# Investigating Time Series Classification Techniques for Rapid Pathogen Identification with Single-Cell MALDI-TOF Mass Spectrum Data

Christina Papagiannopoulou[1]✉, René Parchen[2], and Willem Waegeman[1]

[1] Department of Data Analysis and Mathematical Modelling, Ghent University,
Ghent, Belgium
{christina.papagiannopoulou, willem.waegeman}@ugent.be
[2] BiosparQ B.V., Leiden, the Netherlands
parchen@biosparq.nl

**Abstract.** Matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry (MALDI-TOF-MS) is a well-known technology, widely used in species identification. Specifically, MALDI-TOF-MS is applied on samples that usually include bacterial cells, generating representative signals for the various bacterial species. However, for a reliable identification result, a significant amount of biomass is required. For most samples used for diagnostics of infectious diseases, the sample volume is extremely low to obtain the required amount of biomass. Therefore, amplification of the bacterial load is performed by a culturing phase. If the MALDI process could be applied to individual bacteria, it would be possible to circumvent the need for culturing and isolation, accelerating the whole process. In this paper, we briefly describe an implementation of a MALDI-TOF MS procedure in a setting of individual cells and we demonstrate the use of the produced data for the application of pathogen identification. The identification of pathogens (bacterial species) is performed by using machine learning algorithms on the generated single-cell signals. The high predictive performance of the machine learning models indicates that the produced bacterial signatures constitute an informative representation, helpful in distinguishing the different bacterial species. In addition, we reformulate the bacterial species identification problem as a time series classification task by considering the intensity sequences of a given spectrum as time series values. Experimental results show that algorithms originally introduced for time series analysis are beneficial in modelling observations of single-cell MALDI-TOF MS.

**Keywords:** MALDI-TOF MS · single-cell spectrum · single-ionization-event · classification · machine learning methods · bacterial species identification · time series

# 1 Introduction

In the diagnostics of infectious diseases, matrix-assisted laser desorption/ ionization-time-of-flight mass spectrometry (MALDI-TOF-MS) is used to identify the causative organism of an infection as a first step in establishing an antibiotic therapy. Owing to its ease of use, its reliability and the low cost of ownership, the introduction of MALDI-TOF-MS revolutionized the diagnostics of infectious diseases during the last decade [4]. Specifically, this technology, applied to bacteria, generates a mass spectrum exhibiting peaks of a limited number of (household) proteins (ribosomal proteins) and peptides produced by the organism. Since the extracted structure of these proteins depends on the species, the MALDI mass spectrum is used as a signature enabling identification of microorganisms up to species level.

For a highly reliable classification result, a significant amount of biomass is required in MALDI. However, for most samples used for diagnostics of infectious diseases, the amount of the obtained biomass is often very low. As a consequence, amplification of the bacterial load is required by culturing. Furthermore, since there is almost no part in the human body that does not contain some form of a natural flora consisting of microorganisms, a sample 'harvested' from a patient sample will generally contain a mixture of organisms, optionally including the organism responsible for the infection. Since interpreting samples containing bacterial mixtures is still in an experimental phase [25], for routine diagnostics, isolation of the causative organism is required as well. Thus, even though the time required for a MALDI based identification is extremely short, in terms of seconds or minutes, since it is dominated by the culturing process, the time-to-result is still in the order of multiple hours (over-night culture) to days. Furthermore, for choosing the culture medium/conditions and for choosing the colony(ies) to identify, a-priori knowledge on the cause of the infection is required. If the MALDI process can be applied to individual bacteria, it would be possible to circumvent the need for culturing and isolation, accelerating the whole process.

Separating a patient sample into a stream of individual cells is possible using the technology developed by [26], originally aimed at dispensing individual eukaryotic cells into well plates. In another work, researchers [24] developed an aerosol TOF MS, able to apply MALDI to individual aerosol particles and demonstrated that recognizable spectra could be accumulated from spectra of large numbers of aerosol particles containing pure proteins only. By combining the cell dispensing technology introduced by [26], with the aerosol TOF technology of [24], the desired single-cell capability can be realised. BiosparQ in the Netherlands developed an instrument, called Cirrus D20, together with the appropriate protocols that is able to produce an information rich signature of bacteria based on accumulated spectra. In this paper, we evaluate this single-cell MALDI-TOF MS methodology and we demonstrate the use of the single-particle spectra for the application of pathogen (bacterial species) identification.

The classification of single-cell bacterial fingerprints is not a trivial process even for human annotators. Thus, MALDI-TOF single cell spectrum analysis should be combined with statistical and machine learning methods. Previous studies have focused on the statistical analysis of accumulated spectra (i.e., mass spectra formed by averaging multi-cell mass spectra) [14, 21, 11, 5]. For instance, in the work of [11], machine learning algorithms, such as SVMs [7] and RFs [6], have been exploited for bacterial identification from accumulated MALDI-TOF mass spectra. For species identification, machine learning methods have been also successfully applied on other representations coming from flow cytometry [2, 19] and Raman spectroscopy [22, 17, 23] data. In this work, we focus on the analysis of MALDI-TOF single-cell spectra for rapid species identification using machine learning techniques. Unlike previous studies, e.g., [21, 11], instead of only applying general purpose machine learning techniques, we also experimented by framing the problem as a time series classification task. In particular, by mapping mass-over-charge (M/Z) ratios to the time axis, we consider the sequences of the different intensities in a spectrum as time series values. This way, standard time series classification methods [1] can be applied. To the best of our knowledge, this is the first time that machine learning approaches and time series classification methods are being applied on single-cell MALDI-TOF data.

