Pipeline design to identify key features in prognosis biomarker analysis using a real lung cancer dataset

María Gabriela Valdés[1⋆], Xavier Rafael-Palou[1⋆], Ivan Galván-Femenía[2⋆], Xavier Duran[2], Jun Yokota[2], Ricard Gavaldà[3], Rafael de Cid[2⋆⋆], and Vicent Ribas Ripoll[1⋆⋆]

[1] Eurecat. Technology Centre of Catalonia, Av. Diagonal 177, 9th floor, 08018 Barcelona (Spain),
gabriela.valdes, xavier.rafael, vicent.ribas @eurecat.org
[2] IGTP-PMPPC. Institut Germans Trias i Pujol (IGTP) - Programa de Medicina Predictiva i Personalitzada del càncer. GCAT Lab.,
igalvan, xduran, jyokota, rdecid @igtp.cat
[3] Universitat Politècnica de Catalunya. Facultat d'Informàtica de Barcelona, Edifici B6 del Campus Nord C/Jordi Girona Salgado, 1-3 08034 Barcelona (Spain),
gavalda@cs.upc.edu

Abstract. During the last decade, the interest to apply machine learning algorithms to genomic data has significantly increased for a variety of bioinformatics applications. Analyzing this type of data entails tackling difficulties related to high-dimensionality and class imbalance for knowledge extraction and identifying important features. In this study, we propose a general framework to tackle those challenges by stacking different machine learning algorithms and techniques to choose the best configuration as the final model to be used for classification and identify relevant SNPs (e.g. single nucleotide polymorphism). We test and compare the machine learning framework presented in this short paper in a real data-set and compared with the standard state-of-the-art Genome Wide Association approach.

Keywords: GWAS, SNP, machine learning, classification, feature selection

1 Introduction

Lung cancer (LC) is a malignant lung tumor characterized by uncontrolled cell growth in tissues of the lung. If left untreated, it can spread beyond the lung tissue and cause metastasis into nearby tissue or other parts of the body. There are two main types of lung cancer: small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC). Treatment and long-term outcomes depend on

⋆ co-author
⋆⋆ co-corresponding

the type of cancer, the stage (degree of spread), and the person's overall health. LC is commonly treated with surgery, chemotherapy, and radiotherapy. NSCLC is sometimes treated with surgery, whereas SCLC usually responds better to chemotherapy and radiotherapy. Moreover, LC is an important cause of admission to the ICU after surgery and is also considered an important co-morbidity in critical care.

Array based genotyping and sequencing technologies has enabled genome-wide characterization of the effect of DNA variation (e.g. single nucleotide polymorphisms (SNPs)) on diseases and their complex traits. Many studies have identified genetic variants, which affect the risk of multiple diseases. However, a single variant confers a small risk with low prediction power. To identify individuals at high risk it is appropriate to consider mixed models with genetic and epidemiological data. The huge number of genetic variants available from OMIC data, and the single SNP approach of the classical genome wide association studies (GWAS) analysis present the limitation of not being able to identify potential interactions. This analysis is challenging because of the high dimensionality of genomic data (up to millions of SNPs), the relatively small number of analyzed individuals and the uneven proportion of individuals belonging to each class (class imbalance). This is why the interest of applying machine learning algorithms to array-based SNP data has recently become so popular [2] [3] [4] [7] [5].

In cancer therapy there are clear evidences that response to treatment is a complex trait where genomic and environmental factors play overlapping roles as modifiers. In this scenario, GWAS analysis is commonly used to investigate the predictive role of genomic variants in response to treatment in a case-control design.

Our main goal in this short study is to identify a reduced set of SNPs with the highest prognostic value in a cohort of lung cancer patients treated with platinum-based chemotherapy. We accomplish this task by designing a specific framework/pipeline to deal with the multi-dimensionality problem using different feature selection techniques, followed by the application of sampling techniques to deal with class imbalance, and, finally, make predictions using machine learning classification methods [8]. We also propose different metrics to rank and select a relevant subset of SNPs. Ranked SNPs are then compared against the p-value based rank from the standard state-of-the-art method used in GWAS, which uses single-SNP logistic regression.

