Variable Selection for Multiclass Imbalanced Microarray Data using Random Forests Quantile Classifier
BST 640 Final Project

Wei Zhang

November 29, 2020

Abstract

Variable selection for classification of cancer microarray dataset is challenging for the big p and small n data settings and the multi-class imbalanced data problem. This project proposes a variable selection method that can handle the multi-class imbalanced microarray datasets using a random forest quantile classifier (RFQ). We first derive the multi-class RFQ and embed it into the backward elimination selection framework (RFQ-VS). Two criteria are used in the variable selection. Two cancer datasets application, Leukemia and Brain, demonstrate the method and show good performance in both class error and overall error with small-sized genes. We also show that RFQ-VS has comparable performance to other random forest variable selection methods.

1 Introduction

Variable selection for microarry data is a common task in classification of cancer type. Several challenges are addressed when facing the classification of cancer microarray datasets:

1. The $p \gg n$ problem. Typically, the number of genes, p, is much larger than the number of microarray observations, n. There could be thousands of the genes with less than a hundred of samples in a microarray dataset.

2. The multi-class imbalanced data problem. Cancers like leukemia or brain cancer have multiple subtypes with extremely low sample sizes of some particular subtypes.

There are methods in dealing with each of the above problems using random forest. For the first problem, Diaz-Uriarte and Alvarez de Andre [**1**] proposed a backward elimination strategy with random forest. Ishwaran et al. [**2**] described a variable selection method based on the minimal depth of trees. Recently, Genuer et al. created a R package, `VSURF`, selecting variables with a random forest stepwise forward strategy [**3**]. For the second problem, data-level methods such as balanced random forest (BRF) [**4**] downsampling the majority class to the minority class. In the algorithmic level methods, a random forest quantile classifier was proposed by O'Brien and Ishwaran [**5**]. The methods for directly handling variable selection in the multi-class imbalanced data settings are still worth to explore.

This project aims to explore the variable selection methods with multi-class imbalanced data. Limited by the time of the project, currently our main approach is implementing the multi-class random forest quantile classifier with the backward elimination variable selection method. Section 2 describes the implementation of the combined method. Two microarray datasets are selected to apply the combined method in section 3. Moreover, we conduct comparisons on the performance of the combined method and other variable selection methods with the two datasets (Section 4). Section 5 gives a discussion of our findings. The `R` package mainly used for this project is the `randomForestSRC` package [**6-8**].

2 Methodology

In this section, we describe the implementation of the combined method. For RFQ, the original method was conducted in the two-class imbalanced data setting. Therefore, we first implement RFQ into a multi-class data setting. Then for the variable selection, we use the backward elimination framework.

2.1 Multi-class RFQ

In the paper of the RFQ method, the authors mentioned that the RFQ could perform multi-class classification with Friedman's one-vs-one approach [**5**, **9**]. The RFQ is applied to 2 of the K classes data for $\binom{K}{2}$ times. For every two-class in the K classes, the q^*-classifer is:

$$\delta_{q^\star}^{(kl)}(x) = \mathbf{1}_{\{\frac{p_k^{(kl)}(x)}{\pi_k^{kl}} \geq \frac{p_l^{(kl)}(x)}{\pi_l^{(kl)}}\}}$$

where

- $p_k^{(kl)}(x) = P(Y = k|X = x)$, $p_l^{(kl)}(x) = P(Y = l|X = x)$, and $p_k^{(kl)}(x) + p_l^{(kl)}(x) = 1$
- $\pi_k^{(kl)} = P(Y = k)$, $\pi_l^{(kl)} = P(Y = l)$, and $\pi_k^{(kl)} + \pi_l^{(kl)} = 1$

The one-vs-one K classifier then becomes:

$$k_{q^\star}(x) = \arg \max_{1 \leq k \leq K} \sum_{l=1}^{k} \mathbf{1}_{\{\frac{p_k^{(kl)}(x)}{\pi_k^{kl}} \geq \frac{p_l^{(kl)}(x)}{\pi_l^{(kl)}}\}}$$

Since the two-class RFQ algorithm is already implemented in the function `imbalanced` to the `randomForestSRC` package in R, we directly apply the multi-class RFQ approach to the function `imbalanced`.

