The Effect of Different Classifiers on Recursive Cluster Elimination in the Analysis of Transcriptomic Data

Nurten Bulut
Department of Computer Engineering
Abdullah Gul University
Kayseri, Turkey
0000-0002-1895-8749

Burcu Bakir-Gungor
Department of Computer Engineering
Abdullah Gul University
Kayseri, Turkey
0000-0002-2272-6270

Bahjat F. Qaqish
Department of Biostatistics
University of North Carolina at Chapel Hill
NC, Chapel Hill, USA
0000-0002-5154-2059

Malik Yousef
Department of Information Systems,
Galilee Digital Health Research Center
Zefat Academic College
Zefat, Israel
0000-0001-8780-6303
malik.yousef@gmail.com

2023 Innovations in Intelligent Systems and Applications Conference (ASYU) | 979-8-3503-0659-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASYU58738.2023.10296645

***Abstract*—Gene expression data with limited sample size and a large number of genes are frequently encountered in genetic studies. In such high-dimensional data, identification of genes that distinguish between disease states is a challenging task. Feature selection (FS) is a useful approach in dealing with high dimensionality. Support Vector Machines Recursive Cluster Elimination (SVM-RCE) is a technique for FS in high-dimensional data. The SVM-RCE approach has been utilized for identification of clusters of genes whose expression levels correlate with pathological state. A key step in SVM-RCE is the use of an SVM classifier to assign an area under the curve (AUC) score to each gene cluster based on its ability to predict class labels. In this study, we investigate the use of alternative classifiers in the cluster-scoring step. Specifically, we compare Support Vector Machines, Random Forest, XgBoost, Naive Bayes, and linear logistic regression. In addition to AUC score performance evaluation, the algorithms are compared in terms of the number of selected genes at different levels of clustering and in terms of the running time.**

Keywords— Recursive Cluster Elimination, Feature Selection, Clustering, Gene Expression Data Analysis.

I. INTRODUCTION

DNA microarrays and RNA sequencing have been recognized as potentially valuable techniques in the identification and categorization of diverse medical conditions such as cancer. Gene expression datasets typically contain a small number of samples and tens of thousands of genes, which makes the classification a challenging task. The identification of genes that differentiate between healthy individuals and patients is a challenging task in the analysis of high-dimensional data. The primary objective of biomarker identification in transcriptomics data analysis is to identify a minimal set of predictive genes while simultaneously maximizing the classification accuracy.

The challenge of categorizing biological specimens was expounded upon by Jain et al [1]. They presented a new hybrid method for cancer classification. Their method integrates Correlation-based Feature Selection (CFS) and enhanced-Binary Particle Swarm Optimization (iBPSO). They employ a Naive-Bayes classifier along with 10-fold cross-validation to identify a concise group of predictive genes that can correctly classify biological samples of cancers with two or more classes. Their method demonstrated enhanced precision in classification and gene selection across most of the data sets they analyzed [1].

Aziz et al [2] proposed a method that integrates fuzzy backward feature elimination (FBFE) and independent component analysis (ICA) for the purpose of selecting independent DNA microarray data components. Their aim was to improve the effectiveness of Support Vector Machine (SVM) and Naive Bayes (NB) classifiers, while also reducing their computational cost. Aziz et al. presented a comparative analysis between the proposed method and Principal Component Analysis (PCA), a commonly employed feature extraction technique. The comparison was carried out on five distinct DNA microarray datasets. Their method demonstrated enhanced classification efficacy with a reduced gene count in comparison to PCA. The utilization of Receiver Operating Characteristic (ROC) analysis offered an optimal methodology for the selection of genes that are appropriate for both classifiers.

Dashtban et al. [3] have proposed a novel method for gene selection within the binary feature selection domain, drawing inspiration from biological mechanisms. The approach is intended to efficiently optimize multiple objectives simultaneously. The Bat Algorithm, which is a metaheuristic algorithm for global optimization, is improved through the integration of multi-objective operators, local search tactics, and a random walk scheme founded on social learning principles. The study employs a hybrid model that integrates the Fisher criterion and a multi-objective version of the bat algorithm for binary feature selection. This model is utilized to examine three commonly used microarray cancer datasets. The aim is to identify informative genes. The experimental findings have revealed novel combinations of informative biomarkers that exhibit associations with other relevant studies.

