Overcoming Challenges of Reproducibility and Variability for the *Clostridioides difficile* typification

Lucía Bravo Antón†, Alejandro Guerrero-López†‡*, Carlos Sevilla-Salcedo†, Mario Blázquez-Sánchez‡,
David Rodríguez-Temporal‡, Belén Rodríguez-Sánchez‡, Vanessa Gómez Verdejo†‡
†Department of Signal and Communication Theory, Universidad Carlos III de Madrid, ES28911 Spain
‡Instituto de Investigación Sanitaria Gregorio Marañón, ES28040, Spain
*Corresponding author: alexjorguer@tsc.uc3m.es

***Abstract*—The implementation of Matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) mass spectrometry has had a profound impact on clinical microbiology, facilitating rapid bacterial identification through protein profile analysis. However, the application of this technique is limited by challenges related to the reproducibility and variability of spectra, particularly in distinguishing closely related bacterial strains, as exemplified by the typification of *Clostridioides difficile* ribotypes. This thesis investigates the integration of Machine Learning algorithms to enhance the robustness and accuracy of MALDI-TOF spectra analysis. The aim is to compare traditional classifiers in order to gain insight into how spectral variability affects their performance in typification. Furthermore, specific data augmentation techniques for MALDI-TOF spectra are designed to enhance the classification of *C. difficile* ribotypes, to alleviate the inherent variability in MALDI-TOF spectra, and to address the issue of limited sample sizes. The results demonstrate that these methods can significantly enhance the accuracy of classification of *C. difficile* strains, thereby rendering MALDI-TOF a more reliable tool in clinical diagnostics.**

***Index Terms*—MALDI-TOF, Machine Learning (ML), variability, Clostridioides difficile, Peak Information Kernel (PIKE), Random forests (RF), data augmentation, typification, clinical microbiology.**

I. INTRODUCTION

Matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) mass spectrometry (MS) has revolutionized the field of clinical microbiology in the last decade. In the first case, this technique allows rapid and accurate bacterial identification by analyzing their protein profiles in a matter of minutes, unlike traditional testing methods, which can take several days to provide the samples and require specialized personnel trained in molecular biology, including Pulsed Field Gel Electrophoresis (PFGE) and Polymerase Chain Reaction (PCR) Ribotyping [1]. The spectra analysis is based on a visual comparison of a limited number of attributes, including peak height and area under the peak of ribosomal proteins, which are consistently present and useful for differentiating between species. Nevertheless, when antibiotic susceptibility or different types of the same bacterial species are to be detected (typification) this analysis cannot be done visually, but requires ribotyping of the samples to fully characterise them. Recently, Machine Learning (ML) has been employed in conjunction with MALDI-TOF for the purpose of accelerating and reducing the cost of characterisation [1] [2].

This paper will focus on the typification of *Clostridioides difficile*, a gram-positive, anaerobic bacterium that forms spores and is transmitted via the faecal-oral route. This specie is the primary cause of antibiotic-associated infectious diarrhea in hospitals [3] [4], due to certain strains of *C. difficile* can produce toxins that lead to intestinal damage, playing a vital role in the disease's development and progression [5] [6]. Numerous ribotypes (RTs) of *C. difficile* have been identified, some of which are associated with increased pathogenicity and hospital outbreaks [7], as Ribotype B1/NAP1/027 (RT027) [8] [9], which strains are considered hypervirulent due to the hyperproduction of toxins [4]. Other RTs, as hypervirulent as RT027, are the RT078 and RT181, which have recently been reported as highly prevalent and emerging in European countries [10]. One of the primary challenges in hospital is the classification of samples extracted from patients, particularly in instances where some RT spectra are scarce. This may impede the ability to make meaningful comparisons between samples, as it is necessary to know the RT of the patient in order to determine whether the patient should be isolated.

In addition to the lack of samples, especially for hypervirulent RTs, multiple acquisitions of MALDI spectra have extensive variability because such spectra often contain peaks that are inconsistent in their presence or absence and show great variability in intensity [11] [12]. A primary concern regarding the characterization of intact bacteria using MALDI-TOF MS is reproducibility, which stems from the fact that microorganisms express different proteins under different environmental conditions. Therefore, it can be reasonably deduced that the profiles of bacteria may vary depending on the medium used for extraction because different extraction protocols may yield varying amounts and types of proteins, as well as varying time periods or even laboratories in which the spectra have been obtained. In particular, certain studies have investigated the reproducibility within individual labs and across multiple labs [13] [14]. These studies demonstrate that the performance of the laser is dependent on a number of factors, including the intensity of the laser beam, the calibration of the machine, and the maintenance of the equipment. In addition, the expertise of the machine operator personnel is a crucial element to ensure optimal results [12].

In this paper the aim is to investigate the effect of this variability for *C. difficile* typing when MALDI-TOF spectra are processed with ML models, with two main objectives in mind. First, investigate the impact of differences in time of sampling, the medium, and the machine used for acquisition on variability and how this variability affects typification when using different classification methods. The second objective is to establish a robust ML approach that enables classification independent of the spectral acquisition when new samples are introduced. Furthermore, it should be reproducible when samples of certain strains are scarce.

