Primal-Dual for Classification with Rejection (PD-CR): A Novel Method for Classification and Feature Selection. An Application in Metabolomics Studies.

David Chardin
Nice Sophia Antipolis University
Michel Barlaud (barlaud@i3s.unice.fr)
Nice Sophia Antipolis University
Olivier Humbert
Nice Sophia Antipolis University
Fanny Burel-vandenbos
Nice Sophia Antipolis University
Thierry Pourcher
Nice Sophia Antipolis University
Valerie Rigau
University of Montpellier

Research Article

Keywords: Supervised classification methods, omics studies, PD-CR, Partial Least Squares Discriminant Analysis (PLS-DA)

Posted Date: January 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-150506/v1

Version of Record: A version of this preprint was published at BMC Bioinformatics on December 1st, 2021. See the published version at https://doi.org/10.1186/s12859-021-04478-w.

RESEARCH

Primal-Dual for Classification with Rejection (PD-CR): A novel method for classification and feature selection. An application in metabolomics studies.

David Chardin [1,2], Olivier Humbert [1,2], Caroline Bailleux [1,4], Fanny Burel-Vandenbos [5], Valerie Rigau [6], Thierry Pourcher [1*] and Michel Barlaud [3*]

Full list of author information is available at the end of the article

Abstract

Background: Supervised classification methods have been used for many years for feature selection in metabolomics and other omics studies. We developed a novel primal-dual based classification method (PD-CR) that can perform classification with rejection and feature selection on high dimensional datasets. PD-CR projects data onto a low dimension space and performs classification by minimizing an appropriate quadratic cost. It simultaneously optimizes the selected features and the prediction accuracy with a new tailored, constrained primal-dual method. The primal-dual framework is general enough to encompass various robust losses and to allow for convergence analysis. Here, we compared PD-CR to two commonly used methods : Partial Least Squares Discriminant Analysis (PLS-DA) and Random Forests. We analyzed two metabolomics datasets: one urinary metabolomics dataset concerning lung cancer patients and healthy controls; and a metabolomics dataset obtained from frozen glial tumor samples with mutated isocitrate dehydrogenase (IDH) or wild-type IDH.

Results: PD-CR was more accurate than PLS-DA and Random Forests for classification using the 2 metabolomics datasets. It also selected biologically relevant metabolites. PD-CR has the advantage of providing a confidence score for each prediction, which can be used to perform classification with rejection. This substantially reduces the False Discovery Rate.

ConclusionThe confidence score provided with PD-CR adds considerable value to the prediction as it includes a metric that is implicitly used by every physician when they make a medical decision: the probability to make the wrong choice. So far, one of the main obstacles to the use of machine learning in medicine resides in the fact that it is harder to trust the decision of a machine learning method than that of a physician when it comes to health issues. We believe that providing a confidence score associated to the decision would make these new tools more convincing if used in routine clinical practice.

1 Introduction

Among the different omics fields, metabolomics is the most recent and provides new insights for a global study of biological systems. Metabolomics is a rapidly

growing and promising field of research in biology and healthcare. Metabolomics approaches are based on the determination of the levels of different small molecules or metabolites in biological samples (tissue, cells, serum, urine...). Interestingly, ever since the early metabolomics studies, supervised classification methods have been used for the analysis of the related datasets. One of the initial aims of metabolomic studies was to establish useful biomarkers, indicative of specific physiological states or aberrations. The challenge now is to understand the mechanisms by which changes in the metabolome are implicated in different phenotypic outcomes in a complex systems biology approach [1, 2].

Most metabolomics studies generate complex multivariate datasets including varying correlations between features and systematic noise. Therefore, multivariate data analysis methods are needed to explore these complex datasets. One of the most frequently used methods for metabolomics analyses is Partial Least Squares-Discriminant Analysis (PLS-DA) [3, 4].

PLS-DA is a chemometric technique used to optimize separation between different classes of samples, which is accomplished by linking two data matrices: X (raw metabolomic data) and Y (class membership). It has the advantage of handling highly collinear and noisy data. Yet, it has some drawbacks and needs to be handled with caution. Indeed it has been reported that PLS-DA can: 1. Lead to overfitting when the number of variables significantly exceeds the number of samples. Indeed, in this setting, the model is likely to lead to accurate classification by chance, based on irrelevant features [5]; 2. Have difficulties when few variables are responsible for the separation between two or more classes and, therefore, require a larger number of variables to achieve a good prediction accuracy [6]; and finally, 3. Lead to an over-optimistic understanding of the separation between two or more classes [7].

Continuous effort is being made to provide new statistical tools to tackle these drawbacks [8]. Some authors use Random Forests [9] as an alternative to PLS-DA for metabolomics studies. Random Forests is based on the bagging algorithm and uses an Ensemble Learning technique. Random Forests creates a large number of decision trees and combines their outputs. Yet Random Forests have significant drawbacks. For instance, they tend to overfit when using noisy datasets. Furthermore, the main disadvantage of Random Forests is their complexity. Indeed, they are much harder and time-consuming to construct, require more computational resources and are less intuitive than decision trees. Furthermore this complexity significantly hampers their interpretability.

Mathematics I3S partner has recently introduced a new tailored, constrained primal-dual method for supervised classification and feature selection [10]. This method has the significant advantage of providing a trustworthy confidence index with each prediction, which we use to define a new classifier with rejection. This is particularly useful in the context of clinical decision making as it diminishes the number of false positive and false negative results. Moreover, we believe this method out-performs other methods in terms of accuracy and feature selection.

Although there are many machine learning methods for feature selection such as support vector machines (SVM) [11], LASSO [12, 13], Discriminant analysis [14],

Proximal methods [15, 16], Boosting [17, 18], we compare here our novel Primal-Dual method for Classification with Rejection (PD-CR) to the state of the art PLS-DA and Random Forests classification methods used in metabolomics studies, available with the popular Metaboanalyst 5.0 (www.metaboanalyst.ca) [19].

2 Methods

2.1 Mathematical background

2.1.1 Robust classification and regression using ℓ_1 centers

Mathematically, classification problems can be described as follows :
Let X be the $m \times d$ data matrix made of m line samples $x_1, \ldots, x_m$ that belong to the d-dimensional space of features.
Let $Y \in \{0,1\}^{m \times k}$ be the matrix of labels where $k \geq 2$ is the number of clusters. Each line of Y has exactly one nonzero element equal to one, $y_{ij} = 1$ indicating that the sample x_i belongs to the j-th cluster. Projecting the data in lower dimension is crucial to be able to separate them accurately.
Let W be the $d \times k$ projection matrix, where $k \ll d$. (Note that the dimension of the projection space is equal to the number of clusters.)
The goal of the supervised classification method is to find the best possible values for the projection matrix W.

