Düzce University Journal of Science & Technology

Research Article

A Novel Ensemble Feature Selection Technique for Cancer Classification Using Logarithmic Rank Aggregation Method

Hüseyin Güney [a *], Hüseyin Öztoprak [b,c]

[a] *Department of Computer Engineering, Bahçeşehir Cyprus University, Nicosia, Northern Cyprus, Turkey.*
[b] *Department of Electrical and Electronics Engineering, Cyprus International University, Nicosia, Northern Cyprus, Turkey.*
[c] *Signal and Technology Research Lab, Cyprus International University, Nicosia, Northern Cyprus, Turkey.*
[*] *Corresponding Author, Email: huseyin.guney@baucyprus.edu.tr*
DOI: 10.29130/dubited.1225446

ABSTRACT

Recent studies have shown that ensemble feature selection (EFS) has achieved outstanding performance in microarray data classification. However, some issues remain partially resolved, such as suboptimal aggregation methods and non-optimised underlying FS techniques. This study proposed the logarithmic rank aggregate (LRA) method to improve feature aggregation in EFS. Additionally, a hybrid aggregation framework was presented to improve the performance of the proposed method by combining it with several methods. Furthermore, the proposed method was applied to the feature rank lists obtained from the optimised FS technique to investigate the impact of FS technique optimisation. The experimental setup was performed on five binary microarray datasets. The experimental results showed that LRA provides a comparable classification performance to mean rank aggregation (MRA) and outperforms MRA in terms of gene selection stability. In addition, hybrid techniques provided the same or better classification accuracy as MRA and significantly improved stability. Moreover, some proposed configurations had better accuracy, sensitivity, and specificity performance than MRA. Furthermore, the optimised LRA drastically improved the FS stability compared to the unoptimised LRA and MRA. Finally, When the results were compared with other studies, it was shown that optimised LRA provided a remarkable stability performance, which can help domain experts diagnose cancer diseases with a relatively smaller subset of genes.

Keywords: *Microarray Data, Ensemble Learning, Aggregation Methods, Support Vector Machine Recursive Feature Elimination (SVM-RFE)*

Kanser Sınıflandırma için Logaritmik Sıra Birleştirme Yöntemini Kullanan Yeni Topluluk Öznitelik Seçim Tekniği

ÖZET

Son araştırmalar, topluluk öznitelik seçiminin (TÖS) mikrodizi veri sınıflandırmasında olağanüstü bir başarı elde ettiğini göstermiştir. Bununla birlikte, yetersiz birleştirme yöntemleri ve optimize edilmemiş ÖS teknikleri gibi konuların kısmen çözülmüş olarak kaldığı görülmektedir. Bu çalışma, TÖS yöntemlerinde özellik birleştirmeyi geliştirmek için logaritmik sıralama birleştirme (LRA) yöntemini önerdi. Ek olarak, önerilen yöntemin performansını geliştirmek için birkaç yöntemle birlikte kullanan

Received: 28/12/2044, Revised: 06/07/2023, Accepted: 06/08/2023

hibrit yöntemler sunulmuştur. Ayrıca, öznitelik seçiminin optimizasyonunun etkisini ölçmek için optimize edilmiş öznitelik seçim tekniğinden elde edilen öznitelik sıralamalarına da önerilen yöntem uygulanmıştır. Hazırlanan deney, beş ikili mikrodizi veri seti üzerinde gerçekleştirilmiş olup, deney sonuçları, LRA'nın ortalama sıra birleştirme yöntemine (MRA) kıyaslanabilir bir sınıflandırma performansı sağladığını ve gen seçim istikrarı açısından MRA'dan daha iyi performans elde ettiğini göstermiştir. Ek olarak, hibrit teknikler, MRA ile aynı veya daha iyi sınıflandırma doğruluğu sağladı ve gen seçim istikrarını önemli ölçüde artırdı. Ayrıca önerilen bazı konfigürasyonlar, MRA'dan daha iyi doğruluk, hassasiyet ve özgüllük performansına ulaştı. Ayrıca, optimize edilmiş LRA, optimize edilmemiş LRA ve MRA'ya kıyasla gen seçim istikrarını önemli ölçüde iyileştirmiştir. Son olarak, sonuçlar diğer çalışmalarla karşılaştırıldığında, optimize edilmiş LRA'nın dikkate değer bir gen se.im istikrarı sağladığı görülmüştür ve bu çalışmanın bu alanda çalışan uzmanların kanser teşhisinde nispeten daha küçük bir gen kümesi kullanarak daha isabetli teşhis koymalarına yardımcı olabileceği vurgulanmıştır.

Anahtar Kelimeler: *Mikrodizi Veri Kümesi, Topluluk Öğrenme, Birleştirme Yöntemleri, Destek Vektör Makineleri, Tekrarlayan Öznitelik Seçimi (SVM-RFE)*

I. INTRODUCTION

With new technologies, many tools, such as microarray technology and medical imaging, are available for obtaining biological data. Since DNA microarray gene expressions are representations of human genes, they are promising for detecting genetic modifications [1]. The analysis of microarray data is a critical task. Therefore, DNA microarrays have been extensively studied to develop diagnosis systems using computational intelligence and power. However, developing an intelligent system for microarray-based cancer diagnosis is challenging due to its characteristics, such as high dimensionality and low sample size.

Microarray-based cancer diagnosis task is a classification problem in the machine learning (ML) domain; therefore, a diagnosis system can be developed using supervised ML algorithms to select the most relevant genes and classify patient data as healthy or cancer with the selected genes. In addition, multivariate feature selection (FS) techniques can help with accurate gene selection since correlation among genes is needed to be considered during the gene selection process.

Feature selection techniques are broadly classified into filter, embedded, and wrapper techniques [2]. Furthermore, ensemble feature selection (EFS) has been widely used for microarray-based cancer diagnosis due to its significant performance for selection stability [3-6]. EFS has partially resolved the instability problem of conventional FS techniques by using resampling techniques to create diversity among the resampled datasets. However, while EFS improves the accuracy and stability of the underlying model, it has a more complex structure than conventional approaches in terms of computational and time complexity. Furthermore, the complexity of the EFS increases in parallel to its ensemble size.

There are two fundamental performance criteria for FS techniques in the context of biomarker gene selection using microarray datasets. Like all other classification problems, one is accuracy to measure how accurately the classification is performed. The accuracy can be further detailed by sensitivity, specificity, and precision measures, all exploited in this study. The second is the stability of the selected features while input data is slightly changed [4]. The stability of gene selection is important mainly for two reasons. First, if the selected features vary greatly, there will be a lack of clarity in designing the final classifier with the minimum number of features to test new samples. Second, domain experts will have difficulty obtaining more information on the nature of the disease due to uncertainty in the relevant genes. Therefore, an accurate and robust model can help domain experts make more confident decisions about cancer diagnosis with less effort, time, and resources.

When considering all FS techniques, embedded FS techniques seem to be the most appropriate for a study that uses EFS. Therefore, this study proposes novel EFS methods to overcome partially resolved issues for this problem domain, such as limited research on aggregation methods and the lack of optimisation of the classifier hyperparameters for classifier-dependent feature selection. In this paper, there are three main contributions. The first contribution is the Logarithmic Rank Aggregation (LRA), which replaces the commonly employed Mean Rank Aggregation (MRA). Second, the First Iteration Framework exploits the initial SVM-RFE ranks and suggests Hybrid Methods. The third and last contribution of this paper is the optimisation of the cost parameter of the SVM in the ranking algorithm of SVM-RFE that is used in EFS.
The remainder of this paper is organised as follows. Section 2 examines related works in the current literature. Section 3 introduces SVM, SVM-RFE, and EFS to provide background information for the proposed methods. Section 4 explains logarithmic, initial SVM-RFE output-based methods, and cost parameter optimisation of SVM-RFE. The results of the experimental procedure are presented in Section 5. The discussion and comparison with conventional methods are mentioned in Section 6. Finally, the conclusion of this study is presented in Section 7.

II. RELATED WORK

This section examines the related works published in the current literature to investigate the studies related to microarray data. Since the proposed method is a machine learning algorithm for microarray-based cancer diagnosis, this section mentions EFS-based ML methods proposed in the literature. The summary of the related works listed in this section are given in Table 1 with their strengths and limitations.

In the article [7], Barbara Pes has conducted extensive experiments to investigate the impact of EFS on the accuracy and robustness of machine learning algorithms for wide datasets. Several FS techniques were implemented in the study, and ensemble methods of these techniques have been evaluated along with the single methods. In [8], feature selectors' reliability assessment was used for developing an EFS method. This study used a classification algorithm to evaluate the selected features by the ensemble method. In [9], filter FS techniques and classifiers-based EFS was proposed for microarray data. In the first step, five filter FS methods were applied for selecting features, and classifiers were used to evaluate the selected features. In the last step, an ensemble of the classifier was developed using simple voting to combine the predictions of each classifier to deliver the final prediction. Ten microarray datasets were used for model evaluation. In the study [10], the authors proposed ensemble SVM-RFE (ESVM-RFE) for classifying microarray data using bootstrap aggregation. It was stated that ensemble learning improved classification performance. The study [11] proposed the multi-filter enhanced genetic ensemble (MF-GE) system as a hybrid gene selection method. It was stated that the proposed method improved classification performance while reducing the selected feature set size. In the article [12], the authors used homogeneous and heterogeneous approaches for developing EFS methods. Therefore, they proposed EFS methods that used one or many FS techniques. For the evaluation of the proposed methods, an SVM classifier was deployed. It was concluded that both approaches achieved comparable performance with other ensemble learning-based FS methods and outperformed non-ensemble FS methods. The article [13] proposed a two-stage algorithm that combined FS and prediction by extending a type of hetero-associative neural network. In the first stage of the design, associative memory was generated, and gene selection was done in the next step. The conclusion was drawn that the proposed method performed similarly to others with reduced computational complexity.

An ensemble method for gene selection from breast cancer microarray data was proposed in [14]. The proposed method was an ensemble filter selection method based on Entropy and SNR evaluation function (EnSNR). The entropy method was used for measuring the uncertainty of a random experiment outcome, and feature discriminative power was measured by SNR. It was shown that EnSNR obtained high accuracy with fewer genes when it was compared to other methods. In [15], a hybrid of filter and

wrapper FS techniques was proposed as an EFS method. In the first step, various FS techniques were used to obtain feature rank links, including Relief, Minimum Redundancy Maximum Relevance (MRMR), and Feature Correlation (FC). In the next step, feature aggregation was done using Fuzzy Gaussian membership. In the final step, the wrapper FS technique, Improved Binary Particle Swarm Optimisation (IBPSO), was applied to select features. After feature selection, an SVM classifier with a non-linear radial basis function was employed for feature subset evaluation. It was mentioned that the proposed method achieved better performance than other FS methods. The research [16] conducted an empirical study to compare the proposed method with the literature. In the study, to develop an EFS method Fast Correlation FS technique was employed. Finally, two stability metrics were used to measure the model's robustness.

Momenzadeh et al. proposed a novel feature selection method using a hidden Markov model (HMM) approach that integrates several FS techniques [17]. The proposed method incorporates five feature selection ranking methods: Bhattacharyya distance, entropy, receiver operating characteristic curve, t-test, and Wilcoxon. Diffuse large B-cell lymphoma, leukaemia and prostate cancer datasets were used for model evaluation, and it was stated that the HMM-based method outperformed Markov chain rank aggregation-based method and single FS techniques. The article [18] proposed the novel FS method, IG-MBKH, which is a combination of Information Gain (IG) and improved binary krill herd (MBKH) algorithms, where IG was used for gathering feature rankings, and MBKH was used for finding the most relevant subset of features. The study showed that IG-MBKH performs better than BKH, MBKH, and several other algorithms regarding the accuracy and the number of selected features.
In [19], a hybrid gene selection technique was developed using filter FS technique robust Minimum Redundancy Maximum Relevancy (rMRMR) and wrapper FS technique Modified Gray Wolf Optimizer (MGWO). In the setup, first, rMRMR was employed to select the most relevant genes, and the Modified Gray Wolf Optimizer (MGWO) algorithm was applied for further gene elimination to finding to smaller gene sets. While MGWO was applied, the TRIZ-inventive solution inspired a new approach to increase the population's diversity to enhance the gene selection process. Nine benchmarking datasets and SVM classifier were used to evaluate the developed method. The results obtained showed the effectiveness of the proposed method. The article [20] presents adaptive hypergraph embedded dictionary learning (AHEDL) model for microarray data classification. In the study, a dictionary was used to learn from the feature space, and then it was reconstructed. After that, l2, l-norm regularisation was used for gene selection. A hypergraph was created and injected to the model to capture the geometrical structure of the data. On the other hand, to solve the optimisation problem, an iterative updating algorithm was designed. The authors state that their method outperformed state-of-the-art methods.

Table 1. *Related Works with Their Strengths and Limitations*

	Method	Strength(s)	Limitation(s)
1	Feature selectors' reliability assessment-based EFS [8]	- Homogeneous and Heterogeneous Filter-based EFS method - Reliability assessment-based aggregation technique	- High computational complexity as an EFS method - Computationally complex classifier-based aggregation method - 10-Fold CV used for model validation
2	Filter-based EFS [9]	- Ensemble of Filters and Classifiers - Cost-effective FS techniques and classifiers	- Gene selection using filter methods - Simple voting aggregation - 10-Fold CV used for model validation
3	ESVM-RFE [10]	- Embedded FS-based Feature Selection - Enhanced version of SVM-RFE	- No optimisation of the FS technique - Weight Sum aggregation
4	MF-GE [11]	- Hybrid Ensemble Method (Filter + Wrapper) - Small gene subset selection	- Computationally Complex due to wrapper and hybrid - Mean and Majority voting aggregation
5	Hybrid EFS Methods [12]	- Hybrid Ensemble Method using Homogeneous and Heterogeneous approaches - Wide range of ensemble aggregation methods	- Computationally Complex due to hybrid approach and embedded FS techniques - No optimisation of Embedded FS techniques - 10-Fold CV used for model validation
6	HAM [13]	- Using an associative memory-based approach and Neural Networks for gene selection	- 5-Fold CV used for model validation - Large set of selected genes
7	EnSNR [14]	- Cost-effective FS Technique - Optimisation is used for gene selection	- Computationally complexity due to wrapper FS - 10-Fold CV used for model validation

8	Hybrid FS (Filter+Wrapper) [15]	- Hybrid Ensemble Method (Filter + Wrapper) - Improved Binary Particle Swarm Optimisation is used for gene selection - Fuzzy Gaussian rank aggregation	- High computational complexity due to wrapper FS technique and SVM-RFE - 10-Fold CV used for model validation
9	Fast Correlation-based EFS [16]	- Cost-effective FS Technique - Several aggregation methods for gene selection	- Classifier-independent FS Techniques - Simple aggregation method - 10-Fold CV used for model validation
10	HMM-based EFS [17]	- Markov chain rank aggregation - Cost-effective FS Techniques	- Classifier-independent FS Techniques - Complex aggregation method
11	IG-MBKH [18]	- Cost-effective FS Technique - Optimisation is used for gene selection	- Computationally complexity technique - 10-Fold CV used for model validation
12	Hybrid FS (rMRMR + MGWO) [19]	- Hybrid FS Method - Optimisation is used for gene selection	- Computationally complexity due to wrapper FS and SVM - 10-Fold CV used for model validation

In the literature, many studies proposed ensemble feature selection techniques for improving microarray data classification performance and gene selection stability. For this reason, this study focuses on the aggregation function of EFS since a more precise feature aggregation method can lead to a better ranking of the features in the ensemble feature list. For this purpose, a logarithmic scaling-based aggregation method was proposed to achieve a more accurate positioning of features in the ensemble feature list. Additionally, some hybrid frameworks were proposed to improve the performance of aggregation method using strengths of two different approaches. On the other hand, the impact of the optimisation algorithm was measured to investigate its effect on the performance of the proposed method. The following section explains background information, and Section 4 provides detailed information about the proposed methods.