Related work

To extract the hidden information necessary for the prediction of antibiotic resistance and the typing of different strains, researchers have implemented ML algorithms to mine more intricate patterns from the spectra. There are studies focus on typification of *C. difficile* as [15], [16] or [17], which evaluates the performance of traditional classifiers to ribotype *C. difficile* with MALDI-TOF MS using only peaks out of the default region of identification, which covers from 2,000 to 20,000 m/z. In [18] authors develop a novel approach for classification from MALDI-TOF mass spectra that exploits all the spectra information through a kernel, the Peak Information Kernel (PIKE), and with a Gaussian Process (GP) classifier, PIKE is capable of performing sample classification with additional confidence estimates in several species and antibiotic resistance prediction scenarios.

Furthermore, studies have employed data augmentation techniques to address the scarcity of spectra. For instance, [19] introduced random noise to intensities and a shift scale for misalignment in other forms of mass spectrometry, excluding MALDI-TOF.

However, all of these studies are limited by the use of ML algorithms that are not specifically adapted to the application problem, and a focus on detecting single peaks (or small subsets of peaks) for the identification purposes. Furthermore, the impact of variability on ML models is not investigated, which motivates us to analyse the effect of this variability on classifiers, with a particular focus on the typification of *C. difficile*.

II. MATERIALS AND METHODS

A. *MALDI-TOF MS spectra*

A total of 30 *C. difficile* samples were employed in the study, with their MALDI-TOF mass spectra extracted over different conditions to be able to carry out the variability analysis. In particular, they were obtained over a three-week period (week 1, week 2, week 3) and in three different media, which are Brucella, *Clostridium* (Clos), and Fenilalanina Agar (FA). In addition, the spectra were obtained in two different hospitals, namely Hospital General Universitario Gregorio Marañón (HGUGM) and Hospital Central de la Defensa Gómez Ulla (CDGU), thus, using different MALDI-TOF MS machines. It should be noted that the spectra obtained from the samples at CDGU were extracted over the three medium and one week, but they were obtained 15 days after the calibration date of the machine. In contrast, at HGUGM, the spectra were taken one week after the calibration date. Figure 1 resumes the acquisition of spectra.

These 30 samples are composed of 10 samples of RT 027, another 10 samples of 181, and the last ten are a combination of samples of RTs 078 (N=3), 106 (N=4) and 165 (N=3).

B. *Preprocessing and Binning*

Each spectrum is represented by a set of measurements, each of which is described by its corresponding mass-to-charge ratio (m/z) and intensity. For the preprocessing, we started applying the conventional open-source software to preprocess MALDI-TOF mass spectra, the MALDIquant R package [20], but using the version in Pyhton code extracted from [21], which consist in five steps: (1) the measured intensity is transformed with a square-root method to stabilize the variance, (2) smoothing using the Savitzky Golay algorithm with half-window-size 10 is applied, (3) an estimate of the baseline is removed in 10 iterations of the SNIP algorithm, (4) the intensity is calibrated using the total ion current (TIC), and (5) the spectra are trimmed to values in a

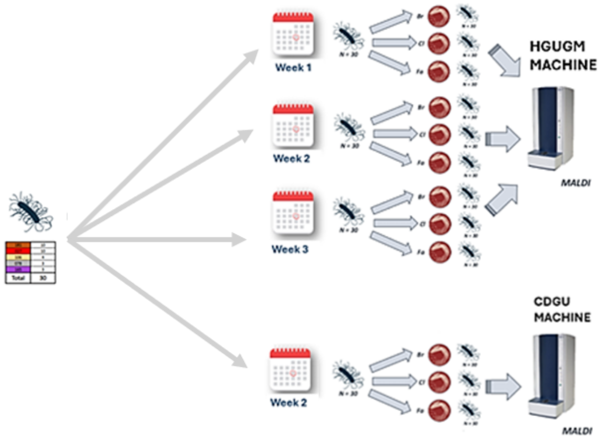


Fig. 1: Spectra acquisition over time and medium from the original 30 samples, using HGUGM and CDGU machines.

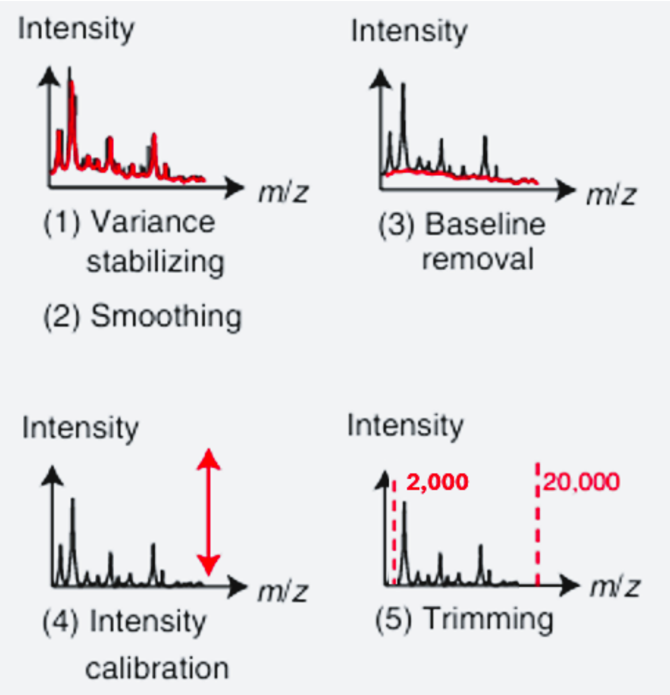


Fig. 2: Summary of the preprocessing steps.

