DEVELOPMENT OF TCA-FS MODEL

Figure 1 shows the proposed framework for selection based on Type Combination approach selection. Here researchers choose benchmark data set (http://featureselection.asu.edu/datasets.php.) (Zhu et al., 2007) to verify the results. The data used, will be filtered properly by standard gene filtering techniques before giving as input to the proposed feature extraction process. The detailed description of the three novel feature extraction techniques(IRFE, RMIC, & UMP) will be discussed in next coming sections. In order to increase the efficiency of the classification, the output of the proposed feature extraction process will be aggregated individually. Selected parameters will be chosen using the concept of feature polymerization. Finally, the performance of the model is measured by analyzing three different classifiers(KNN, SVM, & RF).

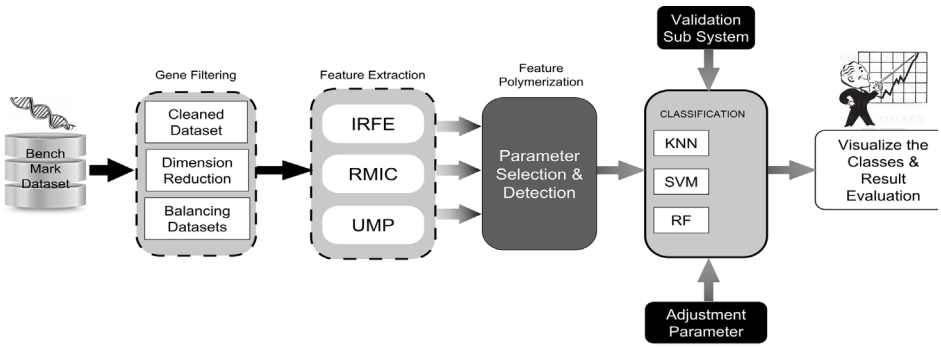
The proposed model uses a collection of indexed, structured FS_1, FS_2, …, FS_t feature subsets, using an IRFE, RMIC and UMP method for selecting features and sorting features according to significance. Researchers use (n-j)/n (A total of 'n' features will be considered) for every j feature in the FSi and achieve the function weight set of the $W_i = \{w_{i1},w_{i2},...,w_{in}\}$. In conjunction with a certain aggregation technique, measure the cumulative weight of each of the elements in set FS_1, FS_2 as an integer average. FS_t, sort the n features by weight and get the final feature sequence W. The top Ge percent features are chosen from the feature sequences for the best feature subset based on the threshold. And the efficiency of the TCA-FS model is tested using many successful classification algorithms using the best feature subset.

The detailed process of the Feature selection based on Type Combination Approach (TCA-FS) involves 5 different phases as below. Input for the model is trained benchmark dataset Genes Ge= $\{x_1, x_2, \ldots, x_{n-1}, x_n\}$. By applying improved-FS algorithms A={IRFE, RMIC, UMP}, & Classifiers C = {KNN, RF, SVM} and the output will be Classified result from Ge.

Phase 1: By referring to the various feature selection algorithm chose the specific feature set that include sets by various component choice calculations.
1: for each Algorithm A_i in {IRFE, RMIC, UMP}.
2: Perform the FS using the algorithm A_i on the dataset Ge.
3: By using the results of A_i sort the features.
4: Sort the feature subset FS_i and return the sorted feature.
Phase 2: From the FS method extract the weight sequence (FS_i).
1: The array of the FS_i in $\{FS_1, FS_2 \ldots FS_t\}$.
2: Perform the for loop on f_j in FS_i.
3: Perform the operation on the w_{ij} using this equation $(n - j) / n$.
4: The generated weights $\{W_1,W_2,\ldots..W_t\}$ from feature set $\{ FS_1, FS_2 \ldots FS_t \}$ are returned.

Phase 3: In this stage, we are going to extract the weighted feature sequence W.
1: Considering the condition, if Arithmetic mean is equal to aggregation strategy.
2: perform the looping on the original feature set FS.
3: $w_i = (w_{1i} + w_{2i} + \ldots + w_{ti}) / t$.
4: if the first condition fails to check for the next condition is geometric mean is equal to aggregation strategy using else if.
5: perform the for loop on the feature set f_i .
6: $w_i = \text{sqrt}(w_{1i} * w_{2i} \ldots * w_{ti})/t$.
7: After performing the weight using step 6 the feature weighted sequence W is returned.
Phase 4: Using the Ge obtain the suitable feature set.
1: By referring weight sequences sort the feature set FS.
2: Equate the FSbest (the first Ge values) to the FS .
3: The obtained best sequence are returned.
Phase 5: Using the KNN, Random forest and SVM perform the classification.
1: Perform the classification Ci on CS and perform the next tasks.
2: The FSbest GE are selected.
3: Learn the classification Ci based on the D.
4: The classification result is returned.

Figure 1. Proposed TCA-FS Model



The researchers concentrated on giving novelty in the feature selection process. A key step in addressing this issue was the selection of genes for removing redundant and irrelevant genes. The three novel feature selection techniques proposed are RMIC, IRFE, & UMP will be discussed in the following sections.