Sharma and Rani [4] have presented a gene selection framework, identified as C-HMOSHSSA, which utilizes a combination of the multi-objective spotted hyena optimizer (MOSHO) and the salp swarm algorithm (SSA). The Salp Swarm Algo-rithm (SSA) is capable of preserving diversity, however, it is hindered by the burden of upholding the requisite information. Conversely, MOSHO computation

necessitates minimal computational resources, rendering it a suitable method for preserving essential data. The algorithm suggested is a hybrid algorithm that combines the features of both SSA and MOSHO to enhance its exploration and exploitation capabilities. The objective is to identify a minimal set of genes while simultaneously optimizing the accuracy of classification. The study involved the training of four distinct classifiers on seven datasets with high dimensions. The subset of features utilized in the training process was obtained through the application of a hybrid gene selection algorithm proposed by the researchers. The findings indicate that the suggested methodology exhibits a notable superiority over the currently available cuttingedge techniques.

Shukla et al. [5] presented a two-step process for removing noisy and duplicated genes using a hybrid model that combines ensemble gene selection (EGS) and the AGA algorithm. The method employs EGS as a filtering mechanism to identify the most highly ranked genes, which will subsequently be subjected to the AGA algorithm. In addition, the AGA algorithm is integrated with Support Vector Machines (SVM) and Naive Bayes (NB) techniques to identify the highest-ranked genes obtained from EGS genes, with the aim of identifying the most informative genes that can effectively facilitate cancer classification. The aforementioned procedure continues until an acceptable level of accuracy is achieved utilizing restricted gene subsets. The results obtained from the experiment indicate that the framework proposed offers supplementary assistance in achieving a substantial decrease in cardinality. Furthermore, it outperforms existing gene selection techniques in terms of accuracy and the identification of an optimal number of genes.

Yousef et al. [6] suggested a novel approach for choosing genes of importance called recursive cluster elimination (RCE) rather than recursive feature elimination (RFE). The SVM-RCE they refer to uses K-means, a clustering approach, to find correlated gene clusters and Support Vector Machines (SVMs), a supervised machine learning classification method, to score (rank) those gene clusters for classification. K-means initially clusters genes. Recursive cluster elimination (RCE) then removes gene clusters that perform poorly in categorization. SVM-RCE finds the clusters of associated genes that differ most between sample classes. Instead of removing genes based on their discriminant weights, using gene clusters improves the supervised classification accuracy of the same data. Then, Yousef et al. [7] suggested a new method referred to as SVM with Recursive Network Elimination (SVM-RNE). They show that integrating network information with SVM-based recursive feature reduction increases performance and biological interpretability. In their next study [8], they adapted the SVM-RCE method to the KNIME in order to make it easier to apply. They make further investigations about the topic. The researchers are engaged in a follow-up investigation of two prior research endeavors, namely SVM-RCE and SVM-RCE-R. SVM-RCE-OPT [10] is a follow-up study that aims to determine the most favorable weights for the scoring function proposed in the SVM-RCE-R study by utilizing optimization methodologies. It has been discovered that the performance of SVM-RCE can be improved in most cases by determining the optimal weights for the scoring function. It has been demonstrated that in certain instances, there is a significant enhancement in performance of up to 10% with regards to accuracy and AUC. Improving the algorithm's performance may enhance the likelihood of extracting a subset of genes that are associated with the class of a microarray sample.

