# Exploring the Stability of Feature Selection Methods across a Palette of Gene Expression Datasets

Zahra Mungloo-Dilmohamud
Department of DT, FoICDT
University of Mauritius
Reduit, Mauritius
z.mungloo@uom.ac.mu

Yasmina Jaufeerally-Fakim
Biotechnology Dept, FoA
University of Mauritius
Reduit, Mauritius
yasmina@uom.ac.mu

Carlos Peña-Reyes
School of Business and Engineering Vaud (HEIG-VD)
University of Applied Sciences Western Switzerland (HES-SO)
Vaud, Switzerland
carlos.pena@heig-vd.ch

## ABSTRACT

Gene expression data often need to be classified into classes or grouped into clusters for further analysis, using different machine learning techniques and an important pre-processing step is feature selection (FS). The aim of this study is to investigate the stability of some diverse FS methods on a plethora of microarray gene expression data. This experimental work is broken into three parts. Step 1 involves running some FS methods on one gene expression dataset to have a preliminary assessment on the similarity, or dissimilarity, of the resulting feature subsets across methods. Step 2 involves running two of these methods on a large number of different datasets to investigate whether the results produced by the methods are dependent on the features of the dataset: binary, multiclass, small or large dataset. The final step explores how the similarity of selected feature subsets between pairs of methods evolves as the size of the subsets are increased. Results show that the studied methods display a high amount of variability in terms of the resulting selected features. The feature subsets differed both inter- and intra- methods for different datasets. The reason behind this is not clear yet and is being further investigated. The final objective of the research, that is to define how to select a FS method, is an ongoing work whose initial findings are reported herein.

## CCS Concepts

• **Applied computing→Life and medical sciences→Bioinformatics**

## Keywords

feature selection; gene expression data; machine learning; stability

## 1. INTRODUCTION

Recent years have seen tremendous progress made in experimental techniques and this has led to a lot of data, in different formats, being produced. Researchers can make use of these data to gain a better understanding of their field and solve many problems. Biomarker discovery is one major area of research. Biomedical investigators have been looking for unique molecular markers associated with disease processes with the aim of detecting diseases early, determining prognosis, monitoring the response of patients to therapy or selecting those treatments most likely to be successful. The data involved here are complex and often contain noise as well as redundancy and can only be processed and analysed computationally. This has resulted in a compendium of tools and algorithms being generated. Machine Learning tools can be used to analyse big data and feature selection (FS) is often a first step in machine learning. Feature selection is the selection of the smallest subset of features that maximally increases the performance of the classification model used by the machine learning tool.

A researcher wishing to analyse his data using machine learning tools faces the problem of selecting an appropriate FS method since a plethora of FS methods are available [19], [9] and [25]. In [19], a meta-review based on a number of pas on FS, the authors have proposed an extended taxonomy for classifying FS methods which is, as close as possible, the current state of the art. The proposed taxonomy has six top-level criteria: *selection management, type of evaluation, training approach, class dimensionality, model linearity* and *additional knowledge required*. Any FS method can be classified into 1 or more of these top-level criteria.

In [25], the authors present a comprehensive list of FS methods and some of these FS methods are listed in Table 1. These methods are Significance Analysis of Microarrays (SAM) [24], LIMMA [23], Rank Product Analysis [5], ReliefF [14], SVM-Recursive Feature Elimination (SVM-RFE) [7, 29], Support-Vector Machine (SVM) [27] and RFE + RF(Random Forest) [15]. SAM is a method where a score is allocated to each gene depending on the change in gene expression relative to the standard deviation of repeated measurements. It uses t-test based statistics. LIMMA incorporates different statistical methods and uses linear models to analyse data. Rank Product Analysis is a non-parametric statistical method and is based on the calculation of rank products. ReliefF, is an extension of the Relief [12, 13] method and is able to deal with incomplete and noisy data as well as multiclass data. SVM-RFE is a popular wrapper approach for variable selection. It is an algorithm for both linear and non-linear kernels and uses a backward elimination procedure to produce a ranked list of features. SVM is a method where a hyperplane is drawn between different points with the objective of finding a plane that has the maximum distance between data points of the different classes. RFE + RF is a method where the RFE has been used as the FS and the RF has been used as classifier.