To sum up, the contribution of this paper is two-fold. Based on the implementation of the MALDI-TOF-MS methodology in a single-cell setting described here, we (i) experimentally prove that the single-cell signatures, produced by this MALDI-TOF-MS implementation, are informative in distinguishing different bacterial species by using machine learning data analysis, and (ii) find that algorithms originally introduced for time series analysis are beneficial in modelling observations of single-cell MALDI. As such, we believe that the use of single-cell MALDI-TOF-MS data combined with an accurate modelling approach comprises a solid framework that strives to solve the problem of fast pathogen identification (in terms of minutes or seconds), revolutionizing current state-of-the-art approaches.

## 2 Materials and Methods

### 2.1 Bacterial Species

The strains used in this study were provided by the Leiden Centre for Applied Bioscience. They are selected from a group of (opportunistic) pathogens, comprising of *Citrobacter freundii* (*C. freundii*), *Citrobacter koseri* (*C. koseri*), *Enterobacter aerogenes* (*E. aerogenes*), *Klebsiaella oxytoca* (*Kl. oxytoca*), often responsible for common and frequent infections such as urinary tract infections. The identity of these strains is established by evaluation of the cultures on a bioMeriéux Vitek MS MALDI instrument.

### 2.2 MALDI-TOF Mass Spectrometry

Performing mass spectrometry of larger molecules, such as proteins and peptides, is not evident. Generally, the amount of energy associated with direct ionization of the analyte exceeds the disintegration energy of the analyte molecules. Since, the information content of the resulting molecular debris is low, a soft ionization technique (such as MALDI), that leaves the molecules intact, is essential.

In MALDI, the analyte (proteins and peptides in case of identification of microorganisms), is co-crystalized with an organic substance generally containing an aromatic ring (the so-called MALDI matrix) and illuminated using a pulsed UV laser [10]. Upon absorption of the UV light by the MALDI matrix, part of the MALDI-matrix/analyte mixture is ablated into a plume comprising of analyte molecules, matrix molecules, molecular debris and clusters of molecules. Through a number of, currently still only partly understood, interactions, during this process, protons are transferred to the analyte molecules (see e.g. [15]). That way, charge separation can be achieved with the minimum amount of energy. An electric field accelerates the ions into a field-free drift region. There, the separation of the ions according to their mass-over-charge (M/Z) ratio is achieved.

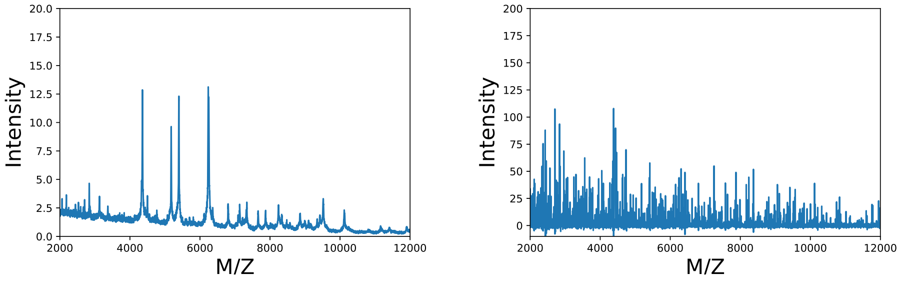
### 2.3 Single-Cell MALDI-TOF Mass Spectrometry

**Implementation of Single-Cell MALDI** To enable application of MALDI to individual bacteria, three aspects of the conventional MALDI logistics need to be changed. Specifically, (i) instead of preparing a large number of cells on a target plate, individual cells need to be made suitable for applying MALDI, (ii) instead of presenting a large number of cells (on a target plate), individual cells need to be presented to the mass spectrometer, and (iii) instead of classifying an accumulated spectrum resulting from a large number of ionization events (applied on a large number of positions within a spot on the target plate), spectra resulting from a single ionization event (applied on a single cell) need to be classified. For the current application, (i) single cells are prepared using a single-cell dispensing technology developed by [26], and (ii) single cells are presented to the mass spectrometer and ionized using an aerosol TOF (ATOF) technology developed by [24]. The current paper concentrates on the third aspect, classifying spectra resulting from single ionization events applied to individual cells.

**MALDI-TOF MS Procedure** In the ATOF mass spectrometer built by BiosparQ BV, each individual particle, which is formed based on a technique described by [26], is illuminated in flight with a pulsed UV laser (337 nm), after which the ions that are produced by the MALDI process are accelerated into the time-of-flight tube. The ions are detected at the end of the time-of-flight tube, using an electron multiplier. For each illuminated particle, the resulting data is recorded as a time series in binary format. The relation between the time-of-flight, and the mass over charge ratio of the ions, $M/Z$, is given by:

$$M/Z = \left(\frac{T_{TOF} - C_2}{C_1}\right)^2$$

The values of the calibration coefficients $C_1$ and $C_2$ are established by calibrating the mass spectrometer using a sample containing a known organism (E-coli,