This study following previous studies elaborate on the components of the SVM-RCE algorithm. SVM RCE introduces an approach, to selecting and classifying genes in gene expression datasets. This method combines the K means clustering algorithm with Support Vector Machines (SVMs) to effectively classify and rank gene clusters. It addresses the challenges of analyzing data, where accurate gene selection is difficult due to limited samples and numerous genes. SVM RCE identifies gene clusters that greatly improve classification performance by merging clustering and classification techniques. One notable aspect of SVM RCE is its process of eliminating features. By evaluating and discarding gene clusters that have impact, on classification accuracy SVM RCE progressively enhances the feature set by retaining only informative gene clusters. Re genes after each elimination step allows for exploring potentially more informative clusters thereby boosting accuracy. This study conducted a comparative analysis to evaluate the effectiveness of SVM-RCE in combination with other algorithms, including Random Forest (RF), XGBoost, Naive Bayes (NB), and Logistic Regression (LR).

II. Materials and Methods

A. Dataset

This study gathered five human gene expression datasets which were procured from NCBI's Gene Expression Omnibus, as documented in [9]. The datasets contained both patient samples and healthy controls. A summary is given in Table 1.

TABLE I. Details about the 5 utilized data sets. The column # of Genes shows the total number of genes. While # of Samples indicate the total number of samples, "Classes" column shows the number of positive samples (patients) and negative samples (healthy controls).

Dataset	# of Genes	# of Samples	Classes
GDS2547	12647	164	positive: 89 negative: 75
GDS3268	44290	200	positive: 129 negative: 71
GDS3646	22186	132	positive: 110 negative: 22
GDS3875	22646	117	positive: 93 negative: 24
GDS5037	41001	108	positive: 88 negative: 20

B. Methods

Assume that D is a gene expression dataset with two-classes, organized so that rows represent samples while columns represent genes. The samples are split into a training set and a test set, with respective data denoted D_train and D_test, respectively. Thus, D_train and D_test have the same number of columns as D.

The SVM-RCE [7,8,9] that is illustrated in Figure 1, requires the user to specify the initial number of clusters k. SVM-RCE utilizes the K-means clustering technique applied on the genes (columns) to cluster them into clusters or groups. Next, for each clusteri i=1,2,..,k its associated two-class sub_dataseti is extracted from the D_train to represent the

identified cluster. The class labels are also extracted. In this stage, we have k two-class sub_datasets that allow us to run an internal cross-validation in order to estimate the performance using a machine learning classifier. The performance of the internal cross-validation could serve as a significant score for the cluster. At this point we have considered accuracy as the performance meas-urement.

Furthermore, Support Vector Machines (SVMs) is employed on each sub_dataseti for scoring the cluster in terms of its ability to classify the two-classes. To compute the score, an internal cross-validation is applied by splitting each sub_dataset into training and testing while this procedure is repeated f times (f is set as 5). The mean of the accuracy is computed and it is assigned as a score to the related cluster.