Sparse learning based methods have received a lot of attention in the last decade because of their high level of performance. The basic idea is to use a sparse regularizer that forces some coefficients to be zero. To achieve feature selection, the *Least Absolute Shrinkage and Selection Operator* (LASSO) formulation [12, 20, 21, 22, 23, 24] adds an ℓ_1 penalty term to the classification cost. An accurate criterion proposed by Witten *et al.* [25] is based on the sum of the square difference (and used in k-means [26]) and can be cast as follows:

$$\|Y\mu - XW\|_F^2 = \sum_{j=1}^{k} \sum_{l \in C_j} \|(XW)(l,:) - \mu_j\|_2^2, \tag{1}$$

where $C_j \subset \{1, \ldots, m\}$ denotes the j-th class, and where the row vector μ_j is the centroid of this class. Therefore, the matrix of centers μ is a square matrix of order k. It is well known that the Frobenius norm is sensitive to outliers. To address this, we have improved the approach by replacing the Frobenius norm by the ℓ_1 norm of the loss term as follows :

$$\|Y\mu - XW\|_1 = \sum_{j=1}^{k} \sum_{l \in C_j} \|(XW)(l,:) - \mu_j\|_1. \tag{2}$$

where $C_j \subset \{1, \ldots, m\}$ denotes the j-th cluster, and where $\mu_j := \mu(j,:)$ is the j-th line of μ. In our method, we simultaneously optimize (W, μ), adding some *ad hoc* penalty to break homogeneity and avoid the trivial solution $(W, \mu) = (0, 0)$.
Using both the projection W and the centers μ learnt during the training step, a

new query x in the test set (a dimension d row vector) is classified according to the following rule: it belongs to the cluster number j^* if and only if

$$j^* \in \arg\min_{j=1,\dots,k} \|\mu_j - xW\|_1. \tag{3}$$

2.1.2 Primal-dual scheme, constrained formulation

To handle features with a high correlation, we consider the convex constrained supervised classification problem,

$$\min_{(W,\mu)} \|Y\mu - XW\|_1 + \frac{\rho}{2}\|I_k - \mu\|_F^2 \quad \text{s.t.} \quad \|W\|_1 \leq \eta, \tag{4}$$

The drawback of the term $\|Y\mu - XW\|_1$ is that it enforces equality of the two matrices out of a sparse set: hence it tunes the parameters to enforce a perfect matching of the training data. To prevent this, we replace the 1-norm with a "Huber function". If $h_\delta(t) = t^2/(2\delta)$ for $|t| \leq \delta$ and $|t| - \delta/2$ for $|t| \geq \delta$, we can replace $\|Y\mu - XW\|_1$ with

$$h_\delta(Y\mu - XW) := \sum_{i=1}^{m}\sum_{j=1}^{k} h_\delta((Y\mu - XW)_{i,j}). \tag{5}$$

We obtain

$$\min_{(W,\mu)} h_\delta(Y\mu - XW) + \frac{\rho}{2}\|I_k - \mu\|_F^2 \text{ s.t.} \quad \|W\|_1 \leq \eta. \tag{6}$$

We can tune a primal-dual method to solve this problem with Algorithm 1 (See [10] and [27] for details)

Algorithm 1 Primal-dual algorithm, constrained case—$proj(V, \eta)$ is the projection on the ℓ_1 ball of radius η

Input: $X, Y, N, \sigma, \tau, \tau_\mu, \eta, \delta, \rho, \mu_0, W_0, Z_0$
for $n = 1, \dots, N$ **do**
  $W_{\text{old}} := W$
  $\mu_{\text{old}} := \mu$
  $W := W + \tau \cdot (X^T Z)$
  $W := proj(W, \eta)$
  $\mu := \frac{1}{1+\tau_\mu \cdot \rho}(\mu_{\text{old}} + \rho \cdot \tau_\mu I_k - \tau_\mu \cdot (Y^T Z))$
  $Z := \frac{1}{1+\sigma \cdot \delta}(Z + \sigma \cdot (Y(2\mu - \mu_{\text{old}}) - X(2W - W_{\text{old}})))$
  $Z := \max(-1, \min(1, Z))$
end for
Output: W, μ

2.1.3 Classification with rejection using a confidence Score for the Prediction (CSP)
False positive (FP) and false negative (FN) results are an important issue for diagnostic tools in medicine. One way to diminish the number of FP and FN results is to use classification with rejection [17, 28] for which classifiers are allowed to report "I don't know". This type of classification enables the incorporation of doubt in the results if the observation x is too hard to classify. Here, we propose to use a confidence score for the prediction (CSP) to devise a classifier with rejection.
In our analysis we only had two clusters with centers μ_1 and μ_2 Lets recall that the predicted label j^* of a sample x is given by

$$j^* \in \arg\min_{j=1,\dots,2} \|\mu_j - xW\|_1. \tag{7}$$

Thus we can compute the distances of sample x to the two centroids, respectively. $d_1 = \|\mu_1 - xW\|_1$ and $d_2 = \|\mu_2 - xW\|_1$ and we propose a confidence indicator for sample x as follows :

$$\rho(x) = \frac{d_1 - d_2}{d_1 + d_2} \tag{8}$$

Thus, the CSP $\rho(x)$ is a value ranging from -1 to 1. The closer the CSP $\rho(x)$ is to +1 or -1 depending on the predicted class, the higher the confidence for the prediction will be.

Thus if ϵ is a given threshold parameter, we can perform classification with rejection by rejecting binary classification for samples with an absolute value of CSP $\rho(x)$ under this threshold. The labels will then be predicted as follows :

$$Label = \begin{cases} -1 & \text{if } \rho(x) < -\epsilon \\ 0 & \text{if } -\epsilon < \rho(x) < \epsilon \\ 1 & \text{if } \rho(x) > \epsilon \end{cases} \tag{9}$$

We can then study the False Discovery Rate (FDR) $FDR = FP + FN$ as a function of parameter ϵ.

2.2 Availability of the method

We implemented PD-CR in python. Functions and scripts are freely available at https://github.com/tirolab/PD-CR.

2.3 Comparison to PLS-DA and Random Forests using 2 datasets

To compare PD-CR to the standard PLS-DA and Random Forests classification methods in terms of accuracy and feature selection, we tested the three methods on two metabolomic datasets named "BRAIN" and "LUNG".

2.3.1 BRAIN dataset

The BRAIN dataset was obtained from a metabolomic study performed by our biological team (TIRO) on frozen samples of glial tumors. The samples were provided by the university hospitals of Nice and Montpellier (France). Metabolite extracts were prepared and analyzed in the TIRO laboratory (Nice, France). With this dataset, the goal was to create a model that accurately discriminated between isocitrate dehydrogenase (IDH) mutated and IDH wild-type glial tumors. This mutation is a key component of the World Health Organization classification of glial tumors [29]. The mutational status is usually assessed by IDH1 (R132H)-specific (H09) immunohistochemistry. Yet this technique can lead to False-Negative results [30], which can only be identified by sequencing. An accurate metabolomic based test able to assess the IDH mutational status could be a promising solution to this problem.

These samples were retrospectively collected from two declared biobanks from the Central Pathology Laboratory of the Hospital of Nice and from the Center of Biological Resources of Montpellier (Plateforme CRB-CHUM). Consent or non-opposition was verified for every participant. For every participant, the IDH mutational status was assessed using immunohistochemistry and pyrosequencing for immunonegative cases.