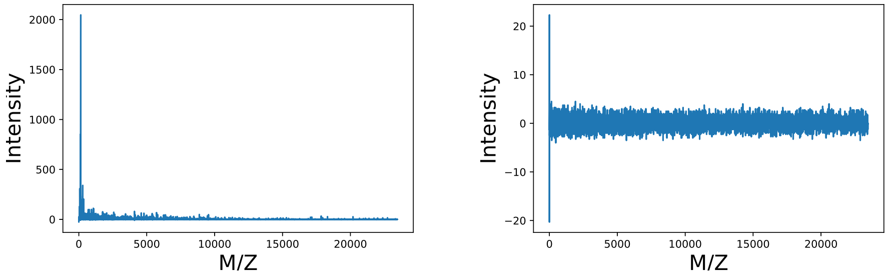
(a) (b)

Fig. 1: Accumulated and single-cell spectra. (a) Accumulated spectrum of multiple single-ionization-event spectra of the species *Kl. oxytoca*, and (b) single-ionization-event mass spectrum of *Kl. oxytoca*. Note that only the mass range above 2000 Da is considered throughout the analysis. See text for more details.

ATCC 25922) and aligning the resulting spectrum with a reference spectrum of E-coli ATCC 25922 recorded on a Vitek MS MALDI mass spectrometer.

**The Structure of Single-Ionization-Event Mass Spectra** Apart from a signature that stems from the analyte molecules, MALDI spectra generally contain signal, or clutter, caused by imperfect ablation of the analyte-MALDI-matrix mixture, leading to (i) charged clusters of molecules and (ii) disintegration of large molecules (and thus leading to charged molecular debris). Note that in single-ionization-event spectra the amplitude of the clutter signal may be of the same order as the analyte signal. However, since the formation of clusters of molecules and molecular debris is a highly stochastic process (while the presence of analyte molecules clearly is not), the expected value of the clutter signal is significantly lower than the one of the analyte signal.

Accumulation of a large number of single-ionization-event spectra will therefore lead to an accumulated spectrum showing high amplitudes at the locations corresponding to the analyte mass molecules and significantly lower amplitudes at intermediate (clutter) locations. By using straightforward baseline correction algorithms, it is possible to remove the clutter contribution from the spectra. However, in single-ionization-event mass spectrum classification, the difference in the clutter and analyte stochastics cannot be used, and a different strategy must be employed. To illustrate the difference between an accumulated signal and a single-cell signal, we depict an accumulated spectrum of the species *Kl. oxytoca* formed by multiple single-ionization-event spectra (Fig. 1a) and a single-ionization-event mass spectrum of *Kl. oxytoca* (Fig. 1b), respectively. From these two figures, it becomes clear that the classification of accumulated spectra is an easier task compared to the classification of single-cell spectra, which is hard even for a human annotator. In our analysis, only the M/Z range above 2000 Da is considered. This is because the well conserved proteins (and peptides) (i.e.,



(a) (b)

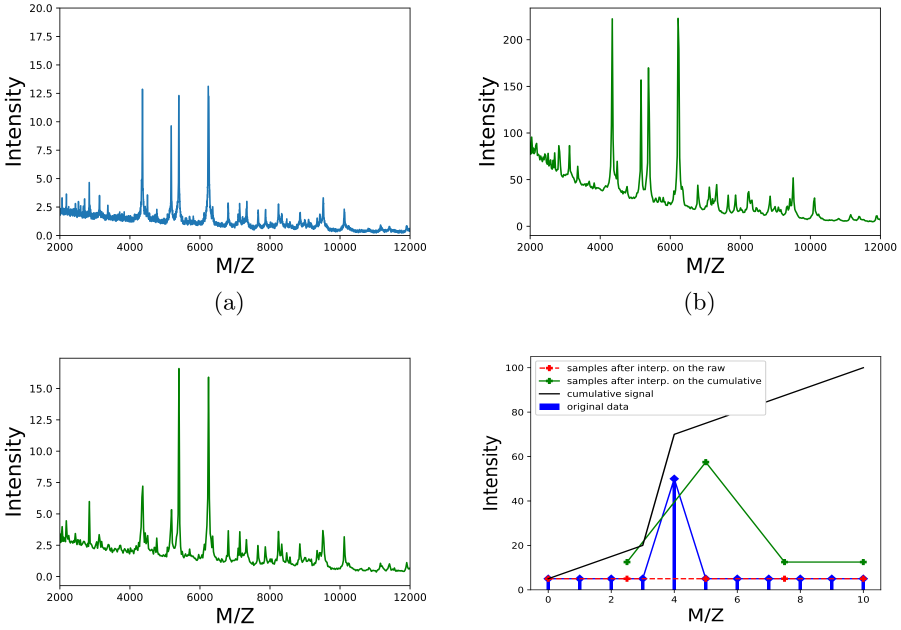
Fig. 2: Informative and non-informative data observations. (a) Informative spectrum of a single-ionization-event of a single *C. freundii* particle, and (b) spectrum of an empty particle (non-informative spectrum).

not depending strongly on the growth medium and growth circumstances) are located in this part of the signal.

### 2.4 Data Pre-processing

Using expression (1) and the associated calibration coefficients the single-particle time series are converted into mass spectra. These single-particle spectra are finally normalized using the Root-Mean-Squared (RMS) voltage of the measurement chain noise.

