Overcoming variability challenges for *Clostridioides difficile* via data augmentation techniques

Alejandro Guerrero-López[a,c], Lucía Bravo Antón[b], Mario Blázquez-Sánchez[b,c], Lucía Schmidt-Santiago[c], Carlos Sevilla-Salcedo[a], David Rodríguez-Temporal[b,c], Belén Rodríguez-Sánchez[b,c], Vanessa Gómez-Verdejo[a,c]

[a]*Signal Processing and Communications Department, Universidad Carlos III de Madrid, Avda. de la Universidad, 30, Leganés, 28911, Madrid, Spain*
[b]*Clinical Microbiology and Infectious Diseases Department, Hospital General Universitario Gregorio Marañón, Dr. Esquerdo, 46, Madrid, 28007, Madrid, Spain*
[c]*Institute of Health Research Gregorio Marañón, Hospital General Universitario Gregorio Marañón, Dr. Esquerdo, 46, Madrid, 28007, Madrid, Spain*

Abstract

Matrix-Assisted Laser Desorption Ionization–Time of Flight (MALDI-TOF) mass spectrometry (MS) has transformed clinical microbiology by enabling rapid bacterial identification. However, its application to more complex tasks, such as *Clostridioides difficile* (*C. difficile*) ribotype (RT) classification, is hindered by data variability stemming from culture media, protein extraction protocols, and laboratory equipment. This study evaluates the impact of such variability on machine learning (ML) models trained for *C. difficile* typing using MALDI-TOF spectra. We analyze the performance of different ML models under diverse experimental conditions and introduce a data augmentation (DA) strategy specifically designed to mimic real spectral variations, including new peak generation, as well as fluctuations in peak intensity and position.

Our results show that while time-related and protein extraction inconsistencies can be effectively addressed, culture media and equipment variations still pose challenges. Importantly, we demonstrate that applying DA to spectra obtained from selective media - requiring between 24-48 hours of incubation– significantly improves classification performance, approaching the results achieved with enriched media which demand over 72-96 hours. By enhancing model generalizability, our approach aims to enable fast and reliable bacterial typing that can be seamlessly integrated into the hospital's

daily routine across diverse clinical settings.

Additionally, we present MALDIDA, an open-source Python library that provides dedicated tools for MALDI-TOF data augmentation.

Keywords: Machine Learning, Data Augmentation, MALDI- TOF spectra, *Clostridioides difficile*, Reproducibility

---

1. Introduction

Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry (MALDI-TOF MS) has revolutionized clinical microbiology in the last decade. This technique allows for rapid and accurate bacterial identification by analyzing their protein profiles within minutes [1], unlike traditional methods, which can take several days and require highly trained personnel in molecular biology methods, including Pulsed Field Gel Electrophoresis (PFGE) and Polymerase Chain Reaction (PCR) Ribotyping [2]. The analysis of MALDI-TOF spectra typically relies on the presence of attributes such as peak height and position, focusing primarily on ribosomal proteins that are species-specific biomarkers [3, 4]. Nevertheless, for more complex tasks like detecting antibiotic resistance or distinguishing between different strains of the same species (bacterial typing), visual analysis is insufficient [5, 6]. So, automatic classification of MALDI-TOF spectra using Machine Learning (ML) models has gained traction in recent years as a means to accelerate and reduce the costs associated with bacterial characterization [2, 7, 8].

In this study, MALDI-TOF and ML methods were evaluated for the automatic differentiation of *Clostridioides difficile* ribotypes (RTs). *C. difficile*, a Gram-positive, anaerobic, spore-forming species transmitted through the fecal-oral route, is the leading cause of antibiotic-associated infectious diarrhea in hospitals [9, 10]. Certain strains of *C. difficile* produce toxins that cause intestinal damage, playing a critical role in disease development and progression [11, 12]. Numerous ribotypes of *C. difficile* have been identified, some of which are associated with increased pathogenicity and hospital outbreaks [13], as ribotype B1/NAP1/027 (RT027) [14, 15], considered hypervirulent due to the hyperproduction of toxins [10]. Other hypervirulent RTs, such as RT078 and RT181, have emerged recently and are reported as prevalent in different European countries [16].

One of the main challenges in automatically classifying clinical samples using MALDI-TOF protein spectra is the limited availability of data [7],

which hampers meaningful comparisons between samples—a crucial step in determining the need for patient isolation. Additionally, MALDI-TOF MS is prone to variability, including inconsistencies in the presence/absence of some peaks and fluctuations in their intensities [17, 18]. This variability is further exacerbated by factors such as growth conditions, protein extraction (PE) protocols, laser intensity, machine calibration, and operator expertise. In fact, this lack of reproducibility is an important concern in bacterial characterization, as studies have shown that these sources of variability affect MALDI-TOF performance both within and between laboratories [19, 20, 18]. Moreover, while enriched media (e.g., Schaedler agar) tend to produce more robust and consistent spectral data, they are impractical for routine laboratory because they first must pass through selective media, which takes at least four days to obtain. In contrast, selective media (e.g., CHROMID® selective agar), although essential for isolate bacteria such as *C. difficile* [21, 22, 23, 24], suffer from greater variability and provide ML models with lower classification performance, as evidenced by lower AUC and accuracy scores in this paper. ML models tend to be biased towards the acquisition specifities of the training domain, providing less accurate results when tested under variable conditions [25]. To overcome these issues, data augmentation (DA) techniques have been underexplored in this field [26] with no satisfactory results, indicating the need for a deeper study of biological variability together to obtain the most adequate DA technique capable of mimicking the actual variability of the spectra.