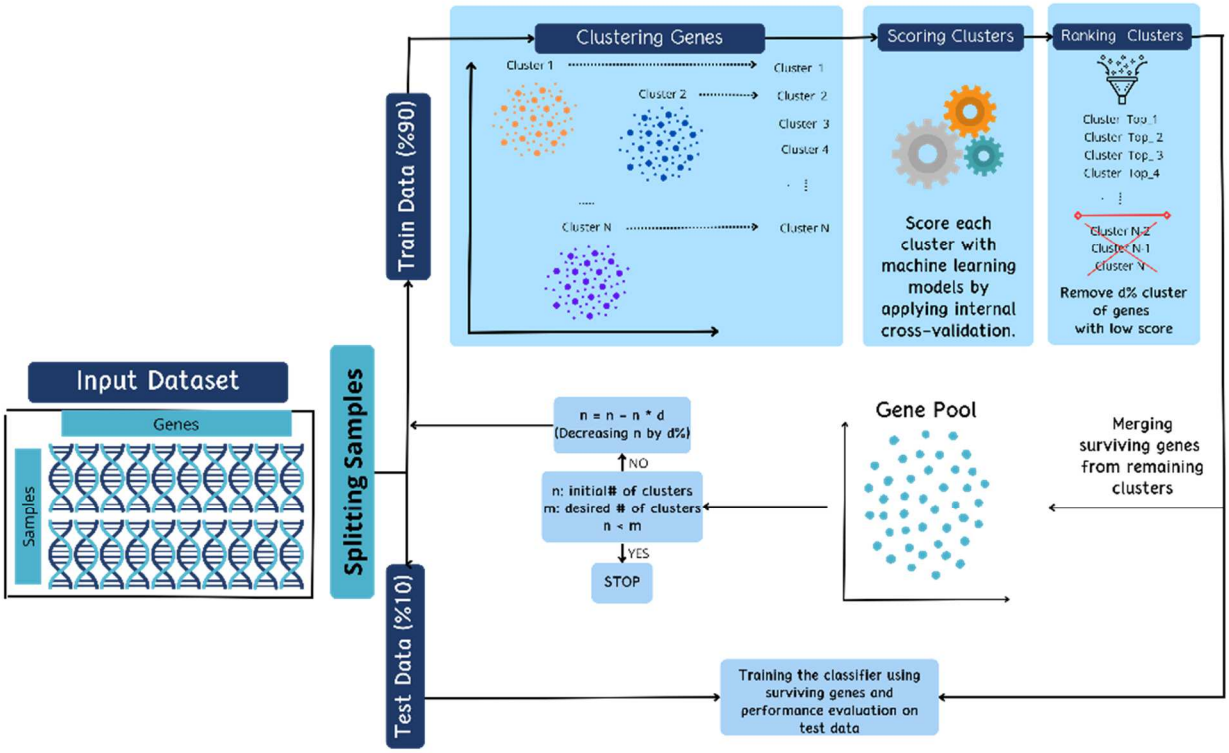

Fig. 1. The flowchart of the SVM-RCE algorithm.

Next step is the elimination stage, where the least significant clusters are removed while retaining the surviving clusters to the next stage of SVM-RCE. The user specifies the amount of cluster to be removed in each stage, let's say d%. In the elimination stage, the genes that are member of the least scored clusters are removed from D_train and D_test to create D_train* and D_test*, respectively (removing the columns representing the genes of the selected clusters). To estimate the performance of SVM-RCE, the SVM is trained on the D_train* and tested on D_test*. In the RCE step, k is reduced and then the whole procedure is repeated until k reaches 1 cluster or other predefined value set by the user.

All these steps mentioned above were carried out on the KNIME platform. Knime is a free and open-source tool that has a user-friendly interface that facilitates the application, helps you create, execute and automate data streams.

In this study, we have analyzed the effect of using other machine learning algorithms than SVM on the classification performance of SVM-RCE, and on the running time. To this end, the machine learning algorithms that have been comparatively evaluated are Random Forest, XGBoost, Naive Bayes, and Logistic Regression.

1) Machine Learning Algorithms.

The machine learning algorithm known as Support Vector Machines (SVMs) was developed by Vapnik [11]. This algorithm has exhibited a significant edge over alternative algorithms since it is applied to a range of classification issues. The employment of this technology has garnered substantial recognition in the realm of bioinformatics, as presented in recent surveys [12]. Linear support vector machines (SVMs) are frequently referred to as SVMs that employ a linear kernel.

The Random Forests (RF) technique, which was introduced by Breiman during the early 2000s [12, 13], has been recognized as a remarkably efficient strategy. This supervised learning methodology was also proposed by Amit and Geman [12], Ho [13], and Dietterich [15]. The methodology employed in this study utilizes the "divide and conquer" approach, whereby the dataset is partitioned into smaller subsets, and a randomized tree predictor is constructed for each of these subsets. The aforementioned predictors are subsequently combined to produce a comprehensive pre-diction model.

XGBoost is the name of a software package that utilizes extreme gradient boosting techniques. The efficacy and extensibility of the gradient boosting framework, as originally formulated by Friedman [16], are noteworthy. The package mentioned above includes a capable linear model solver and a tree learning algorithm. The system possesses the capacity to

incorporate diverse objective functions, such as regression, classification, and ranking, among others. The package was designed with a focus on extensibility, affording users the ability to easily define their own objectives and achieve greater flexibility.