As a first preprocessing step, we remove particles (observations) that do not contain any information. To do so, we calculate the variance of each particle. If the calculated variance is low, there are no intensities captured by the ionization procedure. In Fig. 2b, an observation of an empty particle is depicted. In this paper, we demonstrate that the single-cell signals, produced by the aforementioned procedure, form a valid signature of the most common bacterial species. To this end, machine learning classifiers are employed to identify bacterial species from single-cell signals. Following the standard practices of machine learning methods, signals of various species should be aligned in a common feature space (i.e., for the same values of M/Z ratios). This is because intensities for the various species may be measured in different M/Z ratios. Therefore, we employ an interpolation technique as a feature construction approach in order to form a common feature representation for all the studied species. The interpolation procedure followed in our experimental study comprises four steps: (i) M/Z values are defined (these values should be in between the maximum and minimum M/Z values existing in the dataset), (ii) a cumulative spectrum of each individual spectrum is formed, (iii) linear interpolation is performed on the cumulative spectrum, (iv) the interpolated values are differentiated and the final signal is produced. The number of M/Z values (bins) defined in step (i) is a tunable parameter that is optimized during the learning process (model training).



(a) (b)

(c) (d)

Fig. 3: Comparison of the different interpolation approaches. (a) Original accumulated spectrum of *Kl. oxytoca*, (b) accumulated spectrum after linear interpolation on the cumulative spectra, (c) accumulated spectrum after linear interpolation on the original spectra, and (d) toy example of the two interpolation techniques. The original signal is represented with blue bars, the linear interpolated signal on the original spectrum is represented in red color, and the linear interpolated signal on the cumulative spectrum is colored in green. The corresponding cumulative spectrum is drawn in black color.

We also experimented by applying the interpolation method directly on the original spectrum values, resulting in low values for the characteristic peaks of the spectra. Figure 3a shows the mean spectrum of *Kl. oxytoca* observations before the interpolation (raw data), Fig. 3b depicts the same mean spectrum after the interpolation procedure on the cumulative spectrum, while Fig. 3c presents the result of the interpolation on the raw spectrum. The number of bins (M/Z values) selected in these examples (Fig. 3b and 3c) is equal to 1000, while the original signal includes intensities for ∼16000 bins. It is observed that the signals in Fig. 3c have been mostly affected by the interpolation, and the peaks of the average spectrum are not well-formed as in the original data (see Fig. 3a). On the other hand, Fig. 3b shows an interpolated signal that is close to the original one (with respect to the high peaks). Although the intensity scale increases due to the cumulative information that each bin captures, by employing interpolation on the cumulative spectrum, the peaks of the original signal are well-preserved. This is important, since these peaks constitute the informative part of a signal. In Fig. 3d, a toy example of a single-cell spectrum, where a peak is not captured using simple linear interpolation on the original signal, is illustrated. As mentioned above, in single-cell spectra there are no well-shaped peaks (see Fig 2a for an example) and thus, possible spikes in the signal should be preserved.

### 2.5 Machine Learning Methods

We conduct an extensive experimental study on the task of species identification from single-cell MALDI-TOF data. In machine learning, species identification can be formulated as a typical classification task. In mathematical notation, an observation in this task is symbolized as a pair $(\mathbf{x}, y)$, where $\mathbf{x}$ is a feature vector of length $m$, i.e., $(x_1, ..., x_m)$ ($m$ values), and $y$ is a class value. Following this notation, a dataset $D$, in a classification task, consists of $N$ observations with their associated class labels, i.e., $D = \{\mathbf{X}, \mathbf{y}\} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$. The vector $\mathbf{y}$ represents a discrete class variable with $c$ possible values. In our setting, the observations are the various spectra, the intensities are the values of the $m$ features (bins) and the bacterial species the different class values. We evaluate the results of seven machine learning classifiers, four of them are well-known algorithms that have been broadly applied on MALDI-TOF data, namely random forests (RFs), support vector machines (SVMs), logistic regression (LR), and k-nearest neighbor (KNN). The other three algorithms, originally developed for the task of time series classification (described below), are time series forests (TSF), Bag-of-SFA-Symbols (BOSS) and complexity invariant distance (CID).

The first group of models consists of algorithms that can distinguish non-linearly (e.g., RF) and linearly (e.g, LR) separable data. Some of these algorithms are also able to easily handle high-dimensional data (e.g., RFs), while others are not (e.g., KNN). This is the very first time that a predictive modelling benchmarking is performed on this kind of data and thus, an extensive experimental setup of the most well-known predictive models is necessary to set a strong baseline for future studies. Regarding the various models, the RF [6] is the well-known non-linear ensemble algorithm. The SVMs [7] used in this set-up apply a Gaussian RBF kernel to model non-linear boundaries between the different species. The LR algorithm [9] is a simple method that learns linear boundaries between the classes, while the KNN model [8] serves as an intuitive non-linear baseline algorithm.

The second group of models consists of algorithms that are extensively used in the time series classification task. The aim of a time series classification task is the assignment of a given time series to a particular class. As in other classification problems, a classifier is a function or mapping from the input space to the class values. The only difference from the general classification task, described above, is that the feature vector $\mathbf{x}$ of length $m$ is a time series of the same length. In our setting, the observations of the single-cell MALDI-TOF dataset can be seen as time series, since the ion intensities are consecutive values over the M/Z axis (see Fig. 2a). The class values are again the various bacterial species. To the best of our knowledge, time series classification methods have not yet been investigated on MALDI-TOF data. As such, in the next paragraph, the time series classification methods used in this study are briefly introduced.