Samples of brain tumors were analyzed using Liquid Chromatography coupled to tandem Mass Spectrometry (LC-MS/MS) in an unbiased metabolomicss approach. The details of this analysis are available in the Appendix.

2.3.2 LUNG dataset

The LUNG dataset was provided by Mathe *et al.* This dataset includes metabolomics data concerning urine samples from 469 Non-Small Cell Lung Cancer (NSCLC) patients prior to treatment and 536 controls collected from 1998 to 2007 in seven hospitals and in the Department of Motor Vehicles (DMV) from the greater Baltimore, Maryland area. Urine samples were analyzed using an unbiased metabolomics LC-MS/MS approach. This dataset is available from the MetaboLights database (study identifier MTBLS28)

Mathe *et al.* used Random Forests to classify patients as lung cancer patients or controls. The aim was to create a new screening test for lung cancer, based on metabolomics data from urine. Lung cancer is one of the most common cancers and it is well established that early diagnosis is essential for treatment. An efficient screening method based on urinary metabolomics would be of great benefit.

2.3.3 Data Filtering and Preprocessing

Our laboratory performed the LC-MS/MS analysis for the BRAIN dataset. Therefore, we could apply different levels of filtering on this dataset. After processing of the raw data using MZmine 2 software, two types of filtering were applied to the BRAIN dataset, minimal and maximal filtering. The minimal filtering only removed metabolites for which a spike was detected in less than 10 percent of the samples. The maximal filtering removed all unindentified metabolites as well as metabolites that did not have an isotopic pattern. This filtering method is frequently used for metabolomic studies and diminishes the number of noisy features in the dataset. Furthermore, it

diminishes the time necessary for data processing because it diminishes the data volume. Unfortunately, any filtering will necessarily come with a high risk of removing some relevant features. Using the two BRAIN datasets, we aimed to assess how the filtering affected the results of the different classification methods. The LUNG dataset were used without additional normalization or filtering because limited information was available concerning the features in this freely available dataset.

2.3.4 Availability of the data

The datasets are freely available at https://github.com/tirolab/PD-CR.

2.3.5 Comparison to PLS-DA and Random Forests

The data were preprocessed as follows: i) Log-transformation for the following main benefits: Reducing heteroscedasticity and thus the biais on regression, Tansforming multiplicative noise into additive noise, ii) Zero mean and Scaling [31].

PD-CR [10] was compared to PLSDA[32] and Random Forests (by default 100 trees)[9] using the sklearn python package.

We computed the accuracy of the 3 classification methods for the two metabolomics datasets using 4-fold cross-validation (Script "PD-CR vs PLSDA and RF" on https://github.com/tirolab/PD-CR). The selected metabolites were analyzed and compared between methods for the metabolomics datasets.

For PD-CR, we plotted the histograms of the CSP $\rho(x)$ and the probability distribution function (PDF) as well as the False Discovery Rate ($FDR = (FP + FN)/total$) and the rate of rejected samples ($RRS = rejectedsamples/totalsamples$) depending on epsilon (the CSP threshold) ("ScriptrhoComputing on https://github.com/tirolab/PD-CR).

3 Results

The characteristics of the two metabolomics datasets are presented in Table 1 . The LUNG dataset included a large number of patients (a little over 1,000) with an equivalent number of features (a little under 3,000) and the BRAIN dataset included a smaller number of patients (88) with a much higher number of features. While obtaining metabolomics data concerning as many patients as there are in the LUNG dataset is remarkable, the number of patients in the BRAIN dataset is closer to the number of patients in most metabolomics studies.

Dataset	No. of Samples	No. of features	Sample type
Lung	1005	2944	Urine
Brain	88	25,287	Glial tumor tissue

Table 1 Overview of the datasets.

3.1 LUNG:

	Random Forests	PLS-DA	PD-CR
Accuracy %	75.10	77.79	80.38

Table 2 Accuracy using 4-fold cross validation: comparison of algorithms on LUNG dataset

As shown in Table 2 PD-CR outperformed PLS-DA and Random Forests by 3.5% and 5.2% respectively and is the only classification method with an accuracy higher than 80% for the LUNG dataset.

Even though an accuracy of 80.38% may be high enough to consider using our PD-CR method and urinary metabolomics for the screening of lung cancer, Figure 1 shows that the accuracy may be even higher if the CSP is taken into account and if it is used to perform classification with rejection. Indeed, in Figure 1 the top left shows the histogram of the CSP and the top right the kernel probability distribution function (PDF). We can see that healthy controls and cancer patients are predicted with an equal high confidence. On the bottom left the False Discovery Rate ($FDR = (FP+FN)/totalsamples$) decreases as the confidence score threshold increases, but as shown in the bottom right, the rate of rejected samples ($RRS = rejectedsamples/totalsamples$) increases.

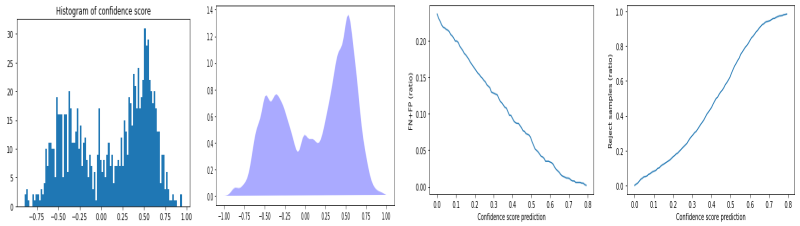

Figure 1 *Confidence Score for the Prediction (CSP) on the Lung dataset. Top Left: Histogram of the CSP, Top right: Kernel density estimation; Bottom left $FDR = FP + FN$ as a function of CSP, Bottom right: rate of rejected samples as a function of CSP.*

RF	PLSDA	PD-CR
MZ 264.1215224	MZ 264.1215224	MZ 264.1215224
MZ 656.2017529	MZ 126.9069343	MZ 126.9069343
MZ 441.1613664	MZ 613.3595637	MZ 170.0605916
MZ 584.2670695	MZ 170.0605916	MZ 243.1004849
MZ 247.0970455	MZ 243.1004849	MZ 308.0984878
MZ 486.2571336	MZ 308.0984878	MZ 613.3595637

Table 3 Top 6 features selected by Random Forests, PLSDA and PD-CR on dataset **LUNG**

RF	PLSDA	PD-CR
MZ 264.1215224	MZ 264.1215224	MZ 264.1215224
MZ 656.2017529	MZ 126.9069343	MZ 308.0984878
MZ 441.1613664	MZ 170.0605916	MZ 126.9069343
MZ 584.2670695	MZ 613.3595637	MZ 613.3595637
MZ 247.0970455	MZ 243.1004849	MZ 243.1004849
MZ 486.2571336	MZ 486.2571336	MZ 247.0970455
MZ 308.0984878	MZ 308.0984878	MZ 332.0963401
MZ 204.1345526	MZ 561.3432022	MZ 441.1613664
MZ 247.1384435	MZ 94.06574518	MZ 94.06574518
MZ 447.10803	MZ 269.1280232	MZ 561.3432022

Table 4 Top 10 features selected by Random Forests, PLSDA and PD-CR in the LUNG dataset

As shown in Table 4, PD-CR selected "MZ 264.1215224" for a molecular ion at m/z 264.1215224 and "MZ 308.0984878" for a molecular ion at m/z 308.0984878 as the top two features.
These features "MZ 264.1215224" and "MZ 308.0984878" most likely correspond to

creatine riboside (expected m/z value in the positive mode: 264.1190; mass error: 10 ppm) and N-acetylneuraminic acid (expected m/z value in the negative mode : 308.0987; mass error: 1 ppm), respectively. These two metabolites were described by Mathé *et al.* [33] as the two most important metabolites to discriminate between lung cancer patients and healthy individuals using Random Forests on metabolomic data from urine samples. Indeed, these two metabolites were significantly higher in the urines of lung cancer patients, as shown in Figure 2.