The Naïve Bayes (NB) [17] algorithm is widely recognized as a prominent data mining technique utilized for classification purposes. The statement posits the likelihood of a novel instance being assigned to a particular category, predicated on the supposition that all characteristics are mutually exclusive, given the category. The rationale behind this proposition is driven by the necessity to compute the probabilities of multiple variables based on the available training data. Empirically, a majority of attribute value combinations are either absent from the training dataset or occur in inadequate frequencies. Hence, the precise evaluation of every pertinent multivariate probability cannot be deemed dependable. The conditional independence assumption of Naïve Bayes enables it to overcome this predicament. Despite the stringent assumption of independence, the naïve Bayes classifier exhibits remarkable proficiency in numerous practical scenarios. Logistic regression (LR) is a statistical model for binary and binomial outcomes that expresses the log odds of a certain outcome as a function of explanatory variables [18].

III. Results and Discussion

The present study utilized five gene expression datasets (GDS2547, GDS3268, GDS3646, GDS3875, and GDS5037) to compare five different approaches. The performance was obtained through 10 iterations using Monte Carlo Cross Validation (MCCV). During each iteration, 90% of the available data was allocated for the purpose of training, while the remaining 10% was reserved for testing. Subsequently, the average of all performance metrics over different iterations are computed. The initial SVM-RCE algorithm is comparatively evaluated with alternative RCE algorithms, namely Random Forest, XgBoost, Naive Bayes, and Logistic Regression.

The evaluation criteria employed in our study consists of the execution time, area under the curve, and the count of the surviving genes in the stage of two-clusters and 1 final cluster. Each metric is assessed based on the mean value of the cross validation iterations and averaged over five datasets. The execution time refers to the mean duration per iteration, denoting the quantity of time (measured in hours) necessary for a single cycle. The algorithm's successive rounds involve the removal of clusters with the lowest scores, with each iteration resulting in the elimination of 10% of such clusters. As a result, the number of genes that persist after removal decreases. A reduction in the quantity of genes leads to a corresponding reduction in the number of clusters. (90, 80, 70, 60, 50, 40, 30, 20, 10, 5, 2, 1). The scenario illustrated in Figure 4 portrays a circumstance in which the quantity of extant clusters is two.

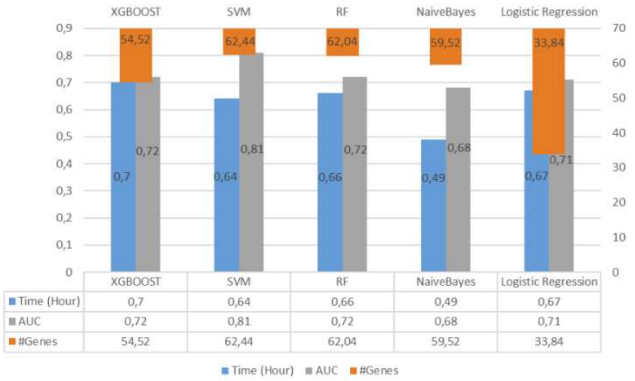

Fig. 2. Evaluation of different classifiers in RCE algorithm in terms of their execution time, area under the curve (AUC) values and number of surviving genes for 2 clusters.

SVM produces higher AUC values in the final two clusters, which are about 9% higher than the compared classifiers. Despite the improved AUC values, the number of surviving genes is slightly higher than others. In addition, all techniques except Naive Bayes classifier take about the same amount of time per repetition.

For the top 2 clusters, the number of surviving genes is minimal in the LR approach, which makes LR a better alternative in terms of defining biomarkers, However, as shown in Figure 2, the AUC value of LR is 10% lower than the SVM classifier. In general, LR technique is very comparable to other classifiers with the exception of Naive Bayes method in terms of the running time; and with the exception of SVM in terms of the AUC value.

For the top 2 clusters, the most effective computational method is the Naive Bayes. This approach beats the competitors by 17% in terms of speed. Except for the LR approach, this method, in exchange for quickness, generates comparable AUC values and selects comparable numbers of genes.