Apart from the multitude of methods available, FS methods may produce different results for the same data [1, 4, 8]. Therefore the

*ICBBE '19*, November 13–15, 2019, Shanghai, China
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7299-2/19/11…$15.00
DOI: https://doi.org/10.1145/3375923.3375938

robustness or stability of the methods is a concern. Some attempts have been made to solve this robustness problem and various criteria to assess the stability of results have been proposed [2, 11, 16, 17, 20–22, 26]. Some of these robustness metrics are *Robustness Index* [22] *Significantly self-consistent selections* [17], *Kuncheva Stability Index* [16], *Pearson's correlation coefficient* [2, 11], *Spearman Rank Correlation Coefficient* [2, 11] and *Tanimoto distance metric or Jaccard Index* [2, 11]. It should be pointed out that often, in literature, robustness and similarity have been used interchangeably and have been calculated using different approaches.

In our research, a robust FS method has been defined as one which is able to find stable subsets of features while enabling a good performance of the classifier or the predictive model. A simple and intuitive way to assess the robustness or stability is to measure the difference/similarity between features present in various subsets obtained under different starting conditions.

## 2. MATERIALS AND METHODS

### 2.1 FS Methods and Gene Expression Datasets

Since the goal of the work is to study the stability (repeatability) of the resulting feature subsets, it is important to compare a relatively large spectrum of feature selection methods with different datasets. We would like to emphasize that the aim is not to compare/analyse classic performance measurements such as accuracy, sensitivity, etc. but to concentrate, instead, on the lists of biomarkers (feature subsets) produced by various methods.

A simple and intuitive approach to assess the robustness/consistency of feature selection is to compare the resulting subsets from different methods on the same dataset. Hence, the need to study different methods with same/different datasets. Each case is to be further investigated in terms of the dataset specificity and feature selection method's way of choosing features to find out what caused the selected feature subset to be similar or different.

Table 1 shows the selected feature selection methods categorized according to 4 top-level criteria [1], and Table 2 lists the selected datasets that were studied. The methods were implemented either using R or Python or ran on Weka [28].

As seen in Table 2, the datasets selected were quite diverse with the classes being binary or multiclass, sample sizes ranging from 16 to 190 and number of features ranging from around 7,000 to around 55,000. The methods that were run with the datasets were either binary methods or multiclass methods. When using binary methods on a multiclass classification problem, the typical strategy is to decompose it into a series of binary ones, and to generate an importance statistic for each feature on each binary problem. This issue was handled by the methods themselves.

**Table 1. Feature selection methods selected for performing analyses**

| FS Methods | Selection Management | Class Dimensionality | Model Linearity | Model Representation |
|---|---|---|---|---|
| SAM [24] | Filter | Multiclass | Linear | Statistical Based |
| LIMMA [23] | Filter | Multiclass | Linear | Statistical Based |
| Rank Product Analysis [5] | Filter | Binary | Non Linear | Statistical Based |
| ReliefF [14] | Filter | Multiclass | Non Linear | Similarity Based |
| SVM-RFE [7, 29] | Ensemble | Binary | Linear/Non Linear | Kernel |
| SVM [27] | Embedded | Binary | Linear | Kernel |
| RFE + RF [15] | Wrapper | Multiclass | Non Linear | Trees |

**Table 2. Datasets used for running the FS methods**

| Datasets | Source | Number of Samples | Number of Features | Classes |
|---|---|---|---|---|
| Leukemia 1 (GOLUB) | [6] | 72 | 7129 | Binary |
| Leukemia 2 | [3] | 104 | 22283 | Binary |
| Leukemia 3 | [10] | 190 | 22277 | Multiclass |
| Breast cancer 1 | [3] | 50 | 22283 | Multiclass |
| Prostate cancer 1 | [10] | 22 | 22153 | Multiclass |
| 9-cancers Dataset | [3] | 174 | 54674 | Multiclass |
| MELAS 1 | [3] | 16 | 22214 | Binary |
| Ulcerative colitis 1 | [3] | 16 | 54675 | Binary |
| Pregnancy Stages Dataset | [3] | 48 | 33297 | Multiclass |

The selected methods are also quite varied, using different model representations and are either deterministic or non-deterministic. Non-deterministic methods produce different subsets of candidate biomarkers at each run for the same data [9]. To handle this variability, non-deterministic methods have been run multiple times and a consensus list of selected features has been created based on the different lists obtained across the runs. This is performed, according to the type of outcome of the feature selection algorithm, as follows:

For simple lists of features – the number of times a feature is selected across runs is assigned as a score to each selected feature as shown in equation (1) below. F represents the number of features in all in one list and L represents the number of times the method has been run or the number of lists to be averaged.

$$score(f) = \sum_{l=1}^{L} \sum_{i=1}^{F} [f_i = f] \qquad (1)$$

For lists with a score – a new score is computed based on the number of times a feature is selected across runs—occurrence(f) computed as in equation (1)—and on the median score of the feature over the multiple runs—median(f). Finally, a coefficient of 1.2 is applied to the occurrence so as to cater for cases where the median is the same and to give an advantage to features that appear more often.

$$score(f) = 1.2 * occurrence(f) + median(f) \qquad (2)$$

For ranked lists – a new score is calculated using the inverse of the median of the ranks, since the smallest rank means the best position.

$$score(f) = 1.2 * occurrence(f) + \frac{1}{1 + median(f)} \qquad (3)$$

The aim is therefore to give more importance to the features that are selected many times with a given algorithm. In case of features with the same occurrence, their rank or score is then considered. So, more-frequent features are considered as more important than those which are less frequent although ranking/scoring better.

## 2.2 Investigating the Effect of using Different Fs Methods on the Same Dataset

For the purpose of this experiment, 7 feature selection methods were run with the Golub dataset (Leukemia 1) and the first 500 selected features, regardless of their position in the features-list, have been considered for the similarity comparisons. The methods were run using default values. The aim of this step is to provide a preliminary assessment on the similarity, or dissimilarity, of the resulting feature subsets across the different methods. For non-deterministic methods, the method was run multiple times and average lists were considered.

## 2.3 Investigating the Effect of Dataset Properties on 2 Different Methods

The second part of the experiment consisted of finding out whether only the FS method and its underlying properties impact on the feature subset produced when running a specific method or whether there are other criteria that also impact on the results. The impact of the method has already been looked into in Experiment 1. Since the size of the datasets range from a few thousands to tens of thousands and the data can be binary or multiclass, it was decided to explore the effect these properties on the selected subset. Therefore 2 very different methods: RFE (Wrapper method) and LIMMA (Filter method) were run across the different datasets listed in Table 2. The aim is to find out if all datasets follow the same trend for a method and even across methods and to try to correlate the results produced by the methods with the features of the dataset. Here RFE, being non-deterministic, was run multiple times and average lists were considered.

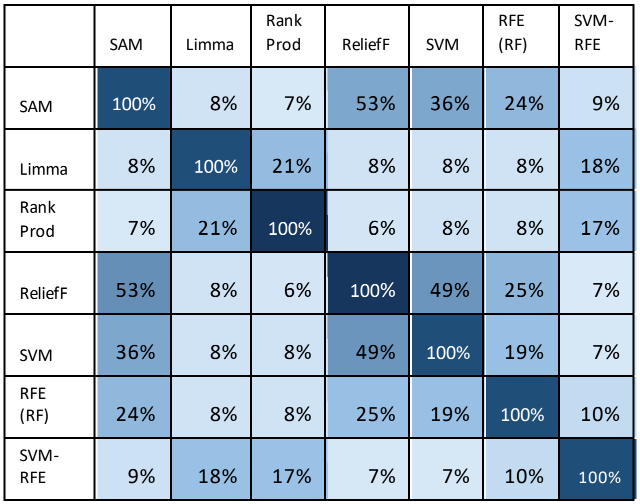
## 2.4 Investigating the Size of the Feature Subset on the Similarity Inter-methods

When performing feature selection, some methods automatically choose the optimum size of the feature subset while others require the user specify the size of the feature subset. So here, the similarity between pairs of methods as the size of the feature subset is increased has been studied. Multiple pairs of methods were selected and studied at subset sizes 20, 50, 100, 250, 500, 1000, 2500, 5000 and 7129 (full dataset). Four (4) methods having different selection management and model representation were selected. They are SAM (filter and statistical-based), ReliefF (filter and similarity-based), SVM (embedded and kernel) and RFE (wrapper and trees). For non-deterministic methods, the method was run multiple times and average lists were considered.