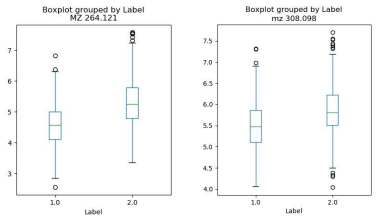

Figure 2 Boxplot MZ 264 and MZ 308

3.2 **BRAIN:**

3.2.1 Minimally filtered dataset :

	PLS-DA	Random Forests	PD-CR
Accuracy %	87.5	88.68	93.18

Table 5 Accuracy using 4-fold cross validation: comparison of algorithms in the BRAIN dataset

As shown in Table 5 , PD-CR outperformed PLS-DA and Random Forests by 5.6% and 4.5%, respectively for the BRAIN dataset. For this high dimensional dataset, the number of features (25,287) significantly exceeded the number of samples (88) giving a significant drop in the PLS-DA accuracy.

Furthermore, as shown in Figure 3 the accuracy obtained with PD-CR could be further improved by using the CSP to perform classification with rejection. Indeed, most of the samples were classified with a high CSP and if we apply a CSP threshold ϵ of 0.45, the FDR drops to 0 while only rejecting 10% of the samples. This shows that all the miss-classified samples had a low CSP.

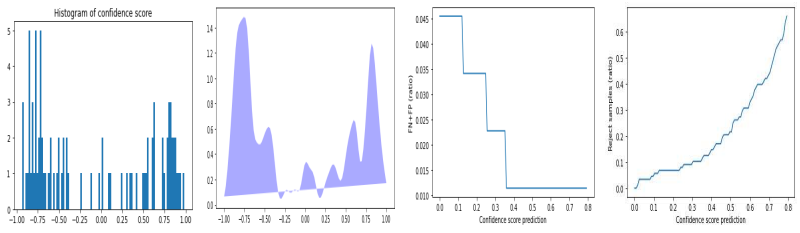

Figure 3 Confidence Score for the Prediction (CSP) on the BRAIN dataset. Top left: Histogram, Top right: Kernel density estimation, Bottom left: $FDR = FP + FN$ as a function of CSP, Bottom right: rate of rejected samples as a function of CSP.

As shown in Table 6, most of the top features selected with the 3 methods correspond to different isotopes and adducts of 2-hydroxyglutarate. Indeed, POS_MZ131.0342,

Random Forests	PLSDA	PD-CR
NEG_MZ147.0867	POS_MZ131.03427	POS_MZ131.0342
POS_MZ133.0384	POS_MZ132.0375	POS_MZ132.0375
POS_MZ166.0713	POS_MZ166.0713	POS_MZ243.9903
POS_MZ228.0182	NEG_MZ147.0288	POS_MZ166.0713
POS_MZ132.5234	NEG_MZ148.0321	NEG_MZ147.0288
POS_MZ173.0306	NEG_MZ149.0329	NEG_MZ148.0321
POS_MZ219.0082	POS_MZ171.0265	POS_MZ123.5181
NEG_MZ215.0168	POS_MZ132.0375	POS_MZ171.0265
POS_MZ171.0265	POS_MZ243.9903	NEG_MZ149.0329
POS_MZ319.0510	POS_MZ123.5181	POS_MZ133.0384

Table 6 Top 10 features selected by Random Forests, PLSDA and PD-CR on the BRAIN dataset with 25287 features

POS_MZ132.0375 and POS_MZ133.0384 all correspond to the [M+H-H2O adduct]+ of 2-hydroxyglutarate with C12, and two C13 isotopes respectively. NEG_MZ147.0288, NEG_MZ148.0321 and NEG_MZ149.0329 correspond to the [M-H]- adduct with C12, and two C13 isotopes respectively. POS_MZ166.0713 corresponds to a [M+NH4]+ adduct. POS_MZ171.02645 corresponds to the [M+Na]+ adduct. POS_MZ243.9903 had the same retention time and chromatographic profile as POS_MZ131.0342, suggesting that it was an unknown fragment or adduct of 2-hydroxyglutarate.
2-Hydroxyglutarate is a well-known oncometabolite produced in high quantities by mutated IDH1/2 in gliomas [34]. It is therefore expected that this compound will have a high weight when classifying mutated vs wild-type gliomas as it should be significantly increased in IDH mutated gliomas (as shown in figure 4).
Here all three methods selected this important feature among a high dimensional dataset (25287 features in this case).

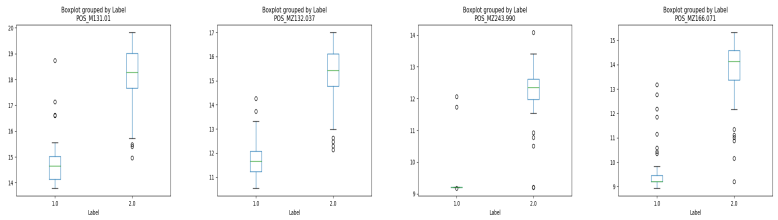

Figure 4 Brain dataset boxplots MZ 131.0342, MZ 132.0375 MZ 243.9903,MZ 166.0712

3.2.2 Comparison to the highly filtered dataset

	PLS-DA	Random Forests	PD-CR
Accuracy %	93.18	92.04	94.31

Table 7 Accuracy using 4-fold cross validation: comparison of algorithms on **BRAIN highly filtered** data set

As shown in Table 7 the accuracies of the different methods were equivalent and very high when using the highly filtered version of the BRAIN dataset. All three methods selected the same top 3 features: POS_M131.034205118815, NEG_M147.028769567142 and POS_M85.0290511756814.

When PD-CR is used on highly filtered datasets, it leads to similar results as with PLS-DA or Random Forests. In contrast, it outperforms these methods when using minimally filtered datasets. In this case, as shown in Table 8 more features

were selected. When using the BRAIN dataset for the IDH-mutated vs wild-type classes, most of these additional features were adducts of 2-hydroxyglutarate and are therefore known to be biologically relevant. The additional features that are not adducts of 2-hydroxyglutarate will be investigated in a future study.

