# Usage of Random Forest machine learning method for classifying gene expression of brain cells

Erik Berg Siebert, Prof. Dr. Edson Emílio Scalabrin
*Pontifical Catholic University of Paraná*
Curitiba, Brazil
eriksie.ebs@gmail.com, scalabrin@ppgia.pucpr.br

***Abstract*—Characterization of gene expressions is a sub-area of health science that has been a constant target by machine learning specialists due to the complexity of its problem. Identifying differences between healthy and ill tissues lead to disease signatures that can help the creation of treatments. The high amount of attributes in gene expression samples brings the curse of dimensionality which causes the degradation to the prediction accuracy of a classifier. We propose the use of Random Forest machine learning method to reduce the effects of the curse of dimensionality using Random Forest for both feature selection method and classification method together with decomposition methods. We also create a simple framework to help us more easily run experiments to evaluate different configuration for the classifier.**

***Keywords—Random forest, high dimensional data classification, decomposition methods, ensemble methods, feature selection, curse of dimensionality.***

## I. Introduction

With the development and the more frequent adoption of machine learning techniques, the field of computational intelligence has gotten close to health science. One of the sub-areas of health science that has been a constant target by machine learning specialists, due to the complexity of its problem, is the characterization of gene expressions. Gene expression is a process in which information from a gene is synthesized into a functional gene, often producing a protein. Previously, the presence of diseases in patience's organism was determined by DNA analysis from extracted tissue or blood samples from the inspected area, but since the technological advances of sequencing, gene expression samples have become the most suitable and affordable option for being a cheaper and faster solution.

Due to the high amount of data of sequenced genomes, these samples became an important analysis tool for solving biological questions, such as classifying tissues. This classification could help identify differences between healthy and ill tissues leading to a disease signature that can help the creation of treatments.

High-throughput data sets, such as gene expressions, suffer from a phenomenon known as *curse of dimensionality* [1] where the increased amount of attributes, followed by low quantities of samples [2], cause degradation to the prediction accuracy of a classifier [3]. Unfortunately, classical prediction methods struggle to classify this data [4], so in order to make the process of classification more treatable, methods need to be adapted to this type of data.

Using dimensionality reduction and decomposition methods, this research proposes the evaluation of different configurations for the Random Forest ensemble method to reduce the effects of the *curse of dimensionality* in the classification of tissues by means of the gene expression to help on the identification of biomarkers of tissues and diseases of brain cells since the method is known to have a good response to data of this nature [5].

In order to automate the evaluation procedure, a simple framework was also developed.

## II. Methods

To deal with the effects of the *curse of dimensionality*, the main classification problem will be broken down into smaller problems and deal with the effects of the *curse of dimensionality* on each of these problems. For that, two decomposition methods, *One-vs-One* and *One-vs-All*, will be implemented and evaluated.

The classification and dimensionality reduction procedures will be approached with Random Forest. These procedures will occur separately, meaning dimensionality reduction happens before than classification. The experimentation process will be more thoroughly described in Section IV.

### A. Random Forest

Random forest [6] is an ensemble learning method that is composed of a multitude of decision trees using averaging to improve the prediction accuracy and fighting over-fitting habits of decision trees. In random forest, each decision tree is built from a sample drawn with replacement from the training set. During the construction of a tree, the node that is chosen to be split is not the best split among all features, but instead is the best split in a random subset of the features. Random forest algorithm can be summarized as below:

Assuming N number of training samples with a feature space of size M and also assuming that random *m* number of features smaller than M is used for each split. Each tree is constructed as follows:

- Choose a random subset of samples out of the original training set as the training set. The rest of the samples are used as the testing set to estimate the error of the tree.
- The tree is fully grown without pruning based on the best split of each tree node with a random subset of features.

When running predictions, the unknown sample is pushed down all trees and a majority vote is casted as the final prediction.

This randomness causes an increase on the bias but a decrease in variance which overcompensates for the bias increase, yielding an overall better model. Although its mechanism appears simple, it relies on many mathematical properties which remain unknown to date.

### B. Decomposition Methods

Many of classification problems involve more than two classes, called multi-class problems. Usually, it is easier to build classifications for two classes since the decision boundaries in this case can be simpler than a multi-class problem. Therefore, decomposition methods, also known as binarization techniques, are used to break down a multi-class problem into multiple two-class problems. The most common, and the ones picked to be implemented in this research, are called "one-vs-one" and "one-vs-all".

- *One-vs-One* divides the problem in all possible combinations between pairs of classes.
- *One-vs-All* divides the problem in such a way in which each class is distinguished from all remaining classes.

As this proposal creates multiple classification problems, an ensemble needs to be created and a voting rule needs to be defined in order to combine the outputs of each classifier.