# 3. RESULTS

## 3.1 Investigating the Effect of using Different Fs Methods on the Same Dataset

Figure 1. shows the similarity matrix obtained when running the FS methods listed in Table 1 using the Golub (Leukemia 1) dataset. The similarity matrix shows the percentage of common biomarkers between the various pair of methods.



| | SAM | Limma | Rank Prod | ReliefF | SVM | RFE (RF) | SVM-RFE |
|---|---|---|---|---|---|---|---|
| SAM | 100% | 8% | 7% | 53% | 36% | 24% | 9% |
| Limma | 8% | 100% | 21% | 8% | 8% | 8% | 18% |
| Rank Prod | 7% | 21% | 100% | 6% | 8% | 8% | 17% |
| ReliefF | 53% | 8% | 6% | 100% | 49% | 25% | 7% |
| SVM | 36% | 8% | 8% | 49% | 100% | 19% | 7% |
| RFE (RF) | 24% | 8% | 8% | 25% | 19% | 100% | 10% |
| SVM-RFE | 9% | 18% | 17% | 7% | 7% | 10% | 100% |

**Figure 1. Similarity Matrix for subset of 500 features for the Golub (Leukemia 1) dataset.**

The results clearly show that the percentage similarity between the subset of selected features is highly variable ranging from 7% to 53% for 500 features with no 2 methods producing identical subsets or completely different subsets. This is an interesting finding since a prior preliminary research work [18] for a similar experiment but for a much smaller subset showed results that were very different. For the smaller subset selection, the methods SAM and LIMMA had produced results that were very similar and so did the methods ReliefF and SVM (percentage similarity of 60%). There were also a number of methods that produced results that were completely different. This change in behavior between subset sizes led us to investigate the behavior of methods as the size of selected subsets increases, hence experimental work 3 (Section 3.3).

It can also be noted that 4 filter methods were short-listed (Table 1) for the current experimental work and although 3 were statistical-based, they selected features that were very different from each other (Rows 1-3) of the matrix. Only methods SAM and ReliefF had more than 50% of selected features common between them and although they are both filter methods one is statistical based and the other is a similarity-based method.

## 3.2 Investigating the Effect of Dataset Properties on 2 Different Methods

Figure 2 shows the percentage similarity between the methods LIMMA and RFE for 9 datasets. To investigate the effect of the datasets, if any, on the subset of features, the methods RFE and LIMMA have been run using 9 datasets having diverse properties.

In the current work, 2 methods having different properties were selected for investigating the impact of the dataset on the selected feature subset. In a prior work [18], 2 methods having similar properties (SAM and LIMMA) had already been investigated and results had shown that multiclass problems seem to favor bigger similarity than binary problems. It was also found that 75% of multiclass datasets led to high correlations (>50%) while only 20% of binary datasets did the same.

One may expect to find that such similarity behaves in the same manner for all gene expression datasets and our results show that this is not necessarily the case. As shown below, as the size of the selected feature subset increases the percentage similarity tends to increase but while for some datasets it increases almost regularly (i.e., MELAS and 9-cancers), for others the similarity increases only after 1000 features (i.e., colitis, prostate, and Leukemia 2). For 2 cases (i.e., Leukemia 1 and Leukemia 3), the lists produced by the 2 methods differ almost completely for small as well as for large subsets. We may also notice from these results that the similarity between these methods is dependent on the target dataset.

## 3.3 Investigating the Size of the Feature Subset on the Similarity Inter-methods

The last part of the experiment was to investigate how the percentage similarity between the methods SAM, SVM, ReliefF and RFE evolves as the subset of features increases. For this experimental work, Leukemia 1 (Golub) dataset was used since it was found in the previous experimental work that the list of features differed completely for both small and large feature subsets.

From the graph of Figure 3, it can be seen that the percentage similarity fluctuates strongly for few features while it increases almost regularly with the number of features in mid-range. Quite logically, the increase of the similarity is steeper when the number of features approaches 100% of the size.

It can also be noted that the methods SAM and ReliefF which are both filter methods shared the highest similarity.