5. IMPLEMENTATION

5.1 Implementation of Proposed Revised Maximal Information Coefficient (RMIC)

On the basis of reciprocal comprehension, the overall information coefficient may be used to accurately describe and evaluate various modes of interaction. The maximum coefficient of information is appropriate in order to find an equal and complete future relationship among pairs of variables in broad sets. For calculating the similarity characteristics, the correlation coefficient between

characteristics is usually used. (Combarro et al., 2005) suggested a linear text categorization application to select the relevant features. The coefficient of Pearson is one of the best-known measurements for relations, as measures are easy to assess and naive. Even Pearson could possibly catch the partnership that was restricted to linear. Various essential links cannot be properly evaluated such as an overlay of characteristics. (Reshef et al., 2011) recently proposed a new relational measure is provided by the Maximum Information Coefficient(MIC). They demonstrate innovatively that MIC is both functional and unfriendly for a wide range of associations.

Revised Maximum coefficient-based screening information examines if there is a linear or other functional connection between the two variables. The RMIC ratio is [0, 1] symmetric and uniform. The variable depends on a high RMIC, while RMIC = 0 defines the variables by two separate variables. Although RMIC appears to suit different kinds of requirements and performs marginally worse than other algorithms like dynamic slicing and t-testing(Jiang et al., 2015), it does encourage potential application in heterogeneous biomedicine data sets with its capacity to process quantitative data.

5.2. RMIC Process: RMIC Removes Redundancy in Features

The process will take input of sample genes with 'k' features and produce output with a subset of features with satisfying performance. As a first, The class labels such as $C = \{C_1, C_2, \ldots. C_3\}$ whereas, C_i belongs to {P, N} and these cases are solved using the binary classification.

Each sample has the k features $< F_1(X), F_2(X), \ldots F_k(X)>$, where F_j is the jth feature. Where Information Features Relevant given by $S = \{F_i \mid RMIC(F_i, C) > t\}$, where t is the threshold pre-set, Redundant features are given by F_i is redundant, if there exists another feature F_j, s.t. $RMIC(F_j,C) > RMIC(F_i,C)$ and $RMIC(F_j,F_i) > RMIC(F_i,C)$ and Criterion of information dominating given by F_j will be maintained if the candidate feature is of highest information relevance in the sub-set $RMIC(F_j, C)$ with variable C and not redundant with selected features.

In the second Step, We use the best first test technique to further increase the amounts of the feature. Our results data show that step 1 selects a subset of characteristics which meet classification efficiency. However, in the previous step which can select dozens or even more than a hundred features that can overcome some large data areas by causing the "big p small n" challenge. The best first search strategy is widely used, so as to further decrease the number of selected functions in a smaller feature subset. The previous step is used for the input of the output sub-set in phase two above and returns the filtered functions. Theoretically, the RMIC is the coefficient of decision (R2). RMIC takes values from 0 to 1, with the logical liberty of 0 and the silence of 1 absolutely.

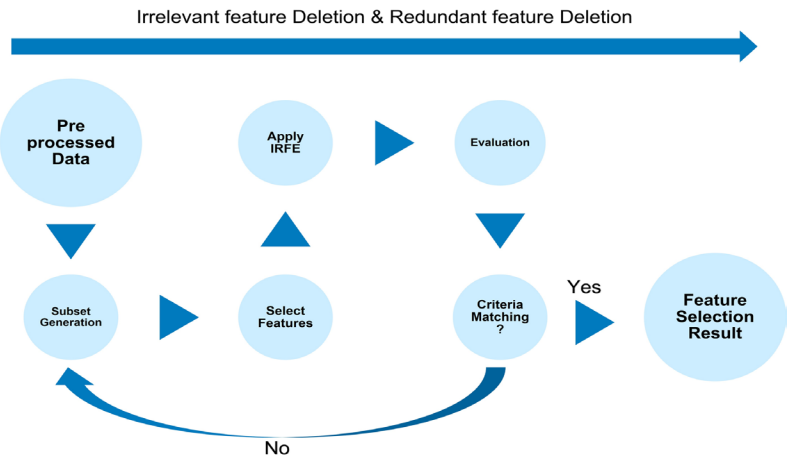
5.3 Implementation of Proposed Improved Recursive Feature Elimination(IRFE)

The data for gene expression are biologically accurate and should therefore not be randomly altered. The approach proposed here is to overcome the class inequality of microarray data without overfitting and lack of knowledge, preserving biological significance. The training data are used for at least three purposes in the current RFE algorithm: predictor selection, model fitting and evaluation of performance. Unless there are large numbers of samples, particularly with regard to the number of variables, one static training package may not satisfy these requirements. Figure 2 demonstrates the possible flow of the design. Microarray gene expression data is part of our input system. In order to ensure noise and in continuity, first data are pre-processed. This approach is then used to create a subset of the dataset. After the evaluation process if criteria match then it will continue to get results from the sub-set generation phase. The appropriate selection methods are then employed in order to select only important features, i.e. the genes by using IRFE.

5.3.1 IRFE: Resampling and External Validation

Cautions should be taken to account for variability caused by the selection of the feature when calculating performance because the feature selection is part of the model building process. For

Figure 2. Process of proposed IR Feature Elimination



examples, RFE may estimate model output with the algorithm. This shows that the resampling was unsuitable for measuring performance, leading to models with poor performance on new samples.