2,000–20,000 Da range. Figure 2 shows a result of all these steps.

After preprocessing, as most previous works [1] [18] [20], we constructed the feature vectors required by the ML methods applying a binning. To perform the binning, the m/z axis is partitioned in the range of 2,000 to 20,000 Da into equal-sized bins of 3 Da size and sums the intensity of all measurements in the sample falling into the same bin. Thus, each sample is represented by a vector of fixed dimensionality, that is, a vector containing 6000 features.

In this study, we proposed a novel binning method in which the size of the bin increases logarithmically in accordance with the decreasing proximity of the peaks as the m/z value rises. The idea is to reduce the feature vector size, but include more information than the normal binning, given the presence of a greater number of peaks at low m/z values. The vector size selected is 1000 bins.

C. *Peak Information Kernel: PIKE*

Peak Information Kernel [18], as mentioned, was specifically developed for MALDI-TOF mass spectra. Given spectra S and S' with m/z values x_i and x'_j, plus intensities λ_i and λ'_j with $i, j = 1...N$, N being the number of bins, we can obtain a closed-form *approximation* as:

$$k_t(S, S') = \frac{1}{2\sqrt{2\pi t}} \sum_i \sum_j \lambda_i \lambda'_j \exp\left(-\frac{(x_i - x'_j)^2}{8t}\right), \tag{1}$$

with a single parameter t, which is responsible for regulating the degree of smoothing, i.e., t controls the width of each peak. To properly compute this kernel, it is necessary to apply an additional normalization step to the mass spectra to ensure that each intensity λ_i, $i = 1...N$, is greater than 1. Else, the product of two intensities will become progressively smaller, resulting in a lower similarity between the two spectra.

This kernel is capable of assessing interactions between different peaks, as evidenced by the examination of the equation, which reveals that the distance between all pairs of peaks is used in its calculation. Consequently, it can be robust to variability introduced in spectra acquisition. In addition, another advantage of this approach is that no feature vectors are required, and the kernel can operate directly on sets of tuples corresponding to a set of peaks. However, it cannot be easily applied to thousands of measurements, as all pairs of peaks are compared [18].

For our study, in order to analyse the behaviour of the PIKE, a similarity metric was proposed. The concept was based on the notion of cosine similarity, which allows for the calculation of this metric between spectra a and b.

$$\text{PIKE-Similarity}_{ab} = \frac{K_{ab}}{\sqrt{K_{aa} K_{bb}}} \tag{2}$$

D. *Data augmentation*

For the first time, Data Augmentation (DA) was proposed as a mean of increasing the number of samples in MALDI-TOF mass spectra, given that the current number may not yield an accurate result, and there are few examples of this technique being applied to different mass spectrometry techniques. However, as proposed, this technique can also be used to reproduce the differences between spectra of a single bacterial sample taking into account the impact of acquisition conditions. This is illustrated in Figure 3, which shows differences between medium, time, and machine acquisition of the same sample of RT 027. A comparison of the time and medium reveals that spectra are similar in terms of the position of ribosomal peaks, those peaks characterizing the species, but in terms of typing, the most interesting peaks are the non-ribosomal ones, as we found a great variation in both intensity and, even, position.

Significant discrepancies are evident when comparing the performance of different machines. It is notable that the CDGU introduces a considerable amount of noise, which is not present to the same extent in the HGUGM.

In order to reproduce these spectra differences, we propose a 3-step data augmentation procedure which will allow us to reproduce the aforementioned variability:

1) Addition of random Gaussian noise with mean equal to zero and exploring different variance values, into a random percentage of peaks. The aim is to recreate the variability in intensities, specially between mediums and weeks.
2) Spectra are then augmented by randomly shifting positions along the m/z dimension by ±3 Da, ±6 Da or ±9 Da. With a given probability, a significant proportion of the peaks remain unshifted, while a smaller random subset exhibits a ±3Da shift, and an even smaller percentage a ±6Da shift, and so on.
3) As final step, noise with a low variance was introduced to peaks with intensities equal to zero. The objective of this step is to reproduce the variations in spectra that originate from different machines. As illustrated in Figure 3, the spectra from CDGU present peaks that are absent in HGUGM.

III. EXPERIMENTS AND RESULTS

Two principal experiments were conducted. The initial study was designed to ascertain the impact of data variability on ML conventional classifiers in variable scenarios, by using spectra from different time acquisition, medium and machine. Secondly, the advantages of the data augmentation process are examined in order to address the impact of this variability.

A. *Analysis of ML model performance in variable scenarios*

The initial experiment was designed to illustrate the variability effect on different classifier methods. To this end, PIKE was integrated with the Support Vector Machine (SVM) method and evaluated in comparison with other conventional classifiers, including Random Forests (RF) and SVM with Radial Basis Function (RBF) kernel and Linear kernel.

In order to establish a baseline for all experiments, the samples from HGUGM of the first week and *Clostridium* medium were selected as training set. Therefore, first experiments were tested in baseline and the following experiments were conducted to analyse the impact of varying time, medium and machine acquisition. In different week scenario, results were analysed by testing during the subsequent two weeks, in *Clostridium* media. Then, Brucella and FA media from the first week were used as the test set, whereas in the subsequent experiment, the spectra from CDGU from the same medium were used for testing.