III. BACKGROUND

In this section, background information is provided to facilitate understanding of the proposed methods, as the underlying classifier and FS techniques of this work, SVM, SVM-RFE, and RCV-based ensemble feature selection, are explained, respectively.

A. SUPPORT VECTOR MACHINE

Due to its outstanding performance as a supervised machine learning algorithm, SVM has been frequently used for classification. Hyperplane construction, margin maximisation, and kernel functions are the powerful characteristics of SVM. Briefly, SVM constructs a hyperspace using features and then finds and draws the hyperplane that separates the instances of classes using margin maximisation, finding maximum distances from each class for the hyperplane. [21,22]. An essential characteristic of SVM is its kernel functions, which allow access to higher space dimensions without explicitly defining the mapping function. The kernel functions of SVM can be divided into two groups, linear and non-linear, where any can be used for classification tasks.

Linear SVM can be optimised in two extreme configurations, hard-margin and soft-margin linear SVM. Hard-margin linear SVM linearly separates both classes, and no instance of any class is left in the region of the opposite class, if possible. On the other hand, in the soft-margin Linear SVM, some instances of any class can be left on the opposite side of the margin. This neglects some outlier instances to find the best hyperplane for class separation, which would help to increase classification accuracy. As a result, it depends on the cost parameter to identify a linear SVM as hard, soft, or in-between [21]. Increasing the cost parameter value narrows the hyperplane and brings the classifier closer to the hard margin and vice versa.

B. SUPPORT VECTOR MACHINE – RECURSIVE FEATURE ELIMINATION

Support Vector Machine Recursive Feature Elimination (SVM-RFE) [23] uses the backward elimination approach, which iteratively eliminates features to find the most relevant (informative) ones. It starts with

the entire feature set and drops the worst feature at each run until all features are ranked, and the surviving feature set becomes empty. SVM-RFE uses the SVM training process to obtain feature weights for feature ranking. After the weight vector is generated, the features are ordered according to their weights, where the feature with the highest weight is considered the best [5]. In the study [6], the elimination process was modified to speed up the FS process, and the elimination percentage parameter, E, and the StopValue parameters were introduced [6]. The elimination percentage parameter, E, eliminates the E percentage of features at each run, and this process continues until StopValue is reached. Once the StopValue is reached, the features are removed one by one, not as a group of features [6].

SVM-RFE [23] is a multivariate embedded FS technique that keeps the original value of the features. That is, it does not transform the features into a new space. It is also faster than wrapper FS techniques [2]. Therefore, as an embedded FS technique, SVM-RFE is suitable for gene selection using microarray datasets. Furthermore, it was observed that, regarding classification performance, SVM-RFE is competitive for the biomarker identification of microarray datasets but suffers from instability of gene selection [5,6].

C. ENSEMBLE FEATURE SELECTION

Ensemble feature selection (EFS) aims to improve the feature selection process by integrating ensemble learning into conventional FS techniques. EFS creates diversity in the feature selection process using data variation, FS technique variation, or both. As a result, EFS leads to better classification performance and stability of feature selection [4,6].

The EFS comprise three main stages. The first stage is to generate several resampled datasets from the original dataset using resampling techniques such as cross-validation or bootstrapping. The second stage includes feature ranking using a minimum one feature selection technique to generate all ranked feature lists. The final stage involves using an aggregation function to aggregate all generated lists into the final list (ensemble) [7]. In the literature, bootstrapping and cross-validation variations were used for the EFS data variation step, e.g., bootstrap-based EFS [3] and repeated cross-validation-based EFS [24].
The mean aggregation method is the most widely used method, which uses the arithmetic mean function to obtain the average of feature ranks in all ranked feature lists [24]. Mean aggregation is simple and can be implemented quickly. However, since it is prone to be affected by outliers, it may not be robust enough for high-dimensional datasets. The lowest aggregation method finds the lowest rank across all lists and assigns it to the ensemble list. In contrast, the highest aggregation method selects the highest rank to construct the ensemble feature list. The lowest and highest aggregation methods also have the same drawback as the mean aggregation, which is prone to be affected by outlier ranks. Several feature selection techniques for microarray datasets based on optimisation techniques have recently been proposed.

IV. PROPOSED METHOD

This section explains the proposed logarithmic rank aggregation method (LRA). Besides, SVM-RFE's first iteration and cost parameter optimisation frameworks are described. A case study is also provided to show the significance of LRA for ensemble feature selection.

A. LOGARITHMIC RANK AGGREGATIONS

The mean aggregate method has a significant weakness, particularly for small datasets (low sample size datasets). It is prone to be affected by outlier feature ranks. In other words, a poorly positioned outlier in a single-ranked feature list can eliminate an elsewhere successful feature from the final list. Stability selection-based methods limit the impact of a single rank on the final list since they assign one and zero to the presence of a feature. However, this approach introduces another problem; a slight change to the

feature subset threshold can change the score of a feature, and thus, this method is also prone to be affected by outliers or high variation in feature ranks.

To overcome the mentioned drawbacks of mean aggregation and stability selection-based methods, Logarithmic Rank Aggregation (LRA) is proposed, which first calculates the log values of feature ranks and then employs mean aggregation to construct the ensemble list. The formulation of LRA is illustrated in Equation 1. The proposed aggregation method is illustrated in Figure 1. We performed a case study that illustrated how different aggregation methods are affected by the outliers and how aggregation methods affect the final positions of the features in an ensemble feature list.

This example was based on the aggregation process within the internal RCV-EFS, i.e., RCV-10-2 twenty folds, training process. The Leukemia dataset was selected for this case study because it was the largest dataset used. The rankings of six features, which held the first position on at least one of the ranked feature lists, are presented in Table 2. The best and worst positions obtained for the selected features across all folds are shown in the table. At the bottom of the table, the final ranks obtained by the MRA and LRA aggregation methods are presented.

It was seen that the selection of the aggregation method dramatically affects the position of the features, which is caused by a few outlier positions. For example, feature 5765, which is out of the top 1% with MRA, is in the top 0.5% with LRA. Similarly, feature 6405, out of the top 2% with MRA, is in the top 0.75% with logarithmic aggregation. Feature Nos. 5765 and 6405, which vary greatly in the ranked feature lists, were selected, and the effects of aggregation methods on each fold of the external RCV were also investigated by recording the ensemble list feature ranks. Figure 2 represents the arithmetic mean and standard deviation values of the obtained feature ranks for both. Feature 5765 and Feature 6405 achieved a lower arithmetic mean value when LRA was used instead of MRA, where the standard deviation was also lower. Smaller arithmetic means indicate that these features are less affected by outlier positions in some of the folds, and smaller standard deviations imply that the feature sets of LRA might be more stable than those selected by MRA.

Formula:

$$f(x_1, x_2, \dots, x_N) = \left(\frac{(\sum_{i=1}^{S} \ln (Fr_i^1)}{S}\right), \dots, \left(\frac{(\sum_{i=1}^{S} \ln (Fr_i^N)}{S}\right)$$

(1)

N = feature size,

S = Ensemble size,

Fr_i = feature k's rank in i^{th} Feature Rank List

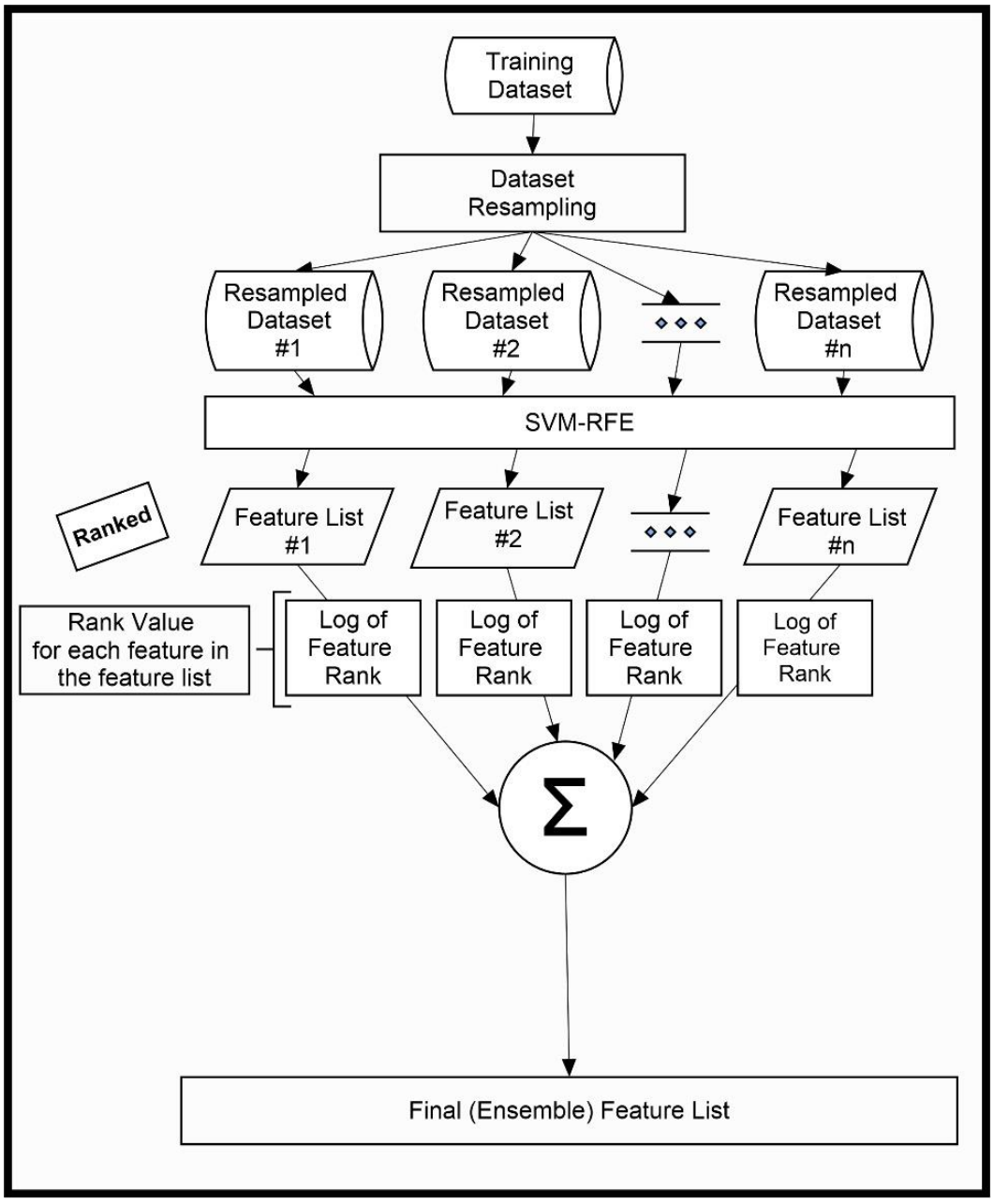

Figure 1. *Log-Based Aggregation Method*

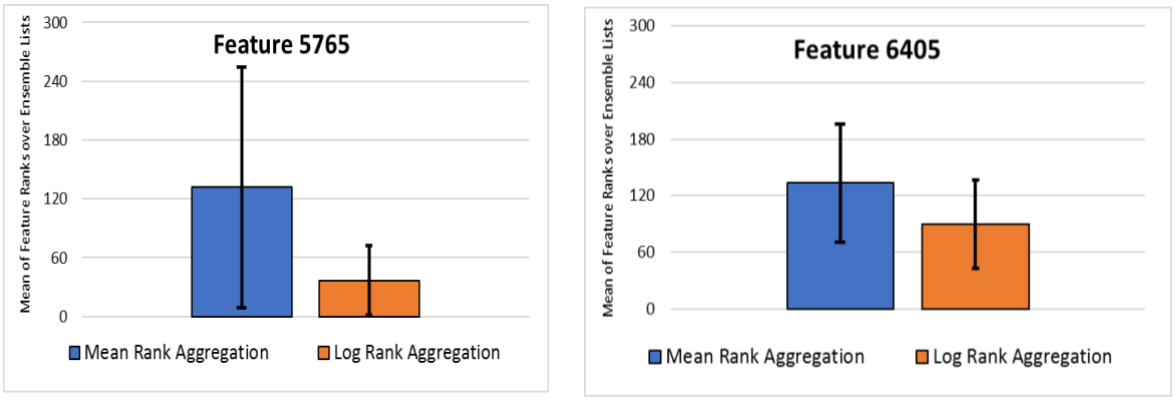

Figure 2. *Calculation of mean and standard deviation values of ranked ensemble lists for two features to measure the effect of aggregation methods on the ensemble lists of EFS.*

One possible shortcoming of the proposed LRA method is its sensitivity to changes in the ranking when the feature performs best in the fold. This can be especially problematic when the signature size is in the medium range since a slight change in the feature performing very well in a single fold should not be decisive. Consequently, a hybrid aggregation method has been developed between the mean aggregation and logarithmic methods. Specifically, a parameter named h is injected into the formulation of LRA to calculate the final rank of the feature. This parameter enables the aggregation function to behave as 'semi-logarithmic'. The formulation of the hybrid model is illustrated in
Equation 2.

