- ORIGINAL ARTICLE -

Predicting Bacterial Antibiotic Resistance using MALDI-TOF Mass Spectrometry Databases with ELM Applications

Predicción de resistencia antibiotica mediante bases de datos de espectrometría de masas MALDI-TOF con aplicaciones de ELM

Vicente Macaya Mejías[1], David Zabala-Blanco[1,3], Xaviera A. López-Cortés[1,2,3,4], Felipe Tirado[1,3], José M. Manríquez-Troncoso[1,4], and Roberto Ahumada-García[1]

[1]*Facultad de Ciencias de la Ingeniería, Universidad Católica del Maule, Talca, Chile*
{dzabala, xlopez}@ucm.cl
[2]*Center for Innovation in Applied Engineering (CIIA), Catholic University of Maule, Talca, Chile*
[3]*Departamento de Computación e Industrias, Facultad de Ciencias de la Ingeniería, Universidad Católica del Maule, Talca, Chile*
[4]*Multidisciplinary Intelligent Data Science Lab, Talca, Chile*

Abstract

Early detection of antibiotic resistance is a crucial task, especially for vulnerable patients under prolonged treatments with a single antibiotic. To solve this, machine learning approaches have been reported in the state of the art. Researchers have used MALDI-TOF MS in order to predict antibiotic resistance and/or susceptibility in bacterial samples. Weis, et al. implemented LR, LightGBM and ANN to study the antibiotic resistance on bacterial strains of *Escherichia Coli*, *Staphylococcus Aureus*, and *Klebsiella Pneumoniae*. Despite promising results, the models have not achieved perfect accuracy, specifically when the classes are unbalanced. On the other hand, Extreme Learning Machine (ELM) is a training algorithm for forward propagation of single hidden layer neural networks, which converges much faster than traditional methods and offers promising performance along with less programmer intervention. In this way, this study introduced improved ELMs, including two weighted ELMs proposed by Zong, and the SMOTE technique in order to create new synthetic samples of the minority class. After heuristic optimization of ELM hiperparameters, results demonstrated 85% in accuracy and 85% in geometric mean for the classification problem in the case of weighted ELM 1 subject to the SMOTE technique of oversampling.

Keywords: Antibiotic Resistance Prediction, MALDI-TOF Mass Spectrometry, Machine Learning in Medicine, Extreme Learning Machines, Weighted ELM

Resumen

La detección temprana de la resistencia a los antibióticos es una tarea crucial, especialmente en el caso de pacientes vulnerables sometidos a tratamientos prolongados con un único antibiótico. Para resolver este problema, se han utilizado métodos de aprendizaje automático. Los investigadores han utilizado MALDI-TOF MS para predecir la resistencia y/o susceptibilidad a los antibióticos en muestras bacterianas. Weis, et al. aplicaron LR, LightGBM y ANN para estudiar la resistencia a los antibióticos en cepas bacterianas de Escherichia Coli, Staphylococcus Aureus y Klebsiella Pneumoniae. A pesar de los prometedores resultados, los modelos no han logrado una precisión perfecta, concretamente cuando las clases están desequilibradas. Por otro lado, Extreme Learning Machine (ELM) es un algoritmo de entrenamiento para la propagación hacia delante de redes neuronales de una sola capa oculta, que converge mucho más rápido que los métodos tradicionales y ofrece un rendimiento prometedor junto con una menor intervención del programador. De este modo, este estudio introdujo ELMs mejorados, incluyendo dos ELMs ponderados propuestos por Zong, y la técnica SMOTE para crear nuevas muestras sintéticas de la clase minoritaria. Tras la optimización heurística de los hiperparámetros del ELM, los resultados demostraron un 85% de precisión y un 85% de media geométrica para el problema de clasificación en el caso del ELM ponderado 1 sujeto a la técnica SMOTE de sobremuestreo.

Palabras claves: Enter previous key words or phrases in alphabetical order in Spanish, separated by commas.