This paper examines how such variability impacts the performance of ML models for *C. difficile* typing. The study has two key objectives: (1) to assess how factors such as culture media, PE methods, incubation time, and instrument affect the performance of commonly used ML algorithms and (2) to develop a robust ML approach capable of delivering consistent classifications, even when dealing with limited or newly acquired spectra obtained through different protocols. For the latter aim, DA techniques have been implemented, an innovative approach in the field of MALDI-TOF MS. These techniques artificially enhance the diversity of training data by applying transformations such as random noise addition or spectral shifting, thereby improving model robustness and performance in classification tasks. To facilitate the use of this DA strategy, we provide MALDIDA, an open-source Python library that offers dedicated tools for MALDI-TOF data augmentation. A graphical abstract summarizing the study's objectives and methodology is presented in Figure 1.

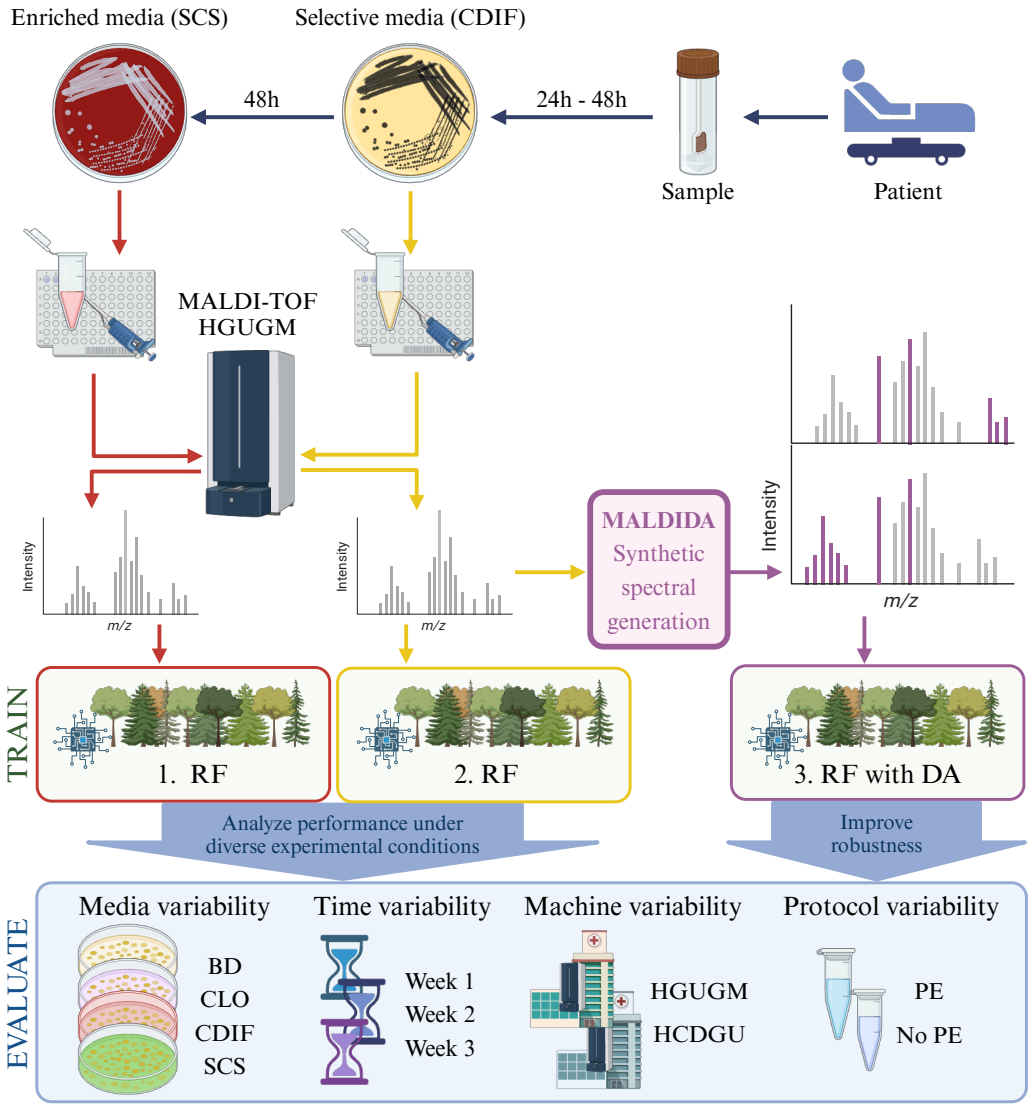


Figure 1: Overall workflow. Samples of six ribotypes of *C. difficile* were initially grown on the selective CHROMID® agar (CDIF) for 24–48h, followed by a second culture on enriched Schaedler agar (SCS). Mass spectra were then obtained via MALDI-TOF from both media, with and without protein extraction. Three classification scenarios were trained: (1) a Random Forest (RF) model on SCS spectra, (2) an RF model on CDIF spectra, and (3) an RF model on CDIF spectra augmented with our proposed data augmentation (DA) techniques. Each model was subsequently evaluated for robustness against various sources of variability, including culture media, intra-hospital variability, instrumentation, and experimental protocol.

2. Materials and methods

2.1. ***MALDI-TOF spectra***

To evaluate how different conditions affect the MALDI-TOF MS spectra of bacterial isolates, we analyzed 60 strains. These strains consisted of 10 isolates from each of the following six rybotipes: RT023, RT027, RT078, RT106, RT165, and RT181. For each strain, MALDI-TOF spectra were repeatedly extracted under varying laboratory conditions, including intrahospital variability -extracted in three different weeks- (Time), types of culture medium (Media), protein extraction protocols (PE) and mass spectrometers used in different hospitals (Hosp). Moreover, for biological reproducibility, each isolate is extracted by triplicate (replicates) following the same procedure. That means that, we have 180 mass spectra (3 replicates per each one of the 60 strains) for each condition analyzed in this study.

