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

Lucía Bravo Antón[1], Alejandro Guerrero-López*[1,3], Lucía Schmidt-Santiago[1], Carlos Sevilla-Salcedo[1], Mario Blázquez-Sánchez[3], David Rodríguez-Temporal[3], Belén Rodríguez-Sánchez[3], Vanessa Gómez-Verdejo[1,2]
[1]Department of Signal and Communication Theory, Universidad Carlos III de Madrid, ES28911 Spain
[2]Signal Processing Group, Gregorio Marañón Health Research Institute, Madrid, 28040 Spain
[3]Department of Clinical Microbiology and Infectious Diseases, Gregorio Marañón Health Research Institute, Madrid, 28040 Spain Email: alguerre@pa.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 differentiation of *Clostridioides difficile* ribotypes. This study evaluates the integration of Machine Learning algorithms to enhance the robustness and accuracy of MALDI-TOF spectra analysis. For this purpose, we compared traditional classifiers to gain insight into how spectral variability affects their performance in bacterial typing. Furthermore, specific Data Augmentation techniques for MALDI-TOF spectra were designed to reduce 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 especially when facing variability in the sample acquisition process. This highlights that the combination of domain adaptation with machine learning (ML) could make MALDI-TOF a more versatile and reliable tool in clinical diagnostics.**

***Index Terms*—Machine Learning, Data Augmentation, MALDI- TOF spectra, *Clostridioides difficile*, Reproducibility**

I. Introduction

Matrix-Assisted Laser Desorption Ionization–Time of Flight (MALDI-TOF) mass spectrometry (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 area under the curve, 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. Ad-

ditionally, 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 a major concern in bacterial characterization, as studies have shown that these sources of variability impact MALDI-TOF performance both within and across laboratories [18]–[20]. To overcome this issue, Data Augmentation (DA) techniques have been underexplored in this field [21] without satisfactory results indicating that the need of a deeper biological variability study together to most adequate DA technique able to mimic the actual spectra variability.

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, the instrument affect the performance of commmonly 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, Data Augmentation techniques [21] 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.

II. Materials and Methods

A. *MALDI-TOF spectra*

To evaluate how different conditions affect the MALDI-TOF MS spectra of bacterial isolates, we have analyzed 60 strains focusing on different variables: inter-week variability, culture media, sample processing protocols and mass spectrometers at different hospitals.

For inter-week variability and culture media effects, the samples were evaluated over three weeks using the direct smear protocol on four culture media: CHROMID® *C. difficile* agar (Ch), Schaedler agar (Sc), Brucella agar (Br) and Clostridium agar (Cl). The use of a PE protocol was also compared to the direct smear method on the same media including, for the Ch media, an additional evaluation of the incubation time impact with an incubation time of 24 hours, in comparison to the usual 48 hours of incubation.

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). In order to test hospital variabilities, all isolates were processed using the direct smear method on Ch at another hospital, Hospital Central de la Defensa Gómez Ulla (CDGU), performed with its MALDI Biotyper® sirius System (Bruker Daltonics), i.e., different from the HGUGM one, but with same MALDI-TOF MS technology.

These 60 strains were composed of 10 isolates of each of the six classes, namely RT023, RT027,078, RT106, RT165 and the last ten from RT181.

B. *Preprocessing and Binning*

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

$$M = \{p, i\} \quad \text{where} \quad p, i \in (\mathbb{R}^+)^{18,000}$$

where p denotes the positions (in m/z, measured in Daltons) and i denotes the corresponding intensities, both constrained to positive real numbers, with a fixed length of $18,000$ after trimming. For the preprocessing, we started applying the conventional open-source software MALDIquant R package [22] to preprocess MALDI-TOF MS, but using the version in Python code extracted from [23], which consists of five steps: (1) the intensity is transformed with a square-root method to stabilize the variance, (2) a smoothing using the Savitzky-Golay algorithm with half-window-size 10 is applied, (3) a baseline estimation is removed, this estimation is computed with 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 2,000–20,000 Da range. Figure 1 shows a result of all these steps.

After preprocessing, following previous works [2], [24], 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 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 was represented by a vector of fixed dimensionality containing 6000 features.