Table 2. *Feature rankings obtained from the Leukemia dataset*

Internal Fold #	Feature **1779**	Feature **1882**	Feature **2128**	Feature **1699**	Feature **5765**	Feature **6405**
1.	6	36	2	37	85	17
2.	9	**1**	8	**396**	64	395
3.	9	**1**	8	258	42	1628
4.	9	**1**	32	56	65	4
5.	2	14	**258**	23	24	17
6.	**30**	32	4	97	169	**3511**
7.	5	19	31	66	**1**	119
8.	**1**	6	54	8	695	46
9.	9	**39**	24	**1**	65	2326
10.	6	2	14	214	48	5
11.	9	**1**	24	82	2	824
12.	14	**1**	106	237	185	44
13.	5	17	213	296	3	10
14.	6	18	**1**	24	1868	2194
15.	4	13	78	82	**1**	47
16.	**1**	3	18	47	**3316**	31
17.	3	**1**	7	34	78	1473
18.	2	3	10	313	60	**1**
19.	**1**	3	32	73	2	28
20.	8	**1**	5	3	1986	83
MRA	**1**	**3**	**10**	**23**	**116**	**195**
LRA	**2**	**1**	**6**	**26**	**22**	**38**

* The bold values represent the best and worst ranks of each feature.

B. FIRST ITERATION SVM-RFE FRAMEWORK

It has been established that SVM-RFE is a highly effective feature selection technique in terms of accuracy. However, filter methods perform better than SVM-RFE in terms of stability in many classification setups [25]. That can be linked to the simplicity of the filter methods, where within each training fold, features are ranked only in one iteration. Although SVM-RFE starts with the same feature set in each training fold, the algorithm may progress differently, and the final iterations in each training fold can be performed with very diverse feature sets. It is also known that single-run SVM-RFE achieves

better feature selection stability than multi-run SVM-RFE; however, its classification performance is worse [5].

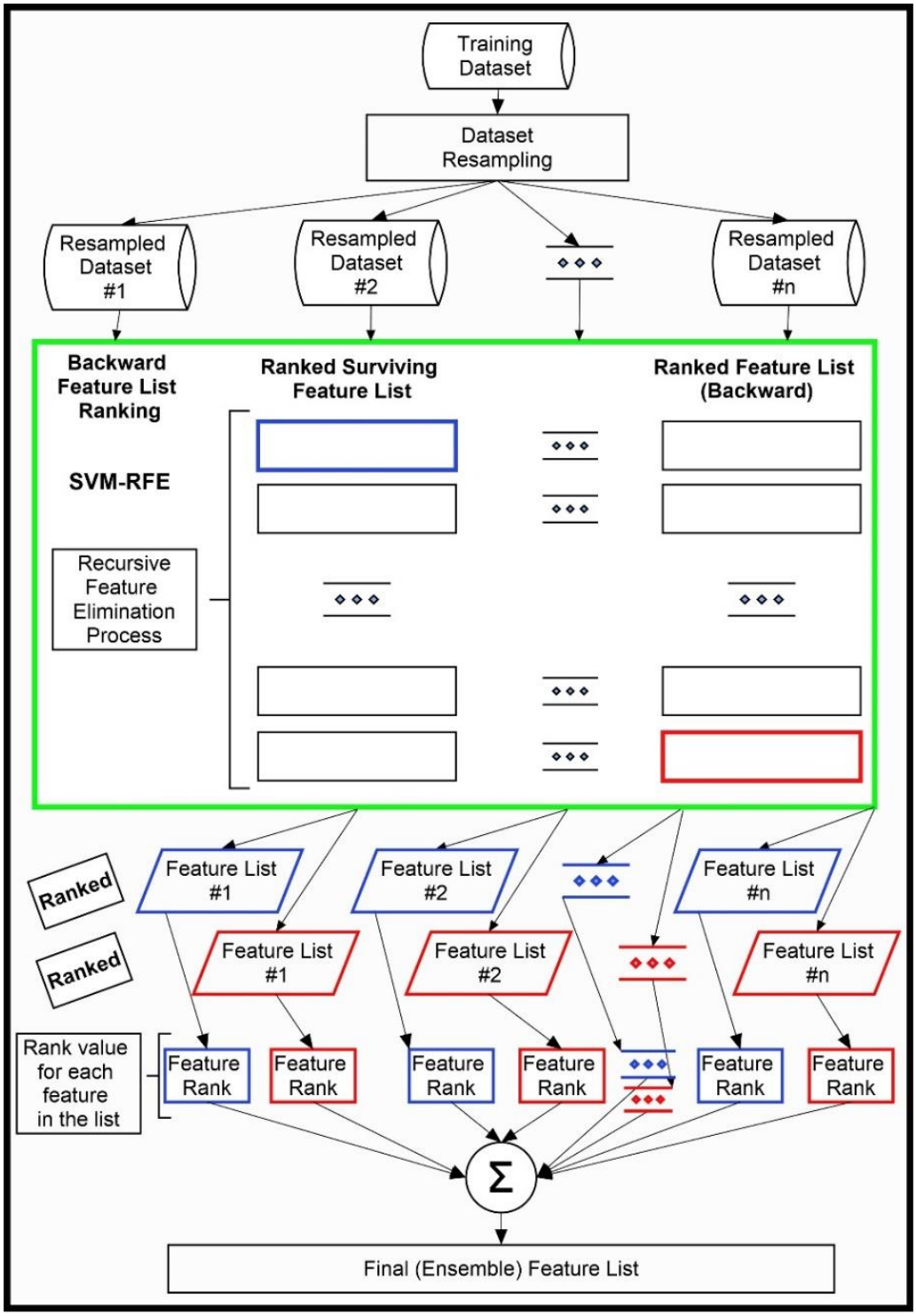

Figure 3. *Hybrid Aggregation Method*

Formula:

$$f(x_1, x_2, \dots, x_N) = \left(\frac{(\sum_{i=1}^{S} Fr_i^1 + (h * FI\_Fr_i^1)}{S}\right), \dots, \left(\frac{(\sum_{i=1}^{S} Fr_i^1 + (h * FI\_Fr_i^1)}{S}\right) \quad (2)$$

N = Feature size,

S = Ensemble size,

Fr_i = Feature k's rank in i^{th} Feature Rank List

FI_Fr_i = Feature k's rank in i^{th} Feature Rank List of 1^{st}_SVM-RFE based method

h = constant-value parameter to define the ratio of a particular feature rank for the hybrid method (to balance the weight of the rank value of FI_SVM-RFE method for the hybrid aggregation method).

We hypothesise that the feature ranks obtained in the first iteration of SVM-RFE can be unified by the conventional iterative SVM-RFE so that the advantages of both methods can be utilised. This is achieved using two different methods. The first method aggregates the final and first iteration rankings of RCV-20 SVM-RFE in an equally balanced way (Figure 3). The second method uses the final rankings and log values of the first iteration rankings with varying degrees of balance between them using the h parameter (Equation 2) described in Section 2.

C. SVM-RFE COST OPTIMISATION

SVM optimisation, specifically cost parameter (*C)* optimisation, has been extensively studied in various classification domains. However, the cost-parameter optimisation of the SVM algorithm employed in SVM-RFE and the SVM-RFE ensemble has been less studied. It has been well established that the *C* parameter affects the trade-off between training and generalisation accuracy of a linear SVM classifier. Choosing the *C* value in the higher range results in higher training accuracy but leads to a higher risk of overfitting. Choosing a low *C* might minimise the risk of overfitting. However, this might also lead to underusing the training data's potential, also known as the underfitting problem.

It is plausible that in the FS phase, this parameter has an important impact on the final selected feature set, and thus, on the performance of the classifier. More specifically, since *C* is effectively a regularisation parameter, it can affect the adaptation to the training data, and therefore, the trade-off between training and generalisation errors of the classifier. This, in turn, may not only determine the accuracy of the testing but also has an important impact on the stability of the feature selection. Therefore, it is sensible to experimentally study the cost parameter's effect on the feature selection performance.

Choosing an appropriate optimisation algorithm is the first concern of optimising SVM-RFE. Avoiding both overfitting and underfitting being caused by optimisation is another concern. The third concern about optimisation is preserving the separation of the training/testing dataset, which is critical for classification tasks to avoid classification bias.

One commonly used approach in optimisation is to perform optimisation outside the feature selection and classification processes on a small partition of the training dataset into two sections. In this approach, the whole process is repeated for each cost value to find the best one. We did not utilise this approach for two main reasons. Firstly, this method requires the already computationally complex EFS process to be repeated as many times as the number of experimental values of the parameter. Second, some training data should be kept only to validate the experimented value of the optimised parameters. That is, the classification algorithm and feature selection cannot use this section of the training dataset. Therefore, we injected an optimisation within a two-stage approach to perform an ensemble feature selection process, first optimising the FS technique and then using it for EFS. The first stage involves searching for FS's most fitted cost parameter using all training samples. The second stage will be the same as the ensemble learning described in Section 2 with the predetermined cost parameter.

For the optimisation process within each training fold, the selection of the features of the ensemble is mimicked by downgrading the entire 10×2 RCV ensemble feature selection process to a 5-fold CV. The 5-fold CV is appropriate for the search for cost parameters since it provides an acceptable trade-off

between generating a sufficient number of data sets to find the best cost value for the target model and the computational complexity of the optimisation algorithm. After data resampling is performed, the next step is to employ SVM-RFE to test the performance of each of the predefined cost parameter values in classification performance and gene selection stability. The stability obtained by the tested cost parameter can be calculated directly by the five training folds without testing the classifier. On the other hand, accuracy can only be measured by testing the classifier. Therefore, accuracy is obtained by averaging the results of all training and testing pairs. This process is depicted in Figure 4. The formulation for finding the best cost value is given in Equation 3.

