MSclassifR: an R Package for Supervised Classification of Mass Spectra with Machine Learning Methods

Alexandre Godmer[1,2,*], Yahia Benzerara[1], Emmanuelle Varon[3], Nicolas Veziris[1,2], Karen Druart[4], Renaud Mozet[5], Mariette Matondo[4], Alexandra Aubry[2,6], Quentin Giai Gianetto[4,7,*]

[1] AP-HP.Sorbonne université, Hôpital Saint-Antoine, Département de Bactériologie, Paris, France

[2] Sorbonne Université, INSERM, U1135, Centre d'Immunologie et des Maladies Infectieuses, Cimi-Paris, Paris, France

[3]*Laboratory of Medical Biology and National Reference Centre for Pneumococci, Intercommunal Hospital of Créteil, 94000 Créteil, France*

[4] Institut Pasteur, Université Paris Cité, Proteomics Platform, Mass Spectrometry for Biology Unit, UAR CNRS 2024

[5] UFR de Mathématiques et Informatique, Faculté des Sciences - Université Paris Cité

[6] AP-HP, Sorbonne Université, Hôpital Pitié-Salpêtrière, Centre National de Référence des Mycobactéries et de la Résistance des Mycobactéries aux Antituberculeux, Paris, France

[7] Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics HUB

[*] Corresponding author(s).

E-mail: alexandre.godmer@aphp.fr (Godmer A), quentin.giaigianetto@pasteur.fr (Giai Gianetto Q).

Abstract

MSclassifR is an R package that has been specifically designed to improve the classification of mass spectra obtained from MALDI-TOF mass spectrometry. It offers a comprehensive range of functions that are focused on processing mass spectra, identifying discriminant *m/z* values, and making accurate predictions. The package introduces innovative algorithms for selecting discriminating *m/z* values and making predictions. To assess the effectiveness of these methods, extensive tests were conducted using challenging real datasets, including bacterial subspecies of the *Mycobacterium abscessus* complex, virulent and avirulent phenotypes of *Escherichia coli*, different species of Streptococci and nasal swabs from individuals infected and uninfected with SARS-CoV-2. Additionally, multiple datasets of varying sizes were created from these real datasets to evaluate the robustness of the algorithms. The results demonstrated that the Machine Learning-based pipelines in MSclassifR achieved high levels of accuracy and Kappa values. On an in-house dataset, some pipelines even achieved more than 95% mean accuracy, whereas commercial system only achieved 62% mean accuracy. Certain methods showed greater resilience to changes in dataset sizes when constructing Machine Learning-based pipelines. These simulations also helped determine the minimum sizes of training sets required to obtain reliable results. The package is freely available online, and its open-source nature encourages collaborative development, customization, and fosters innovation within the community focused on improving diagnosis based on MALDI-TOF spectra.

Keywords: [MALDI-TOF MS]; [Machine Learning]; [R package]; [microbiology]

Key points:

- MSclassifR is a comprehensive R package enabling the construction of data analysis pipelines for the precise classification of mass spectra.
- Our R package contains an innovative method for variable selection from random forests, which delivered excellent results on real data.
- In-depth analysis of various machine learning-based pipelines using our package allowed us to make conclusions about the optimal m/z selection and prediction methods depending on the size of the training dataset.
- Using a publicly available dataset of mass spectra obtained from various MALDI-TOF instruments across different countries, MSclassifR is able to build robust pipelines capable of adapting to different instruments in an automatic way.
- When tested on an in-house dataset, MSclassifR pipelines consistently outperformed a commercial software in terms of prediction accuracy.

Introduction

In clinical microbiology, the accurate identification of pathogens is crucial for proper patient management. Matrix Assisted Laser Desorption Ionization - Time of Flight (MALDI-TOF) mass spectrometry has become widely adopted for pathogen identification in biological cultures due to its speed, affordability, and user-friendly nature [1]. This technique involves comparing the mass spectrum obtained from ionizing intact proteins of a viral or bacterial culture with reference mass spectrum profiles. Commercial software provided by spectrometer manufacturers uses algorithms to match the measured mass spectrum with the spectra of known microorganisms. While this technique is effective in distinguishing many species, it often falls short in differentiating spectrally close organisms, such as subspecies within the *Mycobacterium abscessus* complex [2] [3] or species from the *Enterobacter cloacae* complex [4] [5].

To overcome these limitations, several studies have explored advanced machine learning (ML) techniques to fully exploit the information contained in mass spectra and accurately classify them into distinct categories, such as detecting resistant or virulent microorganisms in closely related mass spectra [6] [7] [8]. However, many of these studies focus on a single ML method [9] [10] [11] [12] and others lack the provision of corresponding computer code, making reproducibility of results challenging [8].

While open-source R packages like MALDIquant and MALDIrppa already offer well-documented methods for pre-processing MALDI-TOF mass spectra [13] [14], additional steps are necessary to classify the pre-processed mass spectra, including the selection of discriminative peaks and an optimal mass spectra classification model. These steps require advanced knowledge and specialized R packages for variable selection and ML methods, such as the caret and mixOmics R packages, which further demand expertise in parameter optimization [15] [16]. On the other side, non-coders have access to all-in-one software solutions, but they are often expensive, do not

provide the computer code used, and provide limited user control over parameters (e.g. [17] [18]).

In response to these challenges, we have developed MSclassifR, a comprehensive package that enables the construction of data analysis pipelines for the precise classification of mass spectra. Within this article, we introduce the functionalities offered by MSclassifR and assess their effectiveness. We conducted tests using challenging real datasets and created additional datasets with varying sizes from them to evaluate the performance of data analysis pipelines built with MSclassifR.