C. *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 2 for an RT027 sample — reveals significant spectral variability. This includes peaks that appear or disappear, changes in intensity, and shifts in positions. Such variability can

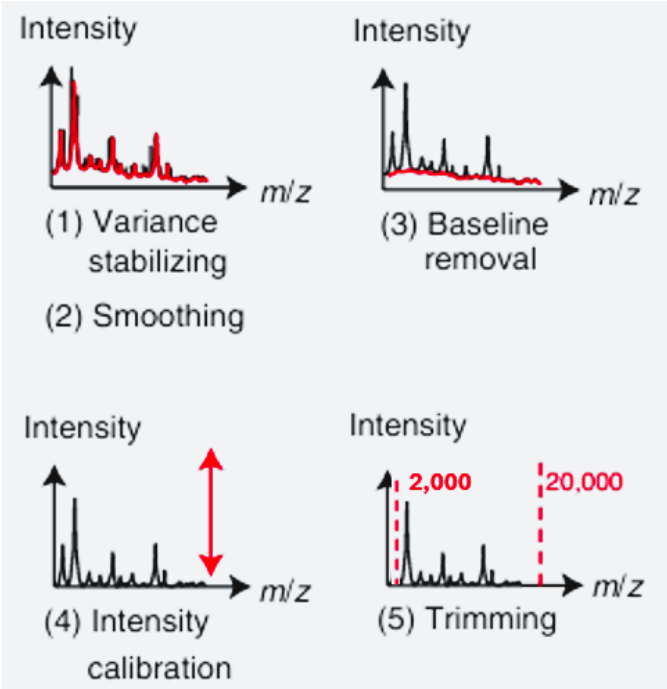


Fig. 1: Summary of the preprocessing steps.

lead to overfitting in learning models if they fail to capture these variations during training, 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 where this variability may not be fully represented, we propose simulating these variations through DA.

In particular, in order to reproduce these spectra differences, we propose a 3-step DA procedure:

1) Intensity variability: To recreate the variability in intensities, specially between mediums and weeks, we first add random Gaussian noise with zero mean and a given variance v_1 to a random percentage (p_1) of peaks.
2) Shifting: randomly shifting positions along the m/z dimension by ± 3 Da, ± 6 Da or ± 9 Da. With a given probability (p_{0Da}), a significant proportion of the peaks remain unshifted, while with(p_{3Da}) the peaks exhibit a ± 3 Da shift, with (p_{6Da}) a ± 6 Da shift, and, finally, with(p_{9Da}) a ± 9 Da shift where $p_{3Da} > p_{6Da} > p_{9Da}$.
3) Machine variability: As final step, zero mean gaussian noise with low variance v_3 was introduced to spectra regions with intensities equal to zero. The objective of this step is to reproduce the spectra variations caused by the used of different machines. As illustrated in Figure 2, the spectra from CDGU present peaks that are absent in HGUGM.

D. *Experimental Setup*

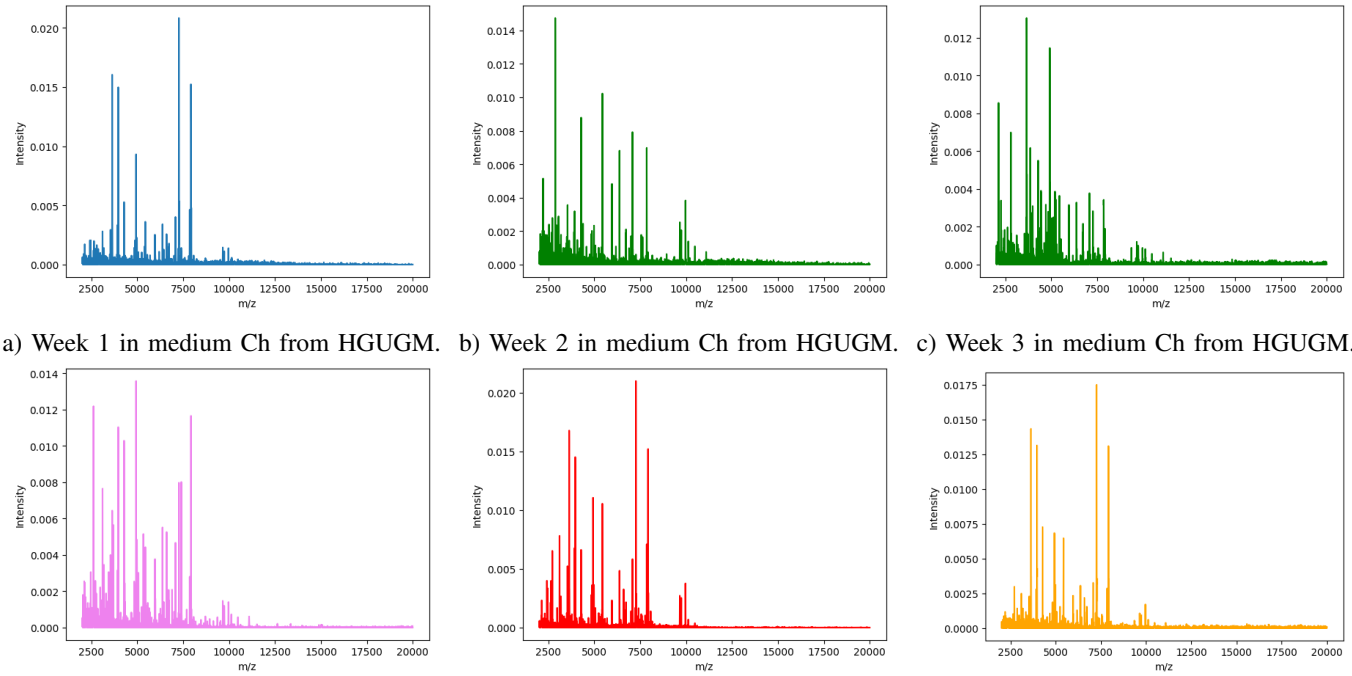
nf spectral variability on several conventional ML classifiers, using spectra from different acquisition times, PE protocols, media, and machines. Second, we evaluate the advantThenf the DAhow can mitigate the impact of this variability.