### C. Dimensionality Reduction

Dimensionality reduction techniques, done by feature selection methods, have the goal of finding an optimal set of features that should ideally possess the following characteristics [7]:

- The quantity of features of the subset should be minimal and sufficient to accurately predict the class of unknown samples,
- The subset of features should improve the prediction accuracy of the classifier run on data containing only these features rather than on the original dataset with all the features,
- The resulting class distribution, given only the values for the selected features, is as close as possible to the original class distribution given all feature values.

Random forest has shown advantages on handling high dimensional data and assigning importance scores to features on gene expression data [8] and therefore has been chosen as the feature selection method in this research. In order to find the right amount of features that suffices the conditions of a feature selection method, a simple framework written in *Python* using the machine learning library *sklearn* has been developed and will be described on Section VI.

## III. Dataset

The dataset used for this experiment contains 109 samples with 65,989 features split into 7 classes. Each class is part of a specific brain region and has a specific condition. It is distributed as shown in Table 1.

TABLE I.

| ***Class ID*** | ***Brain region*** | ***Condition*** | ***Train/Test set size*** |
|---|---|---|---|
| 0 | DLPFC | Bipolar Disorder | 23/4 |
| 1 | DLPFC | Control | 25/6 |
| 2 | PolyA | DM1 | 5/1 |
| 3 | PolyA | DM2 | 5/1 |
| 4 | PolyA | Control | 5/1 |
| 5 | BA9 | Parkinson's | 14/3 |
| 6 | BA9 | Control | 13/3 |

Classes with the *control* condition are tissues that are not affected by any disease. The test set size has been based on the class with the lowest amount of samples.

## IV. Framework

In order to evaluate different configurations of Random forest for classification and feature selection, and the decomposition methods, a simple framework was developed in *Python* using the machine learning library *sklearn*. The framework follows the steps bellow:

*1)* Read the configuration file which holds important information used throghout the experiment, such as decomposition method, number of classes, dimensionality reduction rules and set file paths;

*2)* Configure the pair of classes and dataset for each classification based on the decomposition method;

*3)* For each classification problem:

*a)* run Random forest and list columns in descendent order of importance;

*b)* run Random forest again using only the first *n* most important columns and iterate, with step *s*, until *m* most important columns is reached;

*c)* on each iteration, test for model precision and keep track of the model with the best f1 score;

*d)* return the list of columns that gave the best f1 score;

*4)* Run prediction on the training set;

*a)* On *one-vs-one* experiments, the output of the ensemble is the class which had the majority of votes of all classifiers.

*b)* On *one-vs-all* experiments, the output of the ensemble is the class which had the highest probability of being the correct according to the probability given by all classifiers.

*5)* Plot results such as precision score of each classifier; amount of columns used by each classifier; and the confusion matrix according to the brain region and condition of the sample; and a confusion matrix according to the brain region alone.

## V. Results

Before we show the differences of using Random Forest as feature analysis tool, the framework has been configured to include all columns in the creation of the ensemble. According to the output of the framework, the results of each model are shown in Fig. 1 for decomposition method *one-vs-one* and Fig. 2 for decomposition method *one-vs-all*.

Now, when the framework is configured to evaluate the F1 precision scores on the models, limiting to 5 to 50 columns, we see big differences when compared with the

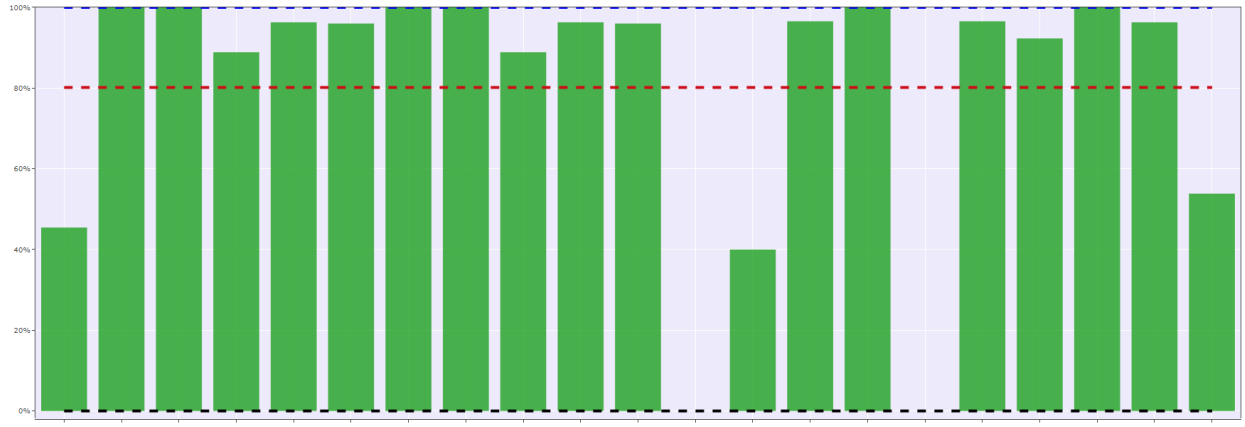


Fig. 4. F1 precision scores for each model built with decomposition method *one-vs-one*. Highest score of 100%, lowest of 0% with a mean F1 score of 80%.

