Research Article ISSN 2639-9458

Microbiology & Infectious Diseases

Integrating Machine Learning with MALDI-TOF MS for Advanced Anthrax Diagnosis and Surveillance

Rocca María Florencia[1,2,*], Motter Andrea[3], Etcheverry Paula[1], Noseda Ramón[4], Combiesses Gustavo[4] and Prieto Mónica[1]

[1]*Servicio Bacteriología Especial, Instituto Nacional de Enfermedades Infecciosas, ANLIS Dr. Carlos G. Malbrán, Argentina.*

[2]*Red Nacional de Espectrometría de Masas aplicada a la Microbiología Clínica ReNaEM, Argentina.*

[3]*Unidad Operativa de Contención Biológica, UOCCB, ANLIS Dr. Carlos G. Malbrán, Argentina.*

[4]*Laboratorio Azul, Buenos Aires, Argentina.*

[*]Correspondence:
Rocca María Florencia, Servicio Bacteriología Especial, Instituto Nacional de Enfermedades Infecciosas, ANLIS Dr. Carlos G. Malbrán, Argentina.

Received: 07 Dec 2024; **Accepted:** 29 Dec 2024; **Published:** 08 Jan 2025

Citation: Rocca María Florencia, Motter Andrea, Etcheverry Paula, et al. Integrating Machine Learning with MALDI-TOF MS for Advanced Anthrax Diagnosis and Surveillance. Microbiol Infect Dis. 2025; 9(1): 1-5.

ABSTRACT

Anthrax disease, caused by Bacillus anthracis, is a zoonotic disease with significant epidemiological implications. This study demonstrates the application of machine learning algorithms and MALDI-TOF mass spectrometry for rapid identification and profiling of B. anthracis and the related species Bacillus cereus in Argentina. Our results validate the efficacy of these techniques in creating a local database of peptide fingerprints, providing a foundation for robust surveillance and diagnosis in public health laboratories. Statistical analyses confirm species-specific biomarkers, supporting the development of accessible screening protocols. This approach highlights the potential for MALDI-TOF MS in anthrax diagnostics and lays groundwork for future expansions in pathogen profiling.

Keywords

Bacillus anthracis, MALDI-TOF, Machine learning, Diagnosis, Biomarkers, Surveillance.

Introduction

Anthrax disease, also known as carbuncle, is a zoonosis caused by *Bacillus anthracis*, a gram-positive, spore-forming member of the bacillus cereus group. Jadhav et al. [1] previously discussed the challenges posed by the genetic and proteomic similarities among these species, which complicate accurate identification.

It primarily affects livestock and wild animals all over the world, although it represents occupational risks for humans who come into contact with infected animals or contaminated animal products [2,3]. The disease can present in cutaneous, respiratory, or gastrointestinal forms, with rapid diagnosis essential for controlling its spread. In Argentina, Anthrax is a notifiable disease under National Law 15465, and recent outbreaks have underscored the need for improved diagnostic and surveillance techniques, particularly given the public health risk associated with bioterrorism agents [4]. Current diagnostic methods rely on a characteristic dark, hard, painless lesion accompanied by the particular epidemiology, followed by laboratory identification; traditional culture, biochemical assays, serology tests and PCR (https://wwwn.cdc.gov/nndss/conditions/anthrax).

However, MALDI-TOF MS has emerged as a powerful tool for microbial identification due to its speed, reliability, and cost-effectiveness [5]. Coupling MALDI-TOF with machine learning models allows for the classification of peptide mass fingerprints, enabling a robust differentiation of *B. anthracis* from closely related species, such as *B. cereus* sensu lato, as previously described by

Lasch et al. From the National Reference Laboratory, one of our missions consists in the design and validation of new diagnostic tools and the transference to small laboratories in endemic regions. This, added to the clinical and epidemiological importance of the detection of *B. anthracis* and the recent emergence of severe infections by *B. cereus group* that produces anthrax toxin, justified the development of our research. This study aims to establish a peptide profile database for *B. anthracis* and *Bacillus cereus* in Argentina and to develop machine learning models for rapid, accurate and safe screening technique.