For time variability and media effects, the samples were evaluated over two weeks using the on-plate extraction protocol on four culture media: CHROMID® *C. difficile* selective agar (CDIF, bioMérieux), enriched Schaedler agar (SCS), enriched BD BBL *Brucella* agar (BD, Becton Dickinson) and selective *Clostridium* agar (CLO, bioMérieux). The use of a off-plate extraction (PE) protocol was also compared with the off-plate extraction protocol on the same medium for the selective CDIF media.

All these analyses were conducted exclusively at the Hospital General Universitario Gregorio Marañón (HGUGM) using MALDI-TOF MS technology with an MBT Smart MALDI Biotyper (Bruker Daltonics, Bremen, Germany). To test hospital variabilities, all isolates were processed using the off-plate extraction protocol in CDIF at another hospital, Hospital Central de la Defensa Gómez Ulla (HCDGU), performed with its MALDI Biotyper® sirius System (Bruker Daltonics), that is, different from HGUGM, but with the same MALDI-TOF MS technology.

2.2. ***Preprocessing and binning***

Each MALDI-TOF MS spectrum is represented by a set of k tuplse of mass-to-charge ratio and intensity:

$$M = \{p_k, i_k\} \quad \text{where} \quad p_k, i_k \in \mathbb{R}^+$$

where p_k denotes the position of the k-th peak (in m/z, measured in Daltons) and i_k denotes its corresponding intensity, both constrained to positive

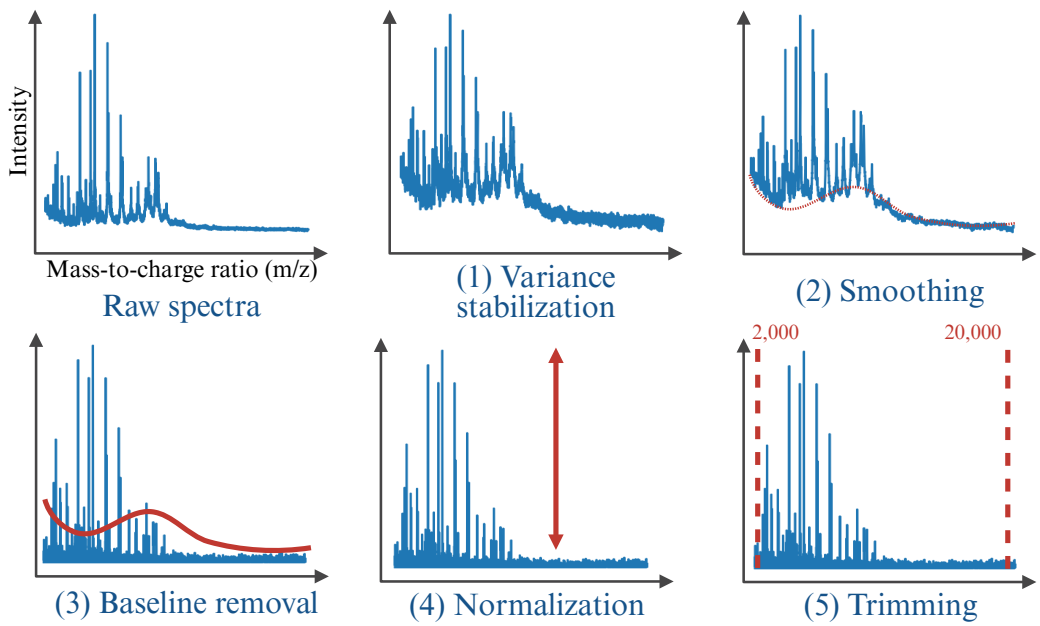
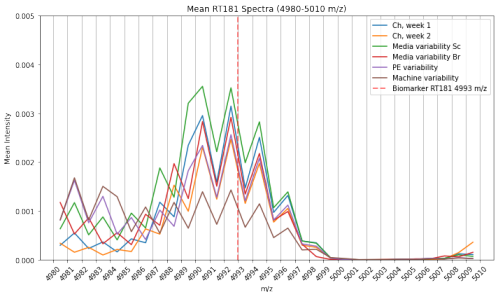


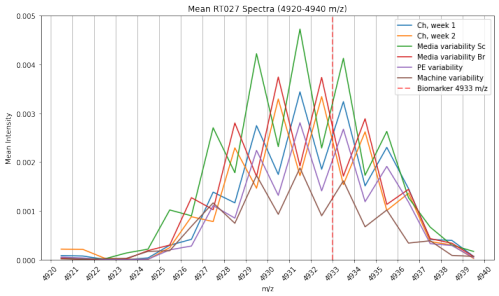
Figure 2: Summary of the preprocessing steps.

real numbers. For the preprocessing, we started applying the conventional open-source software MALDIquant [27] to preprocess MALDI-TOF MS, but using the version in Python code extracted from [28], which consists of five steps: (1) a intensity transformation with a square-root method to stabilize the variance, (2) a smoothing using the Savitzky-Golay algorithm with half-window-size 10, (3) a baseline removal where the baseline is estimated with 10 iterations of the SNIP algorithm, (4) an intensity calibration using the total ion current (TIC), and (5) a spectra trimming to values in the 2,000–20,000 m/z range. Thus, each MALDI-TOF is a set of $k = 18,000$ tuples of positions (p_k) and positions (i_k). Figure 2 shows the result of all these steps.