Notably, our findings highlight the pivotal role of mass-over-charge (*m/z*) value selection in achieving accurate classifications. We have devised an innovative statistical method for variable selection, taking advantage of variable importances derived from Random Forests (RF) and employing a mixture model that distinguishes between discriminant and irrelevant features. This unique approach has led to superior performance in terms of accuracies and Cohen's Kappa coefficient values compared to alternative feature selection methods used before a prediction model.

Methods

The MSclassifR package offers R functions that concerns three main topics [*Figure 1A*]: (i) pre-processing of mass spectra, (ii) selection of discriminative *m/z* values, (iii) estimation of a classification model to predict the category of a mass spectrum from shortlisted *m/z* values.

Pre-processing of mass spectra methods

Many of the algorithms already available in the MALDIquant and MALDIrppa packages are implemented in MSclassifR [13] [14]. For example, the smoothing step can be performed by using the "Savitzky Golay" algorithm or the "Undecimated Wavelet Transform" (UDWT) algorithm [19] [20] ; baseline subtraction can be performed using the "statistics-sensitive non-linear iterative peak-clipping" (SNIP) algorithm [21], the TopHat algorithm [22], or the Convexfull algorithm [23]; intensities

123 can be normalized using the "Total Ion Current" (TIC) algorithm or the "Probabilistic
124 Quotient Normalization" (PQN) algorithm [24]; mass spectra can be aligned using the
125 "Locally Weighted Scatterplot Smoothing" (LOWESS) method [25] or by employing
126 linear, quadratic, or cubic functions [26]. Finally, the peak selection step can be
127 repeated using the "Median Absolute Deviation" (MAD) method [27] or the "super
128 smoother" method [28]. However, we have defined a pipeline by default in MSclassifR
129 including several algorithms that are widely used for processing MALDI-TOF MS
130 spectra [29] [*Figure 1A*] . It consists of the following steps: (i) square root intensity
131 transformation, (ii) spectrum smoothing with UDWT algorithm, (iii) baseline
132 processing with SNIP algorithm, (iv) intensity calibration using TIC algorithm, (v)
133 spectrum alignment with the LOWESS method and selection of peaks with signal-to-
134 noise ratio (S/N).

135 **Selecting discriminant *m/z* values**

136 The selection of discriminant *m/z* values is analogous to the variable selection problem
137 in statistics. It is crucial to choose the right variables to avoid incorporating irrelevant
138 ones into prediction models, which can lead to inaccurate predictions. Additionally,
139 variable selection helps reduce the dataset size, which is advantageous from a
140 computational time perspective. MSclassifR provides several methods for variable
141 selection, which can be categorized into two groups [30]:

142 (a) the first aims to identify a minimum number of variables that yield the highest
143 possible performance criterion (e.g., Accuracy or Cohen's Kappa), often resulting in
144 the elimination of correlated variables;

145 (b) the second aims to identify all variables with potential predictive power (e.g. without
146 deletion of correlated variables).

147 On one hand, the first group is commonly employed in ML literature [30] where the
148 primary objective is to achieve the highest accuracy with the smallest set of variables.
149 In omics sciences, this approach is used to identify a small set of potential biomarkers

150 (e.g. genes, proteins, metabolites) associated with a specific biological condition. A
151 popular method of this type of approach is "Recursive Feature Elimination" (RFE)
152 which consists of recursively selecting smaller and smaller sets of variables. This type
153 of method typically involves setting a specific number (or range for that number) of
154 variables to select.

155 On the other hand, the second group enables the filtration of variables that lack
156 predictive power (e.g. background noise) and retains only those with predictive
157 capabilities. For example, this approach is used to extract the set of genes or proteins
158 that exhibit differential expression between different biological conditions. Thus, this
159 type of method adapts to the dataset to select all the predictive variables without
160 specifying the number of variables to select.

161 In addition to existing methods, we have implemented an original approach to identify
162 all variables with predictive power using RF. This method involves estimating the
163 distribution of variable importances for non-discriminant (useless) variables from
164 negative importances or simulated non-discriminant variables [*Figure 1B*].
165 Subsequently, a mixture model is assumed, where the cumulative distribution of
166 variable importances for all variables (F) is a combination of the cumulative
167 distributions for useless variables (F_u) and discriminant variables (F_d), with π
168 representing the proportion of useless variables in the dataset:

169 $$F(x) = \pi \times F_u(x) + (1 - \pi) \times F_d(x)$$

170 By estimating the distribution of useless variables ($F_u(x)$), we can determine its
171 quantile values (x_q) and compute $\epsilon_q = F(x_q)/q$ for a given quantile (q) of useless
172 variables [*Figure 1C*]. For high quantile values of q, it is expected that $F_u(x_q)$ tends
173 to 1. Thus, $\epsilon_q = \pi + (1 - \pi) \times F_d(x_q)$ for such quantiles when $F_d(x_q)$ is not null. It
174 is assumed that F_d becomes zero when the variable importance is small. If this
175 assumption also holds, ϵ_q tends to π. The minimum value of the ϵ_q corresponds to an

estimated proportion of useless variables in the dataset, i.e. $\tilde{\pi} = \min_q \epsilon_q$ [*Figure 1C*], enabling us to estimate a number of discriminant variables (N_d) as floor (N$\times$ $(1 - \tilde{\pi})$), where N represents the total number of variables and "floor()" is the integer value function. Finally, the N_d variables with the highest variable importances are selected as the set of discriminant variables. Here, the π estimation is analogous to finding the proportion of true null hypotheses in the context of multiple hypothesis testing with the difference that we use the variable importances instead of p-values. A major difference lies in the absence of a reliable theoretical framework for estimating the distribution of variable importances for useless variables, whereas an assumption of independence between the tests and therefore a uniform distribution of p-values under the null hypothesis can be assumed in the multiple testing context [31]. It is therefore necessary to estimate the distribution of useless variables either by generating such variables or using the negative importance variables and assuming that $F_u(x)$ is a more or less symmetrical distribution with respect to zero. Both approaches are possible through our "SelectionVar()" function of our package ("*nbf*" argument). Because variable importances can be estimated with different approaches using RF, two ways of estimating them are implemented in "SelectionVar()": "mda" for the "mean decrease in accuracy" variables importances, and "cvp" for "cross-validated permutation" variables importances [32].