Identified (495 features)	Large (25287 features)
POS_M131.0342	**POS_MZ131.0342**
NEG_M147.02882	POS_MZ132.0375
POS_M85.0291	POS_MZ243.9903
POS_M149.0450	POS_MZ166.0713
NEG_M112.0220	**NEG_MZ147.0288**
POS_M154.0864	NEG_MZ148.0320
NEG_M171.0847	POS_MZ123.518
NEG_M320.0627	POS_MZ171.0265
POS_M113.0350	NEG_MZ149.0329
POS_M147.1170	POS_MZ133.0384

Table 8 Top 10 features selected by PD-CR in the highly and minimally filtered versions of the **BRAIN** dataset

4 Discussion

Machine learning methods are of particular interest for metabolomics studies and are being used increasingly for other omics studies. Herein we introduce a new primal-dual method for supervised classification and feature selection. To our knowledge, a primal-dual method had never been used in this way. We compare this method to two of the most frequently used methods: PLS-DA and Random Forests, on two metabolomics datasets. For metabolomics, PD-CR appears to be more accurate than both methods while selecting biologically relevant features and providing a confidence score for each prediction. An important upside associated with the inclusion of a confidence score for each prediction is that it enables classification with rejection.

We believe that this confidence score is of great value, particularly for applications in medicine. Metabolomics approaches are of particular interest for medical applications. They could be used in routine clinical practice as they are relatively inexpensive and can be performed rapidly compared to proteomics, transcriptomics or genomics analyses. More and more studies suggest [35, 33], that metabolomics associated to classification methods are very promising tools for individual personalized medicine. To use metabolomics in routine clinical practice it is paramount to obtain robust, rapid and trustworthy classification methods. The confidence score provided with PD-CR adds considerable value to the prediction as it includes a metric that is implicitly used by every physician when they make a medical decision: the probability to make the wrong choice. So far, one of the main obstacles to the use of machine learning in medicine resides in the fact that it is harder to trust the decision of a machine learning method than that of a physician when it comes to health issues. We believe that providing a confidence score associated to the decision would make these new tools more convincing if used in routine clinical practice.

When comparing methods with the minimally filtered and the more filtered versions of the BRAIN dataset, the results obtained using the PLS-DA method appeared to be more impacted than those of the Random Forests and PD-CR. Indeed, the accuracy significantly decreased when the less filtered dataset was used. For this reason, several strategies are commonly used to reduce the number of features in

metabolomics datasets. Features can be filtered according to the number of detected peaks in all samples, the correct identification of the compound (using the most common adduct) or the presence of isotopes... Working with filtered data has some advantages, including the fact that it appears more biologically relevant to work on less noisy and more reliable data. However, filtering also has some important drawbacks, the most important being the high risk of removing interesting metabolites from the dataset. As shown in this work, PD-CR can be applied to both minimally filtered and highly filtered metabolomics datasets.

As it has been previously reported, when designing prediction models, some methods may lead to a more accurate model for a specific dataset while others may be more adapted with other datasets [36]. Indeed, even though we can discuss which machine learning method is the best, most often, researchers try out several different machine learning methods on their metabolomics datasets and report the results of the most accurate one. This process has even been automated by some authors [37]. PD-CR is an advanced method, based on recent development in convex optimization and we believe it should be considered by researchers when designing prediction models for metabolomics studies.

While prior metabolomic studies did not necessarily focus on validating which features the prediction models relied on, it is now admitted that to be trustworthy a model must be based on biologically relevant features and must therefore be interpretable [38]. indeed, interpretability of machine learning methods [39] is crucial to assess if selected features are biologically relevant. PD-CR offers a straightforward, reliable metric based on the weights of each feature in the model (matrix W).

Conversely, non-linear methods such as Random Forests and the linear method PLS-DA do not provide such easily interpretable metrics. For Random Forests, the Mean Decrease Impurity (MDI) is usually the default metric for variable importance [40]. It is computed as a mean of the individual trees' improvement in the splitting criterion produced by each variable. For PLS-DA, the Variable Importance for the Projection (VIP) score is often used. The VIP score is computed by summing the contributions VIN (variable influence) over all model dimensions. For a given PLS dimension a, $(VIN)^2_{ak}$ is a function of the squared PLS weight w^2_{ak} [41].

While these metrics offer some insight into the importance of each metabolite in the model the weights provided with PD-CR offer truly interpretable quantitative information concerning the importance of each metabolite for the model.

Furthermore, relevant feature selection is necessary for a correct understanding of the biological mechanisms underlying classification. It is well established that when expressed, mutant IDH 1/2 reduces 2-oxo-glutarate to 2-hydroxyglutarate [42]. It was therefore expected for 2-hydroxyglutarate to be a feature of importance as was the case when using PD-CR on the BRAIN dataset for the classification of IDH-mutated vs wild-type gliomas. As the biologically relevant features are known in advance, the BRAIN dataset is a good testing set for this new method. Furthermore, as we described, the features selected with PD-CR in the LUNG dataset are identical to the ones described by Mathé *et al.* in their original study, which also validates the accurate feature selection performed by PD-CR.

5 Conclusion

Herein we propose a new primal-dual method for feature selection and classification with rejection (PD-CR). To our knowledge, the primal-dual method has never been used in such fashion. PD-CR includes a sparse regularization factor which can is particularly appropriate for high dimensional sparse datasets such as metabolomics or other omics datasets.

We highlight the two main results. First, PD-CR is more accurate than PLS-DA and Random Forests and leads to the selection of biologically relevant features. Second, our method provides a confidence score for each prediction and allows classification with rejection, which is of significant interest for medical applications.

6 Appendix : Obtaining metabolomic data for the BRAIN dataset

6.1 Sample preparation

A fragment of 100mu of each frozen specimen was used for the metabolomic analysis. These fragments were prepared for an unbiased Liquid Chromatography-Mass Spectroscopy (LC-MS) analysis by applying the following procedure :
1. Frozen tissues were placed in microcentrifuge tubes and ground in 1mL of cold methanol (LC-MS grade, Merck Millipore, Molsheim, France) using pestles.
2. Homogenized samples were incubated overnight at -20°C then centrifuged at 15000g for 15 minutes.
3. The supernatants were removed and dried using a SpeedVac concentrator (SVC100H, SAVANT, Thermo Fisher Scientific, Villebon-sur-Yvette, France).
4. The lyophilized samples were resuspended in 180μL of 50:50 acetonitrile-H2O mix (LC-MS grade, Merck Millipore) prior to LC-MS/MS analyses.

6.2 LC-MS analysis

Liquid chromatographic analysis was performed using a DIONEX Ultimate 3000 HPLC system (Thermo Fisher Scientific). The mass spectrometry analysis was carried out on a Q Exactive Plus Orbitrap mass spectrometer (Thermo Scientific) with a heated electrospray ionization source, HESI II, operating in both positive and negative modes (See [35] for more details).

6.3 Metabolomic profiling

Metabolomic profiling was performed using MZmine (Version 2.39) [43]. The data obtained from positive and negative ionization modes were analyzed separately. The results obtained with each polarity were combined.