Furthermore, the classifiers were trained using the Leave-One-Out (LOO) method, which means that all baseline samples are used for training except for one that is reserved for testing in baseline experiment, and depending on the scenario, this test sample is extracted from the corresponding new week, medium, or machine.

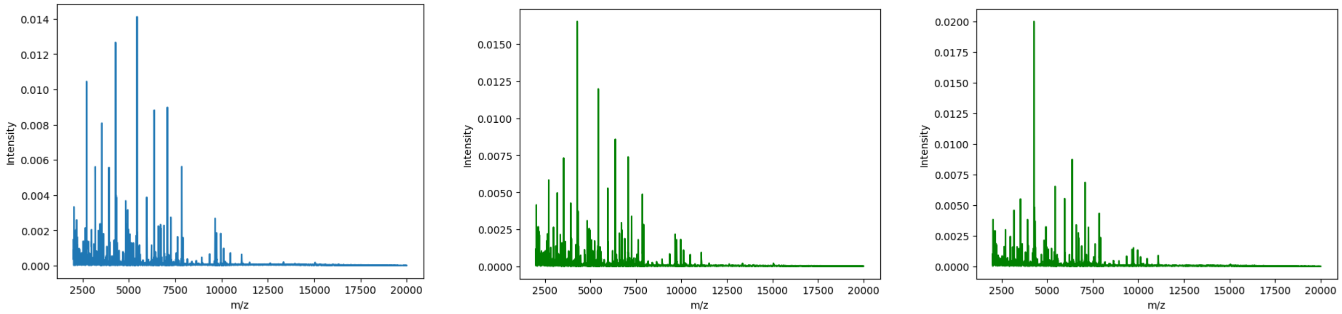
The objective is to solve a multilevel classification problem comprising three categories: RT 181, RT 027, and the remaining RTs, which have similar spectra, as the last class. The hyperparameters were validated through a five-fold Cross-Validation technique. In the case of the RF method, the number of estimators was subjected to evaluation within a range of 50 to 300, and when the RBF kernel was employed, the evaluation was conducted between distinct values of gamma, spanning a logarithmic range from 0.001 to 1000. The SVM hyperparameter C was varied between 0.01 and 100. Furthermore, a range of values for t were validated in the PIKE case, including 0.01, 0.1, 0.25, 0.5, 1, 2, 4, 8, and 16.

Table I collates the findings of all the aforementioned experiments, where the accuracy percentages are obtained using preprocessed spectra with distinct binning methods and for a size of 1000 when using logarithmic binning.

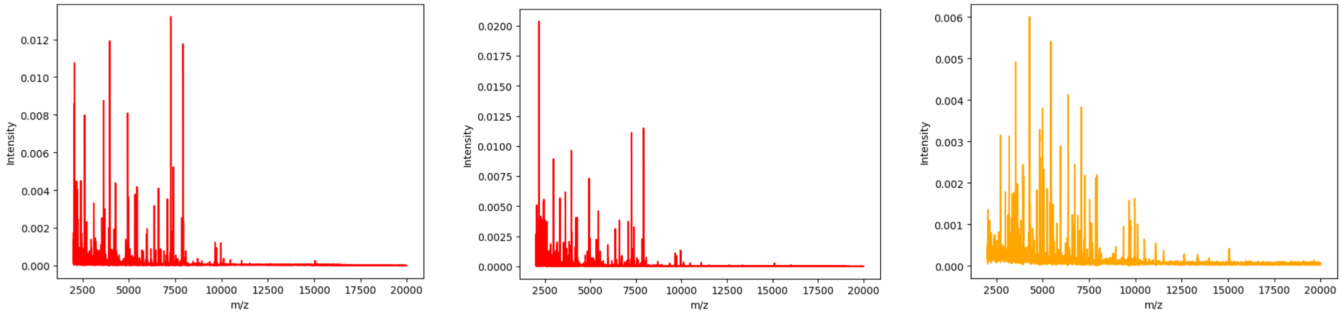
To analyse in detail the performance of the PIKE, examples of similarity matrices are presented in Figure 4, which illustrate the degree of similarity between first week and the remaining weeks, using t equal to the validated value of 1.

B. *Data augmentation experiments*

In order to combat the degradation suffered by ML models due to variability in the MALDI-TOF spectra



a) Week 1 in medium Clos from HGUGM. b) Week 2 in medium Clos from HGUGM. c) Week 3 in medium Clos from HGUGM.



d) Week 1 in medium Br from HGUGM. e) Week 1 in medium FA from HGUGM. f) CDGU in medium Clos.

Fig. 3: Differences between spectra in different weeks (subfigures (b) and (c)), mediums (subfigures (e) and (f)), and machines (subfigure (d)) for an sample of 027.

				TIME		**MEDIUM**		**MACHINE**
		BININNG	**BASELINE**	**WEEK 2**	**WEEK 3**	**BR**	**FA**	**CDGU**
RF		**Standard**	76.67	74.52	78.18	86.67	91.67	36.67
		Log	76.67	73.33	78.28	75.33	86.28	33.33
SVM	**LINEAR**	**Standard**	53.33	46.67	53.33	36.67	43.33	23.33
		Log	70	56.67	36.67	46.67	33.33	16.67
	RBF	**Standard**	70	46.67	36.67	46.67	36.67	23.33
		Log	83.33	50	40	42.33	46.67	16.67
	PIKE	**Standard**	93.33	36.67	50	26.67	30	26.67
		Log	90	36.67	33.33	46.67	43.33	20

TABLE I: Accuracy results: spectra with different binning methods, standard and logarithmic binning. Data set used for training was collected during the first week, *Clostridium* medium and HGUGM, and the testing phase was conducted during the subsequent weeks (using *Clostridium* medium), mediums (extracted from first week) and machines (same medium).