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 [25]: (i) a Support Vector Machine (SVM) with most common kernels such as linear kernel and Radial Basis Function (RBF) and with the Peak Information Kernel (PIKE), which was specifically developed for MALDI-TOF MS data [24]; (ii) K-Nearest Neighbors (KNN); (iv) a Random Forest (RF) ; and (v) a Light Gradient Boosting Machine (LGBM). The hyperparameters of these models were validated through a 5-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, and a range of neighbors for KNN between 3 and 20.

Two baseline models were defined: (1) **Baseline Ch** model trained on week 1, medium Ch, no PE and extracted by HGUGM's machine; (2) **Baseline Sc** similar to Baseline Ch but using the Sc medium.

Each one of these models was tested on an outer Leave-One-Out (LOO) strategy to assess variability performance. Five scenarios were proposed where, for each case, one unique variable was changed: a) no variability at all, i.e., test data has identical conditions as train data; (b) time variability, i.e., test data was analysed in different weeks (week 2 and week 3), (c) media variability, i.e., test data was analysed in three different media (Br, Cl and Sc/Ch -always different to the baseline-); (d) PE variability, i.e., direct PE or 24h incubation PE were used in test data; and (e) machine variability, i.e., test data was analysed on different hospital and machine.

It is important to note that the left-out isolate was never included in the training data and was used solely for testing purposes. This LOO process was repeated for each isolate in the dataset, ensuring that every sample was used as a test case in turn. This provided mean



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

d) Week 1 in medium Sc from HGUGM. e) Medium Ch with PE from HGUGM. f) Medium Ch from CDGU.

Fig. 2: Differences between spectra for the sample sample of RT027 in different weeks (sub-figures (b) and (c) in green), culture media (sub-figure (d) in pink), with PE (sub-figure(e) in red) and machines (subfigure (f) in yellow).

and standard deviation results for each model under each testing scenario.

To evaluate the DA performance, the best baselines from above was replicated using DA techniques to extend the training pool.

III. Results

A. Variability analysis regarding ML performance

Tables I and IV present the results of **Baseline Ch** and **Baseline Sc** models, respectively. Additional results for similar baselines using BR and Cl media, **Baseline Br** and **Baseline Cl**, are provided in Appendix A. In all tables, the first row shows the baseline accuracy, obtained when testing the model on samples acquired under identical conditions to the training data (same week, medium, no PE method, and machine). Subsequent rows display the accuracy when only one condition varies: varying sample time acquisition, media, and PE method; besides, for the Ch media, we also varied the machine using the MALDI-TOF equipment from CDGU hospital.

Each column in the tables represents the results for a specific model used to create the baseline. Each cell contains:

- The mean accuracy obtained using the Leave-One-Out (LOO) methodology.
- The Absolute Percentage Error (APE), shown in parentheses, calculated as:

$$\text{APE}_i = \left| \frac{\text{Baseline} - \text{Variability}_i}{\text{Baseline}} \right| \times 100$$

  where Baseline represents the accuracy of baseline model on baseline test samples and Variability_i represents the accuracy of the baseline model on test samples with the i-th condition varied. This indicates the performance increase or decrease (in percentage) for each variable tested.

The final column provides the mean accuracy and APE 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 I shows a consistent performance decline across all tested reproducibility variables.

Changes in time periods resulted in the smallest impact on performance, with APEs of 2.82 for Week 2 and 13.55 for Week 3 across all models. In contrast, switching to different culture media caused the largest performance drop, with a APE of 52.27 for Cl medium. Changes in PE protocols also negatively affected performance, with APEs of 15.97 and 31.88 for the alternative