Algorithm 1 Multi-class RFQ

1: Let N = $\binom{K}{2}$
2: Split the K-class dataset into N 2-class datasets
3: **for** $k = 1$ to N **do**
4: Fit RFQ to the k_{th} dataset
5: Add the k_{th} votes to the classes in the k_{th} dataset
6: **end for**
7: Classify x to the majority vote class

2.2 Variable Selection Framework

For the variable selection framework, we choose to modify the backward elimination for its relevantly simple implementation. Like the original proposed by Diaz-Uriarte [**1**], we iteratively fit random forests, drop 20% of the ranked variables each iteration, and select the variable set that generates the smallest OOB error. The variable importance only calculates at the first iteration and rank for later use. The differences in our algorithm are the following:

1. We calculate the variable importance for the first forest using the original K-class data and drop all variables with importance smaller than 0. Our first thought of the variable importance was calculated directly by the RFQ. However, the variable importance would generate all 0 for large number of variables. Therefore, we replace it to the original random forest.

2. Instead of using the original random forest algorithm, we implement the multi-class RFQ in each iteration after the first iteration.

3. Instead of only selecting the smallest OOB error, we develop another selection strategy by selecting the smallest maximum OOB error of the multi-class OOB error (minimax OOB error). We want to see if the minimax OOB error could prevent selecting a variable set with a low overall error but high in class errors for some of the classes.

4. We do not implement any parameter so far except for the number of trees.

5. All the random forest implementaions are using the `R` package `randomForestSRC`.

We call this algorithm as RFQ-VS:

Algorithm 2 RFQ-VS

Fit random forest and calculate variable importance (VIMP)
Drop variables with VIMP < 0
Select remaining variables to set V
Let $l = \text{length}(V)$
while $l > 1$ **do**
    Fit Multi-class RFQ with variables in V
    Record the overall/max OOB error
    $l = l \times (1 - \text{dropout})$
    $V =$ first l variables in V
end while
Select the set V with the smallest overall/max OOB error

3 Data Experimental Results

The datasets we selected to use in our project are Brain cancer [**11**] and Leukemia [**12**] (See Table 1). In each dataset, the number of genes (p) is hundreds of patients (n). Moreover, as Figure 1 shows, both two datasets suffered from imbalanced data problem. The smallest minority type of the Leukemia (`B-CELL_ALL_MLL`) only contained 17 patients. Only 13 patients in the smallest type of Brain cancer (`normal`). Moreover, in the PCA plot (Figure 2) of Leukemia, the seven types are messed up and not showing any separation in the first two components. Although brain cancer in the PCA plot is more separating by class than in the Leukemia plot, the imbalanced data problem still exists.

This section applies the RFQ-VS method to these two datasets and sees whether this method could select appropriate sets of variables and maintain high predictive accuracy.

3.1 Overall Performance

First, we conducted 3-fold cross-validation for the following models: 1) Random forest without variable selection; 2) RFQ-VS with the smallest overall OOB error strategy; 3) RFQ-VS with the smallest maximum OOB error strategy. Since the method would return the multi-class RFQ with the selected variables, we first compared the results of the selected multi-class RFQ model.

Table 1: Information of the microarray datasets

Dataset	Genes	Patients	Types	Ref.
Brain	20174	130	5	[**11**]
Leukemia	12402	281	7	[**12**]