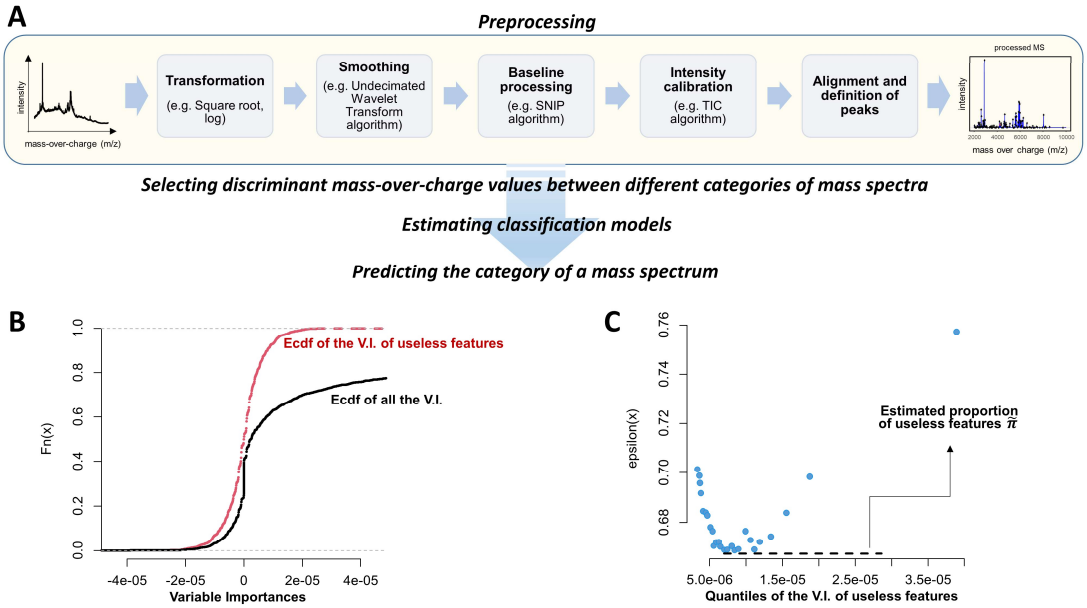

Figure 1. MSclassifR workflow and plots presenting the method developed for thresholding the variable importances of Random Forests

A. Overview of the proposed pipeline in MSclassifR, encompassing mass spectrum preprocessing with various methods, selection of discriminative *m/z*, and the estimation of a classification model to predict mass spectrum categories based on the shortlisted *m/z* values. **B.** Empirical cumulative distribution functions of the variable importances (V.I.) for the useless *m/z* (red) and all the *m/z* (black) estimated by the "SelectionVar()" function with the "mda" method from a real dataset [33] **C.** Estimation of $\tilde{\pi}$ using the same V.I. as in **B.** and the quantities $\epsilon_q = F(x_q)/q$ where q are quantiles of useless variables.

206 **Predicting the category of a mass spectrum**

207 Once we selected *m/z* values, a model can be estimated to predict a category from these
208 shortlisted values. In supervised classification, a commonly used parametric model is
209 multinomial linear logistic regression. It consists in estimating the probability that a
210 mass spectrum belongs to a category from the following model:

211 $$P(MS \in j) = \frac{e^{X\beta_j}}{1 + \sum_{k=1}^{K-1} e^{X\beta_k}}$$

212 where $P(MS \in j)$ is the probability that a mass spectrum belongs to the jth category
213 (among K categories), X are intensities of the mass spectrum measured at shortlisted
214 *m/z* values, and β_j are coefficients related to the jth category. Once all the coefficients
215 are estimated, a mass spectrum is associated to a category by

216 $$MS \in j \text{ if } \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} P(MS \in k) = j$$

217 In addition to multinomial linear logistic regression, popular ML algorithms are
218 available in MSclassifR like RF, neural networks, support vector machines with linear
219 kernels and extreme gradient boosting. Each of these algorithms can be used to estimate
220 a prediction model via the "LogReg()" function. Of note, the "LogReg()" function is
221 able to deal with imbalanced datasets displaying skewed class proportions. This is a
222 characteristic present in many real datasets when some categories are rarer than other.
223 Different approaches are implemented to resample the data before training the
224 algorithms via the "*Sampling*" argument: up-sampling, down-sampling and SMOTE
225 (Synthetic Minority Oversampling Technique). Next, the "PredictLogReg()" function
226 can be used to estimate the probabilities $P(MS \in j)$ from one or several prediction
227 models. In addition to provide the estimated $P(MS \in j)$ of each model in input,
228 "PredictLogReg()" includes two methods to aggregate the predictions of several
229 models: the max-voting method, and the Fisher's combined probability test to merge all
230 the estimated $P(MS \in j)$ of the models in input.