Materials and Methods

Sample Collection and Culture

A total of 12 isolates obtained from various clinical, animal and environmental sources, were culture aerobically on blood agar plates at 37°C for 24 hours, according to biosafety recommendations from CDC. Three isolates recovered from animals were provided by Azul Laboratory; they appear as A in the table below. See Table 1. The rest of the strains correspond to the culture collection of the National Institute of infectious disease. In addition, all of the strains of *Bacillus anthracis* tested positive using PCR targeting the *B. anthracis*-specific pagA, capC genes.

To ensure sample inactivation, 500 µL of 80% Trifluoroacetic acid (TFA) was added in an Eppendorf with 1-2 loops of bacterial material, and the mixture was incubated for 30 minutes at room temperature, resulting in high-quality spectra with minimal background noise, insurance for processing [5]. The samples were checked for sterility during 4 days.

MALDI-TOF MS Analysis

Each sample was prepared for MALDI-TOF analysis by diluting the inactivated culture mixture in an equal volume of ultra-pure water (1:2 dilution). A 1 µl aliquot was placed onto a stainless-steel target plate, by quadruplicate, coated with 1 µl of HCCA (α-Cyano-4-hydroxycinnamic acid) matrix solution. Let air-dried before introduction into the Bruker SIRIUS MALDI-TOF MS system. Spectral acquisition was conducted using the flexControl software (v3.4) with automated peak acquisition settings, generating 12 individual spectra per sample, representing a total of 144 spectra for all the analysis.

Data Processing and Machine Learning Analysis

Database development: the new spectra were generated according to the MALDI BioTyper® standard MSP Creation in FlexControl 3.4 software. Each spectrum was obtained in a positive way linear through 240 laser shots (6 batches of 40 shots; N2 laser frequency: 60 Hz; source voltage ion I: 20 kV; Ion Source II Voltage: 16.7 kV, and range detector masses: 2000-20000 Da). A triple was made reading of each well, generating a total of 12 spectra for each isolate. Table 1.

Spectral data were imported into flexAnalysis (v3.4) and pre-processed using Savitzky-Golay smoothing and baseline correction applying the top hat algorithm. The processed spectra were visually inspected, and statistical analysis was performed to detect mass peaks within the 2-20 kDa range using centroid algorithm (Figure 1).

Table 1. Microorganisms used in the study and information of the MSPs created.

Microorganism	Collection number	Frequency peaks	Peak lists
Bacillus anthracis	CCBE 26-13	70>92%	11
Bacillus anthracis	92411	70>92%	11
Bacillus anthracis	STERN	70>92%	11
Bacillus anthracis (A)	474-23 182814	70>92%	11
Bacillus anthracis (A)	474-23 102271	70>92%	11
Bacillus anthracis (A)	474-23 33900	70>92%	11
Bacillus cereus	CCBE 13-17	70>92%	11
Bacillus cereus	CCBE 134-22	70>92%	11
Bacillus cereus	CCBE 491-24	70>92%	11
Bacillus cereus	CCBE 525-21	70>92%	11
Bacillus cereus	CCBE 641-22	70>92%	11
Bacillus cereus	CCBE 870-18	70>92%	11

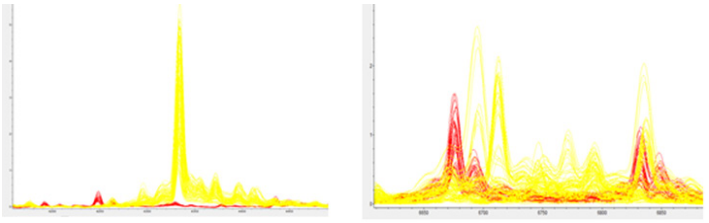

Figure 1: Overlaid view of spectra from *B. anthracis* (red) and *B. cereus* (yellow) in Flex Analysis software v3.4.

Key biomarkers were identified and validated using hierarchical clustering and Principal Component Analysis (PCA), with statistical testing applied to confirm marker significance.

Machine learning models were developed using ClinProTools and Clover Biosoft software to distinguish between *B. anthracis and B. cereus.* Models included genetic algorithms, supervised neural networks, fast classifiers, random forest, achieving high validation and classification rates (100% accuracy in 10-fold cross-validation) in every case.