It is proposed that the Algorithmic steps be confined in an external sampling layer (e.g. 10-fold interval), in order to obtain output estimates which integrate the variance due to functional selection. Algorithm 1 shows a resampling version. Here in each iteration, every sample is separated into test and training data as in step 1 and 2. Tuning will be done on training data and then the sample will be predicted as given in step 3 and 4. The process of resampling is done by calculating the sample importance as in steps 6 to 11. Based on sample importance, calculate the performance using prediction samples. Lastly, in step 14 to 16 estimation of final predictors will be done. Although this offers stronger efficiency predictions, it is more numerical. The first For Loop in the algorithm can easily be paralleled for users with access to machines with multiple processors. Another complication of using resampling is that at every iteration multiple lists of "best" predictors are generated. At first, this may seem like an annoyance but offers an estimation of predictor value more probabilistically than an appraisal focused on a single defined data collection. The best-retained predictors can be defined with a consensus ranking at the end of the algorithm.

Algorithm 1: IRFE with resampling

1. For each sample do
2. Separate test and training data
3. Tune the model based on training data using predictors
4. Prediction of samples
5. Calculate sample importance
6. For each subset size C_i, i = 1 … C do
7. Save C_i most important sample
8. Tune the model based on training data using C_i predictors
9. Prediction of samples
10. Recalculate sample importance
11. end.
12. end.
13. Calculate the performance over C_i using prediction samples
14. Determine the suitable number of predictors

15. Estimation of final predictors
16. Fit the final model based on optimal C_i using the original training set.

5.3.2 Improved Recursive Feature Elimination (IRFE) is Performed by Varying the Step Size

We have suggested IRFE together with variable step size in order to limit the opposite impact on the choice of functions. The first step has been configured and then halved, i.e. the amount of functions to be omitted is halved relative to the original scale. Repeat until the step size is reached. It is explained further from two angles: that is, the phase size of an iteration differs between the width and small and does not change. The number of applications to drop depends on their version, revised law and quantity. Ruggedness also gradually strengthens the process for removing characteristics. Normally there is a large number of genes (features) in the data set with microarray gene expression; very few of them relate closely to the class label. Thus, relatively more initially unrelated genes (features) can definitely be deleted. Therefore, in earlier phases, we can set a quite large stage size in order to minimize the amount of iterations. Slowly raise step size and particular consideration is paid to corresponding applications. This ensures the collection standard of the function.

5.4 Implementation of Proposed Upgraded Masked Painter(UMP)

The name "MaskedPainter" is derived from the algorithm of the computer graphic painter. The painter's algorithm prioritizes the objects to be painted on the basis of overlaps. The MaskedPainter has also given priority to genes based on overlaps in their range of expression. The masked term is because the specifics of the gene are in a certain style known as the gene mask.

5.5. The Upgraded Masked Painter Approach

This strategy calculates the similarity between the groups of each gene. The technique was introduced for small overlaps and the other way around to give lower scores. The consistency of the articulation of time was included. By calculating the precision of the classification given by a classifier, we contrasted improved maskable output with specific techniques of selection. For nearly all cases the improved masked painter is objectively even more effective than other approaches. Tests have been conducted on various gene cardinalities and various classificators. Finally, in the planned technique, the identified genes were checked for their biological relevance.

The H micro-array data is a matrix of expression of the Eq genes. (1) where the gene is each row and the sample is each column. For each sample, the functions of all genes studied are calculated.

$$H = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1M} \\ h_{21} & h_{22} & \dots & h_{2M} \\ \dots & \dots & \dots & \dots \\ h_{N1} & h_{N2} & \dots & h_{NM} \end{bmatrix} \tag{1}$$

Let h_{ij} be measuring gene i for sample j where i = 1,….., N & j = 1,……,M. The domain of class labels is C, and the label Lj of the sample j is given a single value. Certain genes may determine samples from the class since the interval of their expression within that class does not overlap with the intervals of expression of other classes.

Some genes are important to a single class. Some class expressions are concentrated in a little change, which is not the same as those of alternative class expressions. Due to the fact that values for those classes are mainly overlapping, the equal gene cannot be distinguished between classes. The MaskedPainter characterizes each gene first with a mask, which shows clearly, that the gene

can assign samples for training to the right class. The overlap value is a quality index. Genes with fewer similarity are more relevant, as they can easily distinguish the samples. A gene's main class is the class that does not cover the majority of samples. The MaskedPainter describes as the minimum range of genes that will better cover the sample of the training data; the lowest sub-set of genes is the final gene selection when compared with the maximum classification of the category.

For each gene, a gene mask has been calculated. The gene mask indicates what samples are specifically assigned to the correct class. The Sample Range values is a string of 0s and 1s generated by measuring the overlap of the class expression spectrum for each gene. The overlapping score calculation is assigned for each gene to score. It determines the difference between its central intervals of expression. The main interval of expression is the interval of expression obtained by flattening the effect of the outliers, which is defined for the individual classes. A density-based methodology is proposed to measure central expression distances. Each gene is sometimes allocated to a dominant class(dc), further helps in removing redundancy. Max gene subset collection is chosen to provide the strongest training sample through the assessment of genetic masks and overlap scores. With a selfish method, the strongest sampling scope at a small cost of calculation is used. Genes not in the minimum sub-set are divided into the increasing value of the overlap for each dominant class. Top genes are used for the ultimate gene collection, offering the latest gene series.