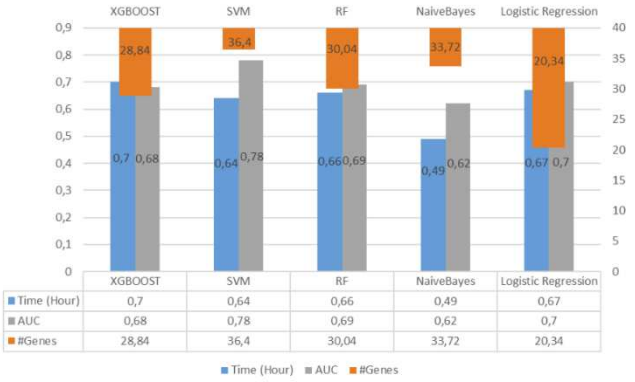

Fig. 3. Evaluation of different classifiers in RCE algorithm in terms of their execution time, area under the curve (AUC) values and number of surviving genes for only 1 clusters.

The Support Vector Machine (SVM) algorithm yields significantly higher AUC values for the top cluster, surpassing alternative techniques by approximately 10%. However, the AUC obtained using the top cluster (0,78) is slightly lower than the AUC of the cumulative top 2 clusters (0,81). Despite the improved AUC values, the quantity of genes that persist is higher in comparison to alternative techniques. In contrast with the cumulative two clusters, SVM effectively removes around 50% of the remaining genes within the ultimate

cluster. However, the LR strategy has consistently demonstrated the lowest number of surviving genes. Furthermore, the duration of each iteration is similar to that of alternative methods, except for the two-cluster technique employed by Naive Bayes.

In the final cluster, the minimal number of surviving genes is determined by LR. Once again, in the top cluster, the AUC of LR (0,70) is slightly lower than that of the SVM (0,78). The application of the LR strategy resulted in the removal of 33% of the remaining genes, while the AUC values remained relatively unchanged. This finding is in contrast with the results of SVM when the top 2 cumulative results are compared with the top cluster results. Apart from the Naive Bayes methodology, the LR technique exhibits a similar level of performance with other methods, albeit with varying execution times.

The management of features in complex gene expression datasets requires the implementation of unconventional computational strategies. The approach employed by our methodology effectively addresses the issue of inherent duplication or correlation among traits. Furthermore, it emphasizes the significance of carefully selecting appropriate grouping measures while exploring the feature space of interest to researchers. The results of our study showed that clusters exhibit elevated values for AUC metrics using the SVM classifier. In future studies, our focus will be on achieving optimal ranking. It is possible to evaluate whether minimizing false positives or false negatives would prove more efficacious in addressing the matter at hand.

IV. Conclusion

In the original version of SVM-RCE the SVM was used for scoring the clusters. In this study we have compared the effect of other machine learning in scoring the clusters. Also we keep track of the number of genes in each cluster in each level of reduction of the RCE procedure. Additionally, we measure the execution time.

The presented results suggest that the employment of the SVM classifier in SVM-RCE procedure yields superior AUC in comparison with alternative methods. Based on the analysis of the number of genes that have been preserved, it is apparent that the utilization of logistic regression leads to more advantageous results. Naive Bayes exhibits superior speed based on its execution time.

In order to ensure the stability of the results, our forthcoming research endeavors will involve the utilization of additional datasets and an increase in the number of iterations. We will further enhance our analyses by exploring various configurations.

References

[1] I. Jain, V. K. Jain, and R. Jain, "Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification," Appl. Soft Comput., vol. 62, pp. 203–215, Jan. 2018, doi: 10.1016/j.asoc.2017.09.038.

[2] R. Aziz, C. K. Verma, and N. Srivastava, "A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data," Genomics Data, vol. 8, pp. 4–15, Jun. 2016, doi: 10.1016/j.gdata.2016.02.012.