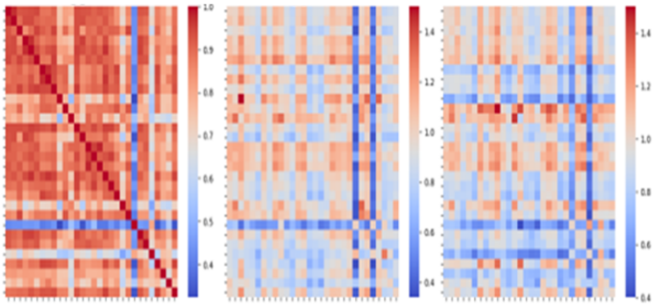


Fig. 4: Similarity matrices for t value of 1, comparing week 1, week 2 and week 3, using spectra from the same medium (*Clostridium*) and with binning of 3 Da equal size binning. The initial figure represent a comparison between week 1 with itself, while the subsequent figures represent a comparison between week one and the remaining weeks.

acquisition process, in this section, we will try to reduce the loss of performance by implementing the DA process described above.

We start with an ablation study of the DA process analyzing the impact of each step over the performance of a RF classifier with the spectra preprocessed using the standard binning of 3 Da equal-size bins. In Step 1, different percentages were evaluated to select the peaks randomly, 25% of the peaks, 50%, 75% or all the spectra, applying different noise standard deviations ([0.5, 1, 2, 5]). In Step 2, two configurations were compared, establishing percentages of 60% / 20% / 10% / 10% or of 50% / 30% / 15% / 5%, for the random shifting process where the configuration means: percentage with no-shifted peaks/ percentage with 3 Da shifted peaks / percentage with 6 Da shifted peaks/ percentage with 9 Da shifted peaks. Finally, in Step 3, three standard deviations were analyzed, 0.005, 0.1 or 0.2.

Figure 5 evaluates the steps of the process with different numbers of generated samples, to develop an optimal DA procedure. First image represents Step 1, where noise were added for a 75% of the peaks, since results comparing percentages did show similar values. Second and third images represent Step 2 and 3, respectively. Furthermore, image (d) is a combination of all the steps establishing a percentage of 75% and standard deviation of 1 for Step 1, a configuration of 50% / 30% / 15% / 5% for Step 2 and a standard deviation of 0.1 for Step 3. In this section of the study, the baseline configuration was once again used as a training set and was tested with all the spectra, independently of the time, medium, and machine. To obtain statistically relevant results, each experiment was conducted 10 times, with the mean and standard deviation calculated for each trial.

From this point onward, the three-step process combination with a number of samples equal to 2000 will be established as DA process.

The subsequent phase of the study will include an examination of the impact of this configuration on the variability inherent to the data, with specific attention paid to the effects of temporal acquisition, the medium through which samples are collected, and the machine used to extract the spectra. The results were then compared with those of the original 30 samples, again using RF as a classifier. Table II compares all possible scenarios with the baseline.

In order to analyse the benefits of the DA, a further experiment was carried out over alternative classifiers, including SVM with RBF kernel and a neural network constituted by three dense layers with ReLU activation, trained with Adam optimisation, a batch size of 32 and 20 epochs. The preceding information is presented in Table III, comparing the performance of classifiers with and without this technique, conducting the training on baseline, and testing all the remaining samples, acquired in different weeks, medium, and machines.

C. *Data augmentation for imbalanced problems*

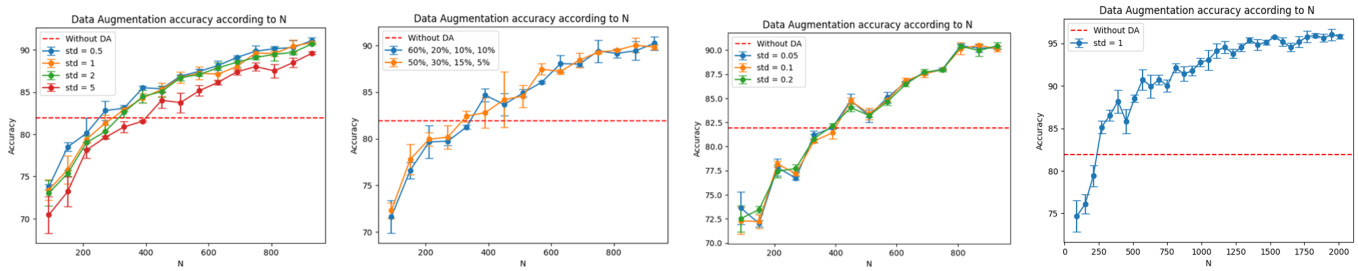
In addition, the process was repeated with the five RTs (Table IV), incorporating additional samples generated through DA for the minority classes, ensuring a balanced distribution across all classes. As a further point of comparison, the results were once again analysed with a sample size of N=30.

In Figure 6, ROC curves and AUC results were obtained using the original dataset and the data augmented dataset in the scenario in which the baseline was designated as the training set and all the spectra as the test set.