1 Introduction

Bacterial antibiotic resistance is considered one of the most significant challenges in global health [1]. De-

spite significant advances in contemporary medicine, this phenomenon continues to be an alarming concern. The ability of bacteria to adapt and develop antibiotic resistance has led to a critical situation, putting the success of medical treatments at risk and compromising the effectiveness of existing medications. This resistance impacts treatment effectiveness, increases the complexity of infections, prolongs hospital stays, and raises healthcare costs. In this context, the search for effective strategies to address this problem has become essential to ensure successful treatment of infectious diseases and preserve long-term public health [2, 3]. This issue has not only involved healthcare professionals. It has also led to interdisciplinary collaborations with engineering and data analysis experts, thus driving significant advances in various areas of medicine.

Early detection of antibiotic resistance is of vital importance, especially in patients with delicate conditions or those who have been exposed to long-term treatments with a single type of antibiotic. The potential of advanced technologies, such as Machine Learning (ML), has been leveraged to address this issue [4, 5]. Recently, the field of medicine has applied ML-based methods to analyze MALDI-TOF (Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight Mass Spectrometer) Mass Spectrometry (MS) data in order to identify different type of biomarkers [6, 7, 8, 9, 10, 11, 4].

Currently, in the fight against antibiotic resistance, techniques utilizing MALDI-TOF MS have been implemented to predict antibiotic resistance or susceptibility in bacterial samples [12]. In this work [13], three key bacterial strains were focused on: *Escherichia Coli, Staphylococcus Aureus*, and *Klebsiella Pneumoniae*. In their research approach, Weis applied three machine learning methods: logistic regression (LR), gradient-boosted decision trees (LightGBM), and a deep neural network classifier (multi-layer perceptron, MLP). Despite promising results, all models reached the threshold of 0.8 ± 0.03 precision in the area under the receiver operating characteristic curve and the area under the precision-recall curve. This challenge is partly attributed to the imbalance between positive (resistance) and negative (susceptibility) values in the databases used, with bacterial resistance being the minority class.

The results obtained in Weis's study demonstrated an excellent ability to predict the antibiotic resistance and susceptibility of the database samples. However, despite the good results obtained in Weis's research, the existence of an imbalance between positive (antibiotic resistance) and negative (antibiotic susceptibility) classes was not taken into account, with the latter being the class that even quintuples the number of cases over the positives, as seen in Fig. 1. The data imbalance leads to unsatisfactory performance of machine learning models, as they tend to favor the majority classes, ignoring or underestimating the minority classes. This disparity in the distribution of samples can severely distort the results, leading to biased and inaccurate models when predicting or classifying the less represented classes [14]. Because this is a constant problem in antibiotic resistance analysis, it is necessary to study different methods for antibiotic resistance analysis or ML techniques for working with imbalanced databases.

Given that class imbalance in antibiotic resistance is a constant, exploring and comparing various strategies to improve these results and ensure more accurate detection is imperative. In this context, the evaluation of two variants of ELM to address the class imbalance problem is presented, applied to the Antibiotic Resistance Information and MALDI-TOF Mass Spectra Database A (DRIAMS-A) used by Weis: Unbalanced 1 and 2 [15]. Additionally, the Synthetic Minority Over-sampling Technique (SMOTE) has been applied to the databases, aiming to counteract the effect of the natural class imbalance: (1) Resistance and (2) susceptibility.

2 Materials and Methods

2.1 Database selection and metrics

Bacterial resistance development continues to be a pressing issue, necessitating the identification of rapid and effective strategies for detecting antibiotic resistance in patients suffering from bacterial pathologies. Focusing on this need, the World Health Organization compiled a list in 2017 of the most important bacteria needing better identification of antibiotic resistance to facilitate the development of new drugs and treatments for affected patients [16]. This list includes three globally common bacteria: *Escherichia coli, Staphylococcus aureus*, and *Klebsiella pneumoniae*, which will be used as sample types in this study. For each species, a relevant antibiotic has been tested based on its clinical use: Ceftriaxone for *E. coli* and *K. Pneumoniae*, Oxacillin for *S. Aureus*. The DRIAMS-A database provides MS data and defines the antibiotic resistance or susceptibility of the different samples for each bacterium.