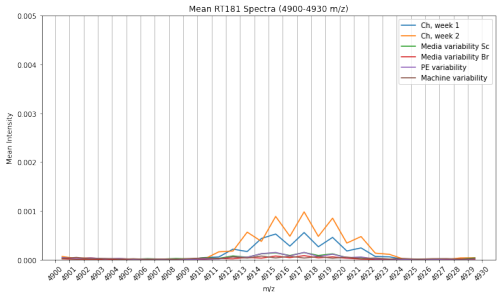
After preprocessing, following previous works [29, 2], we constructed the feature vectors required by the ML methods applying a binning. To perform the binning, the m/z axis was partitioned in the range of 2,000 to 20,000 m/z into equal-sized bins of 3 m/z size and the intensity of all measurements in the sample falling into the same bin were summed up. Thus, each sample was represented by a fixed-dimensionality vector containing 6000 features.



(a) Intensity variability.



(b) Shifting variability.



(c) New peaks variability.

Figure 3: Differences in spectra of the same ribotype due to variability conditions: a) the mean of all RT181 samples where each variable studied offer a different value of intensity peak; b) the mean of all RT027 samples where a well-known biomarker 4933m/z [30, 31] is shifted 2-3Da depending on the variable condition; c) mean of all RT181 samples where a new peak appears only on selective CDIF media.

2.3. ***Data augmentation***

Analyzing the MALDI-TOF MS spectra differences across media, time, PE protocols, and machine acquisition for the same sample — as shown in Figure 3 for an RT027 sample — reveals significant spectral variability. This includes peaks that appear (as shown for Week 2 and Week 3 in CDIF medium in Fig. 3c), changes in intensity (as shown when testing on different hospital machines in Fig. 3a), and shifts in positions (as shown for BD media and Week 2 in Fig. 3b). Such variability can lead to learning models biased to the acquisiton conditions, resulting in poor generalization when faced with acquisition differences in the test set. To improve the robustness of these methods and given the small sample size in which this variability may not be fully represented, we propose simulating these variations through DA.

In particular, we propose to generate synthetic spectra that can reproduce these variations by modifying training data spectra with three techniques:

1. **Intensity variations**: To simulate the variability in intensities observed between different machines and weeks, we perturb a random subset of the peaks. For the k-th peak (p_k) with a positive intensity i_k, we perturb it with probability p_1. Formally, let

$$S \subseteq \{k : i_k > 0\} \quad \text{with} \quad |S| = \lceil p_1 \cdot |\{k : i_k > 0\}| \rceil.$$

Then, the augmented intensity i'_k is defined as

$$i'_k = \begin{cases} \max\{0,\, i_k + \epsilon_k\}, & \text{if } k \in S, \\ i_k, & \text{otherwise,} \end{cases}$$

where $\epsilon_k \sim \mathcal{N}(0, v_1)$ is Gaussian noise with zero mean and variance v_1. In this way, random noise is added to a fraction p_1 of the peaks while ensuring that the intensities remain non-negative following behaviour of Fig. 3a.

2. **Peak Shifting**: To simulate variability due to instrumental conditions, a random fraction p_2 of peaks is selected for shifting. Formally, let

$$S' \subseteq \{1, \ldots, K\} \quad \text{with} \quad |S'| = \lceil p_2 \cdot K \rceil.$$

For each peak k, the new position is defined as

$$p'_k = \begin{cases} p_k + d_k \cdot r_k \cdot \Delta_{\max}, & \text{if } k \in S', \\ p_k, & \text{otherwise,} \end{cases}$$

where $d_k \in \{-1, +1\}$ is a random sign, r_k is the normalized position of the peak (e.g., $r_k = \frac{k-1}{K-1}$), and

$$\Delta_{\max} = \left\lfloor \frac{9}{b} \right\rfloor,$$

with b being the bin size (so that the maximum shift is up to ± 9 Da). Finally, the intensities i_k are re-interpolated on the original p_k grid to yield the augmented intensities following behaviour of Fig. 3b.

3. **New peaks**: To simulate machine variability, we perturb a random fraction p_3 of the zero-intensity locations. Let

$$Z = \{k : i_k = 0\}$$

and choose a subset

$$S'' \subseteq Z \quad \text{with} \quad |S''| = \lceil p_1 \cdot |Z| \rceil.$$

Then, for each $k \in S''$, the augmented intensity is defined as

$$i'_k = \begin{cases} \max\{0,\, i_k + \epsilon_k\}, & \text{if } k \in S'', \\ i_k, & \text{otherwise,} \end{cases} \quad \text{with } \epsilon_k \sim \mathcal{N}(0, v_3).$$

This procedure introduces new peaks by adding low-variance Gaussian noise to regions with zero intensity following behaviour of Fig. 3c. These techniques are implemented in Python and open-sourced in two versions: for reproducibility of this paper can be found here[1] and as a python package (installable via pip) here[2].

2.4. ***Experimental setup***

To illustrate the effect of MALDI-TOF MS variability on different classifier methods, we designed an experiment aimed to classify each isolate into one of six *C. difficile* ribotypes (RTs) (RT181, RT165, RT106, RT078, RT027, and RT023) and we involved several state-of-the-art ML models commonly used in the field [32]: (i) a Support Vector Machine (SVM) with most common kernels such as linear kernel and Radial Basis Function (RBF); (ii) K-Nearest Neighbors (KNN); (iv) a Random Forest (RF) ; and (v) a Light Gradient Boosting Machine (LGBM). The hyperparameters of these models were adjusted using a 5-fold cross-validation technique. In the case of the RF method, the number of estimators was evaluated within a range of 50 to 300, and when the RBF kernel was used in the SVM, the kernel width (γ) was cross-validated, covering 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 neighbors for KNN between 3 and 20 were cross-validated.