231

232 Another approach using parametric models involves conducting multiple linear
233 regressions, one for each category, and selecting the category associated with the
234 regression that has the lowest Akaike Information Criterion:

235 $$MS \in j \text{ if } \underset{k\epsilon\{1,\ldots,K\}}{\operatorname{argmin}} \mathrm{AIC}_k(X = \alpha_k + \sum_{i=1}^{I} \beta_{ik} Z_{ik} + \varepsilon_k) = j$$

236 Here, X represents the intensities of the mass spectrum measured at *m/z* values, Z_{ik}
237 represents the intensities of the ith mass spectrum of the kth category used to estimate
238 the model at the same *m/z* values, ε_k denotes the Gaussian white noise corresponding
239 to the *k*th linear regression with parameters α_k and β_{ik}, and AIC_k represents the
240 Akaike Information Criterion of the *k*th linear regression. This approach has been
241 implemented in an original function called "PredictFastClass()" in MSclassifR. It is
242 important to note that in order to perform the regression, the number *I* of mass spectra
243 used within a category has to be lower than the number of *m/z* values. Additionally, for
244 each category, a Fisher test can be conducted to determine if the β_{ik} coefficients of the
245 linear model are zero, indicating significant correlation between the mass spectrum and
246 at least one of the mass spectra in a category. By performing this test for all categories
247 and selecting the minimum p-value, we can assess if the mass spectrum significantly
248 belongs to at least one of the input categories. If the hypothesis is rejected for all
249 performed regressions, the mass spectrum can be considered as not belonging to any of
250 the input categories.

251 **Experimental datasets used to assess the performance of MSclassifR**

252 We used MSclassifR on openly accessible datasets to assess pipelines for accurately
253 predicting the category of a given mass spectrum. These datasets are of variable sizes:

254 - Dataset 1: 362 nasal swab mass spectra obtained from 312 patients categorized
255 as either SARS-CoV-2 infected (211 mass spectra) or SARS-CoV-2 uninfected
256 (151 mass spectra) from the Nachtigall *et al*. study [33]. These data are available
257 at https://www.ebi.ac.uk/pride/archive/projects/PXD021388.
258 - Dataset 2: 882 mass spectra from 294 strains of *Escherichia coli*, classified into
259 two categories: Shiga-toxin producer (190 strains, 570 spectra) or non-producer
260 (104 strains, 312 spectra) from the of Christner *et al.* study [6]. These data are

261 available at https://zenodo.org/record/4996418.

262 • Dataset 3 (in-house): 1001 mass spectra from 41 clinical strains representing the
263 three subspecies of the *Mycobacterium abscessus* complex: *M.* subsp. *abscessus*
264 (15 strains, 633 spectra), *M.* subsp. *bolletii* (7 strains, 164 spectra), and *M.*
265 subsp. *massiliense* (9 strains, 201 spectra). These data are available at
266 https://zenodo.org/record/5793313.

267 • Dataset 4 (in-house): 1890 mass spectra from 79 clinical strains from
268 *Streptococcus pneumoniae* (60 strains, 1434 spectra) and from the viridans
269 streptococci group including *S. mitis* (11 strains, 264 spectra) and *S.*
270 *pseudopneumoniae* (8 strains, 192 spectra). These data are available at
271 https://zenodo.org/record/8419282.

272 **General design of numerical experiments**

273 In our experiments, we employed the previously discussed datasets and used the mass
274 spectra from each dataset where we separated each into a training set of mass spectra,
275 consisting of 70% of randomly chosen mass spectra, and we used the rest of the mass
276 spectra (30%) to test the prediction models and evaluate the accuracy and the Kappa
277 values. This was repeatedly performed several times to assess the variations of accuracy
278 and Kappa values. For the training datasets, pre-processing of mass spectra was
279 performed similarly on all datasets with the "SignalProcessing()" function of our R
280 package by the following steps: (1) transforming intensities with sqrt method; (2)
281 smoothing with UDWT algorithm; (3) remove baseline with SNIP method; (4) calibrate
282 intensity with TIC method; (5) aligning spectra with LOWESS method and with a
283 tolerance parameter of 5e-04 ppm. Next, intensity peaks of mass spectra were detected
284 using the "PeakDectection()" function of our R package with default parameters
285 (tolerance parameter of 2e-03 ppm, signal-to-noise ratio (SNR) of 3, and
286 halfWindowSize parameter of 11). Mass spectra were normalized according to the
287 maximum intensity value for each mass spectrum. For the testing datasets, the same
288 pre-processing of mass spectra was used except that they were aligned on a reference
289 mass spectrum created from the ones of the training dataset using a tolerance parameter

290 of 5e-04 ppm, a minFrequency parameter of 70% and the "strict" method for the
291 "*binPeaks_method*" argument of the "SignalProcessing()" function. Indeed, the
292 alignment to a reference mass spectrum is important to ensure that the mass spectra in
293 the testing dataset closely match those used during the training of the prediction model.

294 We conducted extensive comparisons by employing the previously discussed datasets
295 2 to 4 and generating subsets of randomly chosen mass spectra of various sizes among
296 them. Each subset has been created by respecting the original proportions of each
297 category in the datasets. To avoid overfitting, we ensured that when multiple replicates
298 of a particular sample were present, they were consistently grouped together within the
299 same subset. Here, our main focus revolved around the training dataset sizes and the
300 advantages of using contemporary ML methods (RF, neural networks, support vector
301 machines with linear kernels and extreme gradient boosting) compared to conventional
302 parametric approaches such as decision rules based on multinomial linear logistic
303 regression or multiple linear regressions, as presented earlier. Additionally, it is
304 important to evaluate the number of mass spectra needed to construct an accurate
305 classifier.

306 # Results and discussion

307 ### Extensive comparison from randomly generated subsets of mass spectra

308 For the datasets (2 to 4) previously described, we randomly drew sets of mass spectra
309 of different sizes. These generated datasets range from 50 mass spectra to 1890 mass
310 spectra. For each of them, a training set consisting of 70% of the mass spectra from the
311 generated dataset and a test set (composed of the remaining 30% of mass spectra) were
312 constituted several times by random draws. For each generated dataset, we evaluated 6
313 different methods available in MSclassifR to select discriminative peaks, which we
314 combined with 6 different methods to predict the different categories of mass spectra
315 on the test sets. All the parameters used for the various functions are in table S4. The
316 36 pipelines were assessed on the test datasets, using both Accuracy and Cohen's Kappa
317 criteria, leading to 8604 assessments (Table S1).