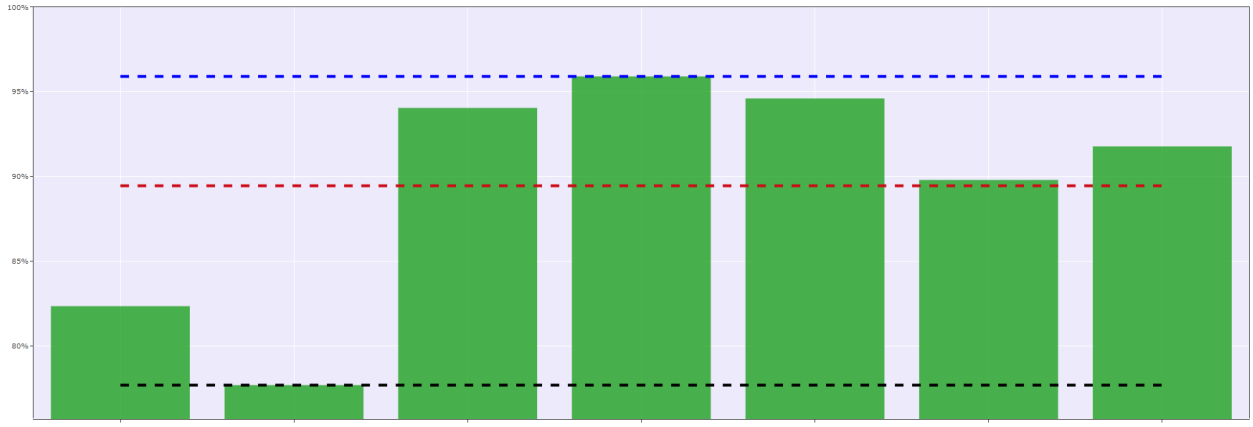


Fig. 3. F1 precision scores for each model built with decomposition method *one-vs-all*. Highest score of 96%, lowest of 78% with a mean F1 score of 89%.

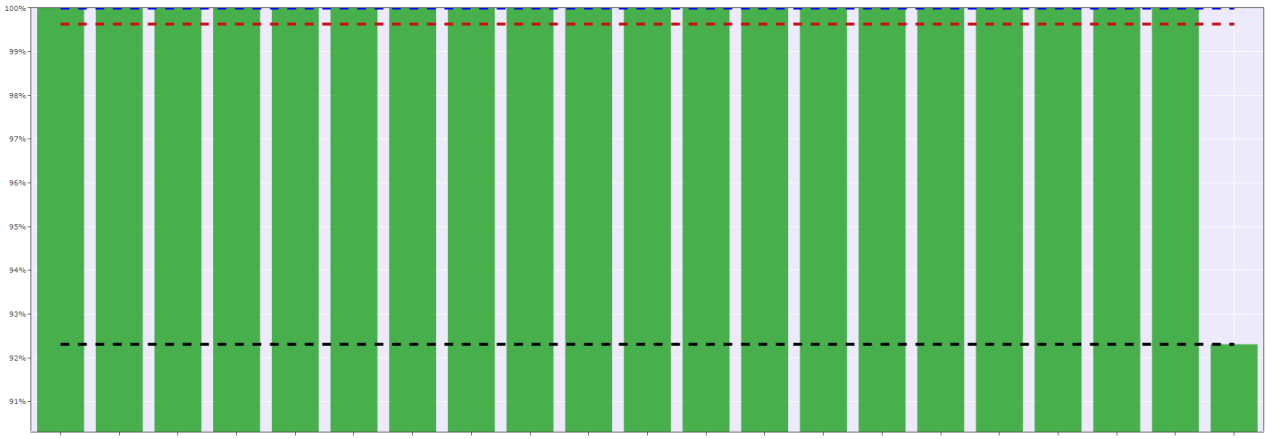


Fig. 2. F1 precision scores for each model built with decomposition method *one-vs-one* using dimensionality reduction. Highest score of 100%, lowest of 92% with a mean F1 score of 100%. Most models used only 5 features where others used up to 25 features.