By considering the Genes as shown in Table 1. Every gene is linked to the overlapping score(os), gene mask (string 0 and 1), and dominant class (dc). For example, the gene Ge4 has a value of 0100101 (this means that the second, fifth and seventh samples are unambiguously classified), a score of 0.77, with a dominant class being class 1. The first gene in the minimal subsets, selected by the greedy technique, is Ge 3, as it has the highest number of sets of 1 bit and the smallest overlap score (the same as Ge 8 and Ge 6). Then, genes with the best additional masks Ge7 selected, Ge1 and Ge4 both of this have the same bit number as 1. Again, because of their lower overlap value, Ge 7 is selected. Finally, the only gene with an extra cap, Ge1, will be used. In this case, there are three genes.

The genes are reported as G3, Ge7, and Ge1 for the minimum gene subset. The majority of genes are categorized and separated by an upward overlap. The gene rank is composed by a round-robin selection of the top genes of each dominant class (e.g., Ge 4 for class 1, Ge 8 for grade 2, Ge 5 for grade 3, Ge 2 for grade 1, Ge6 for grade 2 etc.), results show that five to six Ge values are necessary for performing the investigation. Out of which, the top performing 3 genes will be selected for minimum gene sub-set.

5.6. Computation of the Gene Mask

The class expression interval for equation (1), taking into consideration the gene matrix. Let i be a gene belongs to C-class of M sample. For each gene (one per class). The class expression interval of gene i and class k is as shown in (2):

$$I_{i,k} = \left[\underset{i,k}{min}, \underset{i,k}{max} \right] \tag{2}$$

For the k-classes, the maximum and the minimum values are $min_{i,k}$ and $max_{i,k}$.

$$mask_{ij} = \begin{cases} 1\text{if} \left(h_{ij} \in I_{i,a} \right) \wedge \exists b \neq a \mid h_{ij} \in I_{i,b} \\ 0\text{otherwise} \end{cases} \tag{3}$$

Table 1. Masks, overlap score, & Dominant Class for different Genes

Genes	Masks	OS	DC
Ge1	0001101	0.21	2
Ge2	0100100	0.46	1
Ge3	1100111	0.68	3
Ge4	0100101	0.77	1
Ge5	1000100	1.34	3
Ge6	1100111	1.05	2
Ge7	1010101	0.30	2
Ge8	1110101	0.79	2

The gene mask is an array of M bits, where the number of samples is 'M'. Consider an arbitrary gene bit *i*. Bit *j* of its mask is set to 1 if the corresponding expression value h_{ij} belongs to the class expression interval of a single class, otherwise, it is 0. Formally gene mask is computed as in (3).

5.7. Overlap Score Computation

For every gene i the overlapping scope is going to be computed. We are going to omit the subscript i in the next below formulae for easy readability.

The total expression interval of a gene is defined as the range of its core expression interval limits between the minimum and the maximum. We denote intervals like |W|. The |wt| subinterval is described more precisely as an interval between two ends, as shown in (4). Subinterval is marked with |wt|. Subintervals with larger class overlap, therefore, make the os more effective.

$$os = \sum_{t=1}^{T} k(t) \frac{m_t}{M} \frac{|w_t|}{|W|} \tag{4}$$

In equation (4), where T is the number of subintervals, mt is the number of samples in t, M is the total number of samples. The function k(t) assesses and defines a series of classes which overlap in the subinterval t as in (5).

$$k(t) = \begin{cases} |C_t| & \text{if } |C_t| \geqslant 2 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where |Ct| is the number of classes belonging to the subinterval *t*.

5.7.1. Minimum Gene Subset Selection

The genetic masks data are used to recognize the minimum set of genes that accurately categorize the maximum sample number. It also allows for the elimination of repeated data. The final goal is to define a collection of good quality features surrounding the specimen. The greedy approach(Mahmoud et al., 2014) identifies the gene with the best gene mask in every step, which complements the current global mask. It then adds data to be classified at each stage for the most currently uncovered samples.

Final Gene Selection

A tiny subgroup of genes includes the lowest amount of genes with the largest coverage of the sample as the number 1s of the gene mask decreases. Still, larger genes can either improve the accuracy or be requested from the user in the classification of invisibility data.

6. SINGLE BASE CLASSIFIER

The feature selection is the assembly of basic classificators made up of the various functional areas obtained through feature selection. The workflow for the single base classifier is shown in Figure 3. By providing different training information, which improves selection stability and the performance of the system will be more.

A single base class method is adopted in our model. In this context, filters generate multiple subsets of functions first, and then the cross-sectional strategy combines into a single package(Abeel, et al., 2009). Usually, this arrangement makes a montage subset with greater precision than one filter subset and does not provide a hybrid solution for specific classification systems, but requires many common classification structures

The feature collection help to reduce the effect of large-scale learning algorithms by creating sets from various simple classification systems.