In the following sections, we describe the design of the pipeline-based analysis and the results obtained of its application to a lung cancer data-set. The concluding remarks, as well as future refinements of this approach, are provided in the conclusions section of this paper.

2 Methodology

2.1 Data-set

The study dataset comprises raw data and results from a GWAS analysis conducted at PMPPC (Programa de Medicina Predictiva i Personalitzada del càncer). The data-set comprises a series of 178 patients with advanced NSCLC and measurable disease (at least one target lesion). Subjects are classified according to RECIST (response evaluation criteria in solid tumors) as Non responders (Disease progression) (n=41) and Responder (partial/complete response and stable disease) (n=137). All relevant clinical and sociodemografic variables were included in the analysis. Genome-wide genotypes were generated using arrays-SNP technology using the Infinium HTS Assay, HumanCoreExome-24v1-0 BeadChip, (ILLUMINA, San Diego, CA) at the Genomic Units of PMPPC. A total of 325,762 SNPs remain after systematic quality control on the raw genotyping data (overall call rate of 99.89%). In silico methods were used for genome wide imputation to generate a data-set of 24,873,940 SNPs, from which 10,307,177 SNPs were retained for the association analysis. For the purposes of this analysis we consider only chromosome 12, where several significant GWAS signals were identified. Raw genotype data from 423,929 common variants (hg19 assembly, minor allele frequency > 0.01) and the p-values from uni-variate analysis (additive model) were analyzed.

Genetic variants are encoded using the *additive* model, using 0, 1, or 2 as numerical values [14]. Responders and non responders to treatment are classified as class 0 and 1 respectively. This data-set is split 80-20 into training and test sets resulting in 142 individuals for training (109 from class 0 and 33 from class 1) and 36 for test (28 from class 0 and 8 from class 1). The split is performed in a stratified way, to ensure the same proportion of individuals of each class as in the original data-set.

2.2 Pipeline configuration

To identify combinations of features with significant prediction power, a machine learning pipeline framework has been designed and tested (see Figure 1). The instantiation of this methods and algorithms, unfold a series of different models and their corresponding results.

We first add into the pipeline a feature selection method to deal with the multi-dimensionality problem introduced before. The main idea is to find irrelevant (noisy) or redundant features that do not contribute to the increase of the accuracy/performance of the final classification model, discard those, and keep the relevant ones to move forward in the pipeline process.

Feature selection methods are usually classified into three categories, depending on how they combine the feature selection search with the construction of the predictive model: filter, wrapper and embedded methods. Filter methods work independently of the classifier design, and perform feature selection by

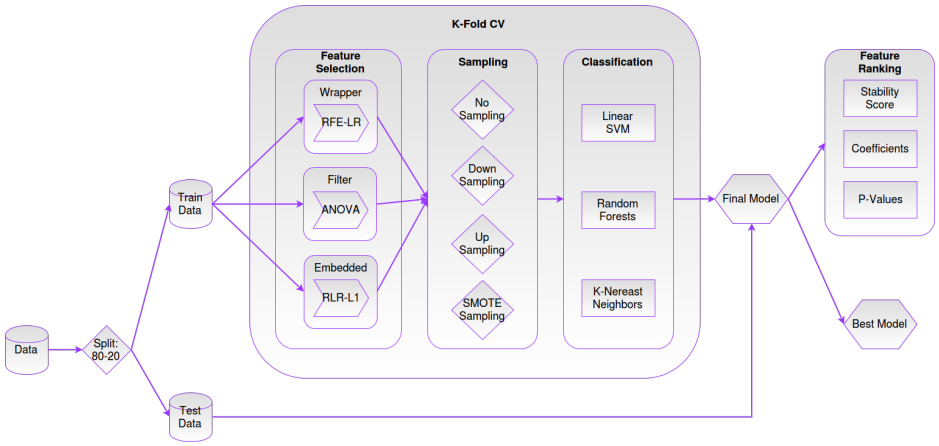

Fig. 1. General Pipeline Framework

looking at the intrinsic properties of the data. In contrast, wrapper and embedded methods perform feature selection by making use of a specific classification model. While wrapper methods employ a search strategy in the space of possible feature subsets, guided by the predictive performance of a classification model, embedded methods make use of the classification model internal parameters to perform feature selection [1]. In our pipeline design we use one of each type of feature selection methods [24] to instantiate this step of the pipeline: ANOVA as a filter method, recursive feature elimination with logistic regression (RFE-LR) as a wrapper method and regularized L1 logistic regression (RLR-L1) as an embedded method.