The DRIAMS-A database has the drawback of being imbalanced, where positive cases (antibiotic resistance) will always be a minority compared to the negatives (antibiotic susceptibility). For further details, refer to Fig. 1.

To construct the datasets used in the models, a matrix was formed using the raw mass spectra, which were subjected to a binning process considering a range of 2,000 Da to 10,000 Da with a bin size of 2 Da to obtain a vector of 4,000 features applicable to our model. The bin size was selected according to previous work[4]. Subsequently, the data were normalized between 0 and 1 and finally divided into 80% for training, and 20% for testing the final model.

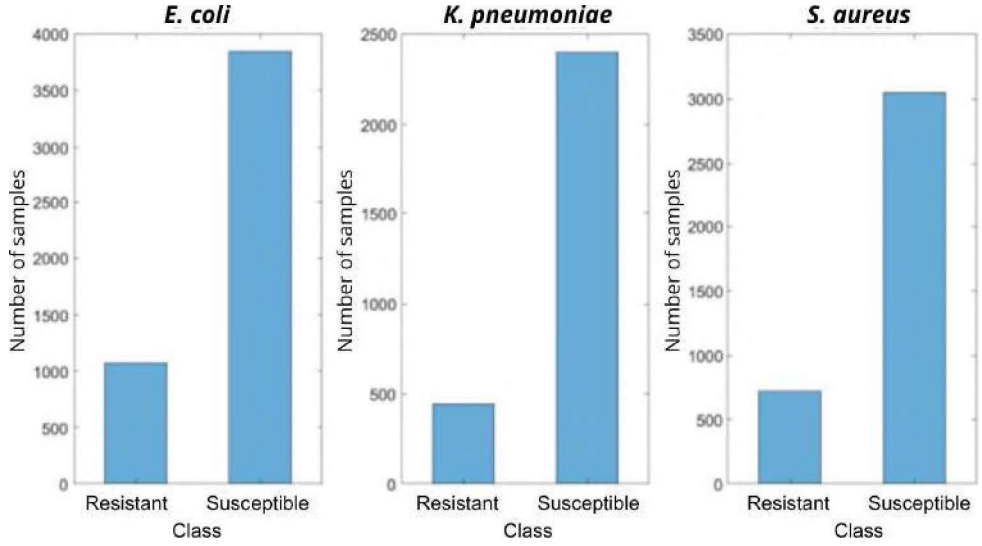

Figure 1: Distribution histogram of *E. coli*, *K. pneumoniae*, *S. aureus* bacteria samples. DRIAMS-A database.

Two metrics have been used to search hyperparameters: the geometric mean (G-Measure), as seen in equation (1), where TP, TN, FP, and FN define true positives, true negatives, false positives, and false negatives, respectively. The second metric corresponds to accuracy, equation (2), considering the main diagonal of the confusion matrices divided by the total samples. To achieve more reliable results, the five-fold cross-validation technique was employed when applying the ELMs to the databases. This involved calculating the average and standard deviation across the five iterations. In addition to accuracy and G-measure metrics, the complexity of the presented methods has been analyzed, applying the Monte Carlo technique with 100 iterations and calculating the average and standard deviation of the iterations.

After the hyperparameter search, the models were evaluated using G-measure, accuracy, sensitivity, and specificity. Sensitivity equation (3), also known as "recall" or true positive rate, measures the proportion of positive cases (antibiotic resistance) correctly identified among the total positive cases. Specificity, equation (4), indicates the proportion of negative cases (antibiotic susceptibility) among the total negative cases.

$$\text{G-Measure} = \sqrt{\frac{TP}{TP+FN} \cdot \frac{TN}{TN+FP}}, \tag{1}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{2}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}, \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN+FP}. \tag{4}$$

2.2 Unbalanced ELM

Conventional ELMs, defined as single-layer feedforward neural networks, appear in their most basic form in equation (5) [17].