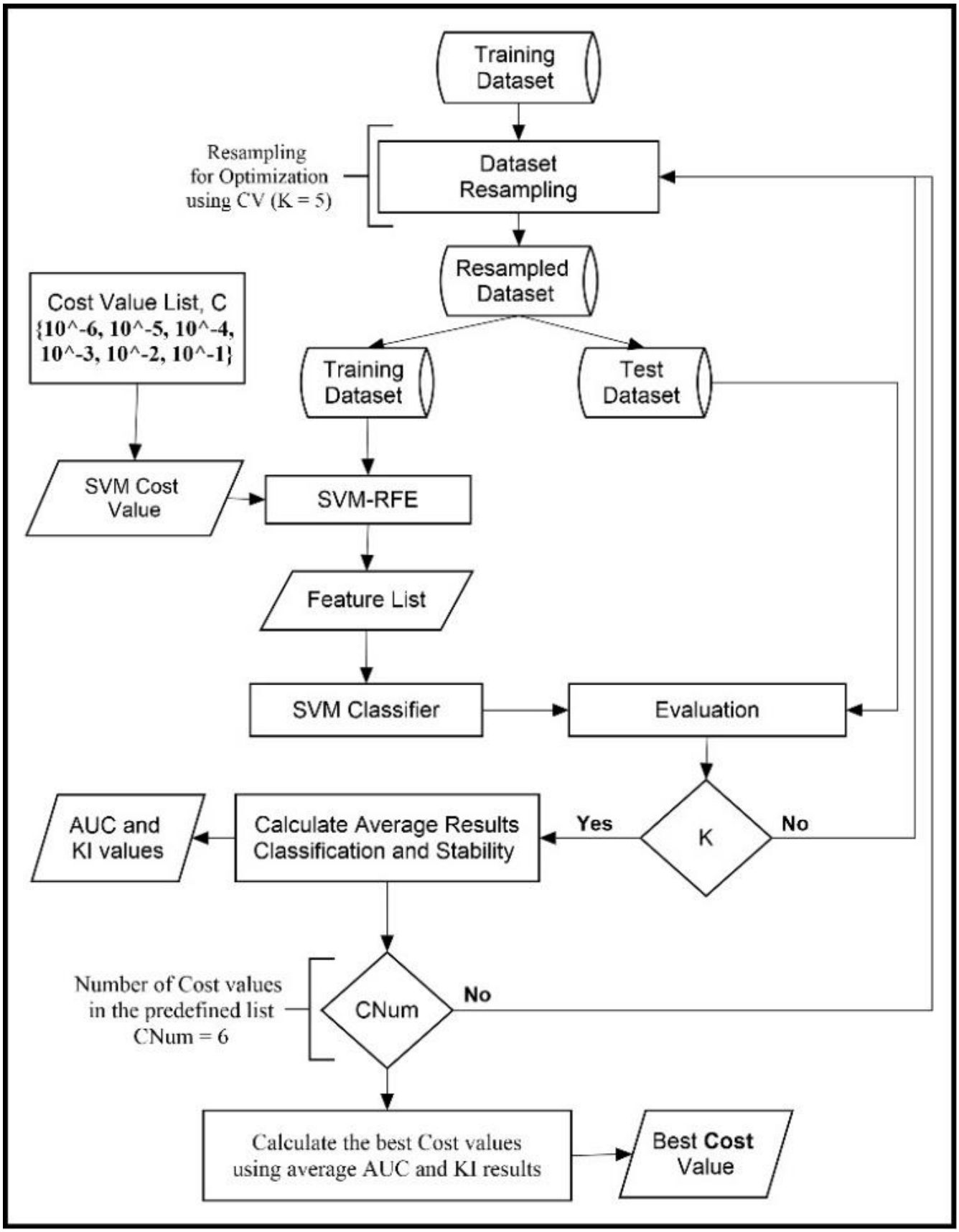

Figure 4. *Flowchart of Cost Parameter Optimisation for EFS.*

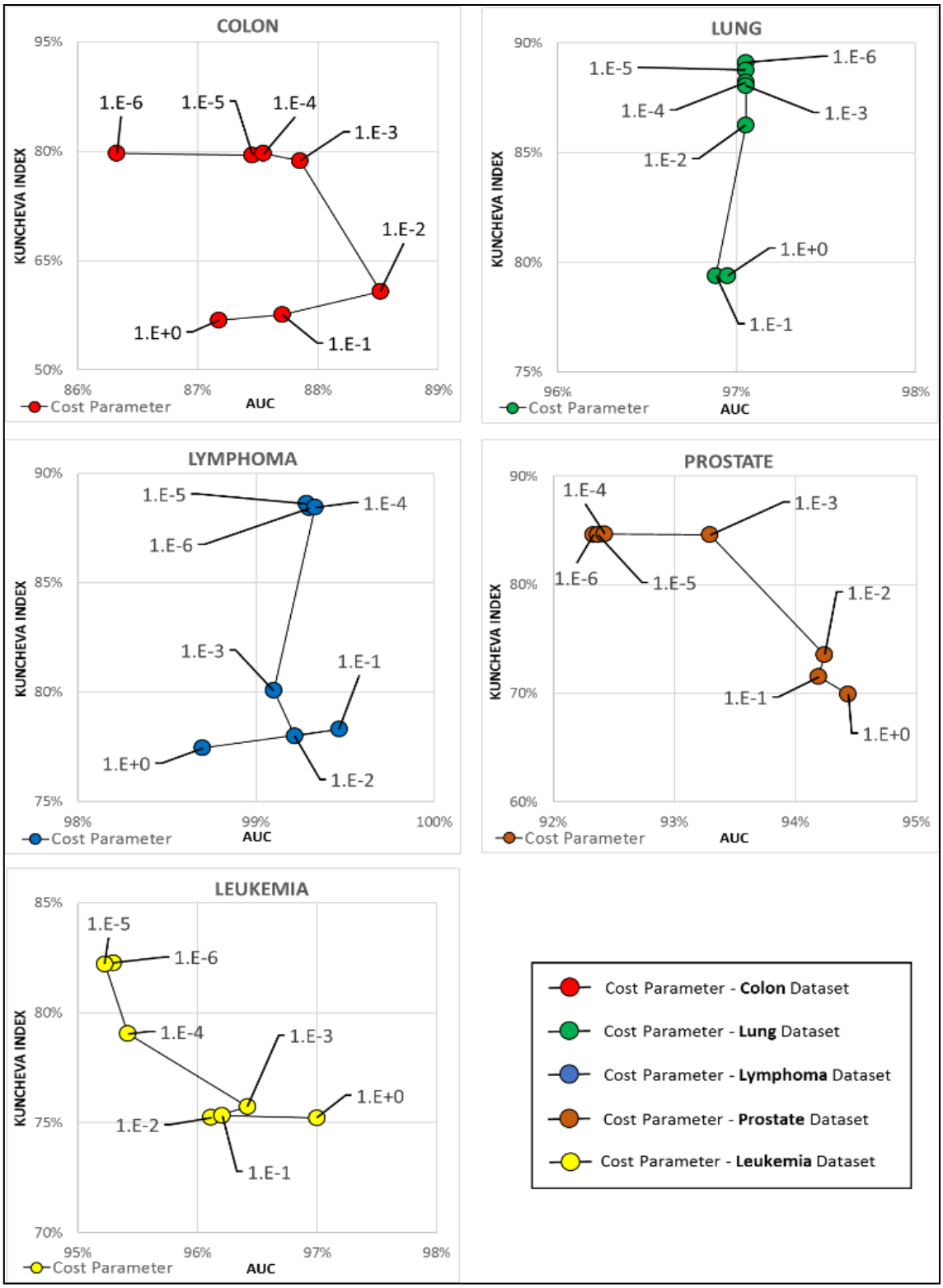

Figure 5. *Measurement of the impact of the cost parameter on the performance of RCV-EFS with ensemble size 20.*

The formula for Finding the **Best Cost Value, C_{Best}:**

$\boldsymbol{C_{Best}} = Max(AVG_{Cost})$, where

(3)

$AVG_{Cost} = ((\alpha * AVG_{AUC}) + ((1 - \alpha) * AVG_{KI})) / 2$

AVG_{AUC} = Average of all AUC values for all folds of CV

AVG_{KI} = Average of all KI values for all folds of CV
α = constant value for balancing the impact of ACC or KI for selecting the best cost parameter.
Figure 5 shows the average accuracy, and the KI results obtained using a range of cost values for all data sets. For all datasets, the cost value affects the KI monotonically. That is, decreasing the *C* value almost invariably results in higher stability. However, the cost parameter affects the trade-off between accuracy and KI in a dataset-dependent manner. For the lung data set, the *C* parameter does not significantly affect the accuracy. For the Colon dataset, the accuracy is affected. Lower performance was achieved when high and low C values were selected. Decreasing the cost parameter has a worse impact on the internal accuracy performance for the prostate dataset. A similar trend is seen within the Leukemia dataset. The impact of the cost parameter on the KI is much more pronounced than its effect on precision.

Additionally, the accuracy is measured within the training fold. Therefore, it should be interpreted cautiously, as the testing error might differ. We have implemented a function that determines the final cost value based on a weighted average of accuracy and KI scores where the total weights of the criteria are fixed as 1. Equation 3 represents the function implemented.

V. EXPERIMENTAL SETUP AND RESULTS

This section presents the testing setup, the used microarray datasets, and the obtained results. The results are also discussed, and the proposed method is compared with the conventional method.

A. MICROARRAY DATASETS

In this study, several publicly available microarray gene expression datasets are used to evaluate the proposed methods to measure performance and compare performance with the conventional method. The Feature (gene) size, sample size, and class distribution of the datasets used are available in Table 3.

Table 3. *List of microarray datasets*

Dataset	Feature Size	Sample Size	Class +	Class -
Colon [26]	2000	62	40	22
Prostate [27]	6033	102	52	50
Leukemia [28]	7129	72	45	27
Lung [29]	3312	156	139	17
Lymphoma [30]	4026	62	42	20

B. TESTING SETUP

To evaluate the performance of the proposed aggregation methods and compare them with the conventional method, the experimental setup of an external RCV was applied. In this setup, to avoid feature selection and classification bias in the results regarding classification performance and gene selection stability, the test dataset has never been seen by either the optimisation, the feature selection

or the training processes. The general outline of the experimental setup is illustrated in Figure 6. As explained in the Introduction, the contributions of this paper are in the cost parameter optimisation and aggregation phases, which are coloured orange and blue in the figure. In the conventional setup, because of its simplicity and lesser tendency to overfit compared to its alternatives, Linear Kernel has been selected as the choice of kernel, and the cost value is set to 1 in a fixed fashion.

Since the sample size of microarray data sets is quite limited, applying multiple training/testing splits for reliable results is necessary. This study used external RCV with ten folds and 25 repetitions (RCV-25-10) [5]. In total, this corresponds to 250 training/testing splits. In addition, internal RCV (RCV-10-2, equivalent to 20 ensembles) is applied wherever EFS is used. For SVM-RFE, E was set at 20%, StopValue was set at 5%, and the SVM-train cost parameter was set to 1.

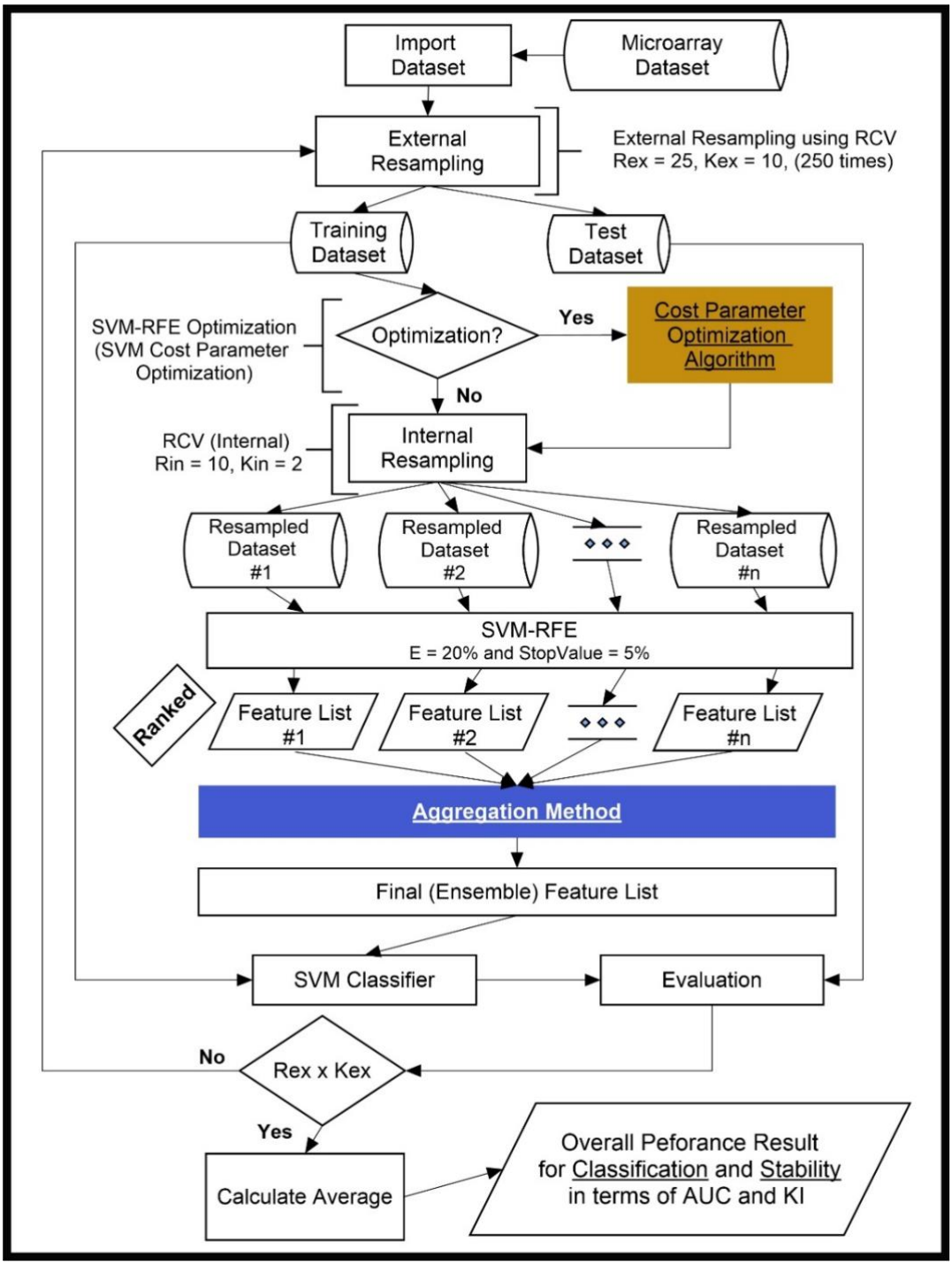

Figure 6. *Experimental Setup.*

Five metrics, area under roc (AUC) [31], accuracy, sensitivity, specificity, precision, and the Kuncheva's index (KI) [32], were used to measure classification performance and gene selection stability, respectively, with the selected percentage of features; 0.25%, 0.50%, 0.75%, 1%, 2%, 5%, 10%, and 25%. The definitions and formulas are given below.

Accuracy: The accuracy of a test is its ability to correctly differentiate the patient and healthy cases. To estimate the accuracy of a test, we should calculate the proportion of true positives and true negatives in all evaluated cases. Mathematically, this can be stated as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Here, TP, FP, TN, and FN are as follows;

- True positive (TP): the number of cases correctly identified as patient.
- False positive (FP): the number of cases incorrectly identified as patient.
- True negative (TN): the number of cases correctly identified as healthy.
- False negative (FN): the number of cases incorrectly identified as healthy.

Sensitivity: The sensitivity of a test is its ability to determine patient cases correctly. To estimate it, we should calculate the proportion of true positives in patient cases. Mathematically, this can be stated as follows:

$$Sensitivity = \frac{TP}{TP+FN} \quad (5)$$

Specificity: The specificity of a test is its ability to determine healthy cases correctly. We should find the proportion of true negatives in healthy cases to estimate it. Mathematically, this can be stated as follows:

$$Specificity = \frac{TN}{TN+FP} \quad (6)$$

Precision is a measure that tells how frequently a patient labelled as positive is actually positive.

$$Precision = \frac{TP}{TP+FN} \quad (7)$$

We are using KI as the stability measure, which is mathematically stated as

$$KI = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} I_C(S_i(k), S_j(k)) \quad (8)$$

Here, I_C is the similarity between the i^{th} and j^{th} feature sets $S_i(k)$ and $S_j(k)$.

C. RANK AGGREGATION METHODS

All rank aggregation methods were built using RCV-EFS [5], with the underlying FS technique, SVM-RFE. The methods presented in this section are *MRA*, *LRA*, *1st_LRA*, and *1st_LRA+MRA* when *h* is set to 0.01 and 0.1. In addition, *MRA+1st_LMRA* is also presented. The obtained AUC and KI values for the mentioned methods are illustrated in Figures 7 and 8, respectively. In addition, Accuracy, Specificity, Sensitivity, and Precision values obtained for the listed methods are presented in the Appendix section.

For the classification performance of Lung, Lymphoma, and Prostate datasets, all techniques achieve similar performances for a wide range of features in terms of accuracy. For Leukaemia and Colon, logarithmic methods perform worse than the *MRA*. *LRA* generally outperforms *MRA* when the feature size is less than 2%. This advantage disappears when the feature size is larger. Furthermore, the hybrid method *MRA+1st_LRA* performs closer to *MRA* or *1st_LRA* depending on the value of *the parameter h.* However, the other hybrid model *1st_MRA+MRA* performs similarly to the *MRA*, indicating that the performance loss in *MRA+1st_LRA* is due to the logarithmic approach rather than using the first iteration ranks. For the colon, *1st_LRA* clearly performs worse than other techniques. *LRA* performs well in the very low range but is worse in the later larger feature sizes. The other techniques perform similarly,

where *1st_LRA +0.01_MRA* slightly performs worse, and *MRA+ 1st _MRA* performs slightly better than other techniques. The picture for leukaemia is similar to the colon, except that *1st_LRA+0.01_MRA* performs well in the low feature range. For lung, all methods perform similarly except for small feature sizes where *1st_LRA* and *1st_LRA+0.01_MRA* perform considerably worse than the others. LRA is the highest for lymphoma, performing from 0.1 to 0.5, where all methods perform virtually 100% afterwards. The prostate is not similar to other datasets in terms of accuracy since *1st_LRA* and *1st_LRA+0.01_MRA* perform well for a wide range of features.

Regarding average stability, *LRA* generally outperforms *MRA* when the feature size is less than 2%. This advantage disappears when the feature set is higher. However, when only the non-hybrid methods are considered, *1st_LRA* clearly outperforms the other methods with a clear margin where the difference is more pronounced for smaller feature sets. This fact seems to be a result of the synergy between the use of the logarithmic approach and the initial ranks together since the *LRA* did not result in such a difference from conventional *MRA*. In terms of accuracy, *1st_LRA* performs less well than *1st_LRA+ 0.01_MRA* in a range of small feature sizes.