To deal with the class imbalance problem of having highly unequal number of individuals for each class of the target column, we propose to add a sampling step to the pipeline. The main objective of using a sampling technique is to adjust the class distribution of a data-set contributing to create a good model using a balanced training data-set [16]. We try down-sampling, up-sampling [18] and SMOTE-sampling [12] as alternatives of instantiation of the sampling step of the pipeline.

Finally, the classifier building step of the pipeline consists of a machine learning supervised model. We try out linear support vector machines (Linear SVM) [10], random forests (RF) [9] and K nearest neighbors (KNN) [11] as classification methods based on very different strategies to discriminate between classes.

In general, the feature selection and sampling steps, should not be decoupled from the process of creating an accurate classification model to avoid over-optimistic estimates of the error [13]. For this reason, we evaluate the whole pipeline process using cross-validation (CV) as a model selection tool. We apply grid-search with cross-validation to find the best combination of parameters for a specific pipeline using a training set, and afterwards having chosen a set of specific parameters (the model with highest CV score), we test the predictive

power of the model with a separate and independent test set, for which sampling has not been applied, preserving the original distribution of the data.

To identify the most important features, we also use the best model found with highest CV score, and measure the stability of the features selected by the feature selection step, by re-training this model several times using S different samplings/shuffles without replacement of 80% of the data of the training set [6]. Each feature of the original data-set will have a stability score associated to it. After re-training the whole pipeline, we add 1 to the stability score of the features that were selected. At the end of this process, features that were consistently selected S times, will have a score of S points. During this process of stability check, we also record other metrics useful to perform a ranking over the SNPs, based on characteristics of specific instantiations of the classification step in the pipeline. For example, if the classifier of the pipeline in analysis is a linear SVM, we save the values of the weights assigned by the algorithm to each feature. In a similar way, for the case of random forests, we keep record of the variable importance metric [15] associated to each feature while using this classification model. Regardless of the ranking measure used, let us refer to these values as coefficients. The sign of the coefficients can be used to measure the feature effect in the final classification result. A positive coefficient should be interpreted as a contribution to the classifier with no response to treatment and tumor progression (class 1).

After the iterative process is finished, we define the following aggregated metrics: the mean of the coefficients (MC), the mean of the absolute value of the coefficients, (MAC) since some coefficients can be equally important as positive and negative values, and the scaled version of the previously mentioned measure (MSC), in order to be able to make a comparison between the importance of each feature in terms of it's coefficient measure. We also intend to compare the results obtained of the most important features with the results of the single SNP logistic regression applied using the PLINK tool [17]. The output of applying this method gives a p-value associated to each SNP, and using a *threshold* = 0.01 we can filter the most significant SNPs by selecting those whose nominal value is greater than this *threshold*.

2.3 Functional in silico analysis

Evaluation of the potential functional impact of the identified variants was analyzed by gene annotation on GRch37 (hg19) using the *seq2pathway* R library [23], and GWAS Catalog of the National Human Genome Research Institute (NHGRI) for genome-wide significant matches [21].

3 Results

As stated in the previous section, the whole pipeline model was validated using 5-fold cross-validation and using the F1 weighted measure as scoring function. We use the latter scoring function due to the nature and distribution of the data,

since we know a priori that classes are imbalanced and we want to give equal importance to the precision and recall of both classes. The tuning of parameters associated to each step was performed using a grid-search. The different parameters tried are shown in Table 1. The C parameter in RLR-L1 refers to the inverse of regularization strength. In the case of Linear SVM it refers to the penalty parameter of the error term. In both of the latter cases, the smaller the values, the stronger the regularization. The *n_estimators* parameter in RF refers to the number of trees in the forest and *n_neighbors* in KNN is the number of neighbors to take into account in the neighbors voting step of the classifier.