$$f(x_j) = \sum_{i=1}^{N_o} \beta_i g(w_i \cdot x_j + b_i), \tag{5}$$

in the equation, x_j represents the j-th input data, and $\beta_i = [\beta_{i1}, \beta_{i2}, ..., \beta_{iu}]^T$ is the output weight matrix connecting the output layer to the hidden layer. The activation function is denoted by "g," while w_i is the weight vector between the input node and the i-th hidden node. N_o indicates the number of hidden neurons. For simplification, the equation can be described as follows:

$$H\beta = T. \tag{6}$$

in equation (6), T is the target data matrix with N_o samples, β is the output weight matrix, and H is the output matrix of the hidden layer, specified respectively in equations (7), (8), and (9):

$$T = [t_1, \cdots, t_{N_o}]^T, \tag{7}$$

$$\beta = [\beta_1, \cdots, \beta_N]^T, \tag{8}$$

$$H = \begin{bmatrix} g(w_1 \cdot x_1 + b_1) & \cdots & g(w_L \cdot x_i + b_L) \\ \vdots & \ddots & \vdots \\ g(w_1 \cdot x_{N_o} + b_1) & \cdots & g(w_L \cdot x_{N_o} + b_L) \end{bmatrix}. \tag{9}$$

The solution of minimum norm the least squares is determined analytically to calculate the output weights β using the "generalized" Moore-Penrose inverse H.[18], as seen in equation (10):

$$\beta = H^{\dagger}T \; ; \; H^{\dagger} = (H^T H)^{-1} H^T. \tag{10}$$

This ELM format tends to overfit, which is a significant drawback. To counteract this problem, a regularization parameter, C, is added to better balance empirical risk and structural risk, as detailed in [19].

The mathematical description of parameter C appears in equation (11) :

$$\min \frac{1}{2}\|\beta\|^2 + \frac{1}{2}C\|D\varepsilon\|^2, \tag{11}$$

where ε represents the training error that regulates the parameter C, denoted as $\varepsilon = [\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_{N_o}]$. Additionally, D is defined as the diagonal matrix $D = diag(v_1, v_2, \cdots, v_{N_o})$, which accounts for both empirical risk and structural risk. Zong proposes their weighted ELM in two sections, binary and multiclass classification, of which only the first will be analyzed, given the classification nature of the DRIAMS-A database. For handling imbalance, Zong proposes the integration of a diagonal matrix **W**, associated with each sample training. Finally, for the weighted ELMs, equations (14) and (15) are defined:

$$\text{When N is small: } \beta = H^T\left(\frac{I}{C} + WHH^T\right)^{-1}WT, \tag{12}$$

$$\text{When N is large: } \beta = H^T\left(\frac{I}{C} + H^TWH\right)^{-1}H^TWT. \tag{13}$$

To calculate the weight of the matrix W, an automatic weight generation scheme is defined, as detailed in equations (16) and (17):

$$\text{Weight scheme } \mathbf{W1}\text{: } W_{\#} = \frac{1}{\#t}. \tag{14}$$

$$\text{Weight scheme } \mathbf{W2}\text{: } W_{\#} = \frac{0.618}{\#t}. \tag{15}$$

Where $\#t$ represents the number of samples belonging to a class, another proposed weighting aims to adjust the balance step to a ratio of $0.618 : 1$. This ratio, chosen because it corresponds to the golden ratio (often considered a standard of perfection in nature), helps balance the minority and majority classes. For this study, the weighting without the golden ratio adjustment will be referred to as $W1$ (unbalanced one), and the weighting incorporating the golden ratio will be labeled $W2$ (unbalanced two).