IV. DISCUSSION

The initial introduction of PIKE and similar literature was driven by the need to address typification problems through the reduction of variability, as it employs the utilisation of the entire spectrum in place of a peak selection. In the initial experiment, which utilised a baseline of a single week in a specific medium, PIKE demonstrated superior accuracy compared to traditional classifiers, as RF, with an even greater advantage observed when logarithmic binning was employed.

Nevertheless, when different time acquisitions are compared, the performance of PIKE is found to decrease considerably. Consequently, RF emerges as the optimal option, independently of the binning used. The observed results may be attributed to the PIKE formulation, in



a) **Step 1.** b) **Step 2.** c) **Step 3.** d) **Step 1+2+3.**

Fig. 5: DA process steps: a) using just step 1, with different noise standard deviations, adding proportional noise to a random 75% of the peaks. b) Using step 2 with different combination of percentages for each option: no shifting, a ±3Da shift, ±6Da shift and a ±9Da shift. c) Step 3 with different noise standard deviations. d) All steps. Results compared with accuracy obtained without data augmentation (N=30).

	Baseline	**All Data**	**Week 2**	**Week 3**	**BR**	**FA**	**CDGU**
RF	81.67	68.33	74.52	78.18	86.67	91.67	36.67
RF+DA	100 ±0.00	93.46±0.86	87.33 ±0.81	88.67 ±0.30	99.66±0.17	99.75±0.19	64.44±1.24

TABLE II: Accuracy comparison between scenarios, without DA (N=30) and after applying DA for 3 RTs.

	Baseline	**All Data test**
RF	81.67	68.33
DA+RF	100 ±0.0	93.46 ±0.86
SVM+RBF	56.67	23.33
DA+SVM+RBF	63.24 ±0.35	33.33 ±1.42
NN	33.33	38.33
DA+NN	97.24±0.29	60.8 ±1.39

TABLE III: Comparison between classifiers testing baseline and all spectra.

which smaller peaks are hidden by the intensity of larger ones. It is illustrated in the similarity matrices, where the expected outcome would be a high degree of similarity between different weeks. However, this is not reflected in the observed data, which shows that classes cannot be distinguished. Therefore, this solution is not optimal for typification problems, where the information of all peaks is the fundamental principle underlying an effective classification system.

Furthermore, an examination of the results reveals that the aforementioned conclusions can be derived from all the possible scenarios that introduce variability, where RF is identified as the optimal classifier, as it is less susceptible to it. The extraction of spectra from different mediums is more reproducible than the acquisition of spectra at different times or on different machines, which is particularly susceptible to variability, with results rarely exceeding 40%.

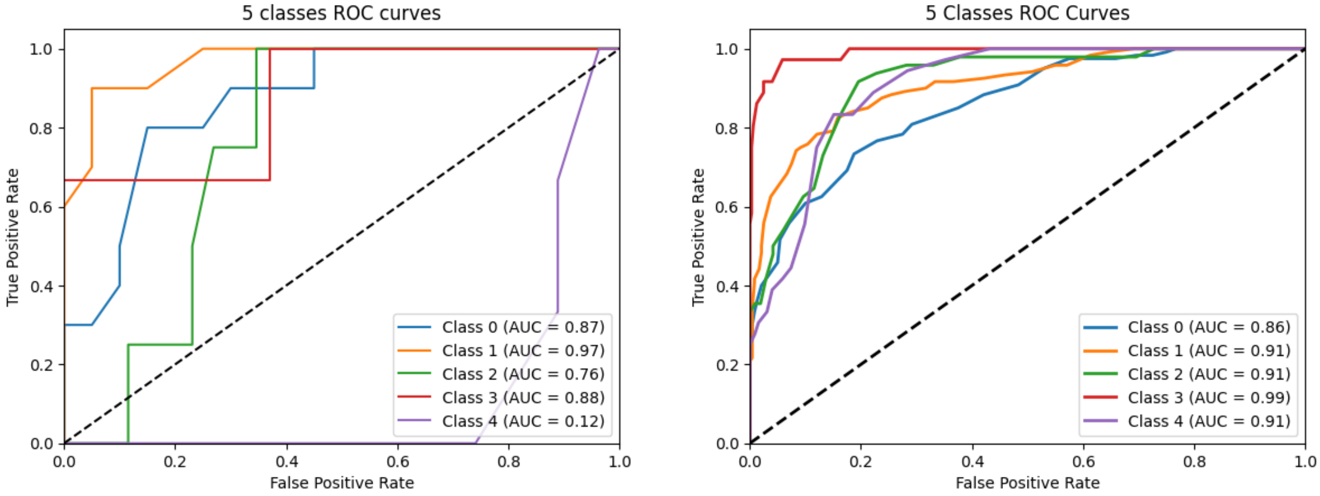
In light of these findings, data augmentation is presented as a promising avenue to enhance the performance of any classifier. To obtain a reliable result, the selected configuration was to apply all the procedures of the DA proposed and a selection of sample size of 2000, since it has been demonstrated that combining all the steps results in a more accurate outcome than would be achieved by applying them individually and the accuracy stabilises at this size.

Once the configuration was established, the results demonstrated that this technique enhances the performance of RF in comparison to the original 30 samples, and RF remains the most effective classifier option for reducing time, medium, and machine acquisition variability. However, in the latter case, the classifier is not as robust. This may be due to the variability introduced by the different ways of machine acquisition or by the different acquisition dates between calibration, as evidenced in Figure 3 where the two spectra are markedly disparate. It is imperative that the spectra must be obtained under identical calibration conditions to guarantee that any observed variability is exclusively attributable to the machines themselves.