			RF	SVM + RBF	SVM + PIKE	SVM + Linear	KNN	LGBM	Mean
Baseline Ch			76.67	83.33	70.00	78.33	71.66	100.00	80.33
Variability	Time	Week 2	75.00 (2.18)	90.00 (8.00)	61.67 (11.90)	81.67 (4.26)	71.67 (0.14)	88.33 (11.67)	78.06 (2.82)
		Week 3	75.00 (2.18)	75.00 (9.99)	56.67 (19.04)	70.00 (10.63)	66.67 (6.96)	73.33 (26.67)	69.44 (13.55)
	Media	Br	46.67 (39.13)	46.67 (43.99)	55.00 (21.43)	56.67 (27.65)	35.00 (51.16)	65.00 (35.00)	50.83 (36.72)
		Cl	56.67 (26.08)	36.67 (55.99)	25.00 (64.28)	30.00 (61.70)	20.00 (72.09)	61.67 (38.33)	38.33 (52.27)
		Sc	58.33 (23.92)	48.33 (42.00)	43.33 (38.10)	51.67 (34.03)	40.00 (44.18)	70.00 (30.00)	51.94 (22.89)
	PE	PE	63.33 (17.39)	78.33 (6.00)	63.33 (9.52)	73.33 (6.38)	50.00 (30.22)	76.67 (23.33)	67.50 (15.97)
		24H PE	60.00 (21.74)	56.67 (31.99)	43.33 (38.10)	58.33 (25.53)	45.00 (37.20)	65.00 (35.00)	54.72 (31.88)
	Hosp	CDGU	56.67 (26.08)	56.67 (31.99)	48.33 (30.95)	53.33 (31.91)	48.33 (32.55)	53.33 (46.67)	52.77 (34.30)
	Mean		61.46 (18.96)	61.04 (26.78)	49.58 (29.17)	59.37 (24.20)	47.08 (34.30)	69.17 (30.83)	57.95 (27.86)

TABLE I: Accuracy percentages of **Baseline Ch** model, i.e., training with week 1, Ch medium, no PE and HGUGM machine. In parentheses, relative differences in terms of APE compared to baseline accuracy.

PE methods tested. Testing samples on equipment from a different hospital resulted in a APE of 34.30, indicating that machine-induced variability significantly affects model performance.

Regarding model robustness, the RF model demonstrated the highest resilience to variability, showing a relatively smaller APE decrease of 18.96 from the baseline. On the other hand, the Light Gradient Boosting Machine (LGBM) achieved the highest global mean accuracy (69.17%) across all conditions. However, it was more sensitive to variability, with a larger APE drop of 30.83. It is worth noting that the baseline model accuracy was 100%.

In summary, the highest variability was observed when switching culture media, while changes in time periods had the least impact. Among the models, LGBM performed the best in terms of mean accuracy, while RF was the most robust to variability.

For the Sc medium (Table IV), the variability results were worse compared to Ch medium. All tested conditions exhibited a APE decrease of at least 26.29, indicating that Sc medium is less robust to variability, making it more susceptible to changes in experimental conditions.

B. Data augmentation experiments

To mitigate the degradation of ML models caused by variability, in this section, we apply the previously described DA techniques.

We start with an ablation study of the DA process analyzing the impact of each step and their parameters over the performance of a RF classifier, given the robustness demonstrated on previous results, when the number of generated DA samples goes from 100 to 2500. In Step 1, we explored 3 values of p_1 (25%, 50%, 75%) and applied different noise standard deviations (v_1) ([0.01, 0.05, 0.1, 0.5, 1]). In Step 2, two configurations of shifting probabilities ($p_{0Da}/p_{3Da}/p_{6Da}/p_{9Da}$ values were compared; concretelly, we established percentages of 60% / 20% / 10% / 10% and 50% / 30% / 15% / 5%. In Step 3, three standard deviations (v_3) were analyzed, 0.01, 0.05, 0.1 or 0.2. Finally, we include a overall evaluation of the DA process including all the steps when $p_1 = 0.75$, $v_1 = 0.05$, $p_{0Da}/p_{3Da}/p_{6Da}/p_{9Da}$ is set 50% / 30% / 15% / 5% and $v_3 = 0.01$. For this study, the RF was trained with **Baseline Ch** and was tested with the LOO scheme using the remaining spectra, independently of the time, medium, PE and machine. To obtain statistically relevant results, each experiment was conducted 100 times, with the mean and standard deviation calculated for each run. The results of this analysis are shown in Figure 3 comparing their performances with that of the RF model without DA.