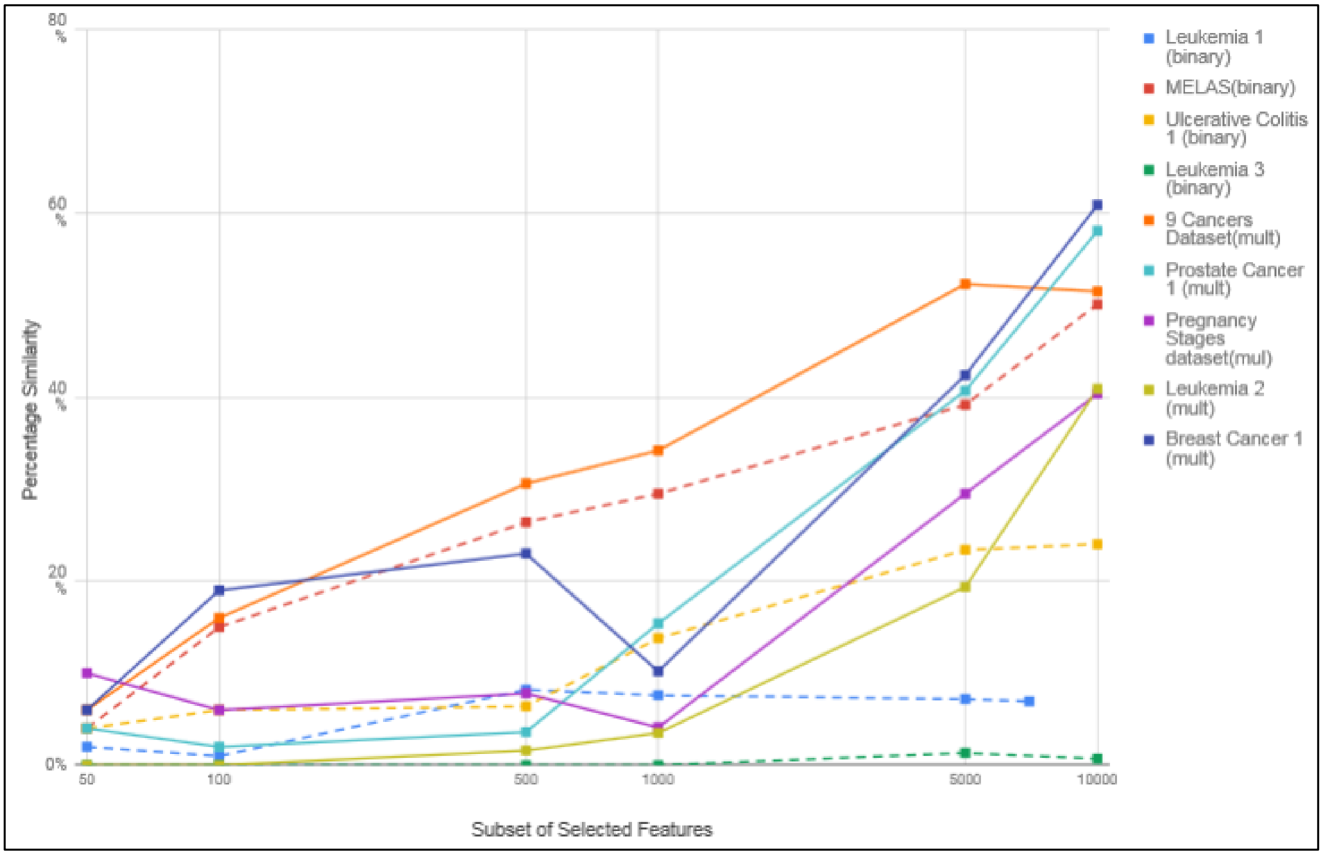


**Figure 2. Comparison of % common features between RFE and LIMMA for 9 different datasets.**

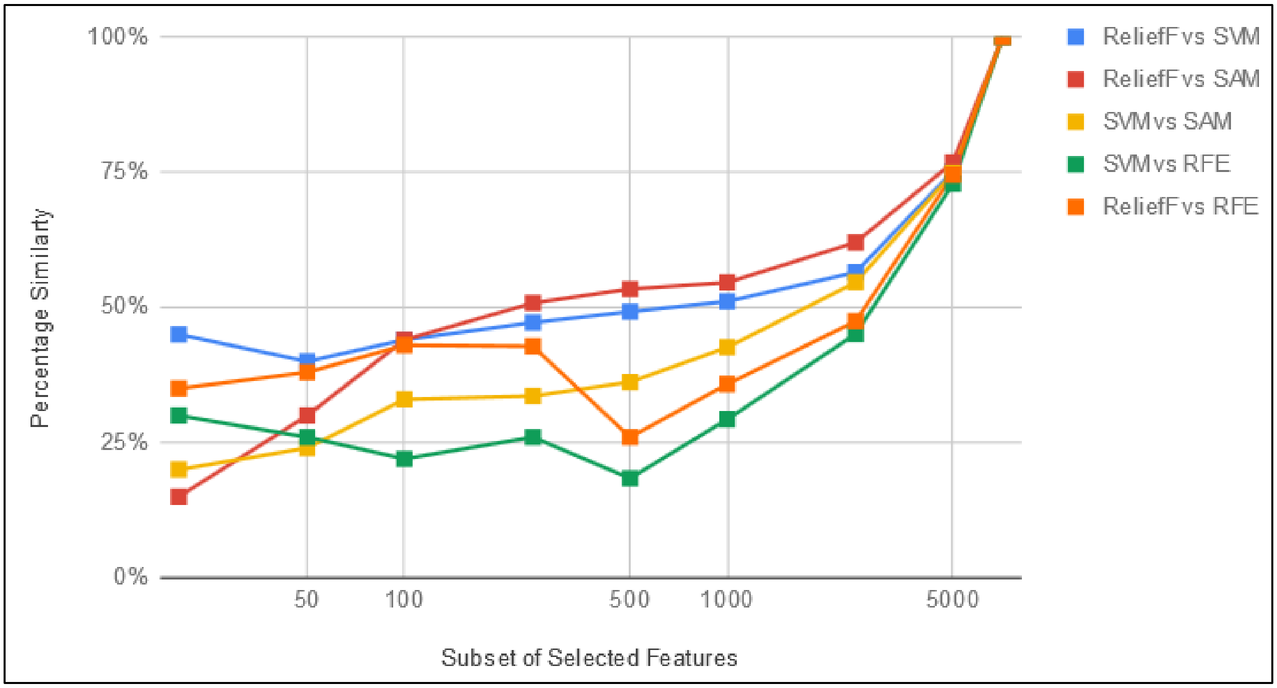


**Figure 3. % Similarity between SVM, ReliefF, SAM and RFE as Feature Subset Increases from 20 to 7129 (Whole Dataset) for the Golub (Leukemia 1) Dataset.**

## 4. CONCLUSION

The results obtained confirm that indeed the FS selected by different FS methods can be very different.

Based on the above observations, a number of questions need to be looked into: firstly whether the feature subsets obtained with a single method are reliable given the high variability when using different methods in experiment 1, secondly whether each method introduces its own bias and finally whether there is a dependency on the dataset, given the high variability when applied to different datasets.

The final objective of the global research being carried is to define how to select FS methods and therefore the current work is ongoing. This work does not provide a definitive guide but we can provide some recommendations: 1) FS methods show some stability between methods and even between datasets around 500 features and 2) similar methods (based on the same selection management and model representation) tend to converge very quickly 3) FS should be applied parsimoniously, i.e too few features should not be selected in a single step, to increase the robustness 4) FS should be performed on the base of several methods instead of a single one. To do this, we should need some kind of "robustness" metrics as those presented in the introduction as well as some kind of list-merging strategy allowing to select the most robust subset.

## 5. ACKNOWLEDGMENTS

None.

## 6. REFERENCES

[1] Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P. and Saeys, Y. 2010. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*. 26, 3 (2010), 392–398. DOI:https://doi.org/10.1093/bioinformatics/btp630.

[2] Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P. and Saeys, Y. 2009. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*. 26, 3 (2009), 392–398. DOI:https://doi.org/10.1093/bioinformatics/btp630.

[3] ArrayExpress < EMBL-EBI: *https://www.ebi.ac.uk/arrayexpress/*. Accessed: 2019-10-09.

[4] Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, a., Beníez, J.M. and Herrera, F. 2014. A review of microarray datasets and applied feature selection methods. *Information Sciences*. 282, (2014), 111–135. DOI:https://doi.org/10.1016/j.ins.2014.05.042.

[5] Breitling, R., Armengaud, P., Amtmann, A. and Herzyk, P. 2004. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*. 573, 1–3 (Aug. 2004), 83–92. DOI:https://doi.org/10.1016/j.febslet.2004.07.055.

[6] Cancer Program Legacy Publication Resources: *http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi*. Accessed: 2019-10-09.

[7] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*. 46, (2002). DOI:https://doi.org/10.1023/A:1012487302797.