It is also important to note that the data augmentation method yield optimal results when applied to all five RTs, as it does with the initial three categories. Therefore, it remains a robust and accurate approach, even when classifying samples in the absence of knowledge regarding the time, medium, or machine, which introduces considerable variability into the data. This

	Baseline	**All Data**	**Week 2**	**Week 3**	**BR**	**FA**	**CDGU**
RF	63.33	55.28	76.67	73.33	81.67	83.33	40
RF+DA	100 ±0.00	90.35±0.89	80.70 ±0.19	76.67 ±1.39	99.17±0.1	99.07±0.08	51.65±5.11

TABLE IV: Accuracy comparison between scenarios, without DA (N=30) and after applying DA for 5 RTs.



a) Roc curves without DA

b) ROC curves with DA

Fig. 6: ROC curves with AUC values, where class 0 is Rt 181, 1 is RT 027, 2 is RT 106, 3 is RT 078 and the last class is RT 165.

is demonstrated by the ROC curves, which achieve an AUC greater than 50%, even for determined RTs that, in the absence of the DA method, did not exceed this percentage.

V. CONCLUSION

The results demonstrate that, under specific conditions of spectra acquisition, typification may be more susceptible, necessitating the use of robust classification methods to ensure reliable predictions. Certain classification techniques, such as PIKE, are effective for other clinical applications, however, they do not perform as well as anticipated for the purpose of typification due to variability in spectra. In this scenario, RF is the optimal choice among several methods, as it is capable of leveraging the full spectrum of available information.

Furthermore, this paper presents a method for accurately classifying different *C. difficile* RTs through the integration of DA with RF. This approach serves to mitigate the variability that is associated with disparate acquisition conditions during the classification process while maintaining robustness to the scarcity of certain strains, allowing typification regardless of the methodology used in the extraction of the spectrum.

Future work

Further investigation is required to identify and mitigate the issue of variability introduced by the use of different machines. This will necessitate the application of diverse preprocessing techniques to address the issue and analyse MALDI-TOF spectra from more than two machines. Additional lines of inquiry will be required to extend these techniques to other bacterial species and clinical applications.

In conclusion, the findings of this study will facilitate the rapid and reliable identification of virulent strains of the pathogen in question, which are responsible for the emergence of new cases and patient relapse. This is of particular importance, as such cases have the potential to result in adverse outcomes for the patient. Furthermore, knowledge will enable the implementation of effective strategies to prevent and control the spread of this species within the hospital setting, thus reducing the risk of its collapse.

REFERENCES

[1] C. Weis, A. Cuénod, B. Rieck, O. Dubuis, S. Graf, C. Lang, M. Oberle, M. Brackmann, K. K. Søgaard, M. Osthoff, K. Borgwardt, and A. Egli, “Direct antimicrobial resistance prediction from clinical maldi-tof mass spectra

using machine learning," *Nature medicine*, vol. 28, no. 1, pp. 164–174, 1 2022. [Online]. Available: https://doi.org/10.1038/s41591-021-01619-9

[2] C. Weis, C. Jutzeler, and K. Borgwardt, "Machine learning for microbial identification and antimicrobial susceptibility testing on maldi-tof mass spectra: a systematic review," *Clinical Microbiology and Infection*, vol. 26, no. 10, pp. 1310–1317, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1198743X20301580

[3] J. Czepiel, M. Dróżdż, H. Pituch, E. J. Kuijper, W. Perucki, A. Mielimonka, S. Goldman, D. Wultańska, A. Garlicki, and G. Biesiada, "Clostridium difficile infection," *European Journal of Clinical Microbiology & Infectious Diseases*, vol. 38, pp. 1211–1221, 2019.

[4] R. Markovska, G. Dimitrov, R. Gergova, and L. Boyanova, "Clostridioides difficile, a new "superbug"," *Microorganisms*, vol. 11, no. 4, p. 845, 2023.

[5] D. A. Leffler and J. T. Lamont, "Clostridium difficile infection," *New England Journal of Medicine*, vol. 372, no. 16, pp. 1539–1548, 2015.

[6] M. Krutova, P. Kinross, F. Barbut, A. Hajdu, M. Wilcox, E. Kuijper, F. Allerberger, M. Delmée, J. Van Broeck, R. Vatcheva-Dobrevska *et al.*, "How to: surveillance of clostridium difficile infections," *Clinical Microbiology and Infection*, vol. 24, no. 5, pp. 469–475, 2018.

[7] K. A. Davies, H. Ashwin, C. M. Longshaw, D. A. Burns, G. L. Davis, M. H. Wilcox, E. S. Group *et al.*, "Diversity of clostridium difficile pcr ribotypes in europe: results from the european, multicentre, prospective, biannual, point-prevalence study of clostridium difficile infection in hospitalised patients with diarrhoea (euclid), 2012 and 2013," *Eurosurveillance*, vol. 21, no. 29, p. 30294, 2016.

[8] M. He, F. Miyajima, P. Roberts, L. Ellison, D. J. Pickard, M. J. Martin, T. R. Connor, S. R. Harris, D. Fairley, K. B. Bamford *et al.*, "Emergence and global spread of epidemic healthcare-associated clostridium difficile," *Nature genetics*, vol. 45, no. 1, pp. 109–113, 2013.