Figure 1: Counts for Different Types of Two Datasets

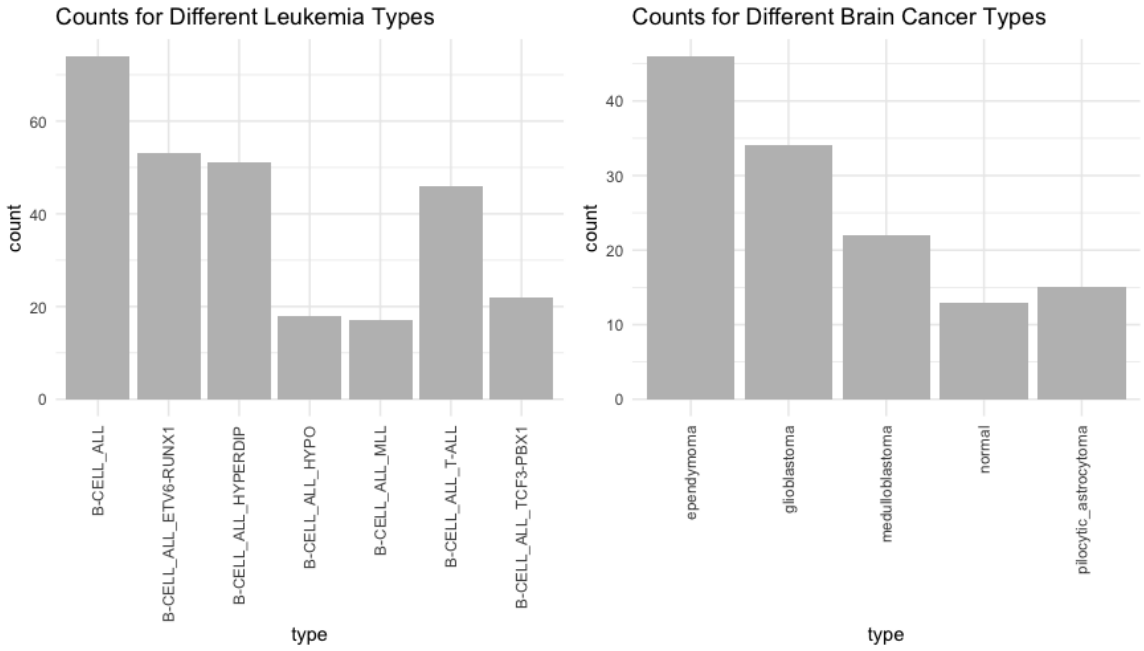


Figure 2: PCA Plots of First Two Components
Left: Leukemia | Right: Brain

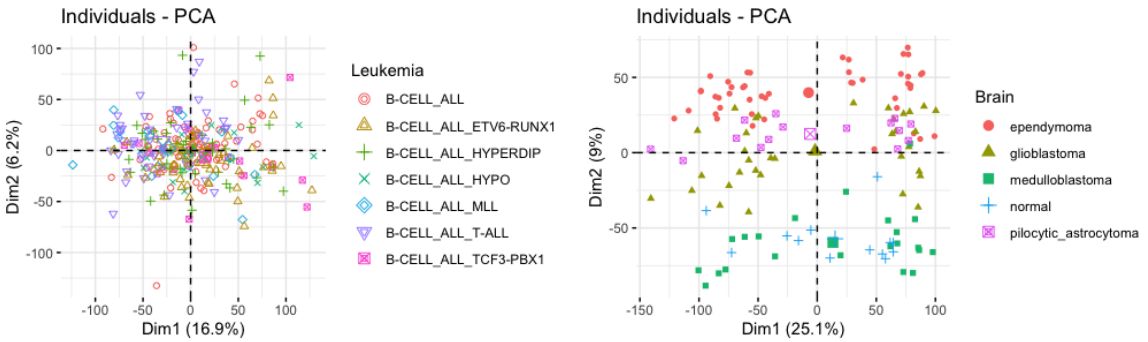


Table 2 shows the overall 3-fold cross-validation error of two datasets with different `ntree` parameters. All other parameters are default settings. Number of selected variables was recorded in each fold from `ntree` = 100, 500, 1000 and 2000. The rightmost column of each table recorded the selection range with the median. Both RFQ-VS methods maintained good performances on the two datasets compared with the original random forest without variable selection. However, the

two methods seem not very sensitive to the number of trees. The cv errors of each method were closed in a different number of trees.