Results

Species Identification and Database Creation

All isolates were identified to the species level using a free MALDI-TOF bioterrorism database (https://spectra.folkhalsomyndigheten.se/spectra/database/bruker.action, Bruker). Then, a local database of peptide fingerprints was generated, marking the first *B. anthracis and B. cereus* fingerprint library in Argentina (Figure 2), which will contribute to national and international surveillance networks [4], because this database has already been transferred to every institution using MALDI-TOF in Argentina and to Microbenet-CDC site on the Argentinian node.

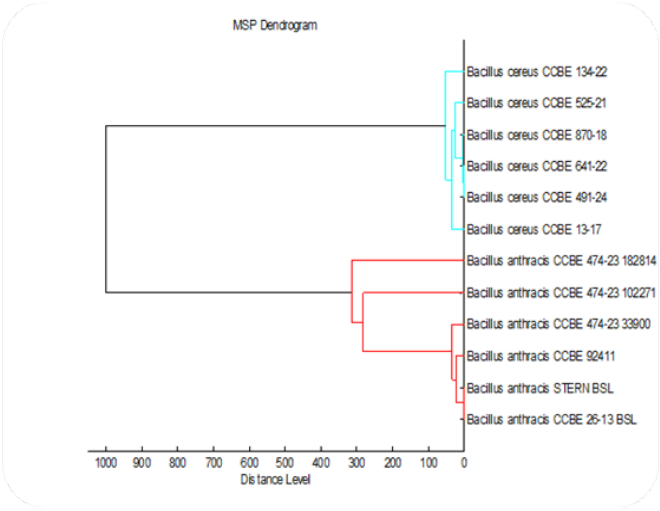

Figure 2: MSPs dendrogram using Maldi Biotyper Compass Explorer v 3.1 software.

Biomarker Identification

The first proteomic analysis on Flex Analysis revealed several probable species-specific peptide markers, validated statistically with Clin ProTools (Table 2). Notably, majority of markers achieved AUC values of 1.0 in ROC curve analyses, indicating strong discriminatory power between *B. anthracis* and *B. cereus*. In particular, two specific peaks (2994 Da and 6262 Da) were statistically significant ($p < 0.05$) and distinctly separated the species groups in dimensional plots (Figure 3).

Table 2: Predictive biomarkers detected for *Bacillus anthracis and Bacillus cereus* strains.

	Bacillus cereus	***Bacillus anthracis***
Predictive BioMarkers	4333 Da	5413 Da
	5886 Da	6675-79 Da
	7163 Da	2994 Da (1°BM PCA)
	9211 Da	3203 Da
	15100 Da	2603 Da
	16000 Da	3450 Da
	6262 Da (2° BM PCA)	4248 Da
		3353 Da

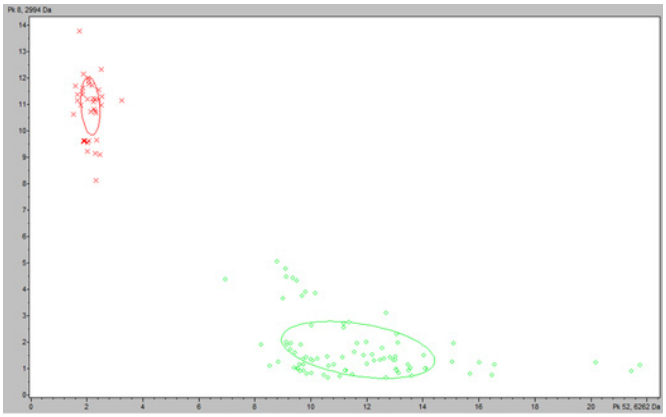

Figure 3: Image of two-dimensional distribution of the spectra not excluded from each class, based on the two best peaks for separation according to the p-value. The values are shown on the x and y axes, are of the peak 2994 Da and 6262 Da respectively, both detected with statistical significance. Clearly the species form two distinct groups (red: *Bacillus anthracis* spectra, green: *Bacillus cereus*).