Acknowledgment
The authors thank Xuchun Zhang for his collaboration to the development of python software during his 2018 internship, Pr Antonin Chambolle and Pr Jean-Baptiste Caillau for fruitfull discussion on Primal-dual method, Dr Jean-Marie Guigonis (Bernard Rossi facility) for the LC-MS analyses and Pr Hugues Duffau and Dr Catherine Goze for validation, and formal analysis.

Funding
This work has been supported by the French government, through the UCAJEDI Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01. Equipment for this study was purchased through grants from the Recherche en matières de Sûreté Nucléaire et Radioprotection program from the French National Research Agency and the Conseil Départemental 06. David Chardin was funded by a grant from GIRCI Méditéranée.

Abbreviations
PLS partial least squares PD-CR Primal dual classification with rejection

Availability of data and materials
We implemented PD-CR in python. Functions and scripts are freely available at https://github.com/tirolab/PD-CR.

Ethics approval and consent to participate
The samples from the BRAIN dataset were retrospectively collected from two declared biobanks.

The authors confirm that all methods were carried out in accordance with relevant guidelines and regulations

The authors confirm that informed consent was obtained from all subjects

The authors confirm that this retrospective study has been approved by institutional ethics committees of the University Hospital of Nice and of the University Hospital of Montpellier

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
DC: Conceptualisation, methodology, formal analysis, investigation, validation, and writing.
OH: validation, formal analysis, and funding acquisition. CB: validation, and formal analysis.
FB-V and VR: validation, and formal analysis.
TP: Conceptualisation, methodology, formal analysis, writing, supervision, resources, project administration, and funding acquisition.
MB: Conceptualisation, methodology, formal analysis, writing, supervision, resources, and project administration.

Authors' information
[1]Transporters in imaging and Radiotherapy in Oncology (TIRO), Direction de la Recherche Fondamentale (DRF), Institut des sciences du vivant Fréderic Joliot, Commissariat à l'Energie Atomique et aux énergies alternatives (CEA), Université Côte d'Azur (UCA), Nice, France. [2]Department of Nuclear Medicine, Centre Antoine Lacassagne,Université Côte d'Azur (UCA), Nice,France. [3]Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis (I3S), Université Côte d'Azur (UCA), Centre de Recherche Scientifique (CNRS), Sophia Antipolis, France. [4]Department of Oncology, Centre Antoine Lacassagne,Université Côte d'Azur (UCA), Nice, France. [5]Central Laboratory of Pathology, University Hospital and Institute of Biology Valrose, Inserm U1091 - CNRS UMR7277, University Côte d'Azur, Nice, France. [6]Department of Pathology and Oncobiology. University Hospital, Montpellier, France. Institute for Neurosciences of Montpellier, INSERM U1051, Montpellier, France

"*Correspondence to barlaud@i3s.unice.fr"

Author details
References

1. Johnson, C.H., Ivanisevic, J., Siuzdak, G.: Metabolomics: beyond biomarkers and towards mechanisms. Nature reviews. Molecular cell biology **17**(7), 451–459 (2016). doi:10.1038/nrm.2016.25. Accessed 2018-06-15

2. Kell, D.B.: Metabolomics and systems biology: making sense of the soup. Current Opinion in Microbiology **7**(3), 296–307 (2004). doi:10.1016/j.mib.2004.04.012. Accessed 2020-04-05

3. Barker, M., Rayens, W.: Partial least squares for discrimination. Journal of Chemometrics **17**(3), 166–173 (2003)

4. Gromski, P.S., Muhamadali, H., Ellis, D.I., Xu, Y., Correa, E., Turner, M.L., Goodacre, R.: A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. Analytica Chimica Acta **879**, 10–23 (2015). doi:10.1016/j.aca.2015.02.012. Accessed 2020-04-05

5. Westerhuis, J.A., Hoefsloot, H.C.J., Smit, S., Vis, D.J., Smilde, A.K., van Velzen, E.J.J., van Duijnhoven, J.P.M., van Dorsten, F.A.: Assessment of PLSDA cross validation. Metabolomics **4**(1), 81–89 (2008). doi:10.1007/s11306-007-0099-6. Accessed 2020-04-06

6. Brereton, R.G.: Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. TrAC Trends in Analytical Chemistry **25**(11), 1103–1111 (2006). doi:10.1016/j.trac.2006.10.005. Accessed 2020-04-06

7. Broadhurst, D.I., Kell, D.B.: Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics **2**(4), 171–196 (2006). doi:10.1007/s11306-006-0037-z. Accessed 2020-04-06

8. Bartel, J., Krumsiek, J., Theis, F.J.: STATISTICAL METHODS FOR THE ANALYSIS OF HIGH-THROUGHPUT METABOLOMICS DATA. Computational and Structural Biotechnology Journal **4**(5), 201301009 (2013). doi:10.5936/csbj.201301009. Accessed 2020-04-06

9. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)

10. Barlaud, M., Chambolle, A., Caillau, J.-B.: Classification and feature selection using a primal-dual method and projection on structured constraints. International Conference on Pattern Recognition, Milan, 6538–6545 (2020)

11. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine learning **46**(1-3), 389–422 (2002)

12. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267–288 (1996)

13. Jacob, L., Obozinski, G., Vert, J.-P.: Group lasso with overlap and graph lasso. In: Proceedings of the 26th International Conference on Machine Learning (ICML-09), pp. 353–360 (2009)

14. Ding, C., Li, T.: Adaptive dimension reduction using discriminant analysis and k-means clustering. In: Proceedings of the 24th International Conference on Machine Learning, pp. 521–528 (2007)

15. Combettes, J.-C. P.L.and Pesquet: A douglas-rachford splitting approach to nonsmooth convex variational signal recovery. IEEE J. Selected Topics Signal Process., 564–574 (2007)

16. Barlaud, M., Belhajali, W., Combettes, P.L., Fillatre, L.: Classification and regression using an outer approximation projection-gradient method. IEEE Transactions on Signal Processing **65**(17), 4635–4644 (2017)

17. Freund, M.Y. Y., Schapire, R.E.: Generalization bounds for averaged classifiers. Annals of Statistics **32**(4), 1698–1722 (2004)

18. Nock, R., BelHajAli, W., Dambrosio, R., Nielsen, F., Barlaud, M.: Gentle nearest neighbors boosting over proper scoring rules. vol. 37, pp. 80–93. IEEE (2015)

19. Xia, J., Psychogios, N., Young, N., Wishart, D.S.: MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Research **37**(suppl$_2$), $652 - -660(2009).doi$: $10.1093/nar/gkp356.https://academic.oup.com/nar/article-pdf/37/suppl\_2/W652/3933058/gkp356.pdf$

20. Hastie, T., Rosset, S., Tibshirani, R., Zhu, J.: The entire regularization path for the support vector machine. Journal of Machine Learning Research **5**, 1391–1415 (2004)

21. Friedman, J., Hastie, T., Tibshirani, R.: Regularization path for generalized linear models via coordinate descent. Journal of Statistical Software **33**, 1–122 (2010)

22. Hastie, T., Tibshirani, R., Wainwright, M.: Statistcal learning with sparsity: The lasso and generalizations. CRC Press (2015)