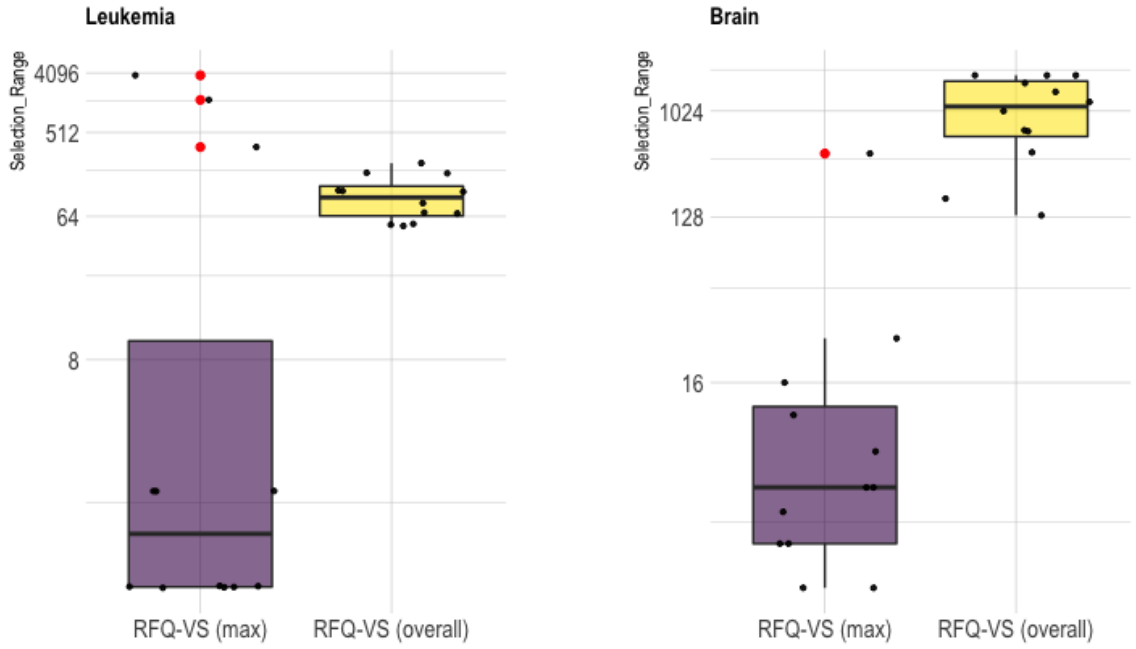
Table 2: Three-fold Cross Validation Error
Left: Leukemia | Right: Brain

Model	100	500	1000	2000	Selection Range
no selection	0.174	0.174	0.167	0.167	20174
overall	0.189	0.192	0.192	0.192	96 (53,216)
max	0.167	0.196	0.195	0.187	3 (3,3716)

Model	100	500	1000	2000	Selection Range
no selection	0.084	0.068	0.061	0.053	12402
overall	0.000	0.000	0.007	0.007	1050 (131,2498)
max	0.023	0.023	0.038	0.015	6 (4,406)

In most cases, minimax RFQ-VS has selected a smaller size of variables than the overall RFQ-VS. However, in both methods, the variation in the size of the selection was large. In the Leukemia dataset, the variables selected by minimax RFQ-VS were varied from 3 to 3,716. And the number of variables selected in Brain datasets with overall RFQ-VS were varied from 131 to 2,498. The boxplots in Figure 3 showed the selection sizes varied from the fold to fold in both datasets.

Figure 3: Selection Range of two RFQ-VS Strategies



3.2 In-class Performance

Next, we compared the overall in-class performance. We selected the models with `ntree` = 1000 and compared the class error for each type of two datasets. In addition, we tested the selected variables on the original random forest model whether the variables selected by RFQ-VS

are reusable for other models.

Since the Leukemia dataset is complicated, and the number of genes is larger than the Brain dataset, the classification of Leukemia data is much more difficult than Brain data. Two types of Leukemia (`B-CELL_ALL_MIL` and `B-CELL_ALL_HYPO`) have only less than 20 observations, and the classification error of `B-CELL_ALL_HYPO` is 1 if using RF without variable selection (See Table 3). However, both RFQ-VS selection methods decreased their class error to 0. Except for the type `B-CELL_ALL_TCF3-PBX1`, other types of Leukemia were well predicted by both RFQ-VS methods. Moreover, the original RF with the variables selected by RFQ-VS are also maintained good performance.

Table 3: Three-fold Cross Validation Class Error for Leukemia

	B-CELL	ETV6-RUNX1	HYPERDIP	HYPO	MLL	T-ALL	TCF3-PBX1	Overall Error
RF (no selection)	0.326	0	0.144	1	0	0.039	0.132	0.167
RFQ-VS (overall)	0.026	0	0	0	0	0	0.702	0.192
RFQ-VS (max)	0.116	0	0	0	0	0	0.681	0.196
RFQ-VS (overall) + RF	0.284	0	0.113	1	0.005	0.039	0.128	0.150
RFQ-VS (max) + RF	0.327	0.042	0.137	1	0	0.021	0.131	0.163

Table 4: Three-fold Cross Validation Class Error for Brain

	ependymoma	glioblastoma	medulloblastoma	normal	astrocytoma	Overall Error
RF (no selection)	0	0.15	0	0.056	0.067	0.061
RFQ-VS (overall)	0	0.026	0	0	0	0.007
RFQ-VS (max)	0	0.053	0	0	0	0.016
RFQ-VS (overall) + RF	0.039	0.163	0	0.056	0.067	0.077
RFQ-VS (max) + RF	0.065	0.245	0137	0.189	0.083	0.139

For Brain data (Table 4), using both RFQ-VS selection methods are well predicted all different types of Brain cancer. The minimax RFQ-VS had higher cv errors but with smaller size of variable sets. RF model with the overall RFQ-VS maintained a good performance, but the RF model with variable selected by minimax RFQ-VS was not good as other models.