In routine laboratories, *C. difficile* is typically detected using selective media, such as CDIF or BD, which facilitate the growth of isolated *C. difficile* colonies. In line with the hospital's standard diagnostic procedures, our baseline model—referred to as the **Baseline CDIF** model—was trained on week 1 data acquired with selective medium CDIF, without PE, and processed on the HGUGM machine. In contrast, the second model—**Baseline**

[1]https://github.com/aguerrerolopez/MaldiMaranon
[2]https://github.com/aguerrerolopez/MALDIDA

SCS— mirrors the **Baseline CDIF** model but uses enriched medium SCS. Although enriched media are designed to promote optimal bacterial growth, they require an initial culture on a selective medium to isolate candidate bacteria, resulting in an additional time of 48 hours before acquiring MALDI-TOF spectra. Therefore, on our first experiment we evaluate how variability affects both media.

We evaluated each model's performance under various variability conditions using an outer Leave-One-Isolate-Out (LOIO) strategy. In this approach, the model was trained on all samples from a chosen baseline domain (either Baseline CDIF or Baseline SCS), while a single isolate was held out for testing. Importantly, all biological replicates of the held-out isolate were selected from one of several distinct conditions, allowing us to assess how well the model generalizes on external data, such as other media, hospital, time, or protocol. These conditions were:

- Baseline condition: The test sample came from the same domain as the training data (Baseline CDIF or Baseline SCS), representing the ideal scenario.
- Time variability: The test sample was selected from a different week (week 2 or week 3), to evaluate the model's robustness to temporal changes.
- Media variability: The test sample was chosen from a different media than the training data, specifically from one of three alternatives (BD, CLO, or SCS/CDIF), ensuring it was always distinct from the baseline.
- PE variability: The test sample varied in terms of the use of direct PE versus no PE.
- Machine variability: The test sample was sourced from a different hospital and generated using a different MALDI-TOF machine.

It is important to note that the isolated that was left out was never included in the training data and was used solely for testing purposes. This LOIO process was repeated for each isolate in the data set, ensuring that each sample was used as a test case in turn.

			RF	SVM + RBF	SVM + Linear	KNN	LGBM	Mean
Baseline CDIF			0.95	0.95	0.88	0.92	0.99	0.94
Variability	Time	Week 2	0.95	0.92	0.78	0.90	0.96	**0.90**
		Week 3	0.93	0.91	0.80	0.91	0.94	**0.90**
	Media	BD	0.86	0.83	0.63	0.71	0.82	0.77
		CLO	0.81	0.73	0.53	0.51	0.87	0.69
		SCS	0.89	0.82	0.63	0.70	0.86	0.78
	PE	PE	0.95	0.95	0.83	0.89	0.95	0.91
	Hosp	HCDGU	0.93	0.87	0.84	0.74	0.88	0.85
	Mean		**0.89±0.05**	0.84±0.07	0.75±0.14	0.75±0.15	**0.89±0.05**	

Table 1: AUC scores of **Baseline CDIF** model, i.e., training with week 1, selective CDIF medium, no PE and HGUGM machine.

3. Results

*3.1. **Variability analysis regarding ML performance***

Tables 1 and 2 present the results of **Baseline CDIF** and **Baseline SCS** models, respectively.

In all tables, the first row shows the baseline AUC, obtained when testing the model on samples acquired under identical conditions to the training data (same week, medium, PE method, and machine). Subsequent rows display the AUC when only one condition varies: varying sample time acquisition (Time), media (Media), PE method (PE), and machine (Hosp). Each column in the tables represents the results for a specific model used to create the baseline. The final column provides the mean AUC across all models, reflecting the overall variability introduced by each condition, regardless of the model used, identifying the most variable condition. The last row summarizes the mean variability for each model, identifying the most robust model in terms of handling variability.

Table 1 illustrates a consistent performance decline across all reproducibility factors. Specifically, changing the time period resulted in an average AUC drop of 4 points for Weeks 2 and 3, whereas alterations in the culture medium caused the largest decrease (up to a 25-point drop when using the selective medium CLO). Variability due to different hospital equipment led to a 9-point reduction in AUC. Despite these degradations, the best-performing models—RF and LGBM—exhibited excellent baseline performance (AUC = 0.99) and high resilience to variability (AUC = 0.89), although both still suffered declines when different media were employed.

In line with these observations, Table 2 shows that the enriched SCS

			RF	SVM + RBF	SVM + Linear	KNN	LGBM	Mean
Baseline SCS			0.99	0.97	0.78	0.91	0.98	0.93
Variability	Time	Week 2	0.97	0.95	0.69	0.86	0.95	0.88
		Week 3	0.96	0.93	0.69	0.82	0.91	0.86
	Media	BD	0.97	0.93	0.70	0.84	0.98	0.88
		CLO	0.94	0.89	0.56	0.68	0.84	0.78
		CDIF	0.94	0.81	0.63	0.67	0.84	0.78
	Mean		**0.96±0.02**	0.90±0.07	0.66±0.08	0.77±0.12	0.90±0.07	

Table 2: AUC scores of **Baseline SCS** model, i.e., training with week 1, enriched SCS medium, no PE and HGUGM machine.

medium produces more robust results, with RF's baseline AUC decreasing only from 0.99 to 0.96 under average variability conditions, while LGBM's AUC remains effectively unchanged (0.90 to 0.90). Nevertheless, enriched media are less practical for routine use since *C. difficile* must first be isolated on selective media (CDIF), thereby extending the total growth time by an additional 24 to 48 hours.