23. Li, J., Cheng, K., Wang, S., Morstatter, F., P. Trevino, R., Tang, J., Liu, H.: Feature selection: A data perspective. ACM Computing Surveys **50** (2016)

24. Ali, A., Tibshirani, R.: The generalized lasso problem and uniqueness. Electronic Journal of Statistics **13**(2), 2307–2347 (2019)

25. Witten, D.M., Tibshirani, R.: A framework for feature selection in clustering. Journal of the American Statistical Association **105**(490), 713–726 (2010)

26. McQueen, J.-B.: Some methods for classification and analysis of multivariate observations. Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability (1967)

27. Chambolle, A., Pock, T.: On the ergodic convergence rates of a first-order primal-dual algorithm. Math. Program. **159**(1-2, Ser. A), 253–287 (2016)

28. Ni, C., Charoenphakdee, N., Honda, J., Sugiyama, M.: On the Calibration of Multiclass Classification with Rejection (2019). 1901.10655

29. Louis, D.N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W.K., Ohgaki, H., Wiestler, O.D., Kleihues, P., Ellison, D.W.: The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathologica **131**(6), 803–820 (2016). doi:10.1007/s00401-016-1545-1. Accessed 2020-04-09

30. Yoshida, A., Satomi, K., Ohno, M., Matsushita, Y., Takahashi, M., Miyakita, Y., Hiraoka, N., Narita, Y., Ichimura, K.: Frequent false-negative immunohistochemical staining with IDH1 (R132H)-specific H09 antibody on frozen section control slides: a potential pitfall in glioma diagnosis. Histopathology **74**(2), 350–354 (2019). doi:10.1111/his.13756

31. van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K., van der Werf, M.J.: Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC genomics **7** (2006)

32. Wold, S., Sjostrom, M., Eriksson, L.: Pls-regression: a basic tool of chemometrics. Elsevier volume 58, issue 2, 109–130 (2001)

33. Mathé *et al*, E.: Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. Cancer research **74**(12), 3259–3270 (2014)

34. Dang, L., White, D.W., Gross, S., Bennett, B.D., Bittinger, M.A., Driggers, E.M., Fantin, V.R., Jang, H.G., Jin, S., Keenan, M.C., Marks, K.M., Prins, R.M., Ward, P.S., Yen, K.E., Liau, L.M., Rabinowitz, J.D., Cantley, L.C., Thompson, C.B., Vander Heiden, M.G., Su, S.M.: Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature **462**(7274), 739–744 (2009). doi:10.1038/nature08617

35. Jing, L., Guigonis, J.-M., Borchiellini, D., Durand, M., Pourcher, T., Ambrosetti, D.: LC-MS based metabolomic profiling for renal cell carcinoma histologic subtypes. Scientific Reports **9**(1), 1–10 (2019)

36. Madsen, R., Lundstedt, T., Trygg, J.: Chemometrics in metabolomics—A review in human disease diagnosis. Analytica Chimica Acta **659**(1), 23–33 (2010). doi:10.1016/j.aca.2009.11.042. Accessed 2020-05-11

37. Leclercq, M., Vittrant, B., Martin-Magniette, M.L., Scott Boyer, M.P., Perin, O., Bergeron, A., Fradet, Y., Droit, A.: Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data. Frontiers in Genetics **10** (2019). doi:10.3389/fgene.2019.00452. Accessed 2020-05-08

38. Zhang, A., Sun, H., Yan, G., Wang, P., Wang, X.: Mass spectrometry-based metabolomics: applications to biomarker and metabolic pathway research. Biomedical Chromatography **30**(1), 7–12 (2016). doi:10.1002/bmc.3453. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bmc.3453. Accessed 2020-05-13

39. Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning (2017). 1702.08608

40. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, : Scikit-learn: Machine Learning in Python. arXiv:1201.0490 [cs] (2018). arXiv: 1201.0490. Accessed 2020-05-11

41. Chong, I.-G., Jun, C.-H.: Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems **78**(1), 103–112 (2005). doi:10.1016/j.chemolab.2004.12.011. Accessed 2020-05-13

42. Losman, J.-A., Kaelin, W.G.: What a difference a hydroxyl makes: mutant IDH, (R)-2-hydroxyglutarate, and cancer. Genes & Development **27**(8), 836–852 (2013). doi:10.1101/gad.217406.113. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. Accessed 2020-05-19

43. Pluskal, T., Castillo, S., Villar-Briones, A., Oresic, M.: MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC bioinformatics **11**, 395 (2010). doi:10.1186/1471-2105-11-395

Figures

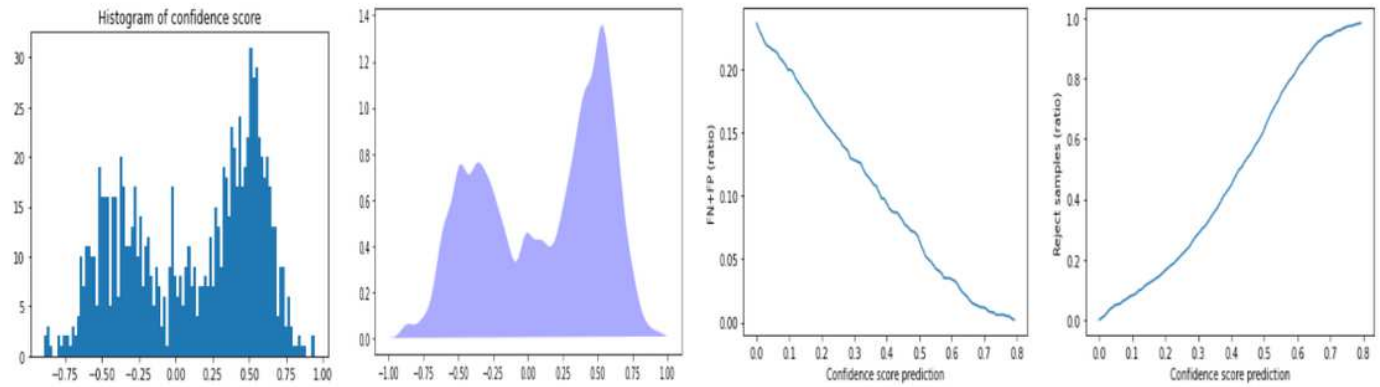

Figure 1

Confidence Score for the Prediction (CSP) on the Lung dataset. Top Left: Histogram of the CSP, Top right: Kernel density estimation; Bottom left FDR = FP + FN as a function of CSP, Bottom right: rate of rejected samples as a function of CSP.