Time series classification algorithms enable some automatic feature construction or filtering of the time series values prior or during constructing the classifier. That way, these methods extract high-level representations for the time series or use similarity metrics for measuring the relatedness between the time series [16, 13, 1]. Time series classification algorithms can be categorized as methods that use: (i) the whole series or the raw data for classification – this family of algorithms mainly consists of one-nearest-neighbour-type (1-NN) classifiers with varying distance measures, (ii) sub-intervals of the raw time series – summary statistics of these sub-intervals are commonly used as discriminative features in this family of classifiers, and (iii) the frequency of the patterns in a given time series – a dictionary of patterns is formed and a histogram for each observation is calculated based on the constructed dictionary by this kind of methods.

In this study, we compare three algorithms, one from each of the aforementioned categories. The complexity invariant distance (CID) [3] algorithm belongs to the first category. This classifier defines the concept of complexity invariance in time series. Intuitively, complex time series are characterized by many peaks and valleys. The distance between pairs of complex time series is frequently greater than the distance between pairs of simple time series (i.e., without many peaks and valleys). A complexity invariant distance measure has been introduced to compensate this phenomenon. Specifically, a distance measure is multiplied by a term that is calculated based on the sum of squares of the first differences of the time series. The Euclidean distance measure has been used as base distance measure from the CID algorithm.

A representative method of the second category is the time series forests (TSF) [12]. This method is similar to the RF model, because it consists of a set of classification trees. Specifically, each tree is trained by using summary statistics (i.e., the mean, standard deviation and slope) of random sub-intervals derived from the times series observations. The calculated summary statistics serve as discriminative features. The classification of a new observation is obtained by majority voting over all trees.

Bag of SFA symbols (BOSS) [20] is an algorithm that belongs to the third category of the methods mentioned above. It starts by creating a dictionary of patterns from the given time series observations. The frequencies of the patterns of this dictionary are used as discriminative features. The different patterns are constructed by using time series sub-intervals in a sliding window setting. Then, the discrete Fourier transform (DFT) is performed on each sub-interval window. Afterwards, the calculated Fourier coefficients are transformed into categorical values (e.g., 'a', 'b', 'c') based on their quantity (i.e., 'high', 'medium', 'low'), and thus, the patterns of the dictionary are formed by the combination of the categorical values of the Fourier coefficients. Finally, each time series observation is represented based on the frequency of the calculated patterns in the time series

Table 1: Predictive performance on the test set in terms of mean accuracy for the compared predictive models. For each model, the optimal value of the parameter number of bins, which is tuned during the training phase, is also reported.

| Model | Number of bins | Mean accuracy |
|---|---|---|
| LR | 9,000 | 0.760 |
| RF | 9,000 | 0.727 |
| SVM | 500 | 0.763 |
| KNN | 500 | 0.633 |
| TSF | 8,500 | **0.832** |
| CIDNN | 500 | 0.666 |
| BOSS | 7,500 | 0.478 |

itself. The classification of a new observation is performed by using an 1-NN-based classifier.

### 2.6 Experimental Setup

In our experiments, we evaluate the predictive performance of the seven aforementioned classifiers (see Sect. 2.5). Specifically, each classifier is assessed based on its ability to distinguish single-cell spectra of four species, namely *C. koseri*, *C. freundii*, *E. aerogenes* and *Kl. oxytoca*. We use the same train/test splits to obtain a fair comparison between the various tested algorithms. In particular, after the removal of the empty particles (see Sect. 2.4), we keep 1000 examples for each species for training and the 268 examples for test (for each species). Parameter tuning is performed in a separate validation set, which is part of the training set. For all the models, the number of bins is considered as a tunable parameter with a tested range [500, 10000] with steps of 500. For the algorithms LR, RFs, SVMs and KNN, the sklearn [18] python implementation is used, while for the CIDNN, BOSS and TSF the java implementation of [1] is evaluated. For the evaluation procedure we report the mean accuracy for all the species and we present the confusion matrix of the best performing algorithms for further discussion (see Sect. 3).

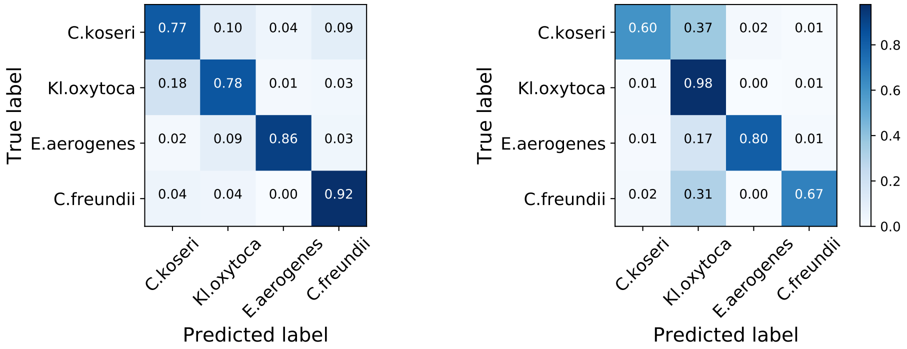
## 3 Results and Discussion

In this section, the performance of the classification algorithms is presented. Table 1 shows the predictive performance of the seven classification algorithms in terms of the mean accuracy (over the four classes). The number of bins (features), which is a tunable parameter for each method, is also noted. The first block of algorithms (i.e., LR, RFs, SVMs, and KNN) consists of well-known methods generally-applied in many applications. On the other hand, the second block of methods (i.e., TSF, CIDNN, and BOSS) comprises models that solve time series classification tasks, as discussed in Sect. 2.5.