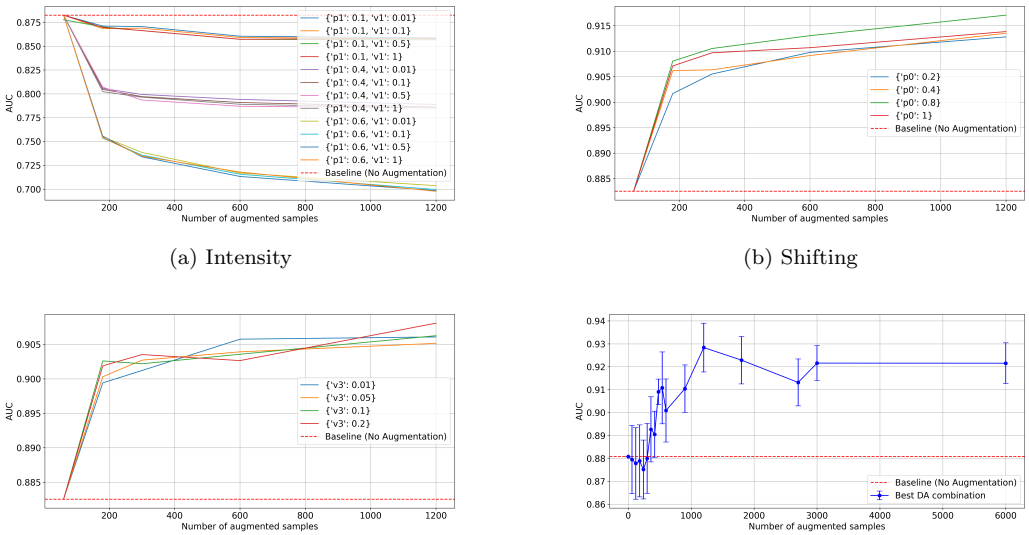
These results emphasize the critical impact of media variability on ML model performance and support the need for tailored methods—such as the data augmentation strategies implemented in our MALDIDA Python library—to enhance model generalizability across diverse laboratory environments.

3.2. ***Evaluation of the DA techniques***

As we proposed different strategies in Section 2.3 to mimic the variations affecting the spectrum, in this section, we begin with an ablation study of the DA process, analyzing the impact of each step and its parameters as the number of generated DA samples increases from 100 to 2500. For this experiment, we used an RF classifier, given the robustness demonstrated on previous results. For intensity variability, we explore 3 values of p_1 (25%, 50%, 75%) and apply different noise standard deviations (v_1) ([0.01, 0.05, 0.1, 0.5, 1]). For shifting, we proposed proportional shifting with $p_2 = [0.2, 0.4, 0.6, 0.8, 1]$. For new peaks, with a $p_3 = 0.15$, three standard deviations (v_3) were analyzed, 0.01, 0.05, 0.1 or 0.2. Finally, we include an overall evaluation of the DA process including all beneficial steps. The models were trained with **Baseline CDIF** and tested with out-of-distribution data following the LOIO scheme. For each isolate left-out, the model is tested over 143 mass spectra: all three biological replicates, regardless of their variability (three times, four

mediums, two protocols and two hospital machines) except by the one kept in the training set.

The results of this analysis are shown in Figure 4, comparing their performances with those of the RF model without DA.



(a) Intensity

(b) Shifting

(c) New peaks

(d) Shifting + new peaks

Figure 4: Ablation study for the DA techniques. Subfigures (a-c) analyze the AUC of a RF trained with the different DA techniques proposed, while Subfigure (d) shows the RF AUC scores obtained by combining the two best DA techniques.

Peak intensity modification, despite its common use with SMOTE in the literature [33, 34, 35, 36, 37], resulted in reduced overall performance (Fig. 4a). In contrast, shifting peak positions led to consistent increases in AUC as more augmented data were added (Fig. 4b), indicating that the model benefits from exposure to this type of variability. Similarly, adding new peaks at zero-intensity regions improved performance by familiarizing the model with random spectral fluctuations. Based on these findings, we selected the two most effective augmentation techniques—shifting and new peaks—for further evaluation. In a final study (Fig. 4d), we tested augmentation sample sizes ranging from 0 to 6000 (100 times the original dataset size of 60) and found that performance improvements plateau after approximately 3000 additional samples (about 30 times the original dataset). Consequently, subsequent experiments were conducted using this final configuration.

			w/ Data Augmentation		w/o Data Augmentation			
			RF	**LGBM**	**RF**	**LGBM**	**RF in SCS**	**LGBM in SCS**
Baseline CDIF			**0.96**	**0.99**	0.95	0.99	0.99	0.98
Variability	Time	Week 2	0.98	0.98	0.92	0.96	0.97	0.95
		Week 3	0.98	0.98	0.93	0.94	0.96	0.91
	Media	BD	0.89	0.88	0.86	0.82	0.97	0.98
		CLO	0.88	0.87	0.81	0.87	0.94	0.84
		SCS	0.91	0.92	0.89	0.86	0.94	0.84
	PE	PE	0.98	0.95	0.95	0.95	-	-
	Hosp	HCDGU	0.93	0.88	0.93	0.88	-	-
	Mean		**0.93 ± 0.04**	**0.93 ± 0.04**	0.89 ± 0.05	0.89 ± 0.06	0.96 ± 0.02	0.90±0.07

Table 3: AUC of **Baseline CDIF** model with DA techniques, i.e., training with week 1, selective SCS medium, no PE and HGUGM machine. It is compared to the same models without (w/o) DA and agaisnt the best model obtained, i.e., a RF trained on enriched SCS medium, no PE and HGUGM machine.

3.3. ***ML model robustness with data augmentation***

In this subsequent experiment, we investigate whether the application of DA techniques to selective CDIF medium can reduce the performance decline of ML models due to variability. Based on the SCS results presented in Table 2, we seek to obtain comparable results in the CDIF medium through the implementation of DA techniques. For this experiment, we evaluated the performance of Random Forest (RF) and LightGBM (LGBM), as these were the models that achieved the best performance in Section 3.1.

To assess how close the selective medium with DA approaches the performance of the enriched medium, Table 3 compares the **Baseline CDIF** model trained with 3000 augmented samples (using shifting and new peak techniques) against the model trained without any DA and also against the enriched SCS medium model.