318 Using this methodology, we have made a number of observations on (i) the impact of
319 the *m/z* selection method on the performances of classification models (ii) the impact
320 of dataset size on the selection of *m/z* values and (iii) the global performances of the
321 used pipelines (*m/z* selection + prediction algorithms).

322 The first observation (i) suggests the importance of performing a variable selection step
323 before estimating a classification model. Regardless of the prediction algorithm used
324 and the dataset size, the "cvp" and "mda" *m/z* selection methods significantly improve
325 both Accuracy and Cohen's Kappa when compared to not selecting *m/z* (e.g. estimating
326 the classification model on all the *m/z*) or using the "SelectionVarStat", "RFERF", or
327 "sPLSDA" methods [*Figure 2A*]. This observation must nevertheless be put into
328 perspective on data sets composed of many mass spectra (more than 1000). In such
329 cases, "sPLSDA" and "RFERF" demonstrate similar performance levels to "cvp" and
330 "mda", while the "SelectionVarStat" method's performance declines with increasing
331 dataset size, leading to performances similar to not selecting *m/z* values [*Figure 2A*].

The second observation (ii) pertains to the number of selected *m/z* values. When there are more mass spectra in a training dataset, the selection methods tend to select a greater number of potentially discriminant *m/z* [*Figure 2B*]. In general, "SelectionVarStat" tends to select the most *m/z* values among the different methods, followed by "mda" and "cvp", while "sPLSDA" and "RFERF" tend to select very few *m/z* compared to the other methods [*Figure 2B*]. Selecting more *m/z* can potentially lead to reduced performance due to errors in mass spectra classification as previously observed with "SelectionVarStat". That is why "sPLSDA" and "RFERF" are relatively robust to the dataset size increasing [*Figure 2A*].

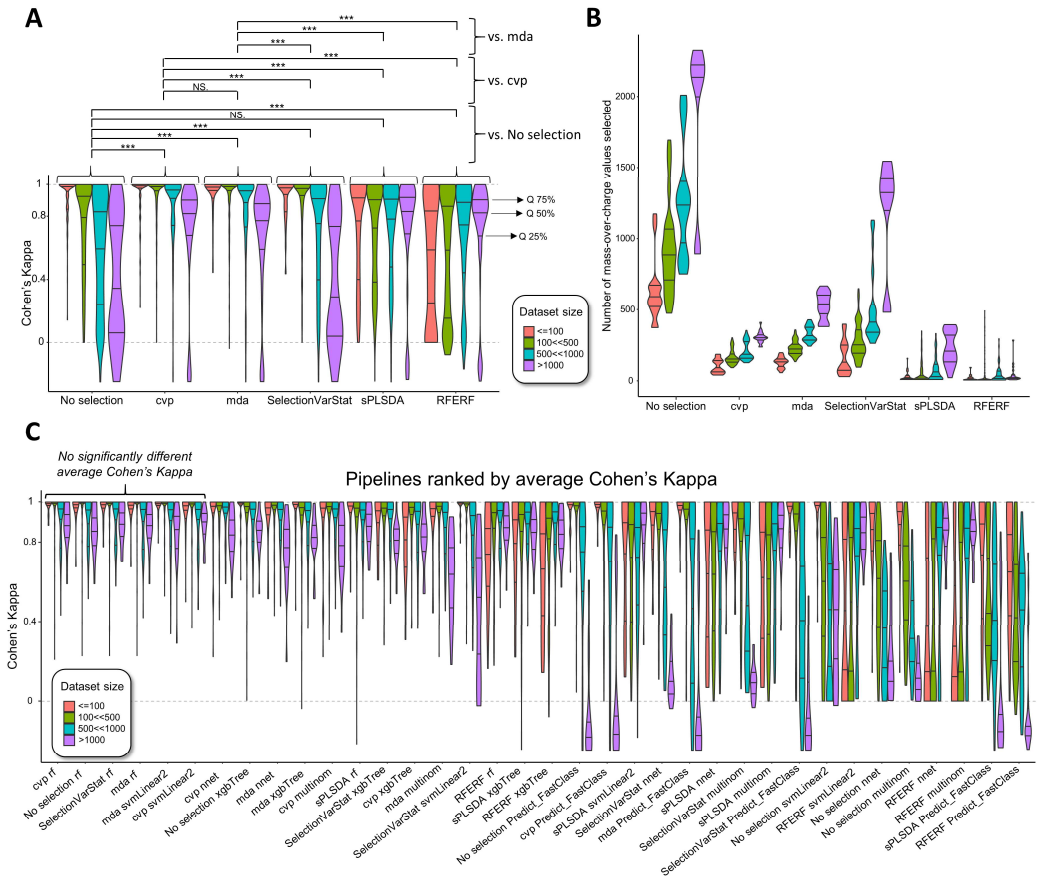

Figure 2. Performances of pipelines implemented in MSclassifR on 239 datasets of various sizes

A. Violin plots of Cohen's Kappa depending on the variable section method used. The dataset sizes are in different colors. **B.** Violin plots of the number of selected *m/z* depending on the variable section method used. The dataset sizes are in different colors. **C.** Violin plots of the Cohen's Kappa depending on the