Principal Component Analysis (PCA)

PCA was performed on the spectral data, showing that the first three principal components accounted for approximately 80% of the variance (Figure 4A, 4B). This allowed clear visualization of the clustering patterns, with *B. anthracis and B. cereus* samples forming distinct groups based on their spectral fingerprints.

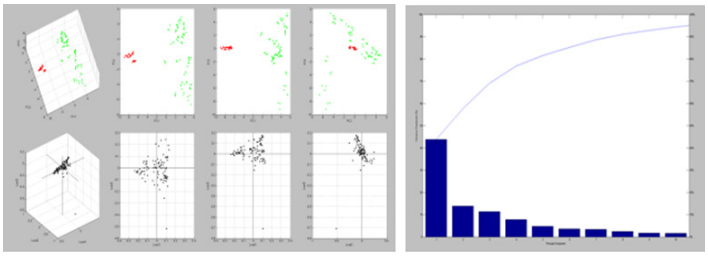

Figure 4A, 4B: Principal component analysis and variance image on ClinProTools software unsupervised clustering.

Machine Learning Models

Three machine learning algorithms (genetic algorithm, supervised neural network, and fast classifier) were applied using Clin ProTools, achieving a 100% accuracy in 10-fold cross-validation. Key parameters for the models included a resolution of 800, baseline correction with top hat (10%), signal-to-noise ratio of 5, and selection of 25 best peaks based on p-value. The models showed high classification rates, with sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) reaching 100% value.

In addition, multivariate analysis with Clover Biosoft confirmed the biomarkers in ROC curves, with minimum AUC values of 0.94 for both categories (Figure 5). A 3D PCA plot showed a clear separation of *B. anthracis and B. cereus* clusters (Figure 6), while hierarchical clustering analysis (HCA) illustrated taxonomical grouping of these species (Figure 7).

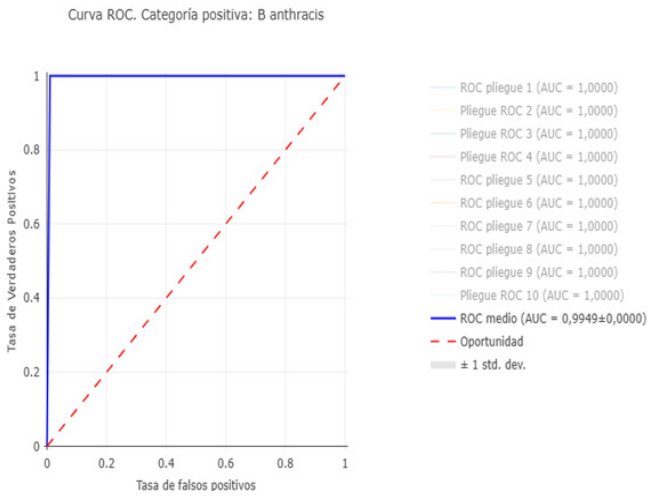

Figure 5: ROC curve obtained from Clover Biosoft, considering *Bacillus*

anthracis as positive category.

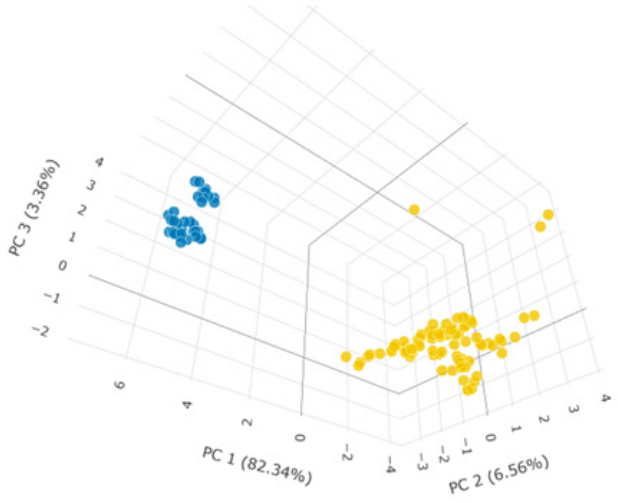

Figure 6: Principal Component analysis of *Bacillus anthracis* spectra (blue) and *Bacillus cereus* spectra (yellow).