Table 1. Parameter evaluated using grid-search and 5-fold cross-validation.

Pipeline Step	Parameter Options
RLR-L1	C = [10, 50, 500, 1000, 1500]
Linear SVM	C = [0.001, 0.01, 0.1, 1, 10, 100, 1000]
RF	n_estimators = [30, 47, 75, 119, 189, 299, 475, 753, 1194, 1892, 2999]
KNN	n_neighbors = [5, 20, 35, 50]

For the training process we obtained the results shown in Table 2. They are ordered in descending order by CV F1 score. For all 36 experiments, we only show the top 5 pipelines with the best performances. There were no significant differences between the top 5 classifiers in CV F1 score (from 0.702 to 0.681). The pipeline with the highest CV F1 score (0.702) consisted of the combination of down-sampling, followed by a regularized L1 logistic regression embedded feature selection (of 5,464 features using C = 1000) and finally using a random forest classifier (with 30 decision trees). The KNN pipeline reached the same CV F1 performance score as RF. However given that the former pipeline did not outperformed the latter in F1 Test score, we chose this one (RF) as the best pipeline given it's better capability to generalize on unseen observations.

The performance of the best pipeline (RF) was adjusted by -0.061 of F1 score in the Test set. This was partially due to the score achieved in Test Recall (0.611) which is affected by the errors made on the class with less number of instances (class 1, non-responders to treatment).

From the results shown in Table 2 it can be seen that RF is the most common classifier, giving better results for this kind of data and complex disease. Previous research shows the good performance of RF using SNP data to predict lung cancer [30] [31] and other complex diseases [25] like multiple sclerosis [26], age-related macular degeneration [27], crohn's disease [28] or rheumatoid arthritis [29]. Also, the top 5 pipelines indicate the use of a sampling techniques to improve the prediction, and up-sampling is the most common technique among them.

As for the feature ranking, we performed $S = 50$ different samplings/shuffles without replacement of 80% of the training set and calculated the stability score.

Table 2. Top 5 results obtained in training process of all possible combinations of instantiations of every step in the general pipeline.

Sampling	Feature Selection	Classifier	CV F1	Train F1	Test F1	Test Accuracy	Test Precision	Test Recall
Down_Sample	Embedded: RLR-L1	RF	0.702	0.849	0.641	0.611	0.700	0.611
Up_Sample	Wrapper: RFE-LR	KNN	0.702	0.751	0.592	0.583	0.601	0.583
Up_Sample	Wrapper: RFE-LR	Linear SVM	0.693	1.000	0.606	0.639	0.577	0.639
Smote_Sample	Filter: ANOVA	RF	0.682	1.000	0.681	0.778	0.605	0.778
Up_Sample	Embedded: RLR-L1	RF	0.681	1.000	0.667	0.750	0.600	0.750

After experimentation and to avoid being too restrictive, we select features with a stability score greater or equal to 45 (see Table 3), along side other coefficient metrics recorded throughout this stability process. The PLINK P-Value column, refers to the associated p-value obtained from single SNP logistic regression using PLINK.

Table 3. Subset of most important SNPs ordered in descending order by stability score.

SNP	Stability	MC	MSC	PLINK P-Value
Variant1	49	1.423e-4	2.360e-3	>1e-2
Variant2	49	5.939e-4	9.851e-3	2.917e-3
Variant3	49	1.589e-4	2.635e-3	>1e-2
Variant4	47	4.839e-4	8.028e-3	>1e-2
Variant5	47	1.065e-3	1.767e-2	4.850e-4
Variant6	46	6.091e-4	1.010e-2	3.101e-3
Variant7	46	1.600e-3	2.655e-2	2.686e-3
Variant8	46	4.328e-4	7.179e-3	>1e-2
Variant9	46	7.398e-4	1.227e-2	>1e-2
Variant10	45	3.058e-4	5.073e-3	3.101e-3
Variant11	45	5.178e-4	8.590e-3	>1e-2
Variant12	45	1.789e-4	2.968e-3	1.543e-3

With the help of the R library *seq2pathway*, we map genome features on the hg19 assembly. Subsequently, using the GWAS Catalog of the National Human Genome Research Institute (NHGRI) [21] and the European Bioinformatics Institute (EMBL-EBI) [22], we were able to search associated traits or diseases to

this genes. The genes in the regions identified in the top-5 models have not been associated directly to cancer, but to lung-related functions (see Table 4).