7. DATA COLLECTION

In order to carry out extensive experiments, we selected three benchmark gene expression microarrays. They are all widely used and accessible online by many researchers in this field. In (https://jundongl.github.io/scikit-feature/datasets.html) there are data collection for leukemia (ALL AML), Central Nervous System, and Prostate GE can be contained and is available in MAT format i.e., matrix format of features. In all of them, problems with class imbalances are common. The information is shown in Table 2.

The current segment discusses an analysis to evaluate the efficiency of the three application selection algorithms by contrasting them which are IRFE, UMP and RMIC technologies. Three micro-array expressions of various sizes are used as seen in Table 2. Performance is evaluated using 3 different classifiers, i.e. kNN, SVM and RF, and an error rate and stability analysis is identified. Table 2 displays the data sets used to do the study. It must be remembered that the optimum efficiency of these classificators depends on a suitable collection of tuning parameters. For all classifiers, the parameter and number of genes to be used is selected from values2,3,4,5,6,7,8,9,10,20,30, and 40 by 10-fold cross validation on the specified data training part. The classifiers were considered with 50 replications of 10-fold cross validation for each sample, gene selection strategy, and a variety of functions choose.

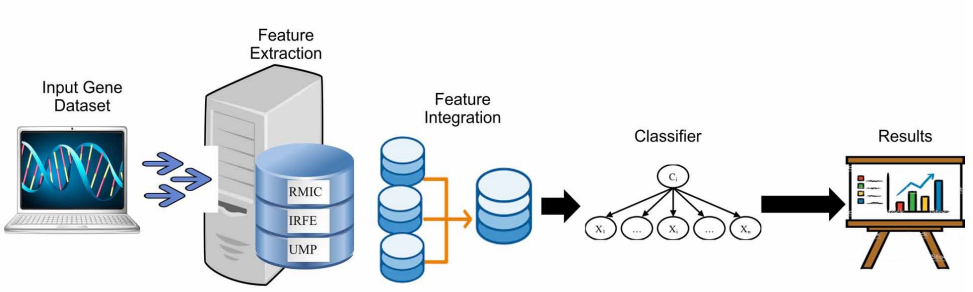
8.RESULTS AND DISCUSSIONS

This approach has been applied by programming to validate the validity of the feature selection process based on the interface sorting suggested in this article. For the experimental purpose the environment chosen are as follows: 8GB RAM with the windows 10 Operating system of 64 bit, the scripting language is Python 3.7 and the processor employed is Intel Core i5-7200 CPU-2.70 GHz.

8.1. Assessment of FS Techniques

In this segment, the execution of FS procedures in three classifiers appears. The discoveries of exactness are based as they were on chosen qualities and demonstrated by algorithms. Classification errors from a variety of FS algorithms on “ALL-AML” are shown in Table 3. Similarly, data collected

Figure 3. Single base classifier framework



for "Central Nervous System" and " Prostate GE". The algorithm and the ranking agents used are different characteristics. The mean classification error rate is determined depending upon the amount of genes defined by the methods for the fifty repetitions of the 10-fold cross validation. Minimum mistake rates for different subsets of features and mean error rates are also shown for all medium errors.

One of the UMP filtering approach data is shown in Table 3. Results are better than other methods in the majority of iterations for all data sets. All three FS strategies provide sensible accuracy in contrast with the other two classifiers; UMP for all gene sub-sets with chosen sizes is the highest. In addition to the cumulative aggregate error in classification and lower error rates, the UMP test will give a better result for each classifier. The UMP strategy is appropriate for kNN, SVM and RF where the genes are five or more. UMP methods for small gene sub-sets are more reliable than other methods.

8.2. Evaluation of Distinctive Classifiers

The output of the three classifiers is analyzed on the three data set in this section. Figures 5, 6 and 7 illustrate the performance concerning classification errors. We have applied three different FS techniques, namely IRFE, RMIC, and UMP on three different data "ALL-AML", "Central Nervous System" and " Prostate GE". Here in Figure 4, 5, and 6; X-axis refers to the selected gene, and Y-axis refers to the error value.

Considering Figure 4, that uses "ALL-AML" dataset, the error rates of IRFE and RMIC techniques appears to be costly for all KNN, RF, and SVM classifiers. At the same time, the UMP technique helped in getting higher precision compared to the other two techniques. Among the three classifiers, RF performance is better compared to other classifiers.

The same observations are made from Figure 5 and Figure 6 where we have used "Central Nervous System" and " Prostate GE" dataset for the same three classifiers. The results also clarify that the performance of the UMP technique is superior with RF classification. This reveals that utilizing all three sorting methods, UMP technique has a minimum error rate, and RF is the strongest performative classifier for all chosen gene collection sizes. In all three filtering strategies, the SVM misbehaved.

Table 2. Description of the Data set used in the model.

Data Set	Feature	Instances	Classes
ALL-AML	7129	72	2
Central Nervous System	7129	60	2
Prostate GE	5966	102	2

Table 3. Average miscalibration rates are given on "ALL-AML" data over all 50 repeats for 10 fold cross-validation on the different genes selected from the algorithms supplied by the KNN, SVM, and RF Classification.