4 Comparisons with Other Variable Selection Methods

Next, we compare the performance of our `RFQ-VS` with the existing random forest variable selection methods. We included the original random forest backward elimination method `varSelRF` [**10**], and the random forest minimal depth variable selection in the ultra-high dimensional settings `RFS-VH` [**2**]. The forward stepwise random forest selection method `VSURF` [**3**] was excluded at this moment for its time consumption. In all methods, we used the default settings with `ntree` = 1000. The 10-fold cross validation was applied to each methods.

Table 5 and Table 6 show that all selection methods, except VarSelRF, maintained good prediction performance in the 10-fold cv error. In the Brain data, both RFQ-VS (max) and RFQ-VS

(overall) had the lowest class and overall cv errors. Moreover, both methods had smaller selection ranges in only (2, 25) with RFQ-VS (max) and (24, 550) with RFQ-VS (overall).

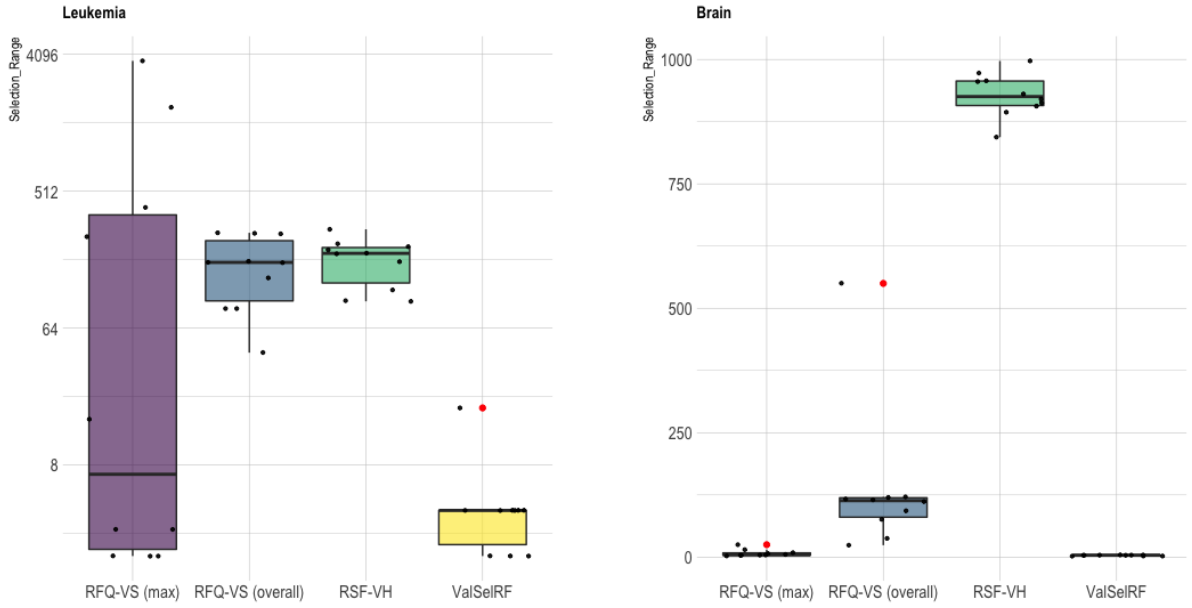
Table 5: Ten-fold Cross Validation Error for Leukemia

	B-CELL	ETV6-RUNX1	HYPERDIP	HYPO	MLL	T-ALL	TCF3-PBX1	Overall Error	Selecion Range
RF (no selection)	0.195	0.050	0.1	1	0.250	0	0.020	0.156	20174
RSF-VH	0.207	0	0.140	1	0.300	0	0.020	0.163	199 (97,287)
VarSelRF	0.58	0.75	0.65	0.95	0.65	0.225	0.227	0.514	4 (2,19)
RFQ-VS (max)	0.076	0	0	0	0	0	0.694	0.192	10 (2,3710)
RFQ-VS (overall)	0	0	0	0	0	0	0.698	0.181	174 (44, 272)