It is seen that the hybrid method, *MRA+1st_LRA*, outperforms *MRA* by a considerable margin, suggesting that aggregation of the first and final ranking is an effective method in terms of stability. However, *1st_LRA* clearly outperforms both *LRA* and *MRA+1st_LRA*, suggesting a synergy between the logarithmic approach and first-iteration ranks. Hybrid methods (*1st_LRA+ MRA)* perform between the two extremes, depending on the parameter *h*.

In terms of average sensitivity, MRA-based methods generally perform better. However, *1st_LRA* performs worse than other techniques for the colon. *LRA* performs in the very low feature range but performs slightly worse than the others in larger feature sizes. For leukaemia, in the minimal feature ranges, *1st_LRA* and *LRA* perform worse than the others. In the higher range, all other methods perform similarly. The picture for the lung and prostate is similar to that for leukaemia. For lymphoma, *1st_LRA* performs well in the very small feature size but degrades later. *LRA* performs in the middle range, and in the high feature range, all methods perform virtually the same.

Regarding average specificity, LRA performs better than other methods in very small feature sizes; however, it performs worse among all other methods. Other methods, except for *1st_LRA*, perform similarly. For the prostate, *1st_LRA* and *1st_LRA+0.01_MRA* are the top-performing methods. For lung and leukaemia, the picture is similar to the average specificity. For Lymphoma, *LRA* and *1st_LRA* perform well. In terms of average precision, hybrid methods perform well in a low feature range, except for the *LRA* and *1st_LRA* methods. For leukaemia and lung, the picture is similar in accuracy. For Lymphoma, *LRA* virtually performs 100%. For prostate, *1st_LRA* and *LRA* perform well. For the colon, except for *1st_LRA*, which performs worse than the others, all methods perform similarly.

Regarding average stability, LRA generally outperforms MRA when the feature size is less than 2%. This advantage disappears when the feature set is higher. However, when only the non-hybrid methods are considered, 1st_LRA clearly outperforms the other methods with a clear margin where the difference is more pronounced for smaller feature sets. This fact seems to be a result of the synergy between the use of the logarithmic approach and the initial ranks together since the LRA did not result in such a difference from conventional MRA. In terms of accuracy, 1st_LRA performs less well than 1st_LRA+ 0.01_MRA in a range of small feature sizes. It is seen that the hybrid method, MRA+1st_LRA, outperforms MRA by a considerable margin, suggesting that aggregation of the first and final ranking is an effective method in terms of stability. However, 1st_LRA clearly outperforms both LRA and

MRA+1st_LRA, suggesting a synergy between the logarithmic approach and first-iteration ranks. Hybrid methods (1st_LRA+ MRA) perform between the two extremes, depending on the parameter h.

In terms of average sensitivity, MRA-based methods generally perform better. However, *1st_LRA* performs worse than other techniques for the colon. *LRA* performs in the very low feature range but performs slightly worse than the others in larger feature sizes.

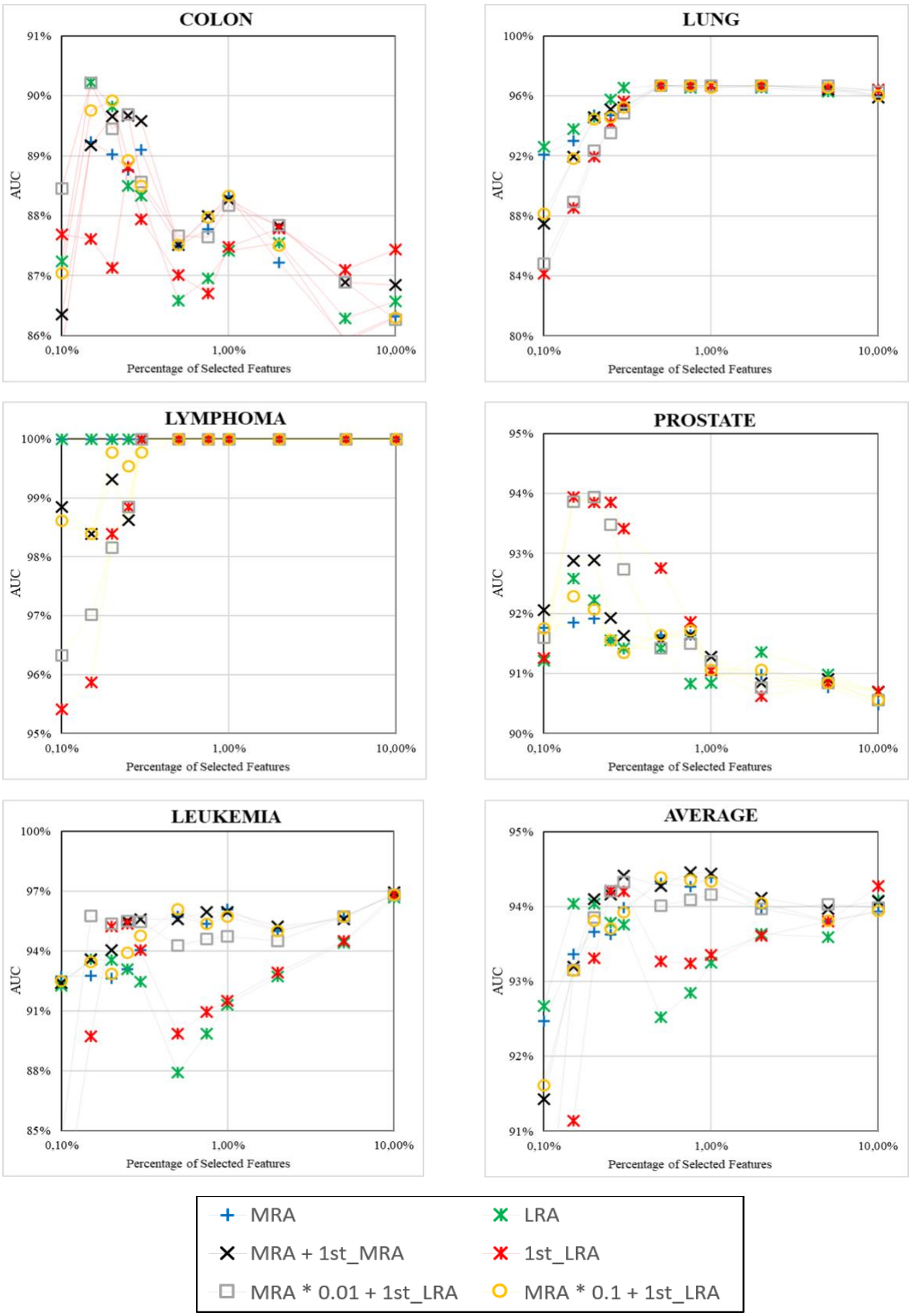

Figure 7. *Classification Performance of the Rank Aggregation Methods in terms of AUC*

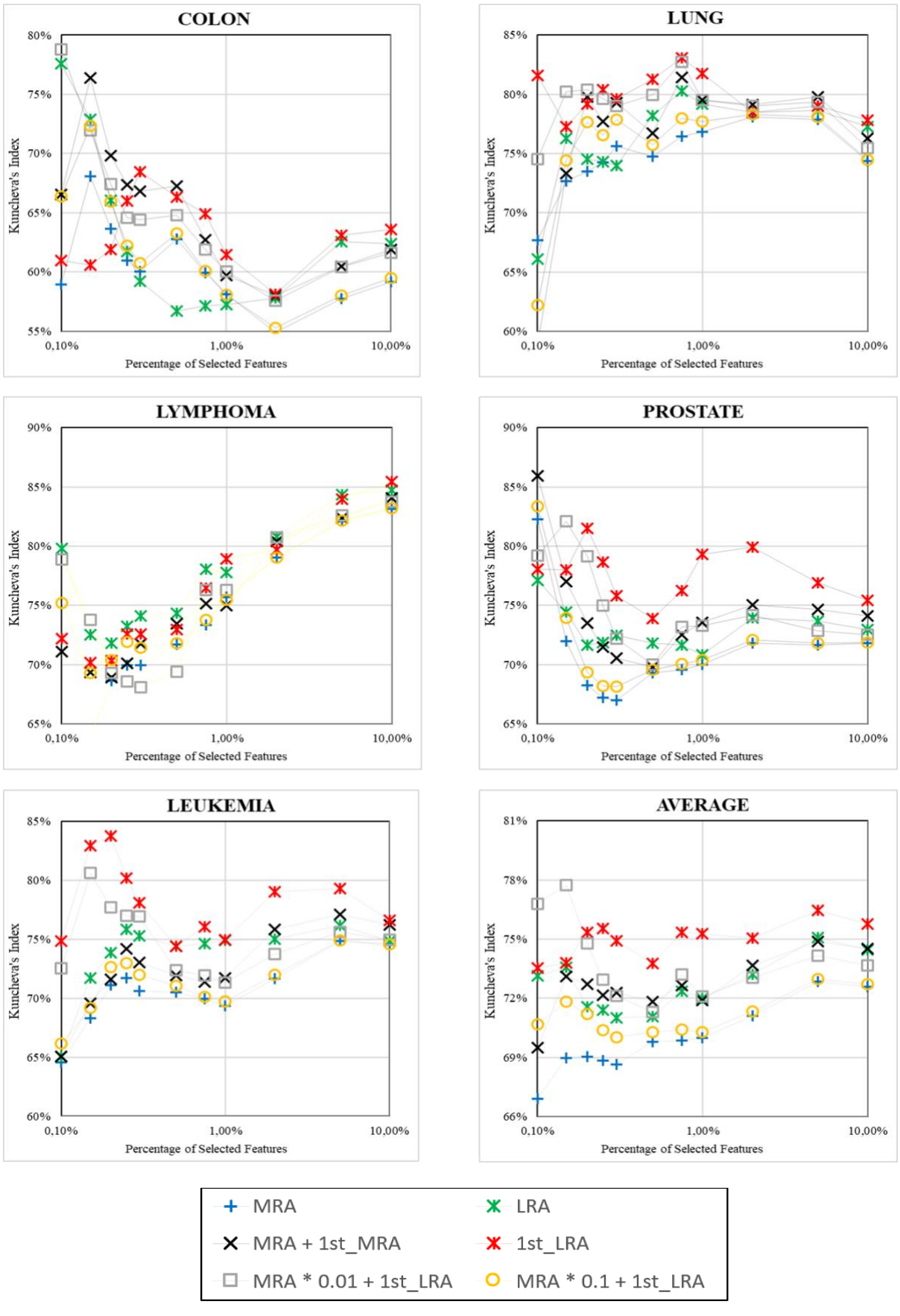

Figure 8. *Gene Selection Stability of the Rank Aggregation Methods in terms of KI*

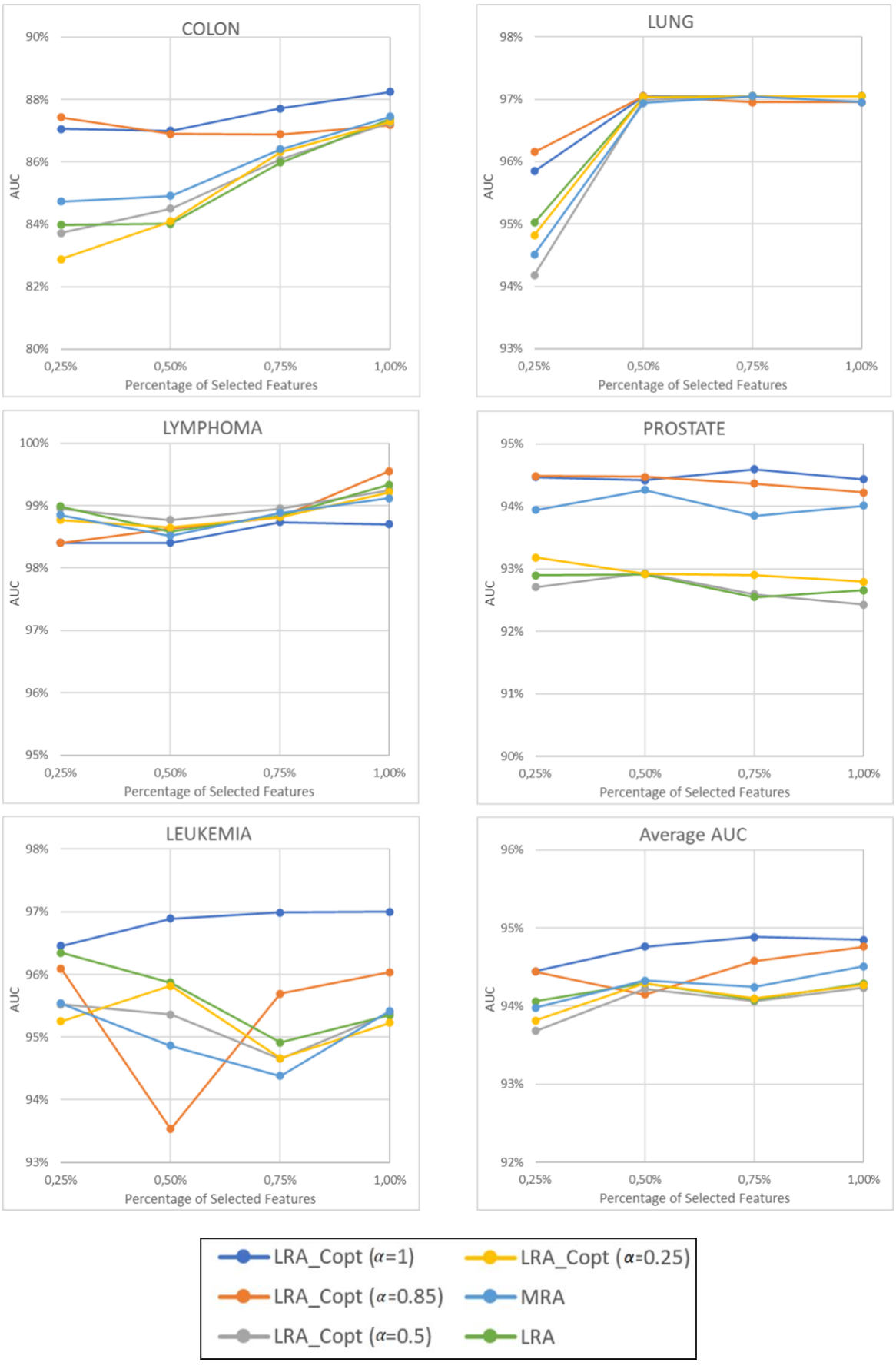

Figure 9. *Classification Performance of the Optimised Log Rank Aggregation Methods in terms of AUC*