Table 3 illustrates the impact of DA on mitigating different sources of variability. For time variability, applying DA increased the AUC from 92 to 98 in Weeks 2 and 3, effectively removing time-induced bias. Similarly, for PE variability, DA improved the AUC from 0.95 to 0.98. In terms of media variability, the baseline RF model yielded AUC values of 0.86, 0.81, and 0.89 for BD, CLO, and SCS media, respectively; with DA, these values increased to 0.89, 0.88, and 0.91. Although DA enhanced model robustness, performance under media variability remains below the baseline AUC of 0.96 observed for other factors. Finally, when assessing variability due to different hospital machines, DA did not affect performance, with both conditions recording an AUC of 0.93.

4. Conclusion

This study demonstrates that variability in culture media, laboratory equipment, and protocols significantly impacts the performance of ML models for rapid *C. difficile* typing using MALDI-TOF data. Our results show that while time-related and protein extraction variability can be effectively mitigated through tailored data augmentation (DA) strategies, challenges persist with culture media and equipment variability. Notably, we demonstrate that applying DA to spectra obtained from selective media substantially enhances model performance, achieving results comparable to those obtained with enriched media. The proposed augmentation methods, particularly when used with RF and LightGBM, improve classification accuracy despite inherent data inconsistencies. Furthermore, our open-source MALDIDA Python library provides the community with dedicated tools for MALDI-TOF DA. Future work will focus on further reducing variability to enhance the generalizability of diagnostic models across diverse clinical settings.

5. Availability and Implementation

The reproducibility code is available at `https://github.com/aguerrerolopez/MaldiMaranon` and we make the data augmentation techniques available as a Python package, named MALDIDA, installed via pip.

6. Acknowledgments

This work is partially supported by grant PID2023-146684NB-I00 funded by MCIN/AEI/10.13039/501100011033 and ERDF/EU and by the project TEC-2024/COM-89. It is also partially supported by 2023-II-PI-02, funded by the Gregorio Marañón Health Research Institute.

References

[1] A. Tran, K. Alby, A. Kerr, M. Jones, P. H. Gilligan, Cost savings realized by implementation of routine microbiological identification by matrix-assisted laser desorption ionization–time of flight mass spectrometry, J. of clinical microbiology 53 (2015) 2473–2479.

[2] C. Weis, A. Cuénod, B. Rieck, O. Dubuis, S. Graf, C. Lang, M. Oberle, M. Brackmann, K. K. Søgaard, M. Osthoff, K. Borgwardt, A. Egli, Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning, Nature medicine 28 (2022) 164–174.

[3] R. J. Pais, F. Sharara, R. Zmuidinaite, S. Butler, S. Keshavarz, R. Iles, Bioinformatic identification of euploid and aneuploid embryo secretome signatures in IVF culture media based on MALDI-ToF mass spectrometry, J. of Assisted Reproduction and Genetics 37 (2020) 2189–2198.

[4] T. Brignoli, M. Recker, W. W. Lee, T. Dong, R. Bhamber, M. Albur, P. Williams, A. W. Dowsey, R. C. Massey, Diagnostic MALDI-TOF MS can differentiate between high and low toxic Staphylococcus aureus bacteraemia isolates as a predictor of patient outcome, Microbiology 168 (2022) 001223.

[5] Q. Mao, X. Zhang, Z. Xu, Y. Xiao, Y. Song, F. Xu, Identification of Escherichia coli strains using MALDI-TOF MS combined with long short-term memory neural networks, Aging 16 (2024).

[6] T.-H. Lin, H.-Y. Chung, M.-J. Jian, C.-K. Chang, H.-H. Lin, C.-M. Yu, C.-L. Perng, F.-Y. Chang, C.-W. Chen, H.-S. Shang, Innovative strategies against superbugs: Developing an AI-CDSS for precise Stenotrophomonas maltophilia treatment, J. of Global Antimicrobial Resistance 38 (2024) 173–180.

[7] C. Weis, C. Jutzeler, K. Borgwardt, Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review, Clinical Microbiology and Infection 26 (2020) 1310–1317.

[8] Y. Zeng, C. Wang, Q. Ye, G. Liu, L. Zhang, J. Wan, Y. Zhu, Machine learning model of imipenem-resistant Klebsiella pneumoniae based on MALDI-TOF-MS platform: An observational study, Health Science Reports 6 (2023).

[9] J. Czepiel, M. Dróżdż, H. Pituch, E. J. Kuijper, W. Perucki, A. Mielimonka, S. Goldman, D. Wultańska, A. Garlicki, G. Biesiada, Clostridium difficile infection, European J. of Clinical Microbiology & Infectious Diseases 38 (2019) 1211–1221.

[10] R. Markovska, G. Dimitrov, R. Gergova, L. Boyanova, Clostridioides difficile, a New "Superbug", Microorganisms 11 (2023) 845.

[11] D. A. Leffler, J. T. Lamont, Clostridium difficile infection, New England J. of Medicine 372 (2015) 1539–1548.

[12] 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 24 (2018) 469–475.

[13] 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 21 (2016) 30294.

[14] 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 45 (2013) 109–113.

[15] 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 27 (2022) 2100704.

[16] M. Kachrimanidou, S. Metallidis, O. Tsachouridou, C. Harmanus, V. Lola, E. Protonotariou, L. Skoura, E. Kuijper, Predominance of Clostridioides difficile PCR ribotype 181 in northern Greece, 2016–2019, Anaerobe 76 (2022) 102601.

[17] M. T. Olson, P. S. Blank, D. L. Sackett, A. L. Yergey, Evaluating reproducibility and similarity of mass and intensity data in complex spectra—applications to tubulin, J. of the American Society for Mass Spectrometry 19 (2011) 367–374.