[3] M. Dashtban, M. Balafar, and P. Suravajhala, "Gene selection for tumor classification using a novel bio-inspired multi-objective approach," Genomics, vol. 110, no. 1, pp. 10–17, Jan. 2018, doi: 10.1016/j.ygeno.2017.07.010.

[4] A. Sharma and R. Rani, "C-HMOSHSSA: Gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods," Comput. Methods Programs Biomed., vol. 178, pp. 219–235, Sep. 2019, doi: 10.1016/j.cmpb.2019.06.029.

[5] A. K. Shukla, P. Singh, and M. Vardhan, "A hybrid gene selection method for microarray recognition," Biocybern. Biomed. Eng., vol. 38, no. 4, pp. 975–991, 2018, doi: 10.1016/j.bbe.2018.08.004.

[6] M. Yousef, S. Jung, L. C. Showe, and M. K. Showe, "Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data," BMC Bioinformatics, vol. 8, no. 1, p. 144, Dec. 2007, doi: 10.1186/1471-2105-8-144.

[7] M. Yousef, M. Ketany, L. Manevitz, L. C. Showe, and M. K. Showe, "Classification and biomarker identification using gene network modules and support vector machines," BMC Bioinformatics, vol. 10, no. 1, p. 337, Dec. 2009, doi: 10.1186/1471-2105-10-337.

[8] M. Yousef, B. Bakir-Gungor, A. Jabeer, G. Goy, R. Qureshi, and L. C. Showe, "Recursive Cluster Elimination based Rank Function (SVM-RCE-R) implemented in KNIME," F1000Research, vol. 9, p. 1255, Jan. 2021, doi: 10.12688/f1000research.26880.2.

[9] T. Barrett et al., "NCBI GEO: archive for functional genomics data sets—update," Nucleic Acids Res., vol. 41, no. D1, pp. D991–D995, Nov. 2012, doi: 10.1093/nar/gks1193.

[10] M. Yousef, A. Jabeer, and B. Bakir-Gungor, "SVM-RCE-R-OPT: Optimization of Scoring Function for SVM-RCE-R," in Database and Expert Systems Applications - DEXA 2021 Workshops, G. Kotsis, A. M. Tjoa, I. Khalil, B. Moser, A. Mashkoor, J. Sametinger, A. Fensel, J. Martinez-Gil, L. Fischer, G. Czech, F. Sobieczky, and S. Khan, Eds., in Communications in Computer and Information Science, vol. 1479. Cham: Springer International Publishing, 2021, pp. 215–224. doi: 10.1007/978-3-030-87101-7_21.

[11] V. N. Vapnik, The Nature of Statistical Learning Theory. New York, NY: Springer New York, 2000. doi: 10.1007/978-1-4757-3264-1.

[12] I. Donaldson et al., "[No title found]," BMC Bioinformatics, vol. 4, no. 1, p. 11, 2003, doi: 10.1186/1471-2105-4-11.

[13] Y. Amit and D. Geman, "Shape Quantization and Recognition with Randomized Trees," Neural Comput., vol. 9, no. 7, pp. 1545–1588, Oct. 1997, doi: 10.1162/neco.1997.9.7.1545.

[14] Tin Kam Ho, "The random subspace method for constructing decision forests," IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 8, pp. 832–844, Aug. 1998, doi: 10.1109/34.709601.

[15] T. G. Dietterich, "Ensemble Methods in Machine Learning," in Multiple Classifier Systems, in Lecture Notes in Computer Science, vol. 1857. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15. doi: 10.1007/3-540-45014-9_1.

[16] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors)," Ann. Stat., vol. 28, no. 2, Apr. 2000, doi: 10.1214/aos/1016218223.

[17] I. Wickramasinghe and H. Kalutarage, "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation," Soft Comput., vol. 25, no. 3, pp. 2277–2293, Feb. 2021, doi: 10.1007/s00500-020-05297-6.

[18] S. Sperandei, "Understanding logistic regression analysis," Biochem. Medica, pp. 12–18, 2014, doi: 10.11613/BM.2014.003.