	KNN Method			SVM Method			RF Method		
Selected Gene	IRFE	RMIC	UMP	IRFE	RMIC	UMP	IRFE	RMIC	UMP
2	0.076	0.045	0.032	0.017	0.013	0.01	0.015	0.004	0.001
3	0.056	0.04	0.09	0.058	0.066	0.044	0.021	0.012	0.0021
4	0.064	0.065	0.056	0.115	0.083	0.0811	0.021	0.0051	0.0031
5	0.043	0.036	0.033	0.218	0.116	0.087	0.02	0.007	0.004
6	0.065	0.055	0.047	0.22	0.113	0.098	0.018	0.006	0.007
7	0.095	0.065	0.081	0.153	0.123	0.068	0.017	0.004	0.007
8	0.09	0.076	0.059	0.12	0.079	0.073	0.013	0.002	0.003
9	0.118	0.089	0.047	0.13	0.064	0.054	0.017	0.005	0.004
10	0.115	0.09	0.047	0.101	0.05	0.055	0.022	0.007	0.003
20	0.11	0.085	0.072	0.098	0.045	0.024	0.028	0.004	0.005
30	0.132	0.095	0.064	0.085	0.033	0.022	0.033	0.008	0.006
40	0.113	0.096	0.035	0.089	0.045	0.0191	0.03	0.003	0.001
50	0.117	0.088	0.042	0.085	0.043	0.025	0.026	0.007	0.001
AVG	**0.091846**	**0.071154**	**0.054231**	**0.114538**	**0.067154**	**0.050785**	**0.021615**	**0.0057**	**0.003631**
MIN	**0.043**	**0.036**	**0.032**	**0.017**	**0.013**	**0.01**	**0.013**	**0.002**	**0.001**

8.3. Stability Analysis

(Lausser et al., 2011) suggested the stability determination test for various screening techniques. A small category of variables or features will be selected regularly and will have several features seldom or never selected at all. The index scoring values vary from one when the variable subset is the same in each m sample, and only all genes are selected once for a functional section. Because index results cannot reach zero, the method would be more stable if higher stability rates were achieved than low stability rates. Table 4 shows the stability results of the 3 FS techniques on Prostate GE data sets. Table 4 shows, of course, that the UMP approach offers the most outstanding stability scores for all dimensions, except the gene subset sizes, i.e. 15 with 0.617, in contrast to the other two models.

8.4 Classification Accuracy Results

Although the α-threshold levels are low for all three groups from 10 to 100 percent, the model's predictive accuracy is contrasted under different conditions. Figure 7. shows the experimental results. This indicates the quality of classification obtained by RF for functional selection sub-sets under specific α limits. The experimental results shown in Figure 7. is the subset of features collected using the UMP selection process; there is better precision than the single system. The middle aggregation approach is among other approaches used in the ALL-AML dataset. More precise results than other current approaches have been achieved through suggested IRFE, RMIC and UMP methods. The value of α in [0.1] and the accuracy of classification vary. Too many functions lead to data redundancy and data noise due to the correlation between different functions, which reduces the performance of the classification.

Figure 4. "AAL-AML" dataset performance with regard to classification errors

Figure 5. "Central Nervous System" dataset performance with regard to classification errors.