Table 4. Some associated traits to genes with lung-related functions.

Associated Trait
Body mass in chronic obstructive pulmonary disease.
Pulmonary function in asthmatics.
Asthma or chronic obstructive pulmonary disease.
Post-bronchodilator lung function in asthma (FEV1).
Post-bronchodilator lung function in asthma (FEV1/FVC).
Post bronchodilator FEV1.
Post bronchodilator FEV1 in COPD.
Post bronchodilator FEV1/FVC ratio.
Post bronchodilator FEV1/FVC ratio in COPD.
Pulmonary function.
Pulmonary function decline.
Lung function (FEV1).
Pre bronchodilator FEV1/FVC ratio.
3-hydroxy-1-methylpropylmercapturic acid levels in smokers.
3-hydroxypropylmercapturic acid levels in smokers.
Exhaled carbon monoxide levels.
MGMT methylation in smokers.
Smoking behavior.
Blood pressure (smoking interaction).
Fibrinogen levels (smoking status, alcohol consumption or body mass index interaction).

Lung cancer is a complex disease and complex diseases are influenced by a combination of multiple genes (possibly scattered across different chromosomes) as well as environmental factors. For this reason, further analysis must be carried out to explore the data from the rest of the chromosomes. Our overall goal is to look for genes globally involved in a defined clinical endpoint, in this case, lung cancer prognosis, and therefore we should search throughout the entire genome. Once we enrich this study with the results obtained from the rest of the chromosomes we will analyze the metabolic pathway associated to these genes to understand the patho-physiology of this type of cancer.

The whole pipeline design in this study was implemented using the *Sklearn* library [24]. *Sklearn* is a simple and efficient open source tool for data mining and data analysis. Several of the implementations of the algorithms used, are designed to be able to run in parallel over multiple cores. All of the experiments were executed on a single computer with 12 CPU cores of 3.30GHz each, and 47GB of RAM memory. Figure 2 shows the computational time efficiency of

the 36 different experiments executed. It is clear that regardless of the sampling technique applied to the data, using RLR-L1 as the feature selection step combined with RF as the classification step, resulted in the experiments that took the longest to finish during model selection, approximately 10 minutes. This makes sense looking at the number of parameter options of both RLR-L1 (5 options for parameter C) and RF (11 options for parameter n_estimators), resulting on 55 different settings and 275 different model fits (because of the 5-fold cross-validation) during the grid-search process. In this case, the "Embedded" type of feature selection took a lot more time than the "Wrapper" type because of the configuration we gave to the RFE method. Previously knowing the huge number of features that we were going to deal with, the method was configured to drop 25% of the features at each iteration, reaching very fast to the final goal of keeping 2000 features.

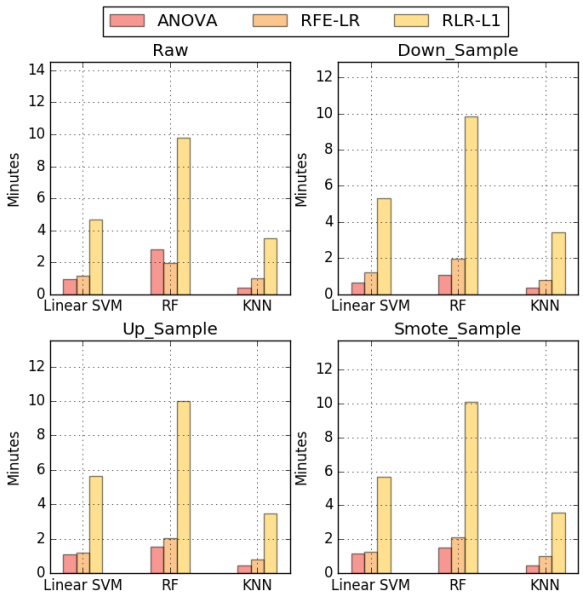

Fig. 2. Computational time efficiency of all 36 experiments during model selection (grid-search + cross-validation).