Overall, the best performing algorithm is the TSF with mean accuracy of 0.832. TSF is considered as a phase dependent classification method. This is due to the fact that it detects the intervals among all observations that are most informative for the classification problem. In our setting, the peaks and valleys, which characterize a bacterial species, are strongly associated with the corresponding M/Z ratios. Therefore, the exact same extracted intervals should be compared across all the observations. This means that comparisons with shifted peaks (or valleys) may lead to incorrect classification results, because spectra of different species may consist of similar signatures (yet not identical). As stated in Sect. 2.5, TSF constructs features from the original time series by calculating summary statistics from arbitrary-sized sub-intervals. This means that it captures information from short-sized intervals till long-sized ones that can even cover entire peaks (and valleys) of the signals. Thus, this combined information (from short and long sub-intervals) comprises a “high-level” representation of the original spectra, imitating the human perception about the spectra. After the feature construction process, TSF selects the features with the maximum discriminative power, by building various weak classifiers (classification trees). This procedure is also combined with majority voting for prediction, ensembling the resulting predictions, a technique that is often beneficial in classification tasks. That way the algorithm is able to explore the discriminative power of more intervals and avoid overfitting.

A high performance ($> 0.72$) is obtained also using other classifiers, such as the SVMs, LR and RFs. These classifiers take combinations of features into account to perform their predictions. This is beneficial in our setting, since combinations of peaks and valleys in the spectra are informative for distinguishing the various bacterial species. Both linear (LR) and non-linear (SVM and RF) models perform similar in this scenario. However, compared to the TSF algorithm, these models are not based on consecutive values (intervals) of a spectrum and thus, are less able to capture information coming from large intervals. On the contrary, TSF encodes the information that the rest of the algorithms do, since it also generates and assesses short intervals (of, e.g., three or four consecutive values). That way, TSF combines information from short and long intervals of the spectrum, and outperforms the other models.

The algorithms that are based on the intuition of the nearest neighbor (NN), i.e., KNN and CIDNN, are outperformed by the aforementioned ones, while the BOSS model gives the worst predictive performance. This result is not surprising, since the NN algorithms are not able to generalize well when the number of features is high (curse of dimensionality). This is the reason why the optimal number of bins is low for these models (i.e., 500) compared to the corresponding number of bins for other models. In addition, the result of the BOSS algorithm is low due to the phase independent features that are extracted during training. As discussed in Sect. 2.5, the BOSS method constructs a dictionary of patterns, and counts the times that these patterns occur in a particular observation. These patterns may occur at any point of the spectrum. Thus, the model counts the



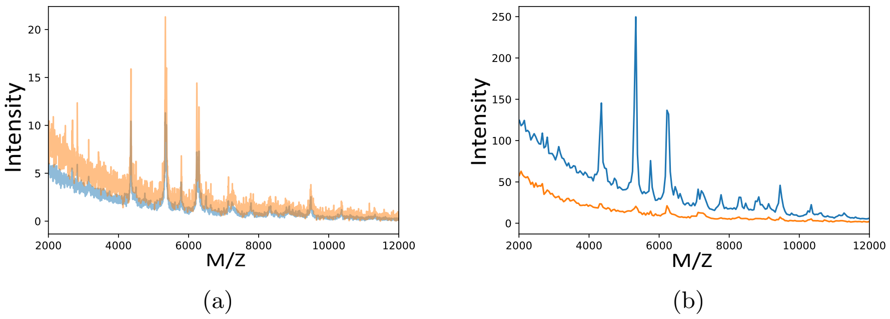
(a) (b)

Fig. 4: Normalized confusion matrices for (a) the TSF, and (b) the SVM algorithm.

presence of the "spectral" patterns by treating them independently from their corresponding M/Z ratios.

In Fig. 4, the confusion matrices of the two best performing methods are presented. Specifically, Fig. 4a and b show the accuracy of the TSF and the SVM model per species, respectively. Overall, both models have high discriminative power for all the species. However, the TSF model performs similar for each of the species, while the SVM model performs really high for two species (*Kl. oxytoca* and *E. aerogenes*) and relatively low for the rest of the species (i.e., *C. koseri* and *C. freundii*). In addition, the spectra of the different species are mostly confused with the spectra of *Kl. oxytoca*. This is especially clear for the *C. koseri* species, for which the model gives the lowest accuracy. The confusion with the *Kl. oxytoca* species can be explained by the fact that the spectra of this species includes peaks with low intensities and thus observations with low peaks (from this or other species) are classified as *Kl. oxytoca* observations. Fig. 5b confirms this conclusion. The mean spectrum of the misclassified *C. koseri* observations is depicted with orange, while the mean spectrum of the correctly classified observations in blue. Most of the *C. koseri* observations have been confused with the *Kl. oxytoca* observations. More specifically, the misclassified *C. koseri* observations are the ones with low intensity values. Therefore, the clearer the signal peaks (high intensities), the better the result of the species classification task for the SVM classifier. Fig. 5a depicts the corresponding misclassified/correctly classified mean of the spectra for the TSF model in orange and blue, respectively. The misclassified observations have higher variance and mean values compared to the correctly classified ones. Differences in variance in these two plots are due to a different number of bins used in the TSF (8500 bins) and the SVMs (500 bins) experiments (tuned in validation set).