347 pipeline (variable selection + prediction algorithm) used. Pipelines are ranked from the highest average
348 Cohen's Kappa (left) to the lowest (right) estimated on all the generated datasets. The dataset sizes are
349 in different colors. **Notes:** cvp: for cross-validated permutation variables importances; mda: mean
350 decrease in accuracy; multinom: logistic regression; nnet: single-hidden-layer neural network; No
351 selection: no *m/z* selection method; Predict_FastClass: multiple regression estimator; RFE-RF: Recursive
352 Feature Elimination with Random Forests; SelectionVarStat: ANOVA test; RF: random forests;
353 sPLSDA: Sparse Partial Least Squares Discriminant Analysis;; svmLinear2: linear Support Vector;
354 xgbTree: eXtreme Gradient Boosting. * *p-value* < 5.10^{-2}, ** *p-value* < 10^{-2}, *** *p-value* < 10^{-3} (estimated
355 using univariate Wilcoxon rank sum test).
356

357 Concerning the global performance of the MSclassifR pipelines (observation (iii)), RF
358 globally outperform other prediction models [*Figure 2C*]. When assessing the
359 effectiveness of variable selection methods in combination with RF, “sPLSDA” and
360 “RFERF” exhibit significantly lower accuracy and Cohen’s Kappa values compared to
361 alternative methods [*Figure 2C*]. Importantly, there is no significant difference in
362 performance when combining RF with the other methods “cvp,” “mda”,
363 “SelectionVarStat”, and “No selection”. These similar performances can be attributed
364 to the inherent functioning of the RF algorithm: it employs a random variable selection
365 process during the construction of each decision tree that can make the selection made
366 upstream obsolete. Interestingly, after performing *m/z* selection with “mda” or “cvp”,
367 support vector machines with linear kernels “svmLinear2” perform similarly to RF and
368 could also be an interesting strategy to apply [*Figure 2C*]. Thus, using “mda” or “cvp”
369 *m/z* selection methods combined with either "rf" or "svmLinear2" often deliver some of
370 the best performances across different dataset sizes, with no significant differences
371 between them. We have also observed that certain strategies perform well on small size
372 datasets but exhibit poor performance on larger ones. Specifically, “Predict_FastClass”
373 associated with “mda”, “cvp”, “SelectionVarStat” or “No selection”, as well as
374 “svmLinear2” combined with “SelectionVarStat” or “No selection”, fall into this
375 category [*Figure 2C*]. While these strategies exhibit competitiveness when confronted
376 with constrained numbers of mass spectra in a training dataset, their performance
377 should be approached cautiously in broader contexts.

378 **Application on another independent dataset**

379 We aimed to determine if the promising pipelines, identified through extensive
380 comparisons, could perform well on a separate dataset (Dataset 1) with its own unique
381 challenges [33]. This particular dataset presents the added complexity of spectra
382 obtained from various MALDI-TOF instruments across different countries. It is worth
383 noting that the original authors applied a calibration function to mitigate inter-
384 laboratory differences in mass spectra, a step we did not replicate in our computational
385 experiments.

In our experiments, we applied the same preprocessing steps to the mass spectra as before. We then rigorously evaluated the models by randomly dividing the spectra into 70% for training and 30% for testing. This process was repeated 15 times, resulting in an average Accuracy of 0.875 and a maximum of 0.925 for the "mda rf" pipeline, which had previously shown good performance (Table S2). Notably, these precision levels closely match those reported in the article related to the dataset, even though we did not perform any calibration. Thus, it shows the ability of MSclassifR to build robust pipelines capable of adapting to different instruments in an automatic way.

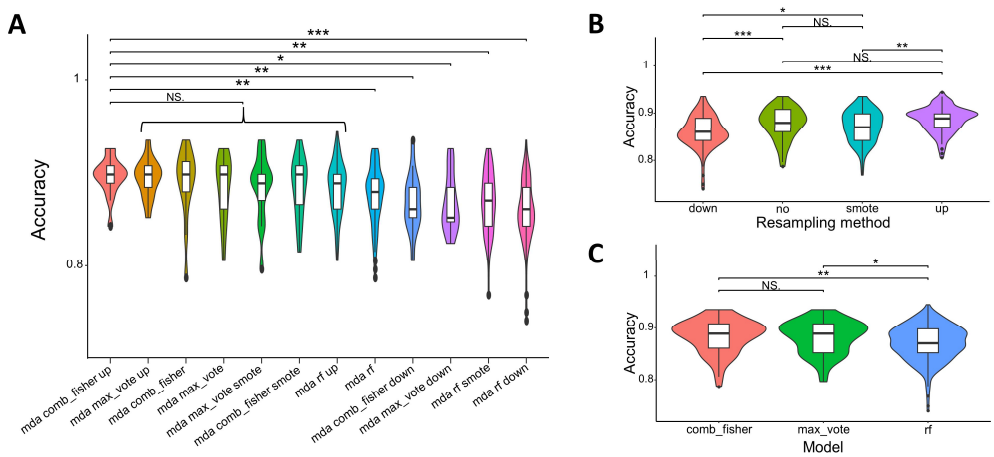

Figure 3. Performances of pipelines implemented in MSclassifR on dataset 1 using 70% of the mass spectra to train the pipelines and 30% to test (15 repeats for each).

A. Violin plots of the Accuracy depending on the pipelines used on Dataset 1. **B.** Violin plots of the Accuracy depending on the resampling method used. **C.** Violin plots of the Accuracy depending on the merging technique used. * *p-value* < 5.10^{-2}, ** *p-value* < 10^{-2}, *** *p-value* < 10^{-3} (estimated using univariate Wilcoxon rank sum test). **Notes:** comb_fisher: combine predictions with Fisher's method; down: down resampling method; max_vote: combine predictions with hard voting system; mda: mean decrease in accuracy; RF: random forests; up: up-resampling method; smote: SMOTE resampling method.