[8] He, Z. and Yu, W. 2010. Stable feature selection for biomarker discovery. *Computational biology and chemistry*. 34, 4 (2010), 215–25. DOI:https://doi.org/10.1016/j.compbiolchem.2010.07.002.

[9] Hira, Z.M. and Gillies, D.F. 2015. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in Bioinformatics*. 2015, 1 (2015).

DOI:https://doi.org/http://dx.doi.org/10.1155/2015/198363 Review.

[10] Home - GEO - NCBI: *https://www.ncbi.nlm.nih.gov/geo/*. Accessed: 2019-10-09.

[11] Kalousis, A., Prados, J. and Hilario, M. 2007. Stability of Feature Selection Algorithms: A Study on High-dimensional Spaces. *Knowl. Inf. Syst.* 12, 1 (May 2007), 95–116. DOI:https://doi.org/10.1007/s10115-006-0040-8.

[12] Kira, K. and Rendell, L. 1992. The feature selection problem: Traditional methods and a new algorithm. *Aaai*. (1992), 129–134. DOI:https://doi.org/10.1016/S0031-3203(01)00046-2.

[13] Kira, K. and Rendell, L.A. 1992. A Practical Approach to Feature Selection. *Proceedings of the Ninth International Workshop on Machine Learning {(ML} 1992), Aberdeen, Scotland, UK, July 1-3, 1992* (1992), 249–256.

[14] Kononenko, I. 1994. Estimating attributes: analysis and extensions of relief. *Machine Learning: ECML-94*. 171–182.

[15] Kuhn, M. 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, Articles*. 28, 5 (2008), 1–26. DOI:https://doi.org/10.18637/jss.v028.i05.

[16] Kuncheva, L.I. 2007. A stability index for feature selection. *International Multi-conference: artificial intelligence and applications*. (2007), 390–395.

[17] Kursa, M.B. 2014. Robustness of Random Forest-based gene selection methods. *BMC Bioinformatics*. 15, 1 (2014), 8. DOI:https://doi.org/10.1186/1471-2105-15-8.

[18] Mungloo-Dilmohamud, Z., Marigliano, G., Jaufeerally-Fakim, Y. and Peña-reyes, C. 2018. A Comparative Study of Feature Selection Methods for Biomarker Discovery. (2018), 2789–2791.

[19] Mungloo-Dilmohamud, Z., Jaufeerally-Fakim, Y. and Peña-Reyes, C. 2017. *A meta-review of feature selection techniques in the context of microarray data*.

[20] Padmanaban, S., Baker, J. and Greger, B. 2018. Feature Selection Methods for Robust Decoding of Finger Movements in a Non-human Primate. 12, February (2018), 1–15. DOI:https://doi.org/10.3389/fnins.2018.00022.

[21] Parmar, C., Grossmann, P., Bussink, J., Lambin, P. and Aerts, H.J.W.L. 2015. Machine Learning methods for Quantitative Radiomic Biomarkers. *Scientific reports*. 5, (Aug. 2015), 13087. DOI:https://doi.org/10.1038/srep13087.

[22] Sarkar, C., Cooley, S. and Srivastava, J. 2015. Robust Feature Selection Technique using Rank Aggregation. *Applied Artificial Intelligence*. 28, 3 (2015), 243–257. DOI:https://doi.org/10.1080/08839514.2014.883903.Robust.

[23] Smyth, G.K. 2004. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*. 3, 1 (2004), 1–25. DOI:https://doi.org/10.2202/1544-6115.1027.

[24] Tusher, V.G., Tibshirani, R. and Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*. 98, 9 (2001), 5116–21. DOI:https://doi.org/10.1073/pnas.091062498.

[25] Venkatesh, B. and Anuradha, J. 2019. A Review of Feature Selection and Its Methods. 19, 1 (2019), 3–26. DOI:https://doi.org/10.2478/cait-2019-0001.

[26] Wang, H. 2011. Measuring robustness of feature selection techniques on software engineering datasets. *Information Reuse and …*. (2011), 309–314. DOI:https://doi.org/10.1109/IRI.2011.6009565.

[27] Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T. and Vapnik, V. 2000. Feature Selection for SVMs. *NIPS* (2000).

[28] Witten, I.H., Frank, E. and Hall, M.A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc.

[29] Yu, Y. 2008. SVM-RFE Algorithm for Gene Feature Selection. (2008).