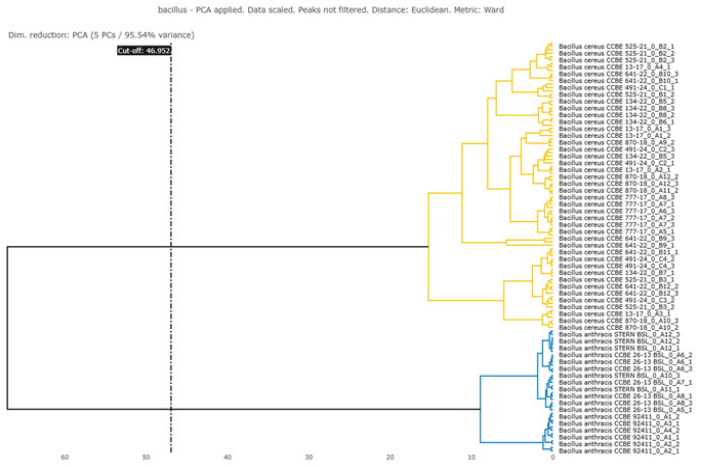

Figure 7: Hierarchical clustering analysis obtained in Clover Biosoft software.

The random forest model, performed with Clover Biosoft, yielded the best theoretical values of performance, after a 10k-fold validation, as shown in Table 3.

The random forest model, performed with Clover Biosoft, yielded the best theoretical values of performance, after a 10k-fold validation, as shown in Table 3.

Table 3: Performance results of Random Forest model in Clover Biosoft software:

REAL	*B. anthracis*	*B. cereus*	% correct
B. anthracis	36 (True Positive)	0 (False Negative)	100
B. cereus	0 (False Positive)	84 (True Negative)	100
	100%	100%	100

Accuracy: 100%, Accuracy: 100%, Sensitivity: 100%, Specificity: 100%, NPV: 100%, PPV: 100%.

This is a distance plot, basically it is a graphical representation that shows the distance between groups of microorganisms based on their mass profiles and allows to analyze similarity relationships or differences of the samples according to their spectral characteristics only (Figure 8).

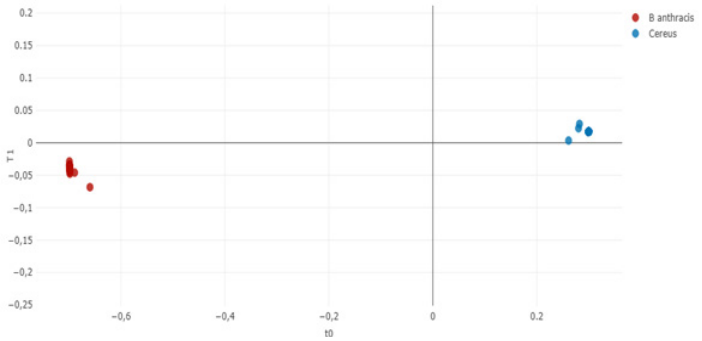

Figure 8: Random forest plot obtained in Clover Biosoft software.

Discussion

The application of MALDI-TOF MS combined with machine learning demonstrates a rapid, accurate, and cost-effective approach for identifying and differentiating *B. anthracis from B. cereus* strains.

As previously cited by Rau et al. [6] exploring the application of MALDI-TOF mass spectrometry for differentiating species within the *Bacillus cereus* group to overcome the taxonomic challenges associated, which includes pathogenic and non-pathogenic species. These methods provide a promising alternative to PCR, especially in resource-limited settings where the cost of reagents and low incidence rates make PCR impractical [2,5].

Eddabra, Azana, and Soumya [7] provides an overview of MALDI-TOF as a transformative tool in clinical microbiology for bacterial identification by reduced reliance on traditional biochemical methods, but they they also highlight limitations such as the requirement for a comprehensive database and the challenge with detecting closely related species. The establishment of a new peptide fingerprint database specific to Argentina is a milestone for local anthrax diagnostics, allowing for greater autonomy and timely responses to potential bioterrorism threats. Key biomarkers identified in this study, including those at 2994 Da and 6262 Da, align with previously reported markers but also offer novel species-specific indicators for local *B. anthracis and B. cereus* isolates [3]. Kampfer et al. [8], reaffirms its contributions to molecular epidemiology by enabling the tracking and monitoring of bacterial outbreaks. They also discuss advancements in database development and its integration with molecular techniques to improve diagnostic accuracy strengthening public health surveillance and pathogen tracking efforts. The integration of artificial intelligence (AI) algorithms with MALDI-TOF MS, demonstrate improved pathogen differentiation, even among closely related species, and the ability to detect emerging or