[18] R. A. Giebel, C. R. Worden, S. Rust, G. T. Kleinheinz, M. Robbins, 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 71 (2010) 149–84.

[19] A. J. Saenz, C. E. Petersen, N. B. Valentine, S. L. Gantt, K. H. Jarman, M. T. Kingsley, 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 13 (1999) 1580–1585.

[20] Olson, Matthew T and Epstein, Jonathan A and Sackett, Dan L and Yergey, Alfred L, Production of reliable maldi spectra with quality threshold clustering of replicates, J. of the American Society for Mass Spectrometry 22 (2011) 969–975.

[21] S. Luk, W. K. To, T. K. Ng, W. T. Hui, W. K. Lee, F. Lau, A. M. W. Ching, A cost-effective approach for detection of toxigenic Clostridium difficile: toxigenic culture using ChromID Clostridium difficile agar, J. of Clinical Microbiology 52 (2014) 671–673.

[22] J. J. Yang, Y. S. Nam, M. J. Kim, S. Y. Cho, E. You, Y. S. Soh, H. J. Lee, Evaluation of a chromogenic culture medium for the detection of Clostridium difficile, Yonsei medical journal 55 (2014) 994–998.

[23] J. H. Chen, V. C. Cheng, O.-Y. Wong, S. C. Wong, S. Y. So, W.-C. Yam, K.-Y. Yuen, The importance of matrix-assisted laser desorption ionization–time of flight mass spectrometry for correct identification of Clostridium difficile isolated from chromID C. difficile chromogenic agar, J. of Microbiology, Immunology and Infection 50 (2017) 723–726.

[24] M. M. Nerandzic, C. J. Donskey, Effective and reduced-cost modified selective medium for isolation of Clostridium difficile, J. of clinical microbiology 47 (2009) 397–400.

[25] A. Cesaro, S. C. Hoffman, P. Das, C. de la Fuente-Nunez, Challenges and applications of artificial intelligence in infectious diseases and antimicrobial resistance, npj Antimicrobials and Resistance 3 (2025) 2.

[26] K. Seddiki, P. Saudemont, F. Precioso, N. Ogrinc, M. Wisztorski, M. Salzet, I. Fournier, A. Droit, Towards CNN Representations for Small Mass Spectrometry Data Classification: From Transfer Learning to Cumulative Learning, bioRxiv (2020). doi:10.1101/2020.03.24.005975. arXiv:https://www.biorxiv.org/content/early/2020/03/25/2020.03.24.005975.full.

[27] S. Gibb, K. Strimmer, MALDIquant: a versatile R package for the analysis of mass spectrometry data, Bioinformatics 28 (2012) 2270–2271.

[28] G. De Waele, G. Menschaert, W. Waegeman, An antimicrobial drug recommender system using MALDI-TOF MS and dual-branch neural networks, bioRxiv (2023). doi:10.1101/2023.09.28.559916.

[29] C. Weis, M. Horn, B. Rieck, A. Cuénod, A. Egli, K. Borgwardt, Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra, Bioinformatics (Oxford, England) 36 (2020) i30–i38.

[30] J. Corver, J. Sen, B. Hornung, B. Mertens, E. Berssenbrugge, C. Harmanus, I. Sanders, N. Kumar, T. Lawley, E. Kuijper, et al., Identification and validation of two peptide markers for the recognition of Clostridioides difficile MLST-1 and MLST-11 by MALDI-MS, Clinical Microbiology and Infection 25 (2019) 904–e1.

[31] M. Blázquez-Sánchez, A. Guerrero-López, A. Candela, A. Belenguer-Llorens, J. M. Moreno, C. Sevilla-Salcedo, M. Sánchez-Cueto, M. J. Arroyo, M. Gutiérrez-Pareja, V. Gómez-Verdejo, et al., Automatic, web-based typing of Clostridioides difficile ribotypes using MALDI-TOF MS, bioRxiv (2024) 2024–01.

[32] X. A. López-Cortés, J. M. Manríquez-Troncoso, J. Kandalaft-Letelier, S. Cuadros-Orellana, Machine learning and matrix-assisted laser desorption/ionization time-of-flight mass spectra for antimicrobial resistance prediction: A systematic review of recent advancements and future development, J. of Chromatography A (2024) 465262.

[33] C. Wang, Z. Wang, H.-Y. Wang, C.-R. Chung, J.-T. Horng, J.-J. Lu, T.-Y. Lee, Large-scale samples based rapid detection of ciprofloxacin resistance in Klebsiella pneumoniae using machine learning methods, Frontiers in microbiology 13 (2022) 827451.

[34] J. Fu, F. He, J. Xiao, Z. Liao, L. He, J. He, J. Guo, S. Liu, Rapid AMR prediction in Pseudomonas aeruginosa combining MALDI–TOF MS with DNN model, J. of applied microbiology 134 (2023) lxad248.

[35] A. Godmer, Q. Giai Gianetto, K. Le Neindre, V. Latapy, M. Bastide, M. Ehmig, V. Lalande, N. Veziris, A. Aubry, F. Barbut, et al., 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 17 (2024) e14478.

[36] C. A. Astudillo, X. A. López-Cortés, E. Ocque, J. M. Manríquez-Troncoso, Multi-label classification to predict antibiotic resistance from raw clinical MALDI-TOF mass spectrometry data, Scientific Reports 14 (2024) 31283.

[37] V. Macaya Mejias, D. Zabala-Blanco, X. A. López-Cortés, F. Tirado, J. M. Manríquez-Troncoso, R. Ahumada-García, Predicting bacterial antibiotic resistance using MALDI-TOF mass spectrometry databases with ELM applications, J. of Computer Science & Technology 24 (2024).