2.3 SMOTE oversampling technique

Oversampling is a technique that balances imbalanced datasets, where some classes have many more samples than others. This is important in machine learning and data mining applications. Imbalanced data can lead to classification problems, especially for minority classes. In the review on "handling imbalanced data using undersampling and oversampling technique" [20], one of the techniques with the best approximation performance and widely used in the literature is the SMOTE technique [21], which has been successful in a variety of areas and has inspired different approaches to addressing the class imbalance problem. Its impact extends to new learning methods and has become an essential reference when working with imbalanced data. SMOTE generates synthetic data for the minority class by taking samples from the nearest neighbors in feature space and creating new weighted data points. This approach helps balance the class distribution by increasing the number of samples in the minority class, thus improving the classification model performance [22]. In the context of this research, the minority class corresponds to samples that exhibit antibiotic resistance; as shown in Fig. 1, the positive class (antibiotic resistance) is, on average, up to 4 times smaller than the negative class (antibiotic susceptibility), which is why it will be sought to increase the number of positive samples synthetically by 250%, aiming for a much more stable database.

3 Results

3.1 Hyperparameter sub-optimization

Effective implementation of different ELMs requires careful tuning of their hyperparameters. An exhaustive search is conducted to identify suboptimal values. The first hyperparameter, regularization (C), is specified as a vector $C = [2^n]$, with n ranging from -20 to 20 in regular increments. The second hyperparameter, the number of hidden neurons (N_o), is defined as a set as a vector of 100 columns, ranging from 100 hidden neurons up to 80% of the total samples in each respective database (*E. coli*, Hidden Neurons max=3900; *K. pneumoniae*, Hidden Neurons max=2280; *S. aureus*, Hidden Neurons max=3000). The third hyperparameter, the activation function, remains constant across the ELMs analyzed, employing the sigmoid function. This function is widely used in machine learning models to map values between 0 and 1, proving especially valuable in binary classification tasks that require probability estimates, such as predicting outcomes in scenarios with two possible results [23].

To compare different ELMs and their results, estimating the suboptimal values of the hyperparameters is necessary. Finding these values is achieved by creating contour plots in each case; the analysis is achieved by locating a combination of parameters that achieves the best results, where the graphs labeled as "Testing" correspond to testing accuracy and "Measure G" corresponds to the geometric measure (equation 1). The approach to sub-optimizing the parameters of all applied ELMs involves visually seeking a value that balances optimal performance and reduced complexity. It has been considered that the hyperparameter C does not add complexity to the system. Hence, the priority is to find a value that ensures good performance without increasing the model's complexity, so the focus will be on finding the fewest number of neurons with the best possible result. This approach allows for identifying a balance point that optimizes the model's performance without overloading it with unnecessary

complexity.

3.1.1 Bacteria *Escherichia Coli*

Fig. 2 and Fig. 3 display contour plots representing the predictions of the ELMs on the resistance of *Escherichia coli* bacteria to the antibiotic Ceftriaxone. Figure 2 pertains to unbalanced ELM 1, and Figure 3 to unbalanced ELM 2. Figure 2 shows that the results for accuracy and geometric measure exceed 0.7 with hyperparameters $C \approx 2^{10}$ and $N_o \approx 1200$. However, it is important to note that unbalanced ELM 1 exhibits erratic and disordered results when the parameter $C < 2^{-5}$ affects both accuracy and geometric measure metrics. Fig. 3 presents the contour plot for unbalanced ELM 2, showing improved performance in both metrics starting from $C \approx 2^{15}$ and $N_o \approx 2000$, achieving scores of 0.6 in both accuracy and geometric measure. Notably, the geometric measure proves to be more demanding than testing accuracy for both ELM configurations. Specifically, unbalanced ELM 1 reaches approximately 0.8 in accuracy, while unbalanced ELM 2 achieves 0.69.