atypical strains. The new researches highlights the transformative potential of AI in microbiological diagnostics, emphasizing its role in strengthening public health surveillance and response [9]. The machine learning models achieved robust classification, confirming that combining MALDI-TOF spectral data with predictive algorithms can significantly enhance species discrimination accuracy. The genetic algorithm, supervised neural network, fast classifier and random forest each reached a 100% cross-validation score, suggesting potential for deployment in clinical microbiology labs with minimal computational resources.

Conclusion

This study highlights the utility of MALDI-TOF MS and machine learning as powerful tools for anthrax surveillance and diagnostic workflows in Argentina. Key findings include: successful establishment of a local *B. anthracis and B. cereus* peptide fingerprint database, contributing to surveillance networks; identification of statistically significant biomarkers with high AUC values, enhancing diagnostic specificity; validation of machine learning models for species classification with perfect sensitivity, specificity, PPV, and NPV in 10-fold cross-validation.

The methodology used in this study for inactivation of highly pathogenic microorganisms, can also be applied to identify microorganisms of other genera that can arrive to the laboratory.

Future Perspectives: Expanding this MALDI-TOF fingerprint database with additional isolates and integrating Fourier-transform infrared (FTIR) spectroscopy could further refine species differentiation, supporting outbreak resolution and clonally studies. Additionally, we have started collaboration with international researchers through spectral data sharing may enhance global anthrax diagnostics without requiring sample exchange. Our next goal will be intending to predict the signals detected as ribosomal specific proteins.

Acknowledgment

Gaston D´Angiolo for the technical support throughout the entire project.

References

1. Jadhav S, Gulhane M, Zinjarde S, et al. MALDI-TOF mass spectrometry for accurate identification of Bacillus anthracis and closely related Bacillus cereus group organisms: A comprehensive review of challenges and advancements. Microb Pathog. 2021; 159: 105132.
2. Pauker VI, Thoma BR, Grass G, et al. Improved discrimination of Bacillus anthracis from closely related species in the Bacillus cereus sensu lato group based on MALDI-TOF MS. J Clin Microbiol. 2018; 56: 01900-01917.
3. Wei J, Zhang H, Zhao F, et al. Strategy for rapid and safe distinction between B. anthracis and B. cereus using peptide mass fingerprints based on MALDI-TOF MS. J Clin Microbiol. 2021; 59.
4. Manzulli V, Rondinone V, Buchicchio A, et al. Discrimination of Bacillus cereus group members by MALDI-TOF mass spectrometry. Microorganisms. 2021; 9: 1202.
5. Lasch P, Beyer W, Nattermann H, et al. Identification of Bacillus anthracis by using matrix-assisted laser desorption ionization-time of flight mass spectrometry and artificial neural networks. Appl Environ Microbiol. 2009; 75: 7229-7342.
6. Rau J, Perz R, Klittich G, et al. MALDI-UP: MALDI-TOF mass spectrometry as a rapid and cost-effective tool for differentiation of Bacillus cereus group species. Syst Appl Microbiol. 2019; 42: 507-513.
7. Eddabra R, Azana S, Soumya F. MALDI-TOF mass spectrometry as a powerful tool for bacterial identification in clinical microbiology: An overview of its benefits and limitations. J Bacteriol Mycol. 2020; 7: 1100.
8. Kämpfer P, Matthews H, Glaeser SP, et al. MALDI-TOF MS in public health microbiology laboratories: identification and molecular epidemiology of bacterial pathogens. Microorganisms. 2020; 8: 1557.
9. Röder G, Brandt C, Proske M, et al. New strategies for MALDI-TOF MS-based identification of pathogens relevant to public health: Incorporating artificial intelligence for enhanced accuracy. Front Microbiol. 2020; 11: 1117.