406 To get the most out of MSclassifR, we assessed its capacity to combine prediction
407 probabilities through ensemble-learning methods like maximum voting and Fisher's
408 method. Simultaneously, we examined the potential enhancements that MSclassifR's
409 resampling techniques could bring to our results. For this purpose, we derived five
410 models from the "mda rf" pipeline and merged them after the application of diverse
411 dataset resampling approaches. Our findings reveal that employing up-sampling and
412 merging with either the Fisher's method or the maximum vote technique leads to a
413 significant improvement in overall performance on average (average Accuracy of 0.895
414 and 0.894 respectively, with maximums of 0.926 and 0.935) [*Figure 3A*]. In contrast,
415 the down-sampling method yielded inferior results when compared to the absence of
416 resampling [*Figure 3B*]. In addition, these ensemble-learning approaches make it
417 possible to reduce the variability of Accuracy, and therefore to obtain more robust
418 results [*Figure 3A-C*].

419 **Comparison with a commercial MALDI-TOF MS system**

420 Finally, we also conducted a comparison between all the pipelines within MSclassifR
421 using dataset 3 [34] and a commercial MALDI-TOF MS system (Microflex® LT
422 instrument (Bruker® Daltonics) IVD MALDI Biotyper®). This dataset comprises mass
423 spectra from three genetically narrowly related subspecies of the *Mycobacterium*
424 *abscessus* complex. The data set is particularly challenging because these subspecies
425 are spectrally very close and difficult to distinguish. The subspecies were characterized
426 using commercial molecular methods [34]. The commercial system provided
427 subspecies information for only 60% of the mass spectra, with a correct identification
428 rate of 62% (Table S3). In comparison, using only 70% of all the 1001 mass spectra to
429 train the pipelines and 30% to test them, this is randomly repeated 5 times, the pipelines
430 of MSclassifR provided a subspecies for all the mass spectra of the test set. The best
431 average accuracy was achieved by the "sPLSDA rf" pipeline, with an average accuracy
432 of 96.4% and a standard deviation of 0.029 from 157.4 selected m/z values in average
433 (Table S1). This highlights the ability of MSclassifR to achieve robust classification
434 accuracy, overcoming some limitations of commercial systems.

Conclusion

MSclassifR is a freely available R package, accessible via the CRAN repository network. It is designed for the automatic classification of mass spectra into predefined categories. In this paper, we have rigorously tested the capabilities of our package on datasets acquired from various sources and instruments, including both publicly available data and MALDI-TOF mass spectra obtained in-house. The results from our investigation reveal that MSclassifR can yield prediction models with excellent accuracy and Kappa values. Furthermore, when we compared MSclassifR predictions to those generated by commercial software on in-house dataset, we observed significant performance advantages for MSclassifR prediction models. Our article also delves into an extensive exploration of diverse pipelines for constructing prediction models with our package. From this exploration, we have drawn some important conclusions:

(a) A selection of discriminant *m/z* values prior to the estimation of the prediction model makes it possible to improve the accuracy of predictions.

(b) The original selection method that we developed and based on thresholding the variable importances of RF leads to better accuracy in predictions, whatever the size of the training dataset.

(c) Overall, the prediction method providing the best average accuracies on the used datasets is the RF algorithm.

(d) Some *m/z* selection methods perform poorly on small size datasets (less than 100 mass spectra) and provide good performances on large datasets (more than 1000 mass spectra).

(e) Some prediction methods provide good performances on small size datasets (less than 100 mass spectra) but perform badly on large datasets (more than 1000 mass spectra).

(f) Using ensemble learning and resampling techniques can provide better and more robust performances.

These outcomes are specific to the datasets and the function parameters we used, and their general applicability to any dataset might vary. Nonetheless, they provide valuable

464 insights into creating reliable prediction models from new coming mass spectra
465 datasets.

466 It should be emphasized that MSclassifR can be used in broader contexts than MALDI-
467 TOF mass spectrometry, for example when differentially analyzing quantified values
468 of analytes (e.g. genes, proteins, etc.) between biological conditions. Notably,
469 MSclassifR can be used for the analysis of liquid chromatography coupled to tandem
470 mass spectrometry (LC-MS/MS) data [35]. This technique can also help accurately
471 distinguish closely related microbial variants, which could promote a deeper
472 understanding of resistance mechanisms to antibiotics [36].

473 In conclusion, MSclassifR offers an attractive solution for biology laboratories seeking
474 to identify pathogens from mass spectra. Its open-source nature fosters collaboration,
475 customization, and innovation within the community dedicated to enhancing diagnostic
476 capabilities using mass spectra. We have provided online vignettes that demonstrate
477 how to use the package effectively. We are committed to regular maintenance and
478 continuous improvement of our package.
479

Data and availability

The MSclassifR R package is available online on the CRAN: https://cran.r-project.org/web/packages/MSclassifR/index.html

Links towards different vignettes illustrating how to use the functions of this package from real data sets are also available from this webpage.

All datasets used in this paper are available online on PRIDE and zenodo websites.