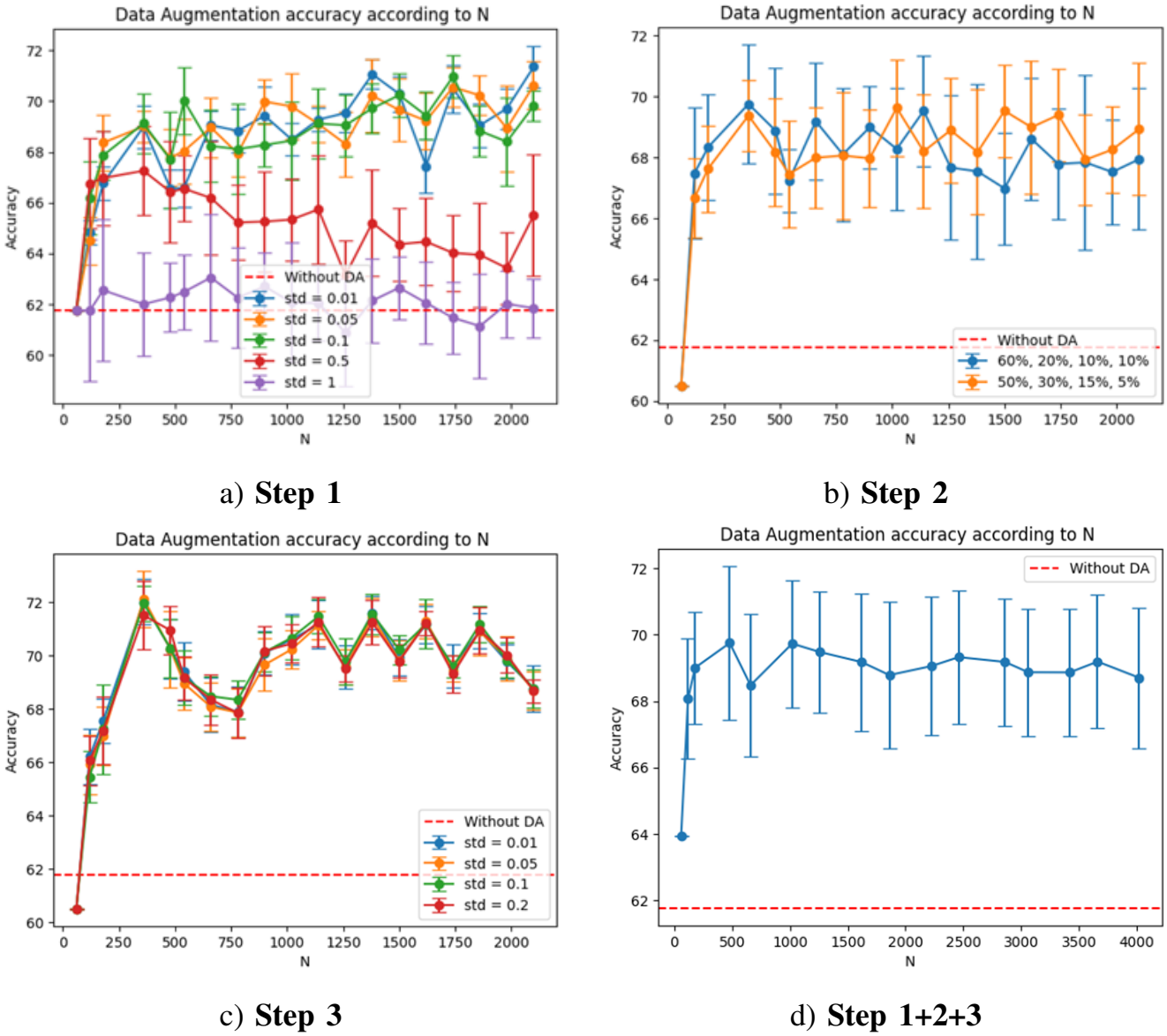
We can observe as in Step 1, i.e., peak intensity variability (see Subfigure 3(a)), the performance increases when peak intensities are slightly modified (v_1 around 0.1 or lower), however higher variations ($v_1 = 0.5$) do not provide significant benefits, and with $v_1 = 1$, the model performs similarly to the baseline without DA. Additionally, for low v_1 values, increasing the number of generated samples N tends to improve the model robustness against sample variability.

For Steps 2 (shifting peaks positions) and 3 (adding random peaks when there is no peak) (see Subfigures 3(b-c)), the models show similar results regardless of the configuration used, which demonstrates strong robustness in selecting the DA process parameters. However, in these two steps, some instability is observed, with improvements and declines in performance as N increases.

Finally, when combining all processes, the model exhibits more stable behavior, achieving approximately

			RF	SVM + RBF	SVM + PIKE	SVM + Linear	KNN	LGBM	Mean
Baseline Sc			91.66	76.66	70.00	71.66	71.66	100.00	80.27
Variability	Time	Week 2	88.33 (3.63)	65 (15.21)	55.00 (21.42)	58.33 (18.60)	30 (58.13)	48.33 (61.67)	57.50 (28.37)
		Week 3	66.67 (27.26)	60 (21.73)	55.00 (21.43)	53.33 (25.58)	36.67 (48.83)	55.00 (45.00)	54.44 (32.17)
	Media	Br	75 (18.17)	70.00 (8.69)	56.67 (19.04)	53.33 (25.58)	35.00 (51.16)	65.00 (35.00)	59.17 (26.29)
		Cl	55.00 (39.99)	35.00 (54.34)	23.33 (66.67)	30.00 (58.14)	20.00 (72.09)	61.67 (38.33)	37.50 (53.28)
		Ch	65.00 (29.08)	43.33 (43.48)	28.33 (59.53)	38.33 (46.51)	70.00 (2.32)	81.67 (18.33)	54.44 (32.17)
	PE	PE	75.00 (18.17)	55.00 (28.25)	36.67 (47.61)	45.00 (37.20)	45.00 (37.20)	45.00 (37.20)	50.28 (37.37)
	Mean		70.83 (20.83)	54.72 (28.26)	42.50 (39.20)	46.38 (35.27)	39.44 (44.96)	59.44 (40.66)	52.22 (34.94)