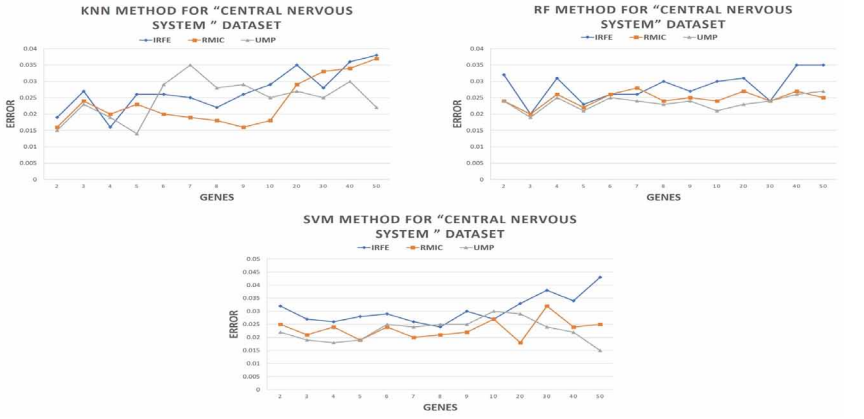


Figure 6. "Prostate GE" dataset performance with regard to classification errors

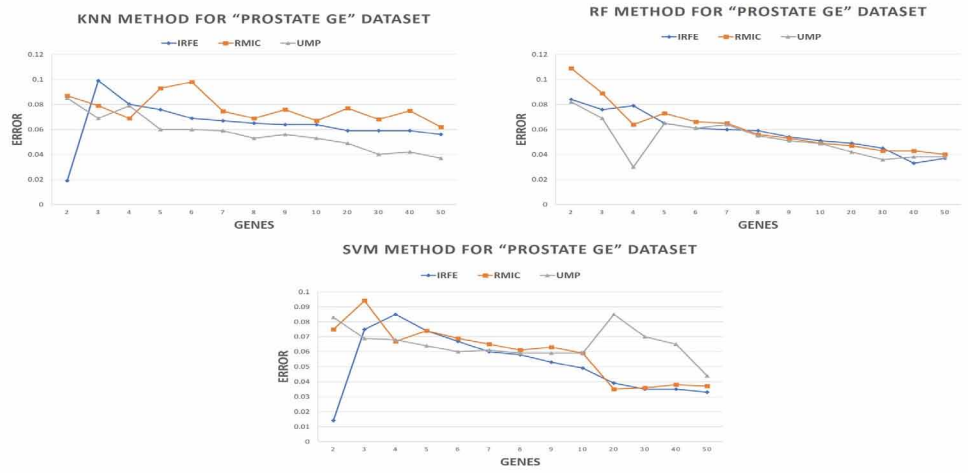
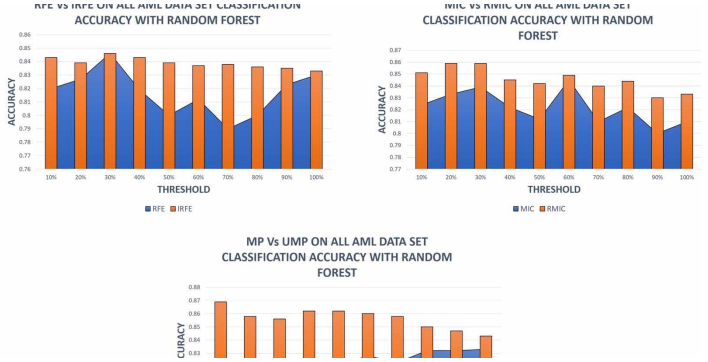


Table 4. Stability values are given for " Prostate GE" data over all 50 repeats for 10 fold cross-validation on the three different FS techniques.

Selected Genes	IRFE	RMIC	UMP
5	0.6	0.567	0.615
10	0.598	0.524	0.669
15	0.61	0.777	0.617
20	0.589	0.797	0.809
25	0.526	0.799	0.804
30	0.616	0.73	0.879
35	0.601	0.625	0.756
40	0.591	0.639	0.76
45	0.599	0.642	0.774
50	0.618	0.638	0.651

Figure 7. Performance evaluation of IRFE, RMIC, and UMP in comparison with existing techniques using Random Forest Classification.



9. CONCLUSION

A comparative study is performed on the various set of supervised classification algorithms. This study helps to identify the algorithm which is most suitable for classifying the gene expression data. Proposed algorithms are validated with (i) The accuracy of novel IRFE in comparison with RFE for different threshold reached 83.89% on an average. Similarly, in comparison with existing MIC and MP, proposed RMIC and UMP, for different threshold reached at 84.52% and 85.65% respectively. The accuracy of the developed Feature selection techniques was shown to reach 84.686% in an average, (ii)The UMP FS offers the highest stability scores for all dimensions, except the gene subset sizes, i.e. 15 with 0.617, in contrast to the other two FS technique. The proposed model shows that the stability of TCA-FS is more when compared to other existing models. Most of the existing models use a single feature selection method. The results show the comparable classification performance of the proposed method, (iii) The proposed Type Combination Approach (TCA-FS) function selection system outperforms other methods with minimum errors in SVM, Random Forests, and KNN classifier.

The UMP in classification reduced error rates would provide a better outcome for each classifier with cumulative median error. UMP methods are more likely to be reliable for small gene sub-sets.

10. FUTURE WORK

The introduction of multivariate selection techniques could be one of the most exciting bio-information careers possible. The advancement, in the well developed collection methods, is a major step towards increasing the vigour of the selected sub-sets of components.

ACKNOWLEDGMENT

This research was supported by Ramaiah Institute of Technology (MSRIT), Bangalore-560054 and Visvesvaraya Technological University, Jnana Sangama, Belagavi -590018.

REFERENCES

Abeel, T., Helleputte, T., Peer, Y. V., Dupont, P., & Saeys, Y. (2009). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics (Oxford, England)*, *26*(3), 392–398. doi:10.1093/bioinformatics/btp630 PMID:19942583

Aouf, M., Sharawi, A., Samir, K., Almotatiri, S., Bajahzar, A., & Kareem, G. (2019). Gene Expression Data For Gene Selection Using Ensemble Based Feature Selection. *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*. doi:10.1109/ICICIS46948.2019.9014722

Apiletti, D., Baralis, E., Bruno, G., & Fiori, A. (2012). MaskedPainter: Feature selection for microarray data analysis. *Intelligent Data Analysis*, *16*(4), 717–737. doi:10.3233/IDA-2012-0546

Bennet, J., Ganaprakasam, C., & Kumar, N. (2015). A Hybrid Approach for Gene Selection and Classification using Support Vector Machine. *The International Arab Journal of Information Technology*, *12*, 695–700.