4 Conclusions

We defined an extensible and flexible framework to build classification pipelines based on machine learning techniques that can be applied to high dimensional and imbalanced SNP data. Using this approach, we were able to identify the

best model configuration and underscore genome variants that could be relevant as prognosis biomarkers in lung cancer patients.

With this general framework we try to fill the gap of commonly used methods that do not take into account cross-validation, class imbalance and the joined predictive power of multiple features to identify relevant features and perform classification tasks. From the set of 36 different experiments done using only SNPs from chromosome 12, the pipeline that performs down-sampling, regularized L1 logistic regression feature selection and random forest classification, was the one with best CV F1 score.

Even if our approach has not been able to find out some of the genome signals (1e-6) identified by PLINK, the annotation analysis of the variants identified showed some correlation with GWAS traits reported for lung/pulmonary functions and smoking interactions, as well as with pathways involved in the carcinogenic process.

We must recognize that the significance of the results presented here are limited by the sample size, the use of data from a single chromosome and the reduced number of epidemiological features of the original dataset. It is specially important to make further experiments using data from the remaining chromosomes in order to do a better evaluation of complex interactions across the genome. Further work can be foreseen in order to improve the performances reported in the results section. Given the nature of the framework described, easily new experiments could be setup to build new pipelines by simply replacing more sophisticated techniques regarding data sampling, feature selection or classification algorithms. In this sense we intend to expand the applied options of the different feature selection techniques, trying mutual information [19] as a filter method, sequential forward selection [20] as a wrapper method and regularized L2 logistic regression as an embedded method. We also intend to expand the number of experiments using other scoring functions during model selection, to give more importance to the recall of class 0, avoiding to predict that a patient's disease does not progress when it really does, and the patient is not responding to the treatment given.

Acknowledgments

IGTP is part of the **CERCA Program / Generalitat de Catalunya**. This research was partially funded by Acción de dinamización MINECO ADE10/00026. RdeCid is granted by the Ramón y Cajal (RYC) Program (RYC-2011-07822).

Co-Authors contributions

- Mrs. Valdés and Mr. Rafael contributed in the definition of the pipeline, methodology, experiments and document edition.
- Mr. Galván-Femenía contributed in the definition of the methodology, experiments and document edition.
- The rest of authors contributed in the methodology and document edition.

References

1. Abeel, Thomas and Helleputte, Thibault and Van de Peer, Yves and Dupont, Pierre and Saeys, Yvan: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics. 26, 392 (2010)
2. Nguyen, Thanh-Tung and Huang, Joshua Z and Wu, Qingyao and Nguyen, Thuy T and Li, Mark J: Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. BMC genomics. 16, 55 (2015)
3. Ban, Hyo-Jeong and Heo, Jee Yeon and Oh, Kyung-Soo and Park, Keun-Joon: Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine. BMC Genetics. 11, 26 (2010)
4. Hemphill, Edward and Lindsay, James and Lee, Chih and Măndoiu, Ion I. and Nelson, Craig E.: Feature selection and classifier performance on diverse biological datasets. BMC Bioinformatics. 15, S4 (2014)
5. Bulinski, Alexander and Butkovsky, Oleg and Shashkin, Alexey and Yaskov, Pavel: Statistical Methods of SNP Data Analysis and Applications. arXiv preprint arXiv:1106.4989. (2011)
6. Haury, Anne-Claire and Gestraud, Pierre and Vert, Jean-Philippe: The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PLOS ONE. 6, 1-12 (2011)
7. Saeys, Yvan and Inza, Iaki and Larraaga, Pedro: A review of feature selection techniques in bioinformatics. Bioinformatics. 23, 2507 (2007)
8. Cho, Baek Hwan and Yu, Hwanjo and Kim, Kwang-Won and Kim, Tae Hyun and Kim, In Young and Kim, Sun I: Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. Artificial intelligence in medicine. 42, 37–53 (2008)
9. Breiman, Leo: Random Forests. Machine learning. 45, 5–32 (2001)
10. Boser, Bernhard E and Guyon, Isabelle M and Vapnik, Vladimir N: A training algorithm for optimal margin classifiers. ACM. 144–152 (1992)
11. Cover, Thomas and Hart, Peter: Nearest neighbor pattern classification. IEEE transactions on information theory. 13, 21–27 (1967)
12. Chawla, Nitesh V. and Bowyer, Kevin W. and Hall, Lawrence O. and Kegelmeyer, W. Philip: SMOTE: Synthetic Minority Over-sampling Technique. Journal of artificial intelligence research. 16, 321–357 (2002)
13. Smialowski, Pawel and Frishman, Dmitrij and Kramer, Stefan: Pitfalls of supervised feature selection. Bioinformatics. 26, 440–443 (2010)
14. Mittag, Florian and Römer, Michael and Zell, Andreas: Influence of Feature Encoding and Choice of Classifier on Disease Risk Prediction in Genome-Wide Association Studies. PloS one. 10, e0135832 (2015)
15. Louppe, Gilles and Wehenkel, Louis and Sutera, Antonio and Geurts, Pierre: Understanding variable importances in forests of randomized trees. Advances in neural information processing systems. 431–439 (2013)
16. Jason Brownlee: Tactics to combat imbalanced classes in your machine learning dataset. http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/ Accessed: 2016-12
17. Purcell, Shaun and Neale, Benjamin and Todd-Brown, Kathe and Thomas, Lori and Ferreira, Manuel AR and Bender, David and Maller, Julian and Sklar, Pamela and De Bakker, Paul IW and Daly, Mark J and others: PLINK: a toolset for whole-genome association and population-based linkage analysis. The American Journal of Human Genetics. 81, 559–575 (2007) https://www.cog-genomics.org/plink2 Accessed: 2016-12