TABLE II: Accuracy percentages of **Baseline Sc** model, i.e., training with week 1, Sc medium, no PE and HGUGM machine. In parentheses, relative differences in terms of APE compared to baseline accuracy.



a) **Step 1**

b) **Step 2**

c) **Step 3**

d) **Step 1+2+3**

Fig. 3: Ablation study for the DA process analyzing the performance of a RF classifier respect to the number of samples generated by DA. Subfigures (a-c) analyze the RF accuracy for the differente DA steps, while Subfigure (d) shows the results of the joint process.

70% accuracy with around 500 generated samples, compared to 62% accuracy for the baseline model without DA.

Tables III, for **Baseline Ch**, and **??**, for **Baseline Sc**, show the impact of the three DA steps to each Variability$_i$ element replicating first experiment with all models.

Variability results are consistent with the observed without DA techniques being the time the less significant variability and the media the most. However, adding DA clearly impacts the overall performance of the models where RF, the most robust to variability, has increased its overall performance from 61.45% to 76.73% for different conditions.

FALTA TEXTO ANALIZANDO LA TABLA (como en la sección anterior)

IV. Discussion

In the initial experiment, which used a baseline of a single week in a specific medium (Tables I, IV, and Appendix A), RF and LGBM achieved the highest baseline accuracies compared to the other classifiers, with RF demonstrating the most robustness under variability conditions across all scenarios.

Regarding the choice of medium, Ch agar provided the best intra-media results, achieving a mean accuracy of 80.33%, and proved the most robust to variability with 57.95% accuracy, outperforming Sc (52.22%), Cl (43.19%), and Br (50.97%).

Nonetheless, the performance drop due to variability is substantial, as evidenced by LGBM's decline from 100% accuracy in identical training and testing conditions to 69.17% under variability. These findings highlight DA as an essential tool to improve performance under such conditions. When DA was applied, RF achieved a mean accuracy of 76.73% across all variability scenarios, outperforming other models.

These results underscore that for developing diagnostic models in microbiology, where variability in conditions such as agar media, protein extraction protocols, MALDI-TOF MS machines, and time schedules is unavoidable, incorporating DA is critical to ensure model robustness and minimize performance degradation.

V. Conclusion

This study demonstrates the critical importance of addressing variability in MALDI-TOF MS data for robust classification of *C. difficile* RTs. The results highlight that classifiers such as RF and LGBM, while effective under controlled conditions, are significantly impacted by variability introduced by factors such as agar media, protein extraction protocols, and machine differences. Incorporating DA proved essential for mitigating this variability, enabling models to maintain performance across diverse conditions and ensuring reliable typification regardless of acquisition methodology.

Moreover, the use of media more robust to variability, such as Ch agar, allows for greater stability in conditions, thereby enhancing the performance of classifiers.

Future work

Future research should focus on further reducing the impact of variability by exploring advanced preprocessing techniques and expanding the analysis to include MALDI-TOF spectra from additional machines and laboratory settings. Extending this approach to other bacterial species and integrating it into broader clinical workflows will further validate its utility. These efforts will ensure even greater accuracy and robustness in strain identification, ultimately facilitating better control of hospital-acquired infections and improving patient outcomes.

References

[1] A. Tran, K. Alby, A. Kerr, M. Jones, and P. H. Gilligan, "Cost savings realized by implementation of routine microbiological identification by matrix-assisted laser desorption ionization–time of flight mass spectrometry," *Journal of clinical microbiology*, vol. 53, no. 8, pp. 2473–2479, 2015.

[2] 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

[3] R. J. Pais, F. Sharara, R. Zmuidinaite, S. Butler, S. Keshavarz, and R. Iles, "Bioinformatic identification of euploid and aneuploid embryo secretome signatures in ivf culture media based on maldi-tof mass spectrometry," *Journal of Assisted Reproduction and Genetics*, vol. 37, pp. 2189–2198, 2020.

[4] T. Brignoli, M. Recker, W. W. Lee, T. Dong, R. Bhamber, M. Albur, P. Williams, A. W. Dowsey, and 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*, vol. 168, no. 8, p. 001223, 2022.