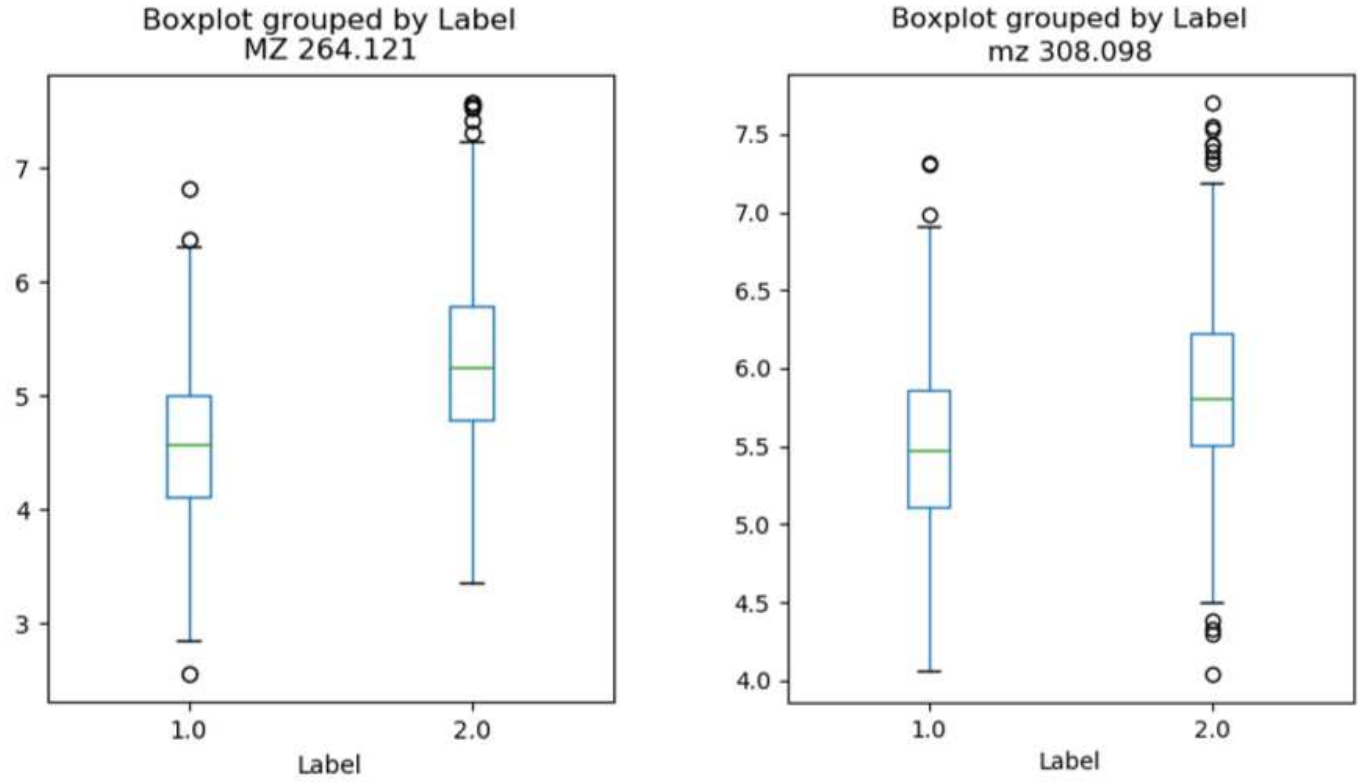

Figure 2

Boxplot MZ 264 and MZ 308

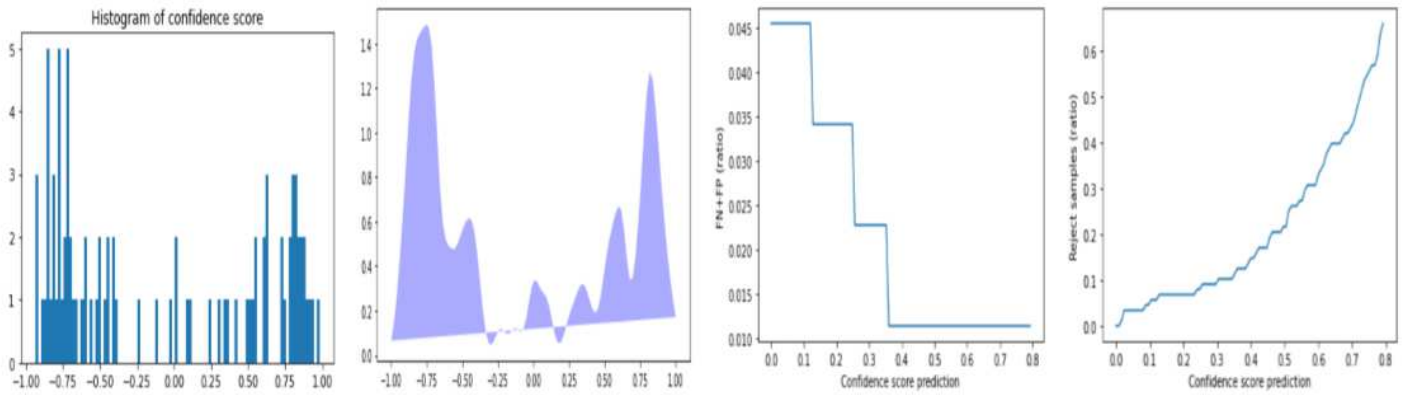

Figure 3

Confidence Score for the Prediction (CSP) on the BRAIN dataset. Top left: Histogram, Top right: Kernel density estimation, Bottom left: FDR = FP + FN as a function of CSP, Bottom right: rate of rejected samples as a function of CSP.

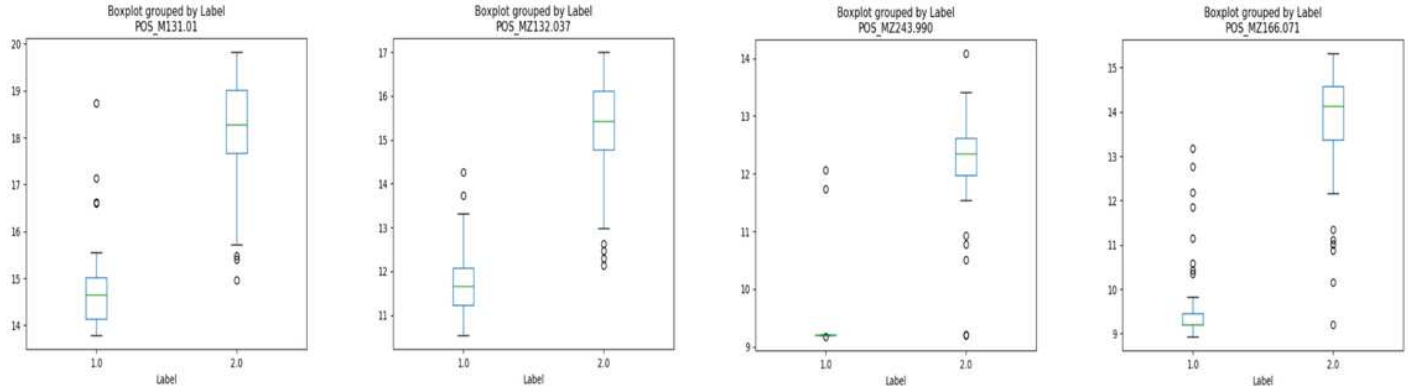

Figure 4

Brain dataset boxplots MZ 131.0342, MZ 132.0375 MZ 243.9903,MZ 166.0712