Note also that for the classification experiments we assume that the quality of the spectra of different species is similar. However, it is known that even between



(a)

(b)

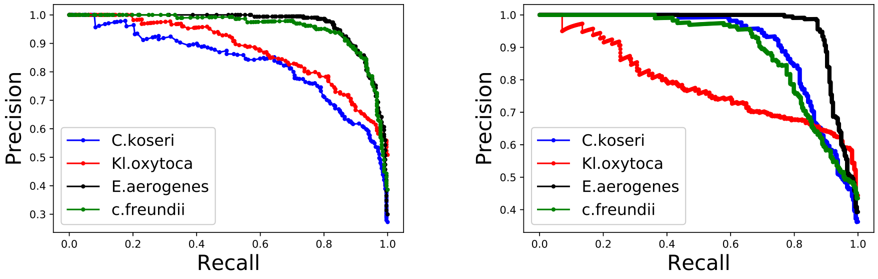
Fig. 5: Accumulated spectrum of the correctly and incorrectly classified particles. The blue accumulated spectrum has been formed by the correctly classified *C. koseri* spectra, while the orange one by the incorrectly classified ones for (a) the TSF, and (b) the SVM models.

successive experiments using the same organism there may be a variation in the intensity of peaks (not the position) in the accumulated spectrum, even when exactly the same protocol has been used. Hence, the difference in performance may be caused by differences in the quality of the spectra. The fact that when using TSF the difference in performance for the different species is less than when using SVN, may indicate that TSF is less sensitive to this type of variation.

From the confusion matrix of Fig. 4b, it seems that the SVM model performs well for the *Kl. oxytoca* species (0.98 accuracy). However, there are many false positive examples that are not taken into account in the calculation of the accuracy metric. The ratio of false positives is incorporated in the estimation of the precision-recall curves, see Fig. 6. Fig. 6 shows the precision-recall trade-off for each of the four species for the TSF (Fig. 6a) and the SVM (Fig. 6b) models. The precision-recall curves for each species have been calculated in a one-versus-all fashion. Fig. 6a shows that for all the species, the precision is above 0.75 for a recall value of 0.8. This means that 80% of the observations (for each species) is identified correctly with accuracy higher than 75%. On the other hand, Fig. 6b illustrates that for the species *C. koseri*, *E. aerogenes*, and *C. freundii*, the SVM model is able to correctly identify more than 75% of the observations with high precision (more than 80%). However, this is not the case for the *Kl. oxytoca* species, where the precision increases (0.8) only when recall drops to 0.3 (or less). This means that the model misclassifies many examples of the other species (i.e., *C. koseri*, *E. aerogenes* and *C. freundii*), when it comes to predict a high ratio of the *Kl. oxytoca* observations.

## 4 Conclusion

In this paper, we described an implementation of a MALDI-TOF MS procedure in a setting of single-ionization-event on individual cells. We demonstrated



(a) (b)

Fig. 6: Precision/recall curves for (a) the TSF, and (b) the SVM models.

the use of the single-cell MALDI-TOF MS data for the application of bacterial species identification. Specifically, we combined the single-cell spectra produced by the described methodology with machine learning algorithms, and we experimentally proved that these signatures are informative in distinguishing different bacterial species. Finally, we formulated the problem of bacterial species classification as a time series classification task and we found that algorithms originally introduced for time series analysis are beneficial in modelling observations of single-cell MALDI-TOF MS. Our conclusions confirm that the use of single-cell MALDI-TOF-MS data combined with an accurate modelling approach comprises a promising and complete framework that gives the green light for fast species identification. The fast response time, which is in terms of minutes or seconds, revolutionizes current time-consuming approaches (due to the dominant culturing time) in pathogen identification related to human infections.

## Acknowledgements

This work is funded by BiosparQ B.V. in the Netherlands. We would like to sincerely thank Dr. Willem van Leeuwen of the Leiden Centre for Applied Bioscience for providing the strains used in this study.

## References

1. Bagnall, A., Lines, J., Bostrom, A., Large, J., Keogh, E.: The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery pp. 1–55 (2016). https://doi.org/10.1007/s10618-016-0483-9
2. Bashashati, A., Brinkman, R.R.: A survey of flow cytometry data analysis methods. Advances in bioinformatics **2009** (2009)
3. Batista, G.E.A.P.A., Keogh, E.J., Tataw, O.M., De Souza, V.M.: CID: an efficient complexity-invariant distance for time series. Data Mining and Knowledge Discovery **28**(3), 634–669 (2014)