Bock, K. W., Coussement, K., & Poel, D. V. (2010). Ensemble classification based on generalized additive models. *Computational Statistics & Data Analysis*, *54*(6), 1535–1546. doi:10.1016/j.csda.2009.12.013

Chuang, L., Yang, C., Wu, K., & Yang, C. (2011). A hybrid feature selection method for DNA microarray data. *Computers in Biology and Medicine*, *41*(4), 228–237. doi:10.1016/j.compbiomed.2011.02.004 PMID:21376310

Combarro, E., Montanes, E., Diaz, I., Ranilla, J., & Mones, R. (2005). Introducing a family of linear measures for feature selection in text categorization. *IEEE Transactions on Knowledge and Data Engineering*, *17*(9), 1223–1232. doi:10.1109/TKDE.2005.149

Ding, Y., & Wilkins, D. (2006). Improving the Performance of SVM-RFE to Select Genes in Microarray Data. *BMC Bioinformatics*, *7*(S2), S12. doi:10.1186/1471-2105-7-S2-S12 PMID:17118133

Gopika, N., & M.E., A. M. (2018). Correlation Based Feature Selection Algorithm for Machine Learning. *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1-3), 389–422. doi:10.1023/A:1012487302797

Haar, L., Anding, K., Trambitckii, K., & Notni, G. (2019). Comparison between Supervised and Unsupervised Feature Selection Methods. *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*. doi:10.5220/0007385305820589

Jeffery, I. B., Higgins, D. G., & Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, *7*(1), 359. doi:10.1186/1471-2105-7-359 PMID:16872483

Jiang, B., Ye, C., & Liu, J. S. (2015). NonparametricK-Sample Tests via Dynamic Slicing. *Journal of the American Statistical Association*, *110*(510), 642–653. doi:10.1080/01621459.2014.920257

Lamba, M., Munjal, G., & Gigras, Y. (2018). Feature Selection of Micro-array expression data (FSM) - A Review. *Procedia Computer Science*, *132*, 1619–1625. doi:10.1016/j.procs.2018.05.127

Lausser, L., Müssel, C., Maucher, M., & Kestler, H. A. (2011). Measuring and visualizing the stability of biomarker selection techniques. *Computational Statistics*, *28*(1), 51–65. doi:10.1007/s00180-011-0284-y

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., & Nowe, A. (2012). A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *9*(4), 1106–1119. doi:10.1109/TCBB.2012.33 PMID:22350210

Lee, C., & Leu, Y. (2011). A novel hybrid feature selection method for microarray data analysis. *Applied Soft Computing*, *11*(1), 208–213. doi:10.1016/j.asoc.2009.11.010

Liu, H., Liu, L., & Zhang, H. (2010). Ensemble gene selection for cancer classification. *Pattern Recognition*, *43*(8), 2763–2772. doi:10.1016/j.patcog.2010.02.008

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, *17*(4), 491–502. doi:10.1109/TKDE.2005.66

Mahmoud, O., Harrison, A., Perperoglou, A., Gul, A., Khan, Z., Metodiev, M. V., & Lausen, B. (2014). A feature selection method for classification within functional genomics experiments based on the proportional overlapping score. *BMC Bioinformatics*, *15*(1), 274. doi:10.1186/1471-2105-15-274 PMID:25113817

Maldonado, S., Weber, R., & Basak, J. (2011). Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences*, *181*(1), 115–128. doi:10.1016/j.ins.2010.08.047

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., Mcvean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., & Sabeti, P. C. (2011). Detecting Novel Associations in Large Data Sets. *Science*, *334*(6062), 1518–1524. doi:10.1126/science.1205438 PMID:22174245

Saqib, P., Qamar, U., Khan, R. A., & Aslam, A. (2020). MF-GARF: Hybridizing Multiple Filters and GA Wrapper for Feature Selection of Microarray Cancer Datasets. *2020 22nd International Conference on Advanced Communication Technology (ICACT)*.

Sun, Z., Huang, D., & Cheun, Y. (2005). Extracting nonlinear features for multispectral images by FCMC and KPCA. *Digital Signal Processing*, *15*(4), 331–346. doi:10.1016/j.dsp.2004.12.004

Vanjimalar, S., Ramyachitra, D., & Manikandan, P. (2018). A Review on Feature Selection Techniques for Gene Expression Data. *2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. doi:10.1109/ICCIC.2018.8782294

Xing, E. P., Jordan, M. I., & Karp, R. M. (2001). Feature selection for high-dimensional genomic microarray data. In *Proceedings of the Eighteenth International Conference on Machine Learning* (*Vol. 1*, pp. 601-608). Academic Press.

Yassi, M., & Moattar, M. H. (2014). Robust and stable feature selection by integrating ranking methods and wrapper technique in genetic data classification. *Biochemical and Biophysical Research Communications*, *446*(4), 850–856. doi:10.1016/j.bbrc.2014.02.146 PMID:24657268

Ye, Y., Wu, Q., Huang, J. Z., Ng, M. K., & Li, X. (2013). Stratified sampling for feature subspace selection in random forests for high dimensional data. *Pattern Recognition*, *46*(3), 769–787. doi:10.1016/j.patcog.2012.09.005

Yu, H., Huang, F., & Lin, C. (2010). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, *85*(1-2), 41–75. doi:10.1007/s10994-010-5221-8

Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., & Liu, H. (2010). *Advancing feature selection research*. ASU Feature Selection Repository.

Zhu, Z., Ong, Y. S., & Dash, M. (2007). Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition*, *40*(11), 3236–3248. doi:10.1016/j.patcog.2007.02.007