18. Guillaume Lemaître and Fernando Nogueira and Christos K. Aridas: Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. CoRR. abs/1609.06570 (2016)
19. Paninski, Liam: Estimation of entropy and mutual information. Neural computation. 15, 1191–1253 (2003)
20. Chandrashekar, Girish and Sahin, Ferat: A survey on feature selection methods. Computers & Electrical Engineering. 40, 16–28 (2014)
21. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, and Parkinson H.: The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research, 2014, Vol. 42 (Database issue): D1001-D1006. https://www.ebi.ac.uk/gwas/
22. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D.: The human genome browser at UCSC. Genome Res. 12, 996-1006 (2002) https://genome.ucsc.edu/index.html
23. Wang, Bin and Cunningham, John M. and (Holly) Yang, Xinan: Seq2pathway: an R/Bioconductor package for pathway analysis of next-generation sequencing data. Bioinformatics, 31, 3043 (2015). https://www.bioconductor.org/packages/release/bioc/html/seq2pathway.html
24. Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 12, 2825–2830 (2011).
25. Chen, Xi and Ishwaran, Hemant: Random forests for genomic data analysis: Genomics. 99, 323–329 (2012).
26. Goldstein, Benjamin A and Hubbard, Alan E and Cutler, Adele and Barcellos, Lisa F: An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. BMC genetics. 11, 49 (2010).
27. Jiang, Rui and Tang, Wanwan and Wu, Xuebing and Fu, Wenhui: A random forest approach to the detection of epistatic interactions in case-control studies. BMC bioinformatics. 10, 565 (2009).
28. Schwarz, Daniel F and König, Inke R and Ziegler, Andreas: On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics. 26, 1752–1758 (2010).
29. Wang, Minghui and Chen, Xiang and Zhang, Meizhuo and Zhu, Wensheng and Cho, Kelly and Zhang, Heping: Detecting significant single-nucleotide polymorphisms in a rheumatoid arthritis study using random forests. BMC proceedings. 3, S69 (2009).
30. Chung, Ren-Hua; Chen, Ying-Erh: A two-stage random forest-based pathway analysis method. PloS one. 7, e36662 (2012).
31. Weissfeld, J. L., Lin, Y., Lin, H. M., Kurland, B. F., Wilson, et al.: Lung cancer risk prediction using common SNPs located in GWAS-identified susceptibility regions. Journal of Thoracic Oncology. 10, 1538-1545 (2015).