4. van Belkum, A., Chatellier, S., Girard, V., Pincus, D., Deol, P., Jr, W.M.D.: Progress in proteomics for clinical microbiology: MALDI-TOF MS for microbial species identification and more. Expert Review of Proteomics **12**(6), 595–605 (2015). https://doi.org/10.1586/14789450.2015.1091731
5. van Belkum, A., Lacroix, B., Veyrieras, J.B., Surre, J., Ramjeet, M., Arsac, M., Perrot, N., Mahé, P., Mailler, S., Chatellier, S., Monnin, V., Girard, V.: Automatic identification of mixed bacterial species fingerprints in a MALDI-TOF mass-spectrum. Bioinformatics **30**(9), 1280–1286 (01 2014). https://doi.org/10.1093/bioinformatics/btu022
6. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (Oct 2001). https://doi.org/10.1023/A:1010933404324, https://doi.org/10.1023/A:1010933404324
7. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (Sep 1995). https://doi.org/10.1023/A:1022627411411, https://doi.org/10.1023/A:1022627411411
8. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory **13**(1), 21–27 (1967). https://doi.org/10.1109/TIT.1967.1053964
9. Cox, D.R.: The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological) **20**(2), 215–232 (1958)
10. Croxatto, A., Prod'hom, G., Greub, G.: Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. FEMS Microbiology Reviews **36**(2), 380–407 (03 2012). https://doi.org/10.1111/j.1574-6976.2011.00298.x
11. De Bruyne, K., Slabbinck, B., Waegeman, W., Vauterin, P., De Baets, B., Vandamme, P.: Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning. SYSTEMATIC AND APPLIED MICROBIOLOGY **34**(1), 20–29 (2011), http://dx.doi.org/10.1016/j.syapm.2010.11.003
12. Deng, H., Runger, G., Tuv, E., Vladimir, M.: A time series forest for classification and feature extraction. Information Sciences **239**, 142–153 (2013)
13. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.: Querying and mining of time series data: experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment **1**(2), 1542–1552 (2008)
14. Hsieh, S.Y., Tseng, C.L., Lee, Y.S., Kuo, A.J., Sun, C.F., Lin, Y.H., Chen, J.K.: Highly efficient classification and identification of human pathogenic bacteria by MALDI-TOF MS. Molecular & Cellular Proteomics **7**(2), 448–456 (2008). https://doi.org/10.1074/mcp.M700339-MCP200
15. Knochenmuss, R.: The Coupled Chemical and Physical Dynamics Model of MALDI. Annual Review of Analytical Chemistry **9**(1), 365–385 (2016). https://doi.org/10.1146/annurev-anchem-071015-041750
16. Liao, T.W.: Clustering of time series data - a survey. Pattern Recognition **38**(11), 1857 – 1874 (2005). https://doi.org/http://dx.doi.org/10.1016/j.patcog.2005.01.025
17. Liu, J., Osadchy, M., Ashton, L., Foster, M., Solomon, C.J., Gibson, S.J.: Deep convolutional neural networks for Raman spectrum recognition: a unified solution. Analyst **142**(21), 4067–4074 (2017)
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

19. Rubbens, P., Props, R., Boon, N., Waegeman, W.: Flow cytometric single-cell identification of populations in synthetic bacterial communities. PLOS ONE **12**(1), 19 (2017), http://dx.doi.org/10.1371/journal.pone.0169754
20. Schäfer, P.: The BOSS is concerned with time series classification in the presence of noise. Data Mining and Knowledge Discovery **29**(6), 1505–1530 (2015). https://doi.org/10.1007/s10618-014-0377-7
21. Schleif, F.M., Lindemann, M., Diaz, M., Maaß, P., Decker, J., Elssner, T., Kuhn, M., Thiele, H.: Support vector classification of proteomic profile spectra based on feature extraction with the bi-orthogonal discrete wavelet transform. Computing and Visualization in Science **12**(4), 189–199 (Apr 2009). https://doi.org/10.1007/s00791-008-0087-z
22. Schmid, U., Rösch, P., Krause, M., Harz, M., Popp, J., Baumann, K.: Gaussian mixture discriminant analysis for the single-cell differentiation of bacteria using micro-Raman spectroscopy. Chemometrics and Intelligent Laboratory Systems **96**(2), 159 – 171 (2009). https://doi.org/https://doi.org/10.1016/j.chemolab.2009.01.008
23. Sevetlidis, V., Pavlidis, G.: Effective Raman spectra identification with tree-based methods. Journal of Cultural Heritage (2018)
24. van Wuijckhuijse, A., Stowers, M., Kleefsman, W., van Baar, B., Kientz, C., Marijnissen, J.: Matrix-assisted laser desorption/ionisation aerosol time-of-flight mass spectrometry for the analysis of bioaerosols: development of a fast detector for airborne biological pathogens. Journal of Aerosol Science **36**(5), 677 – 687 (2005). https://doi.org/https://doi.org/10.1016/j.jaerosci.2004.11.003
25. Yang, Y., Lin, Y., Qiao, L.: Direct MALDI-TOF MS identification of bacterial mixtures. Analytical Chemistry **90**(17), 10400–10408 (2018). https://doi.org/10.1021/acs.analchem.8b02258
26. Yusof, A., Keegan, H., Spillane, C.D., Sheils, O.M., Martin, C.M., O'Leary, J.J., Zengerle, R., Koltay, P.: Inkjet-like printing of single-cells. Lab Chip **11**, 2447–2454 (2011). https://doi.org/10.1039/C1LC20176J