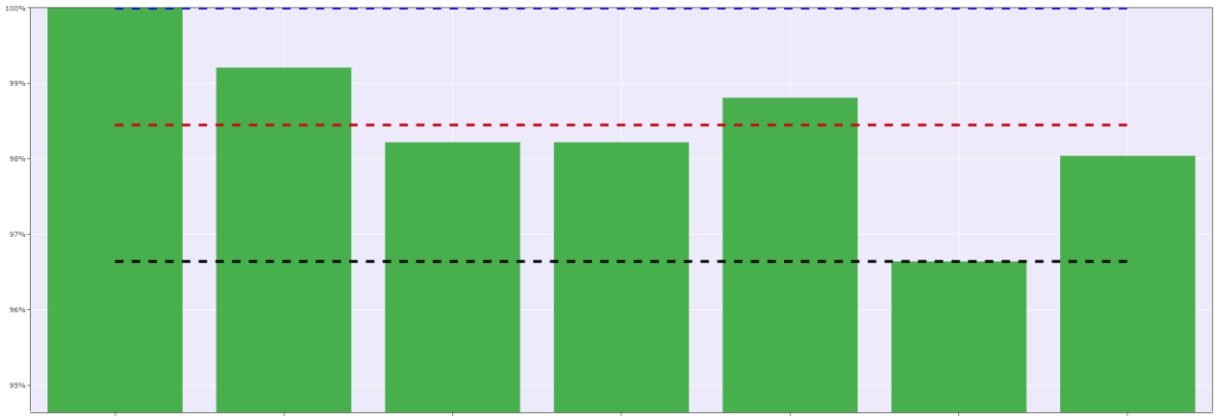


Fig. 1. F1 precision scores for each model built with decomposition method *one-vs-all* using dimensionality reduction. Highest score of 100%, lowest of 97% with a mean F1 score of 98%. Feature usage was varied, ranging from 5 to 45 features.

base results. The results of the improved models are shown in Fig. 3 for decomposition method *one-vs-one* and Fig. 4 for decomposition method *one-vs-all.*

In the case of *one-vs-one* experiment, most classifiers only required 5 features and no more than 25 features were used on the other classifiers. For the *one-vs-all* experiment, classifiers required a varied quantity of features, ranging from 5 to 45 features.

Even though the smaller predictions have been improved, it seems that the overall prediction F1 score has remained too low, but the improved experiments have still gotten a better result than the experiments without dimensionality reduction. Interestingly enough, both experiments, with or without dimensionality reduction, kept a very high prediction score for tissues. The overall prediction F1 score of each experiment can be seen on Table 2.

TABLE II.

| ***Decomposition method*** | ***Dimensionality reduction*** | ***F1 Score (per Condition/ per Tissue)*** |
|---|---|---|
| One-vs-One | No | 40%/95% |
| One-vs-All | No | 35%/95% |
| One-vs-One | Yes | 65%/95% |
| One-vs-All | Yes | 60%/100% |

## VI. CONCLUSION

According to the experiments made, it is possible to see that Random Forest does a good job in reducing the dimensionality, from 65,989 features to below 50 features, and still raising the prediction scores, hence meeting the requirements of a feature selection method.

Overall prediction scores have also shown improvement with the proposed method but still remain too weak to be used in real life gene expression analysis software. Noting that the predictions scores of the individual models has not been well translated into the final ensemble, it possible that further research is needed in the voting method used on each decomposition methods.

Additionally, it is possible that the dataset used in this research falls under the problem of hierarchical classification, where low predictions scores are expected, and therefore should be approached as such which are different from the common flat hierarchy classification normally faced in most classification problems [9].

## REFERENCES


[1] Koeppen. “The curse of dimensionality”. In: Proceedings of the 5th Online WorldConference on Soft Computing in Industrial Applications (WSC5). 2000, pp. 4-8.

[2] Jain, R.P.W. Duin, J. Mao. Statistical pattern recognition: A review. In: Pattern Analysis and Machine Intelligence, IEEE Transactions on 22.1 (2000), pp. 4-37.

[3] Trunk, A problem of dimensionality: A simple example. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 3 (1979), pp. 306-307.

[4] Johnstone. et al., Statistical challenges of high-dimensional data. In: Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 367.1906 (2009), pp. 4237-4253.

[5] Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot. Variable selection using Random Forests. Pattern Recognition Letters, Elsevier, 2010, 31 (14), pp.2225-2236.

[6] L. Breiman. Random forests. Machine Learning, 45(1): 5–32, 2001.

[7] I. Guyon and A. Elisseeff. “An introduction to variable and feature selection". In: The Journal of Machine Learning Research 3 (2003), pp. 1157-1182.

[8] H. Pang et al. “Pathway analysis using random forests classification and regression". In: Bioinformatics 22.16 (2006), pp. 2028-2036.

[9] Silla Jr, Carlos N. and Freitas, Alex A., “A survey of hierarchical classification across different application domains”. Data Mining and Knowledge Discovery, (2011), 22 (1-2). pp. 182-196. ISSN 1384-5810.