Authors' contributions

A.G. and Q.GG. conceived the project. A.G. and Q.GG. designed the algorithms and developed all the programs. M.M., A.A., N.V., E.V. and Y.B. have participated in design of the study and coordination and helped to draft the manuscript. A.G., Q.GG. and K.D performed computational work and analysed the data. All the authors reviewed the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

498 # References

499 1. Hou T, Chuan C, Teng S. Current status of MALDI-TOF mass spectrometry in clinical microbiology.
500 JOURNAL OF FOOD AND DRUG ANALYSIS 2019; 27:404–414

501 2. Brown-Elliott BA, Fritsche TR, Olson BJ, et al. Comparison of Two Commercial Matrix-Assisted
502 Laser Desorption/Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS) Systems for
503 Identification of Nontuberculous Mycobacteria. Am J Clin Pathol 2019; 152:527–536

504 3. Suzuki H, Yoshida S, Yoshida A, et al. A novel cluster of Mycobacterium abscessus complex revealed
505 by matrix-assisted laser desorption ionization-time-of-flight mass spectrometry (MALDI-TOF MS).
506 Diagn Microbiol Infect Dis 2015; 83:365–370

507 4. Pavlovic M, Konrad R, Iwobi AN, et al. A dual approach employing MALDI-TOF MS and real-time
508 PCR for fast species identification within the Enterobacter cloacae complex. FEMS Microbiology Letters
509 2012; 328:46–53

510 5. Godmer A, Benzerara Y, Normand AC, et al. Revisiting Species Identification within the Enterobacter
511 cloacae Complex by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry.
512 Microbiol Spectr 2021; 9:e0066121

513 6. Christner M, Trusch M, Rohde H, et al. Rapid MALDI-TOF Mass Spectrometry Strain Typing during
514 a Large Outbreak of Shiga-Toxigenic Escherichia coli. PLOS ONE 2014; 9:e101924

515 7. Flores-Treviño S, Garza-González E, Mendoza-Olazarán S, et al. Screening of biomarkers of drug
516 resistance or virulence in ESCAPE pathogens by MALDI-TOF mass spectrometry. Sci Rep 2019;
517 9:18945

518 8. Weis C, Cuénod A, Rieck B, et al. Direct antimicrobial resistance prediction from clinical MALDI-
519 TOF mass spectra using machine learning. Nat Med 2022; 28:164–174

520 9. Bruyne KD, Slabbinck B, Waegeman W, et al. Bacterial species identification from MALDI-TOF
521 mass spectra through data analysis and machine learning. Systematic and Applied Microbiology 2011;
522 34:20–29

523 10. Khot PD, Fisher MA. Novel approach for differentiating shigella species and escherichia coli by
524 matrix-assisted laser desorption ionization-time of flight mass spectrometry. Journal of Clinical
525 Microbiology 2013; 51:3711–3716

526 11. Desaire H, Hua D. Adaption of the Aristotle Classifier for Accurately Identifying Highly Similar
527 Bacteria Analyzed by MALDI-TOF MS. Analytical Chemistry 2020; 92:1050–1057

528 12. Asakura K, Azechi T, Sasano H, et al. Rapid and easy detection of low-level resistance to vancomycin
529 in methicillin-resistant Staphylococcus aureus by matrix-assisted laser desorption ionization time-of-
530 flight mass spectrometry. PLOS ONE 2018; 13:e0194212

13. Gibb S, Strimmer K. Maldiquant: A versatile R package for the analysis of mass spectrometry data. Bioinformatics 2012; 28:2270–2271

14. Palarea-Albaladejo J, Mclean K, Wright F, et al. MALDIrppa: quality control and robust analysis for mass spectrometry data. Bioinformatics 2017; 34:522–523

15. Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software 2008; 28:1–26

16. Rohart F, Gautier B, Singh A, et al. mixOmics: an R package for 'omics feature selection and multiple data integration. PLOS Computational Biology 2017; 13:1–19

17. . Clover Data Analysis. Clover MS Data Analysis

18. . Bruker Daltonics ClinProTools 2.0 Download (Free trial)...

19. Savitzky Abraham, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem. 1964; 36:1627–1639

20. Mallat S. A Wavelet Tour of Signal Processing Academic Press, New York. 1998;

21. Ryan CG, Clayton E, Griffin WL, et al. SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. Nuclear Inst. and Methods in Physics Research, B 1988; 34:396–402

22. Sauve AC, Speed TP. Normalization, baseline correction and alignment of high-throughput mass spectrometry data. Proceedings Gensips 2004; 1–4

23. Andrew AM. Another efficient algorithm for convex hulls in two dimensions. Information Processing Letters 1979; 9:216–219

24. Dieterle F, Ross A, Schlotterbeck G, et al. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics. Anal. Chem. 2006; 78:4281–4290

25. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. Journal of the American statistical association 1979; 74:829–836

26. Eilers PH, Marx BD. Flexible smoothing with B-splines and penalties. Statistical science 1996; 11:89–121

27. Leys C, Ley C, Klein O, et al. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of experimental social psychology 2013; 49:764–766

28. Friedman JH. A variable span smoother. 1984;

29. Stanford TE, Bagley CJ, Solomon PJ. Informed baseline subtraction of proteomic mass spectrometry data aided by a novel sliding window algorithm. Proteome Science 2016; 14:19

563 30. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using
564 random forest. BMC Bioinformatics 2006; 7:3

565 31. Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series
566 B (Statistical Methodology) 2002; 64:479–498

567 32. Janitza S, Celik E, Boulesteix A-L. A computationally fast variable importance test for random forests
568 for high-dimensional data. Advances in Data Analysis and Classification 2018; 12:885–915

569 33. Nachtigall FM, Pereira A, Trofymchuk OS, et al. Detection of SARS-CoV-2 in nasal swabs using
570 MALDI-MS. Nat Biotechnol 2020; 38:1168–1173

571 34. Godmer A, Aubry A, Gianetto QG, et al. Spectral database of the subspecies of the Mycobacterium
572 abscessus complex (MALDI-TOF Mass Spectrometry). 2021;

573 35. https://agodmer.github.io/MSclassifR_examples/Vignettes/Vignettemsclassifr_DAv3.html

574 36. Roux-Dalvai F, Gotti C, Leclercq M, et al. Fast and accurate bacterial species identification in urine
575 specimens using LC-MS/MS mass spectrometry and machine learning. Mol Cell Proteomics 2019;
576 18(12): 2492-2505

577