[9] V. F. Viprey, G. L. Davis, A. D. Benson, D. Ewin, W. Spittal, J. J. Vernon, M. Rupnik, A. Banz, F. Allantaz, P. Cleuziat *et al.*, "A point-prevalence study on community and inpatient clostridioides difficile infections (cdi): results from combatting bacterial resistance in europe cdi (combacte-cdi), july to november 2018," *Eurosurveillance*, vol. 27, no. 26, p. 2100704, 2022.

[10] M. Kachrimanidou, S. Metallidis, O. Tsachouridou, C. Harmanus, V. Lola, E. Protonotariou, L. Skoura, and E. Kuijper, "Predominance of clostridioides difficile pcr ribotype 181 in northern greece, 2016–2019," *Anaerobe*, vol. 76, p. 102601, 2022.

[11] M. T. Olson, P. S. Blank, D. L. Sackett, and A. L. Yergey, "Evaluating reproducibility and similarity of mass and intensity data in complex spectra—applications to tubulin," *Journal of the American Society for Mass Spectrometry*, vol. 19, no. 3, pp. 367–374, 2008, pMID: 18207417. [Online]. Available: https://doi.org/10.1016/j.jasms.2007.11.012

[12] R. A. Giebel, C. R. Worden, S. Rust, G. T. Kleinheinz, M. Robbins, and T. Sandrin, "Microbial fingerprinting using matrix-assisted laser desorption ionization time-of-flight mass spectrometry (maldi-tof ms) applications and challenges." *Advances in applied microbiology*, vol. 71, pp. 149–84, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:44707766

[13] A. J. Saenz, C. E. Petersen, N. B. Valentine, S. L. Gantt, K. H. Jarman, M. T. Kingsley, and K. L. Wahl, "Reproducibility of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for replicate bacterial culture analysis," *Rapid Communications in Mass Spectrometry*, vol. 13, no. 15, pp. 1580–1585, 1999. [Online]. Available: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0231%2819990815%2913%3A15%3C1580%3A%3AAID-RCM679%3E3.0.CO%3B2-V

[14] M. T. Olson, J. A. Epstein, D. L. Sackett, and A. L. Yergey, "Production of reliable maldi spectra with quality threshold clustering of replicates," *Journal of the American Society for Mass Spectrometry*, vol. 22, no. 6, pp. 969–975, 2011, pMID: 21953038. [Online]. Available: https://doi.org/10.1007/s13361-011-0097-9

[15] A. Godmer, Q. Giai Gianetto, K. Le Neindre, V. Latapy, M. Bastide, M. Ehmig, V. Lalande, N. Veziris, A. Aubry, F. Barbut, and C. Eckert, "Contribution of maldi-tof mass spectrometry and machine learning including deep learning techniques for the detection of virulence factors of clostridioides difficile strains," *Microbial Biotechnology*, vol. 17, no. 6, p. e14478, 2024, e14478 MICROBIO-2024-099-RA.R1. [Online]. Available: https://enviromicro-journals.onlinelibrary.wiley.com/doi/abs/10.1111/1751-7915.14478

[16] A. Calderaro, M. Buttrini, B. Farina, S. Montecchini, M. Martinelli, M. C. Arcangeletti, C. Chezzi, and F. De Conto, "Characterization of clostridioides difficile strains from an outbreak using maldi-tof mass spectrometry," *Microorganisms*, vol. 10, no. 7, 2022. [Online]. Available: https://www.mdpi.com/2076-2607/10/7/1477

[17] A. Candela, D. Rodriguez-Temporal, M. Blázquez-Sánchez, M. J. Arroyo, M. Marín, L. Alcalá, G. Bou, B. Rodríguez-Sánchez, and M. Oviaño, "Analysis of high-molecular-weight proteins using maldi-tof ms and machine learning for the differentiation of clinically relevant clostridioides difficile ribotypes," *bioRxiv*, 2024. [Online]. Available: https://www.biorxiv.org/content/early/2024/06/20/2024.06.18.599569

[18] C. Weis, M. Horn, B. Rieck, A. Cuénod, A. Egli, and K. Borgwardt, "Topological and kernel-based microbial phenotype prediction from maldi-tof mass spectra," *Bioinformatics (Oxford, England)*, vol. 36, pp. i30–i38, 07 2020.

[19] K. Seddiki, P. Saudemont, F. Precioso, N. Ogrinc, M. Wisztorski, M. Salzet, I. Fournier, and A. Droit, "Towards cnn representations for small mass spectrometry data classification: From transfer learning to cumulative learning," *bioRxiv*, 2020. [Online]. Available: https://www.biorxiv.org/content/early/2020/03/25/2020.03.24.005975

[20] S. Gibb and K. Strimmer, "MALDIquant: a versatile R package for the analysis of mass spectrometry data," *Bioinformatics*, vol. 28, no. 17, pp. 2270–2271, 07 2012. [Online]. Available: https://doi.org/10.1093/bioinformatics/bts447

[21] G. De Waele, G. Menschaert, and W. Waegeman, "An antimicrobial drug recommender system using maldi-tof ms and dual-branch neural networks," *bioRxiv*, Sep. 2023. [Online]. Available: http://dx.doi.org/10.1101/2023.09.28.559916