[5] Q. Mao, X. Zhang, Z. Xu, Y. Xiao, Y. Song, and F. Xu, "Identification of escherichia coli strains using maldi-tof ms combined with long short-term memory neural networks," *Aging*, vol. 16, 06 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, and H.-S. Shang, "Innovative strategies against superbugs: Developing an ai-cdss for precise stenotrophomonas maltophilia treatment," *Journal of Global Antimicrobial Resistance*, vol. 38, pp. 173–180, 2024.

			RF	SVM + RBF	SVM + PIKE	SVM + Linear	KNN	LGBM	Mean
Baseline Ch w/o DA			76.67	83.33	70.00	78.33	71.66	100.00	80.33
Baseline Ch			100±0.00	98.33 ±0.00	88.72±1.03	95.78±0.44	92.22±1.65	100±0.00	95.84
Variability	Time	Week 2	90±0.78 (10.00)	94.17±1.54 (4.23)	77.5±3.82 (12.64)	93.67±1.25 (2.20)	78±2.56 (15.42)	90.67±2.62 (9.33)	87.33 (8.06)
		Week 3	81.67±2.30 (18.33)	84±2.91 (14.57)	72.67±2.13 (18.09)	80±2.36 (16.47)	72±3.71 (21.92)	77.83±3.34 (26.17)	78.02 (17.87)
	Media	Br	65.00±0.65 (35.00)	57.50±2.39 (41.52)	55.33±1.45 (0.60)	53.33±2.94 (5.89)	39.33±2.38 (12.37)	52.33±4.78 (19.49)	53.80 (43.86)
		Cl	70±0.76 (30.00)	48.83±2.24 (50.34)	39.5±1.67 (55.64)	41.33±2.67 (56.84)	27.83±1.67 (69.82)	50.08±3.59 (49.92)	46.26 (51.73)
		Sc	83.33±0.56 (16.67)	49.17±3.96 (49.99)	57.33±2.49 (35.38)	52.17±1.67 (45.53)	44±3.96 (52.28)	60.42±3.80 (39.58)	57.73 (39.75)
	PE	PE	81.67±1.89 (18.33)	80.83±2.5 (17.79)	80.33±2.56 (9.45)	82±1.94 (14.38)	62.17±2.79 (32.58)	82.5±5.36 (17.5)	78.25 (18.35)
		24H PE	75.5±4.54 (24.50)	59.67±1.80 (39.82)	58.67±1.80 (33.87)	59.33±1.86 (38.05)	52.17±1.83 (43.42)	67.08±7.19 (32.92)	62.07 (35.23)
	Hosp	CDGU	66.67±2.78 (33.33)	57.17±3.34 (41.85)	63.33±3.16 (28.61)	55.67±3.67 (41.87)	43.38±3.08 (52.96)	65.83±3.83 (34.17)	58.67 (38.77)
	Mean		76.73 (23.27)	66.42 (32.45)	63.08 (28.89)	64.68 (32.47)	52.33 (43.24)	68.34 (31.65)	65.26 (31.91)
	Mean w/o DA		61.46 (18.96)	61.04 (26.78)	49.58 (29.17)	59.37 (24.20)	47.08 (34.30)	69.17 (30.83)	57.95 (27.86)

TABLE III: Accuracy percentages of **Baseline Ch** model with DA techniques, i.e., training with week 1, Sc medium, no PE and HGUGM machine. In parentheses, relative differences in terms of APE compared to baseline accuracy.

			RF	SVM + RBF	SVM + PIKE	SVM + Linear	KNN	LGBM	Mean
Baseline Sc w/o DA			91.66	76.66	70.00	71.66	71.66	100.00	80.27
Baseline Sc			100.00±0.00	98.61±1.09	79.33±2.57	93.78±1.49	79.97±1.18	100.00±0.00	91.95±1.06
Variability	Time	Week 2	89.17±1.71 (0.95)	83.00±2.21 (27.69)	56.50±1.17 (2.72)	75.00±2.11 (28.58)	68.67±2.96 (12.89)	70.42±5.89 (45.71)	73.79±2.68 ()
		Week 3	83.50±2.52 (25.24)	72.00±2.33 (20.00)	56.00±1.11 (1.81)	68.67±2.34 (28.76)	56.33±1.94 (53.61)	61.42±7.86 (11.67)	66.32±3.02 ()
	Media	Br	89.00±3.59 (18.66)	80.17±1.17 (14.53)	60.83±1.54 (7.34)	75.67±2.71 (41.89)	66.33±2.33 (89.51)	66.92±5.33 (2.95)	73.15±2.78 ()
		Cl	67.67±2.38 (23.04)	54.00±4.55 (54.28)	23.83±1.07 (2.14)	44.67±2.87 (48.90)	38.33±4.59 (91.65)	54.08±1.23 (12.31)	47.10±2.78 ()
		Ch	79.33±4.61 (22.05)	51.33±5.21 (18.46)	37.00±1.63 (30.60)	48.33±3.80 (26.09)	39.93±1.86 (43.81)	47.50±1.11 (41.84)	50.57±3.04 ()
	PE	PE	86.00±1.86 (14.67)	61.50±1.89 (11.81)	41.67±1.83 (13.63)	56.83±4.68 (26.29)	44.17±1.54 (1.84)	61.00±7.10 (7.64)	58.53±3.15 ()
	Mean		82.45±2.78 ()	67.00±2.89 ()	45.97±1.39 ()	61.53±3.08 ()	52.29±2.54 ()	60.22±4.75 ()	61.58 ±2.91 ()
	Mean w/o DA		70.83 (20.83)	54.72 (28.26)	42.50 (39.20)	46.38 (35.27)	39.44 (44.96)	59.44 (40.66)	52.22 (34.94)