Table 6: Ten-fold Cross Validation Error for Brain

	ependymoma	glioblastoma	medulloblastoma	normal	astrocytoma	Overall Error	Selection Range
RF (no selection)	0.045	0.033	0.100	0	0.150	0.052	12402
RSF-VH	0.065	0.058	0.050	0	0.15	0.059	925 (844, 997)
VarSelRF	0.110	0.350	0.717	0.7	0.9	0.414	2 (2,4)
RFQ-VS (max)	0	0.025	0	0	0	0.008	5 (2,25)
RFQ-VS (overall)	0	0.020	0	0	0	0.007	114 (24,550)

Figure 4: Selection Range of Four Variable Selection Methods
Left: Leukemia | Right: Brain



In the more complicated dataset Leukemia, although the overall cv error was well maintained, the selection range of RFQ-VS (max) was not stable as other methods. However, the selection range of RFQ-VS (overall) was more stable. RSF-VH had the best overall cv error with a stable

range of gene selection. However, the performance of one of the smallest size class `B-CELL_ALL_HYPO` in Leukemia was not well predicted. ValSelRF had the smallest variable selection range among all methods, but its performance with selected genes was the worst on both two datasets.

5 Discussion

We have presented a possible variable selection method that embedded random forest quantile classifiers into the backward selection method. Two microarray datasets were applied with the RFQ-VS method and gained good performance. In addition, RFQ-VS (max) tends to select fewer variables, but the selection is not stable when applying to more complicated datasets. In contrast, RFQ-VS (overall) tends to select more variables with a more stable selection range. Still, there is much room for improvement. The method itself still needs many improvements. For example:

1. For the RFQ-VS (max) method, we need to apply some constraints or variable control settings in the model.

2. Parameters tuning and other evaluation criteria should be applied in the future data experiment.

3. More datasets should be experimented in the future.

Moreover, there are several drawbacks with variable selection based on VIMP: 1) VIMP is tied to the type of prediction error used; 2) developing formal regularization for VIMP based methods are very challenging due to its randomness [**2**]. Therefore, applying RFQ to other variable selection methods are much worth to explore. We will enhance the RFQ-VS method and explore the other method that can be applied to the multi-class imbalanced data in future work.

References

[1] Díaz-Uriarte, R., Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7, 3.

[2] Ishwaran, Kogalur, U. B., Gorodeski, E, Minn, A, & Lauer, M (2010). High-dimensional variable selection for survival data, *Journal of the American Statistical Association* 105 (489), 205–217

[3] Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2015). VSURF: An R package for variable selection using random forests. *R Journal*, 7(2), 19–33.

[4] Chen, C., Liaw, A., & Brieman, L. (2004). Using random forest to learn imbalanced data: Technical Report No. 666. University of California, Berkley. Using Random Forest to Learn Imbalanced Data, 110(1–12), 12

[5] O'Brien, R., & Ishwaran, H. (2019). A random forests quantile classifier for class imbalanced data. *Pattern Recognition*, 90, 232–249. https://doi.org/10.1016/j.patcog.2019.01.036

[6] Ishwaran, H. and Kogalur, U.B. (2020). Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC), *R package* version 2.9.3.

[7] Ishwaran, H. and Kogalur, U.B. (2007). Random survival forests for R. *R News* 7(2), 25–31.

[8] Ishwaran, H., Kogalur, U.B., Blackstone E.H. and Lauer M.S. (2008). Random survival forests. *Ann. Appl. Statist.* 2(3), 841–860.

[9] Friedman, J. H. (1996). *Another approach to polychotomous classification* (). Department of Statistics, Stanford University .

[10] Diaz-Uriarte, R. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest 2007. *BMC Bioinformatics*, 8, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-328.

[11] Griesinger, A.M., Birks, D.K., Donson, A.M., Amani, V. et al (2013). Characterization of distinct immunophenotypes across pediatric brain tumor types. *J Immunol.* 191(9):4880-8. PMID: 24078694

[12] Coustan-Smith, E., Song, G., Clark, C., Key, L. et al (2011). New markers for minimal residual disease detection in acute lymphoblastic leukemia. *Blood.* 117(23):6267-76. PMID: 21487112