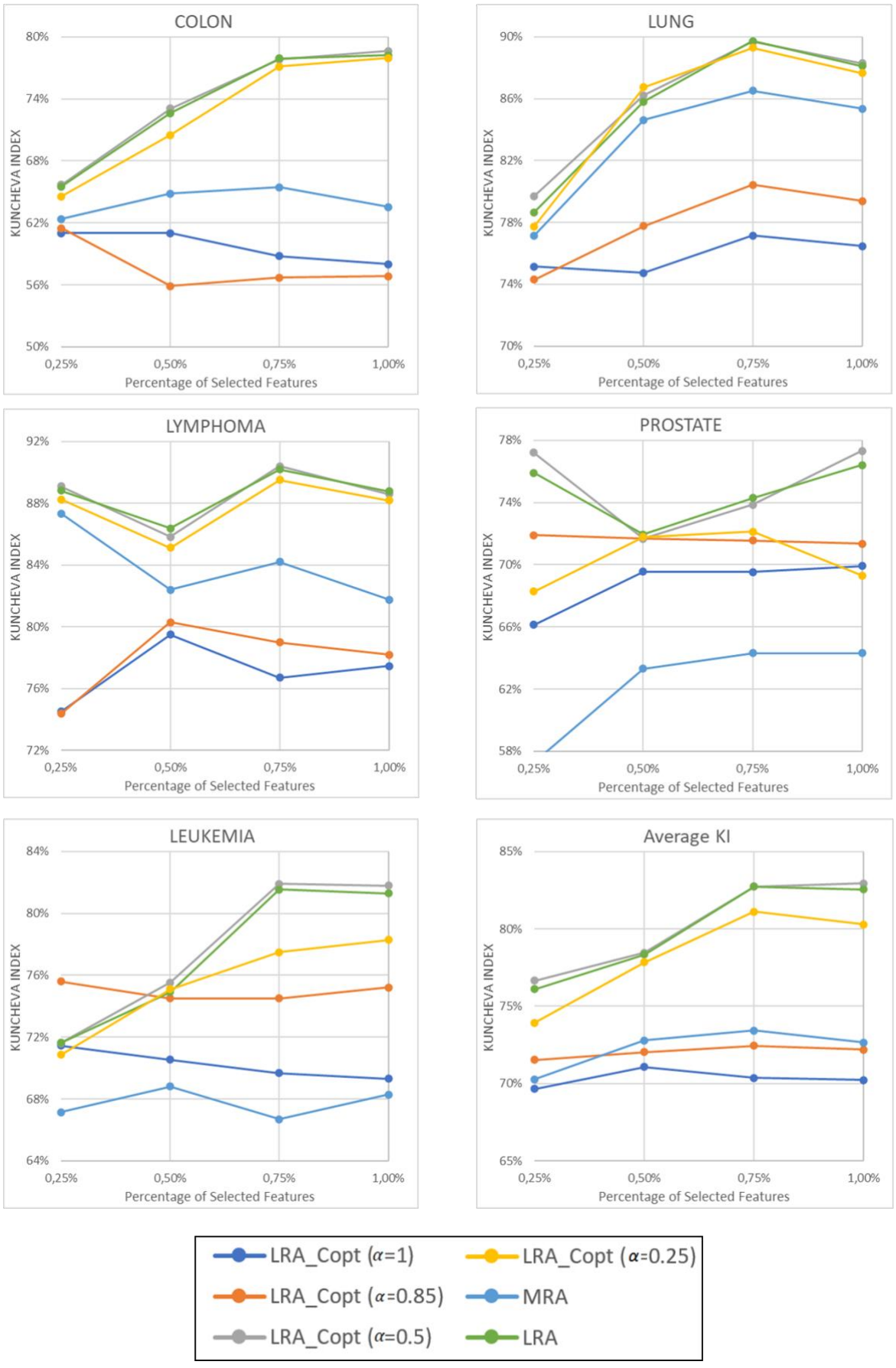

Figure 9. *Gene Selection Stability of Optimised Log Rank Aggregation Methods in terms of KI*

For leukaemia, in the very low feature range, *1st_LRA* and *LRA* perform worse than the others. In the higher range, all other methods perform similarly. The picture for the lung and prostate is similar to that for leukaemia. For lymphoma, *1st_LRA* performs well in the very small feature size but degrades later. *LRA* performs in the middle range, and in the high feature range, all methods perform virtually the same. In terms of average specificity, LRA performs better than other methods in very small feature sizes; however, it performs worse among all other methods. Other methods, except for *1st_LRA*, perform similarly. For the prostate, *1st_LRA* and *1st_LRA+0.01_MRA* are the top-performing methods. For lung and leukaemia, the picture is similar to the average specificity. For Lymphoma, *LRA* and *1st_LRA* perform well. In terms of average precision, hybrid methods perform well in a low feature range, except for the *LRA* and *1st_LRA* methods. For leukaemia and lung, the picture is similar in accuracy. For Lymphoma, *LRA* virtually performs 100%. For prostate, *1st_LRA* and *LRA* perform well. For the colon, except for *1st_LRA*, which performs worse than the others, all methods perform similarly.

D. OPTIMISATION ALGORITHM FOR LRA-EFS

In this subsection, the experimental results of four different settings of optimised LRA, along with non-optimised MRA and LRA, are presented and discussed. The explained optimisation framework is proposed to evaluate and set the best cost value for each fold of the external RCV. The best cost value is selected by obtaining the weighted average of the best accuracy and KI values. Four different α values are selected to measure the impact of the weight on the performance of the proposed algorithm. These AUC and KI values are illustrated in Figures 9 and 10, respectively.

For all datasets, in terms of stability, *LRA_*C_{opt} variations outperform *MRA* with a clear margin. On average, the difference is around 5% when the feature size is 0.25% and 0.5% and jumps to 10% when the feature size is 0.75% and 1%. It is seen that the stability only partially degrades as more weight is given to accuracy. However, as expected, stability decreases to similar levels of *LRA* when α=1. Optimisation with respect to α=1 performs worse than *MRA* in terms of precision. This is possibly linked to two factors: one is the possible overfitting since the cost parameter is optimised within the training data, and the other plausible reason is the downgrading of the actual process to a modest CV process. However, a substantial gain in stability is observed. In terms of sensitivity, specificity, and precision, non-optimised MRA and LRA generally perform better than the optimised schemes except for specific datasets.

VI. DISCUSSION

This section discusses the strength and limitations of this study and compares the proposed method with the current literature. In this paper, we proposed novel aggregation and cost optimisation methods for ensemble feature selection that are proposed to improve classification and gene selection stability for microarray-based cancer diagnosis. To the best of our knowledge, this is the first paper that uses a transformation function for feature aggregation. Another main contribution of this paper is a hybrid model, which uses the initial and final feature ranks. Lastly, a cost-parameter optimisation framework is proposed for microarray datasets. To evaluate the proposed methods and compare the performance of the proposed methods with the conventional method, the experimental evaluation procedure is performed on five publicly available microarray datasets.

MRA has some drawbacks, especially when the signature size is limited; a poorly positioned outlier in a single SVM-RFE fold can eliminate an elsewhere successful feature from the final list. To overcome the mentioned drawbacks of MRA, the LRA is proposed, which calculates the logarithmic values of feature ranks for a particular feature in the list before taking the overall average. Intuitively, this approach is more sensitive to changes in the top-performing range; a slight fall in rank is penalised dramatically. The feature is further

penalised if it is at the end of the range rather than being in the middle of the range; however, only with a limit. This approach might have three critical practical advantages. The first is that obtaining logarithmic values of feature ranks can eliminate the major limitation of mean aggregation by eliminating the impact of bad outliers on a feature. The second is that, unlike stability selection methods, the feature score is not dramatically affected by a slight change if it is close to the decision boundary. Third, modifying the formula of mean aggregation aims to keep its simplicity and computational cost as low as possible while improving the performance. A case study is performed to illustrate how aggregation methods affect the final positions of the features in aggregation. The choice of aggregation method has been proven to impact significantly the position of generally good-performing features associated with a limited number of poor outlier positions. It is seen that the ranks obtained by *LRA* have a lower standard deviation, and *LRA* is more robust against the presence of outliers. The test results show that *LRA* has a comparable but slightly worse performance than conventional *MRA* in terms of precision and significantly improves gene selection stability. Furthermore, this functional modification has no cost in terms of computational complexity.

Methods that exploit the ranks obtained in the first iteration also provide better stability than *MRA*. In particular, the method that utilises both logarithmic aggregation and the first iteration frame at the same time, *1st_LRA*, provides outstanding performance in terms of stability with a generally tolerable decrease in accuracy. Hybrid methods that use the weights obtained in the first and last iterations of SVM-RFE (*MRA+1st_MRA* and *MRA +1st_LRA)* provide good trade-off points between accuracy and stability. 1st*_MRA + MRA* provides a net increase in stability without virtually compromising accuracy. SVM-RFE operates on a fixed set of samples in a single subsampled fold. However, the ranks obtained in the first and final iterations are produced via different conditions due to different feature sets being used. Consequently, obtaining ranks with different feature subpopulations creates diversity analogous to using different sample subpopulations. Therefore, *1st_MRA + MRA* effectively uses an ensemble size practically more than the number of training subsamples, that is, twenty in this study. This is achieved with virtually no increase in computational complexity.

Optimisation within the training folds shows that this parameter has an important impact on the ensemble feature selection phase, like the SVM classifier. In other words, a smaller cost parameter increases the stability of SVM-RFE because this set includes more samples as support vectors. The testing results show that the proposed cost optimisation technique provides remarkable gains in terms of stability with relatively less fall in accuracy. Assigning a higher weight to internal accuracy has a limited positive impact on testing accuracy that can be linked to two factors: one is the possible overfitting since the cost parameter is optimised within the training data, and the other plausible reason is the downgrading of the actual process to a modest CV process. However, a substantial gain in stability is observed.

In summary, the cost parameter has a remarkable impact on the performance of RCV-EFS. It is seen that when the weight of the stability in the algorithm increases, the stability of the features increases as well. The algorithm achieves this usually by reducing the value of the cost parameter, which is equivalent to increasing the impact of regularisation in SVM. It can also be seen that, among the different datasets, the impact of optimisation does not affect the performance in terms of accuracy in a drastic manner. This lack of general effect can be linked to the trade-off between regularisation and generalisation errors in classification problems. Each dataset has a different optimal point in terms of regularisation vs generalisation trade-off; therefore, increasing the regularisation via stability has an individual effect for each dataset.

This subsection discusses the performance of the proposed method in comparison to the studies in the literature that evaluated on the common datasets. The comparisons show that the proposed method have comparable or better classification performance and gene selection stability. It is crucial to note that these comparisons should be interpreted with caution due to the differences in the experimental settings.

Barbara Pes conducted extensive experiments to measure the impact of EFS on classification performance and gene selection stability on various datasets from many problem domains in the study [18]. The present and referred studies have evaluated the performance of the proposed methods on several datasets, including lymphoma. In [18], the classification performance of 0.964 and 0.971 for the mentioned dataset was achieved by SVM and Random Forest classifiers, respectively, where the top 80 genes were selected using SVM-RFE. The proposed method, optimised LRA, achieved 99% AUC with ten features. In addition, the proposed method showed that it is robust since it obtained 90% KI for the top 10 selected features.

In the study, [33] the authors proposed an efficient implementation of linear SVM, improved the recursive feature elimination strategy, and combined them to select informative genes. Experiments were conducted on six frequently used microarray datasets in this field. The results show that the proposed methods have significantly reduced time consumption and obtained comparable classification performance. The common datasets used in the study are Leukemia, Prostate, and Colon. The study's reported results are generally better; however, this could result from the different experimental setups since 5-Fold CV, as the validation method can generate biased results due to insufficient training and test splits. For instance, In the study [5], the classification performance for the Colon dataset with the Repeated CV validation method was around 85% which used SVM-RFE FS and Linear SVM classifier. However, the study [33] achieved 98.75 AUC for the same dataset with the same FS technique and classifier.

Chen et al. [34] proposed a wrapper gene selection approach, WERFE, within a recursive feature elimination (RFE) framework to make the classification more accurate. WERFE employs an ensemble strategy and assembles the top-selected genes in each approach as the final gene subset. By integrating multiple gene selection algorithms, the optimal gene subset is determined by prioritising the more relevant genes selected by each gene selection method. The authors reported 98% classification performance for the prostate dataset with two features, which is better than the results reported in this study. MRMR-HFS [35] is a CFS-inspired filter-based feature selection algorithm that selects features by an ensemble of ranking algorithms by maximising class accuracy and minimising feature-to-feature similarity. The obtained MRMR-HFS is useful for feature subset selection in high-dimensional datasets in terms of various measures. The results reported in [21] were slightly better for the colon dataset than our study and slightly worse for the prostate dataset.

Brahim et al. [8] proposed an EFS approach based on the reliability assessment of feature selectors. It aims to provide a unique and stable feature selection without ignoring the predictive accuracy aspect. A classification algorithm is used as an evaluator to assign confidence to features selected by ensemble members based on their associated classification performance. The common datasets used in the study are Lymphoma, Prostate, and Lung. The proposed method's performance was measured in terms of accuracy using F-measure and stability using Kuncheva's index (KI). The present study reported results in terms of AUC and KI. When the stability of both methods was compared, the results revealed that optimised LRA outperformed the reliability assessment (RAA)-based method, where RAA and LRA achieved stability (KI) 15.66% and 88.80% for lymphoma, 81.94% and 77.30% for prostate, and 85.39% and 88.30% for the lung, with top 1% features, respectively. According to the results, optimised LRA showed less stability than RAA only for the prostate dataset, where default 1st_LRA obtained 80.3% stability.

Venkatesh and Anuradha proposed a hybris ensemble technique using filter and wrapper FS techniques, and the Fuzzy Gaussian membership function was used for feature aggregation. Then, SVM-RBF was used for model evaluation of several microarray datasets, including Colon, Prostate and Lymphoma datasets. For the Lymphoma dataset, the default LRA setting achieved 99.2% with the top 30 features, and the proposed method in [15] achieved 93.33%. In addition, their method obtained 95.89% accuracy with 53 for the prostate dataset, whereas our method had an accuracy of 94.8% with 15 features. On the other hand, 86.5% accuracy was achieved for the

Colon dataset using 20 features by LRA and 94.57% accuracy was reported in the study [15]. It should be noted that data normalisation implementation or model validation can cause such differences in the results. Therefore, the differences in experimental settings can cause differences in the results.