TABLE IV: Accuracy percentages using as baseline spectra of the first week and Sc media, applying DA. In parentheses, percentages of relative differences between results without and with DA. Standard deviations are included.

[7] 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

[8] Y. Zeng, C. Wang, Q. Ye, G. Liu, L. Zhang, J. Wan, and Y. Zhu, "Machine learning model of imipenem-resistant klebsiella pneumoniae based on maldi-tof-ms platform: An observational study," *Health Science Reports*, vol. 6, no. 9, 2023.

[9] 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.

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

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

[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*, vol. 24, no. 5, pp. 469–475, 2018.

[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*, vol. 21, no. 29, p. 30294, 2016.

[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*, vol. 45, no. 1, pp. 109–113, 2013.

[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*, vol. 27, no. 26, p. 2100704, 2022.

[16] 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.

[17] 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

[18] 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

[19] 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

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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.

[25] X. A. López-Cortés, J. M. Manríquez-Troncoso, J. Kandalaft-Letelier, and 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," *Journal of Chromatography A*, p. 465262, 2024.

Appendix A
ML model performance

Tables V and VI present results for Section **??** in mediums Cl and Br. They consist of the accuracy percentage results of the alternative classification methods employed in the mentioned section, including relative differences compared to baseline, in order to facilitate observation of the impact of variability in acquisition on other media.

			RF	SVM + RBF	SVM + PIKE	SVM + Linear	KNN	LGBM	Mean
Baseline Cl			83.33	83.33	71.66	76.67	66.66	100	80.27
Variability	Time	Week 2	65.00 ()	68.33 ()	55.00 ()	68.33 ()	55.00 ()	81.67 ()	65.55 ()
		Week 3	61.67 ()	63.33 ()	43.33 ()	60.00 ()	45.00 ()	56.67 ()	55.00 ()
	Media	Br	23.33 ()	36.67 ()	16.67 ()	18.33 ()	18.33 ()	63.33 ()	29.44 ()
		Ch	23.33 ()	25.00 ()	13.33 ()	16.67 ()	10.00 ()	43.33 ()	21.94 ()
		Sc	28.33 ()	28.33 ()	15.00 ()	23.33 ()	16.67 ()	70.00 ()	30.28 ()
	PE	PE	70.00 ()	61.67 ()	56.67 ()	60.00 ()	26.67 ()	66.67 ()	56.95 ()
	Mean		45.28 ()	47.22 ()	33.33 ()	41.11 ()	28.61 ()	63.61 ()	43.19 ()

TABLE V: Accuracy percentages using as baseline spectra of first week and Cl medium. In parenthesis, relative differences compared to baseline accuracy.

			RF	SVM + RBF	SVM + PIKE	SVM + Linear	KNN	LGBM	Mean
Baseline Cl			86.66	81.66	76.66	78.33	73.33	100.00	82.77
Variability	Time	Week 2	75.00 ()	68.33 ()	58.33 ()	60.00 ()	43.33 ()	61.67 ()	61.11 ()
		Week 3	65.00 ()	56.67 ()	58.33 ()	53.33 ()	43.33 ()	51.67 ()	54.72 ()
	Media	Cl	56.67 ()	26.67 ()	21.67 ()	21.67 ()	15 ()	36.67 ()	29.72 ()
		Ch	66.67 ()	38.33 ()	33.33 ()	33.33 ()	28.33 ()	36.67 ()	39.44 ()
		Sc	68.33 ()	58.33 ()	55.00 ()	58.33 ()	38.33 ()	73.33 ()	58.61 ()
	PE	PE	66.67 ()	61.67 ()	63.33 ()	66.67 ()	50.00 ()	65.00 ()	62.22 ()
	Mean		66.39 ()	51.67 ()	48.33 ()	48.89 ()	36.39 ()	54.17 ()	50.97 ()

TABLE VI: Accuracy percentages using as baseline spectra of first week and Cl medium.In parenthesis, relative differences compared to baseline accuracy.