The study [16] proposes an ensemble fast correlation-based (FCBF) FS method. The proposed method selected the most informative genes from the microarray datasets for cancer classification. According to the reported results, FCBF with the Quarter Scheme aggregation method achieved the best performance. For the common dataset colon, it obtained an accuracy of 83.87% with selected features of an average of 7.6 with a 2.1 standard deviation. LRA achieved 85.50% accuracy for the same dataset with five features. On the other hand, for the Leukemia dataset, one of our hybrid methods (MRA * 0.01 + 1^{st}_LRA) achieved 97.6% with 12 features, where FCBF with the Quarter Scheme aggregation method achieved 98.61% accuracy with selected features of an average of 83.2 with a 4.7 standard deviation.

The study [17] proposed the HMM-based gene selection method using a hidden Markov chain model. The method was evaluated on several datasets, including DLBCL, leukaemia, and prostate datasets. The present study reported 97.6% AUC and 80.6 KI for Leukemia dataset, 94.9 AUC and 78.0 KI for the prostate dataset by the LRA method. The obtained results are better for the prostate and worse for leukaemia. It should be noted that the studies used very different validation methods.

The authors proposed PSO-ENSVM using swarm optimisation, elastic net and SVM classifier [36]. For model evaluation, a 10-fold CV was used. However, our study used repeated CV (RCV, where r = 25 and k = 10) for the same purpose. For the common datasets, Colon, Leukemia, Prostate and Lung, classification performance, 0.85, 0.97, 0.94 and 0.97 were achieved in the study [36], and the present study exhibited similar results with fewer features.

In the study [37], another wrapper feature selection-based method was proposed using cuckoo search with evolutionary operators. This study also used CV-10 for model evaluation. The study reported that their method achieved an accuracy of 98.60% and 95.20% for Leukemia and Prostate, respectively. In the present study, a hybrid method (MRA * 0.01 + 1^{st}_LRA) achieved 97.6% accuracy for leukaemia with 12 features and the same model and 1^{st}_LRA achieved 94.9% accuracy for the Prostate dataset.

In [38], a method was developed to detect Microarray-based Leukemia disease using the ant lion optimisation-based wrapper FS technique. The study reported the accuracy as 90.91% by the ALO attribute selection method. The reported accuracy is lower than the accuracy reported in this study, where a proposed hybrid model achieved 97.6% accuracy for leukaemia with 12 features and default LRA achieved 95.9% accuracy with eight features. As shown, this study outperformed the ALO-based EFS technique. The study [39] employed an artificial neural networks method optimised using Genetic Algorithms. The proposed method was evaluated on the colon dataset, and the reported results suggest that the best performance was 0.832 AUC, which this study achieved higher performance than (default LRA: 90.2% AUC with three genes) ALO-based EFS technique.

In [40], the proposed modified AHP (MAHP) was evaluated using the leave-one-out cross-validation (LOOCV) validation technique, and K was set to 20 for creating training and testing splits. The top five features were selected after feature selection and evaluated with several classifiers, including the SVM classifier, in terms of AUC. MAHP achieved 0.828 AUC when evaluated using the SVM classifier. Our method, LRA, achieved 90.2% AUC with three features when evaluated with the same classifier. Our method achieved 72.9% KI for the same gene set. For the other common dataset Leukemia, MAHP achieved 0.971 classification performance with five features, and a slightly worse performance was achieved (95.7%) by a hybrid method (MRA * 0.01 + 1^{st}_LRA) with 11 features and 80.6% KI.

VII. CONCLUSION

In conclusion, this paper presents contributions in various steps of SVM-RFE-based EFS, particularly for feature aggregation and cost parameter optimisation. The proposed methods resulted in important gains in terms of stability. While some of the proposed techniques showed little compromise in terms of accuracy, others improved stability with virtually no compromise in accuracy. It should be stressed that the gain in stability is cumulative among the different techniques used. Compared to conventional aggregation method MRA, the combined use of the proposed techniques provides approximately a 10% gain in stability with a compromise between 0.5 and 1% accuracy. Accurate and stable artificial intelligence systems help domain experts make more confident decisions for microarray-based cancer diagnosis and provide reliable insight into the disease due to the competitive stability of the selected genes.

Conflict of interest

The authors declare that there are no potential conflicts of interest.

Funding

This research did not receive a specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

VIII. REFERENCES

[1] N. Mahendran, P. M. Durai Raj Vincent, K. Srinivasan, and C.-Y. Chang, "Machine learning based Computational Gene Selection Models: A survey, performance evaluation, open issues, and future research directions," Frontiers in Genetics, vol. 11, 2020. doi:10.3389/fgene.2020.603808.

[2] V. K. Chauhan, K. Dahiya, and A. Sharma, "Problem formulations and solvers in Linear SVM: A Review," Artificial Intelligence Review, vol. 52, no. 2, pp. 803–855, 2018. doi:10.1007/s10462-018-9614-6.

[3] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and Applied Feature Selection Methods," Information Sciences, vol. 282, pp. 111–135, 2014. doi:10.1016/j.ins.2014.05.042.

[4] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with Ensemble Feature Selection Methods," Bioinformatics, vol. 26, no. 3, pp. 392–398, 2009. doi:10.1093/bioinformatics/btp630.

[5] H. Güney and H. Öztoprak, "Microarray-based cancer diagnosis: Repeated cross-validation-based ensemble feature selection," Electronics Letters, vol. 54, no. 5, pp. 272–274, 2018. doi:10.1049/el.2017.4550.

[6] D. Guan, W. Yuan, Y.-K. Lee, K. Najeebullah, and M. K. Rasel, "A review of Ensemble Learning Based Feature Selection," IETE Technical Review, vol. 31, no. 3, pp. 190–198, 2014. doi:10.1080/02564602.2014.906859.

[7] B. Pes, “Ensemble feature selection for high-dimensional data: A stability analysis across multiple domains,” Neural Computing and Applications, vol. 32, no. 10, pp. 5951–5973, 2019. doi:10.1007/s00521-019-04082-3.

[8] A. Ben Brahim and M. Limam, “Ensemble feature selection for High Dimensional Data: A new method and a comparative study,” Advances in Data Analysis and Classification, vol. 12, no. 4, pp. 937–952, 2017. doi:10.1007/s11634-017-0285-y.

[9] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, “An ensemble of filters and classifiers for Microarray Data Classification,” Pattern Recognition, vol. 45, no. 1, pp. 531–539, 2012. doi:10.1016/j.patcog.2011.06.006.

[10] A. Anaissi, M. Goyal, D. R. Catchpoole, A. Braytee, and P. J. Kennedy, “Ensemble feature learning of genomic data using support Vector Machine,” PLOS ONE, vol. 11, no. 6, 2016. doi:10.1371/journal.pone.0157330.

[11] P. Yang, B. B. Zhou, Z. Zhang, and A. Y. Zomaya, “A multi-filter enhanced genetic ensemble system for gene selection and sample classification of Microarray Data,” BMC Bioinformatics, vol. 11, no. S1, 2010. doi:10.1186/1471-2105-11-s1-s5.

[12] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, and A. Alonso-Betanzos, “Ensemble feature selection: Homogeneous and heterogeneous approaches,” Knowledge-Based Systems, vol. 118, pp. 124–139, 2017. doi:10.1016/j.knosys.2016.11.017.

[13] L. Cleofas-Sánchez, J. S. Sánchez, and V. García, “Gene selection and disease prediction from gene expression data using a two-stage hetero-associative memory,” Progress in Artificial Intelligence, vol. 8, no. 1, pp. 63–71, 2018. doi:10.1007/s13748-018-0148-6.

[14] S. Hengpraprohm and S. Jungjit, “Ensemble feature selection for breast cancer classification using Microarray Data,” Inteligencia Artificial, vol. 23, no. 65, pp. 100–114, 2020. doi:10.4114/intartif.vol23iss65pp100-114.

[15] B. Venkatesh and J. Anuradha, “A fuzzy gaussian rank aggregation ensemble feature selection method for Microarray Data,” International Journal of Knowledge-based and Intelligent Engineering Systems, vol. 24, no. 4, pp. 289–301, 2021. doi:10.3233/kes-190134.

[16] A. Wang et al., “Stable and accurate feature selection from microarray data with ensembled fast correlation based filter,” 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020. doi:10.1109/bibm49941.2020.9313533.

[17] M. Momenzadeh, M. Sehhati, and H. Rabbani, “A novel feature selection method for microarray data classification based on Hidden Markov Model,” Journal of Biomedical Informatics, vol. 95, p. 103213, 2019. doi:10.1016/j.jbi.2019.103213.

[18] G. Zhang, J. Hou, J. Wang, C. Yan, and J. Luo, “Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm,” Interdisciplinary Sciences: Computational Life Sciences, vol. 12, no. 3, pp. 288–301, 2020. doi:10.1007/s12539-020-00372-w.

[19] O. A. Alomari et al., “Gene selection for microarray data classification based on Gray Wolf optimiser enhanced with TRIZ-inspired operators,” Knowledge-Based Systems, vol. 223, p. 107034, 2021. doi:10.1016/j.knosys.2021.107034.

[20] X. Zheng, W. Zhu, C. Tang, and M. Wang, “Gene selection for microarray data classification via Adaptive Hypergraph Embedded Dictionary Learning,” Gene, vol. 706, pp. 188–200, 2019. doi:10.1016/j.gene.2019.04.060.

[21] S. Raghavendra. N and P. C. Deka, “Support Vector Machine applications in the field of Hydrology: A Review,” Applied Soft Computing, vol. 19, pp. 372–386, 2014. doi:10.1016/j.asoc.2014.02.002.

[22] X. Zhang, D. Qiu, and F. Chen, “Support vector machine with parameter optimisation by a novel hybrid method and its application to fault diagnosis,” Neurocomputing, vol. 149, pp. 641–651, 2015. doi:10.1016/j.neucom.2014.08.010.

[23] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene Selection for Cancer Classification using Support Vector Machines,” Machine Learning, vol. 46(1), pp. 389–442, 2002.

[24] R. Wald, T. M. Khoshgoftaar, and D. Dittman, “Mean aggregation versus robust rank aggregation for ensemble Gene Selection,” 2012 11th International Conference on Machine Learning and Applications, 2012. doi:10.1109/icmla.2012.20.

[25] A.-C. Haury, P. Gestraud, and J.-P. Vert, “The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures,” PLoS ONE, vol. 6, no. 12, 2011. doi:10.1371/journal.pone.0028210.

[26] U. Alon et al., “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays,” Proceedings of the National Academy of Sciences, vol. 96, no. 12, pp. 6745–6750, 1999. doi:10.1073/pnas.96.12.6745.

[27] D. Singh et al., “Gene expression correlates of clinical prostate cancer behaviour,” Cancer cell, vol. 1, pp. 203–209, 2002.

[28] T. R. Golub et al., “Molecular classification of cancer: Class Discovery and class prediction by Gene Expression Monitoring,” Science, vol. 286, no. 5439, pp. 531–537, 1999. doi:10.1126/science.286.5439.531.

[29] G. J. Gordon et al., “Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma,” Cancer Res, vol. 62, pp. 4963–4967, 2002. doi:10.1126/science.286.5439.531.

[30] A. Alizadeh et al., “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling,” Nature, vol. 403, pp. 503–511, 2000.

[31] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006. doi:10.1016/j.patrec.2005.10.010

[32] L. I. Kuncheva, “A stability index for feature selection,” In Artificial intelligence and applications, pp. 421–427, 2007.

[33] Z. Li, W. Xie, and T. Liu, “Efficient feature selection and classification for Microarray Data,” PLOS ONE, vol. 13, no. 8, 2018. doi:10.1371/journal.pone.0202167.

[34] Q. Chen, Z. Meng, and R. Su, “Werfe: A gene selection algorithm based on recursive feature elimination and ensemble strategy,” Frontiers in Bioengineering and Biotechnology, vol. 8, 2020. doi:10.3389/fbioe.2020.00496.

[35] M. K. Ebrahimpour and M. Eftekhari, “Ensemble of Feature Selection Methods: A hesitant fuzzy sets approach,” Applied Soft Computing, vol. 50, pp. 300–312, 2017. doi:10.1016/j.asoc.2016.11.021.

[36] M. Qaraad, S. Amjad, P. El-Kafrawy, H. Fathi, and I. I. M. Manhrawy, “Parameters optimisation of elastic net for high dimensional data using PSO algorithm,” 2020 International Conference on Intelligent Systems and Computer Vision (ISCV), 2020. doi:10.1109/iscv49265.2020.9204218.

[37] M. S. Othman, S. R. Kumaran, and L. M. Yusuf, “Gene selection using hybrid multi-objective cuckoo search algorithm with evolutionary operators for cancer microarray data,” IEEE Access, vol. 8, pp. 186348–186361, 2020. doi:10.1109/access.2020.3029890.

[38] D. Santhakumar and S. Logeswari, “Efficient attribute selection technique for leukaemia prediction using microarray gene data,” Soft Computing, vol. 24, no. 18, pp. 14265–14274, 2020. doi:10.1007/s00500-020-04793-z.

[39] [1] K. Cahyaningrum, Adiwijaya, and W. Astuti, “Microarray gene expression classification for cancer detection using artificial neural networks and genetic algorithm hybrid intelligence,” 2020 International Conference on Data Science and Its Applications (ICoDSA), 2020. doi:10.1109/icodsa50139.2020.9213051.

[40] T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi, “A novel aggregate gene selection method for microarray data classification,” Pattern Recognition Letters, vol. 60, pp. 16–23, 2015. doi: 10.1016/j.patrec.2015.03.018.

IX. APPENDIX

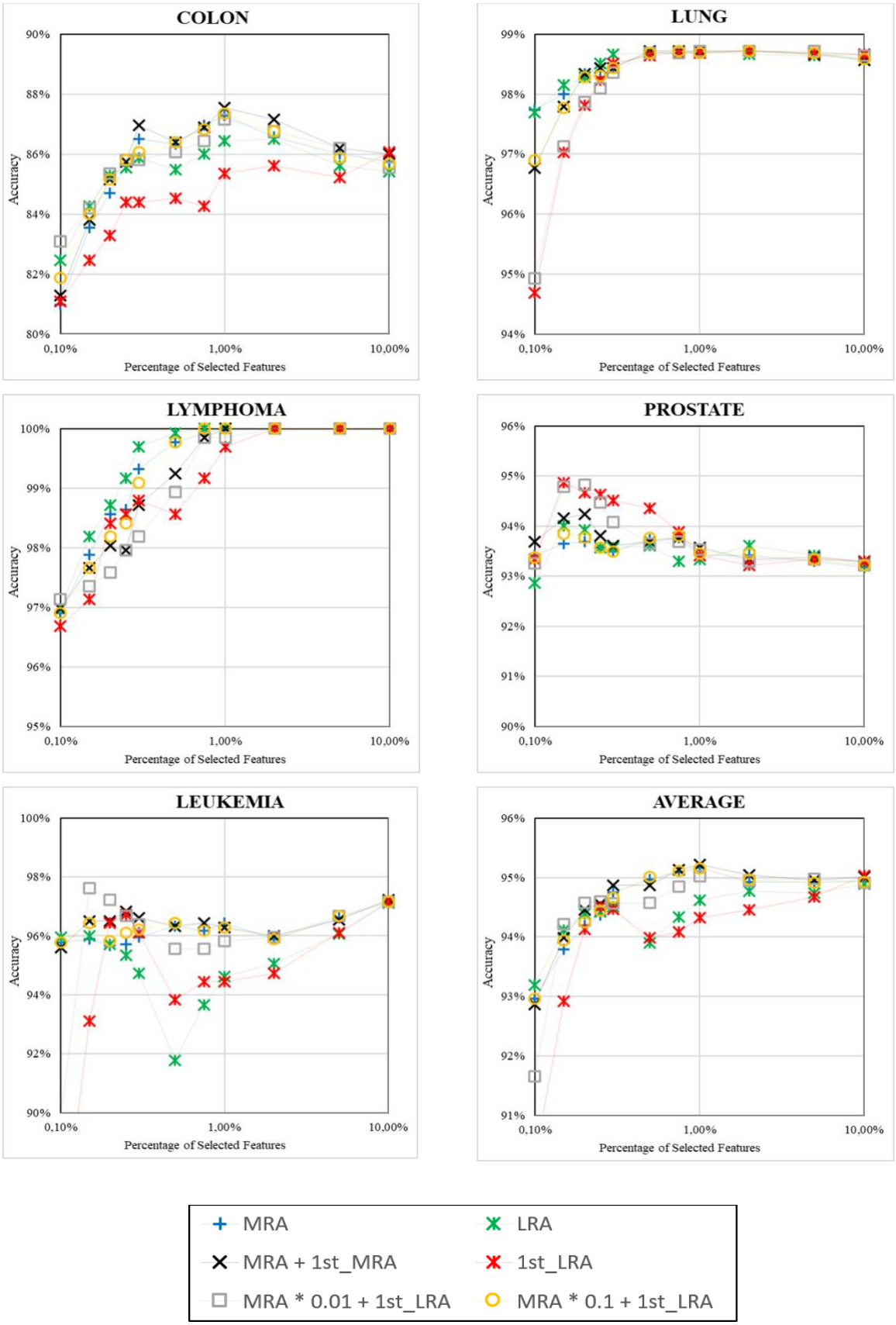

Figure A1. Classification Performance of the Rank Aggregation Methods in terms of Accuracy

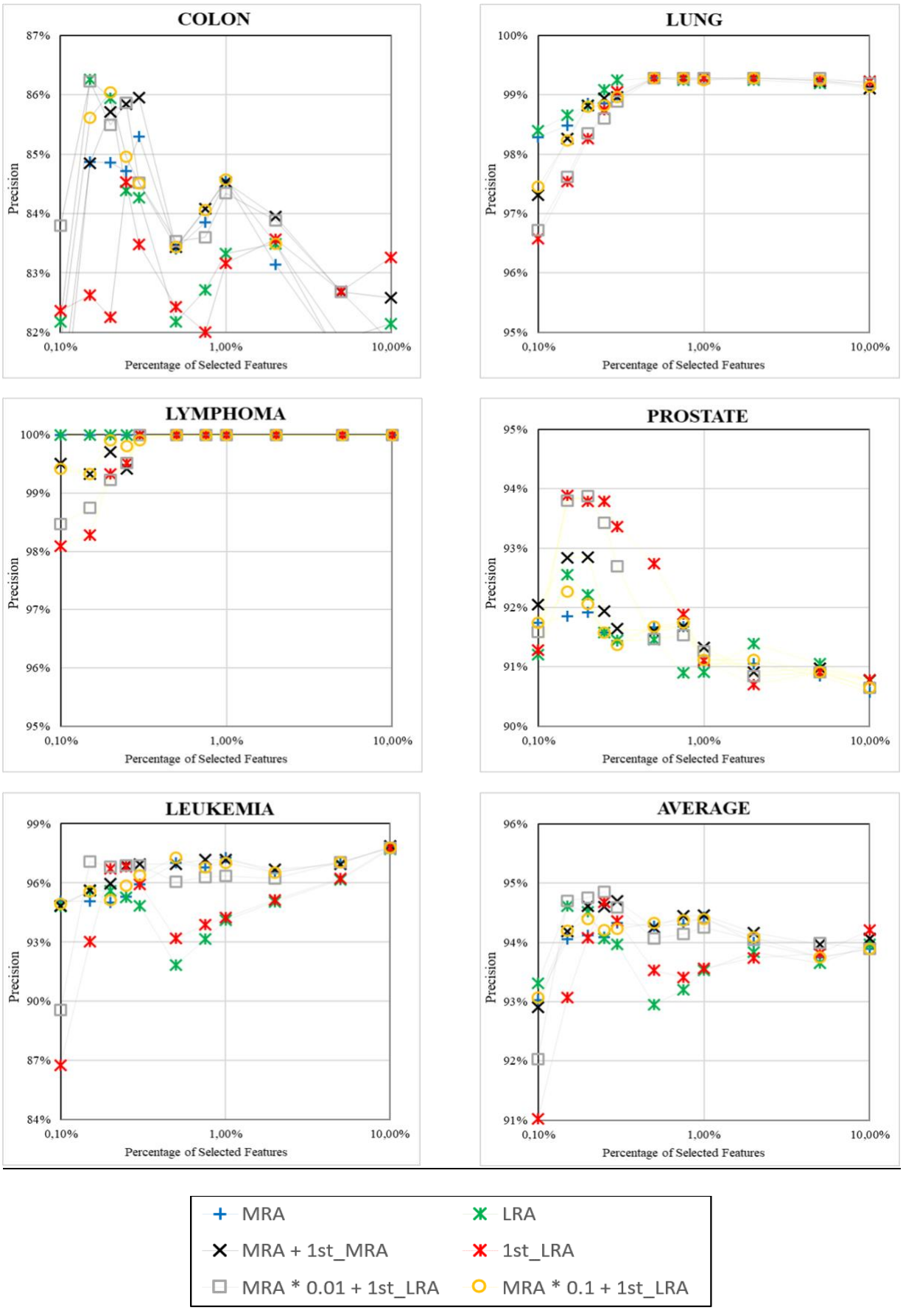

Figure A2. Classification Performance of the Rank Aggregation Methods in terms of Precision

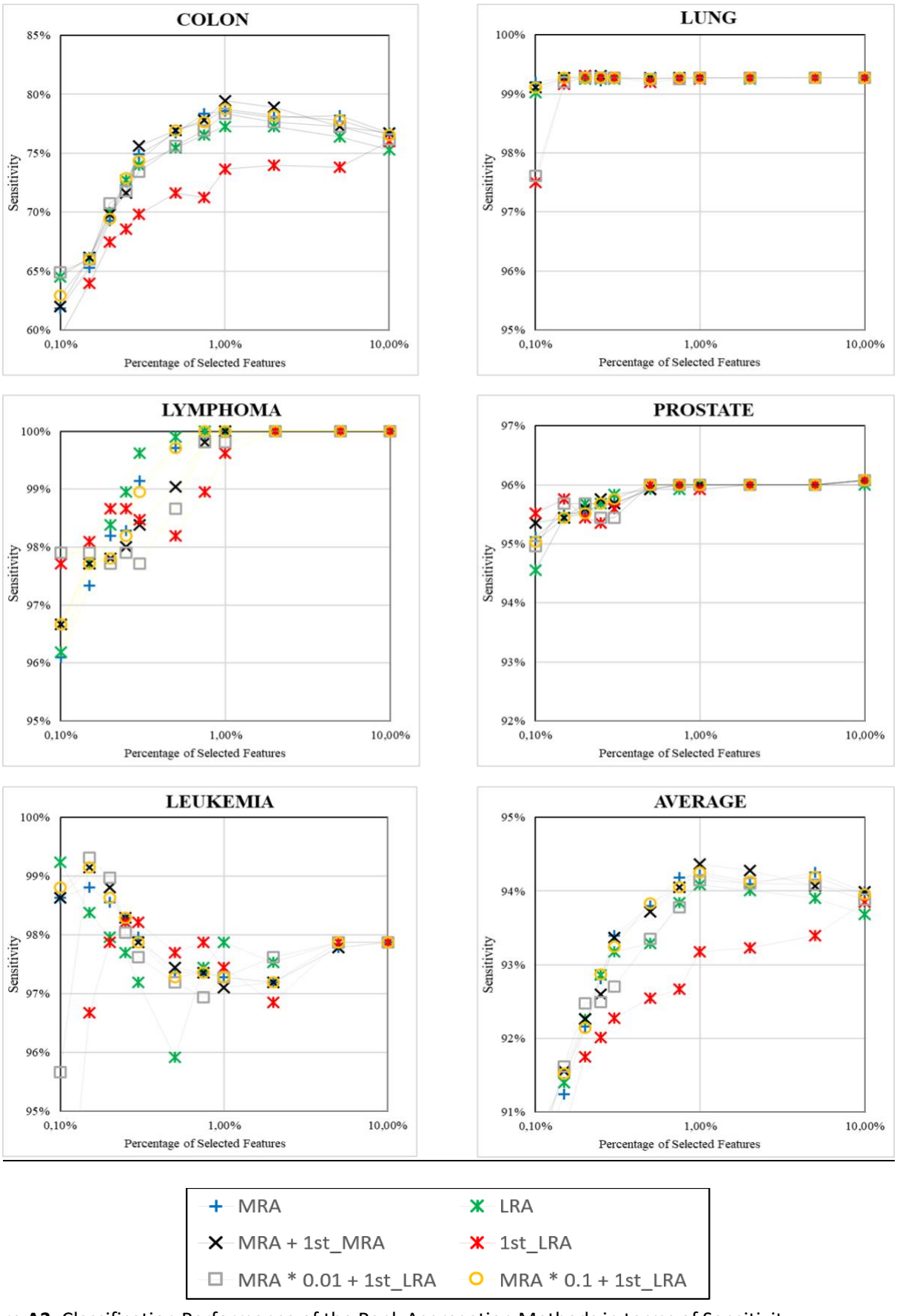

Figure A3. Classification Performance of the Rank Aggregation Methods in terms of Sensitivity

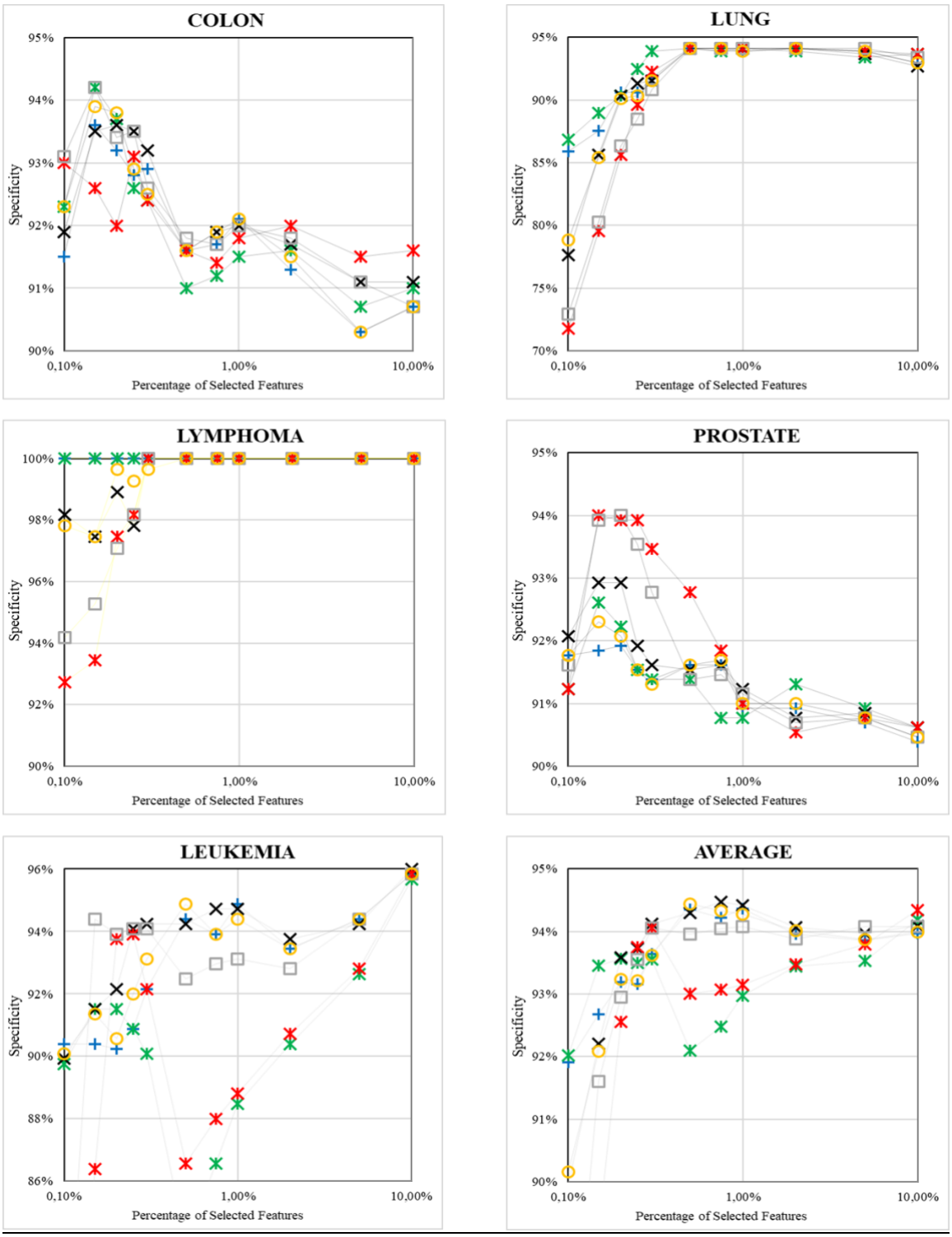

Figure A4. Classification Performance of the Rank Aggregation Methods in terms of Specificity