3.1.2 Bacteria *Klebsiella Pneumoniae*

Similar to earlier observations, Fig. 4 and Fig. 5 illustrate the performance of unbalanced ELM 1 and 2 in predicting the resistance of *K. pneumoniae* bacteria to the antibiotic Ceftriaxone. In Fig. 4, accuracy results around 0.7 are achieved with parameters $C \approx 2^{10}$ and $N_o \approx 500$. However, higher demands are noted for the geometric measure, requiring $N_o \approx 800$ to exceed a value of 0.69. Additionally, the geometric measure exhibits less stability in maintaining its optimal values than the accuracy metric. Fig. 5 shows the performance of the unbalanced ELM 2 algorithm, which generally falls below that of unbalanced ELM 1. The optimal parameter combination for accuracy appears to be $C \approx 2^{15}$ and $N_o \approx 1700$, producing an accuracy of approximately 0.65. Interestingly, the geometric measure reaches a higher value of 0.7 with $C \approx 2^{16}$ and $N_o \approx 2000$. However, this result should be interpreted with caution, as it does not demonstrate a clear linear trend.

3.1.3 Bacteria *Staphylococcus Aureus*

In Fig. 6 and Fig. 7, unbalanced ELM 1 and ELM 2 performance is depicted for *Staphylococcus Aureus*, presenting contour plots highlighting notable differences. Fig. 6 focuses on unbalanced ELM 1, where a significant disparity is observed between the testing metric and the geometric measure compared to previous cases. The testing metric achieves stability with an accuracy of around 0.78, using hyperparameters $C \approx 2^{10}$ and $N_o \approx 1700$. However, the geometric measure displays evident instability at these parameters, indicating that it is more demanding and varies more widely in its values, leading to more noticeable changes in the colors of the plot. The optimal values for the geometric measure are achieved with $C \approx 2^{15}$ and $N_o \approx 2700$, resulting in values fluctuating between 0.7 and 0.8; however, this requires doubling the number of hidden neurons, adding complexity to the algorithm when selecting suboptimal parameters. Fig. 7 examines the performance of unbalanced ELM 2, which again shows lower accuracy in both the testing metric and the geometric measure compared to unbalanced ELM 1. The optimal parameter values for ELM 2 are also $C \approx 2^{15}$ and $N_o \approx 2700$.

3.1.4 SMOTE technique application

Significant performance improvements were noted with unbalanced ELM 1. Given these results, applying the SMOTE oversampling technique to the databases offers an opportunity to reassess the performance of the ELM. Fig. 8 displays the improved results for unbalanced ELM 1 after applying SMOTE to the *E. coli* bacteria database, with noticeable enhancements in the algorithm's metrics and a clearer linearity compared to Fig. 2. The performance begins to optimize with hyperparameters $C \approx 2^{12}$ and $N_o \approx 2000$, achieving up to 0.87 in both accuracy and Geometric Mean. Moreover, Fig. 9 presents the outcomes for unbalanced ELM 1 using SMOTE in the *K. pneumoniae* bacteria database. These results show significant improvements and a more distinct trend than Fig. 4. Optimal performance is observed with hyperparameters $C \approx 2^{16}$ and $N_o \approx 600$, reaching up to 0.9 in accuracy and 0.87 in Geometric Mean.

Similarly, the application of SMOTE to the *S. aureus* database enhances the performance of unbalanced ELM 1, as shown in Fig. 10. The results show notable improvements in the algorithm metrics, showing a clear progression relative to Fig. 6. The best results are achieved with the hyperparameters $C \approx 2^{12}$ and $N_o \approx 2000$, reaching up to 0.87 in precision and geometric mean.

Comparing the results in Table 1 and Table 2, it is evident that unbalanced ELM 1 performs best on the unbalanced database, achieving a maximum accuracy and geometric mean of 0.77 for the *Staphylococcus Aureus* bacteria database. Additionally, unbalanced ELM 1 maintains a lower complexity level than unbalanced ELM 2, which is more complex. In the rest of the paper, complexity associates to the training time. However, as noted in Table 3, using the SMOTE oversampling technique to balance the databases improves performance with unbalanced ELM 1, reaching an accuracy of up to 0.87. This increase in performance comes with a higher complexity in the *E. coli* bacteria database due to a greater number of hidden neurons utilized. Despite this, unbalanced ELM 1 shows better performance and less complexity in the *S. aureus* bacteria database than other models. When applied to a synthesized database, unbalanced ELM 1 signif-

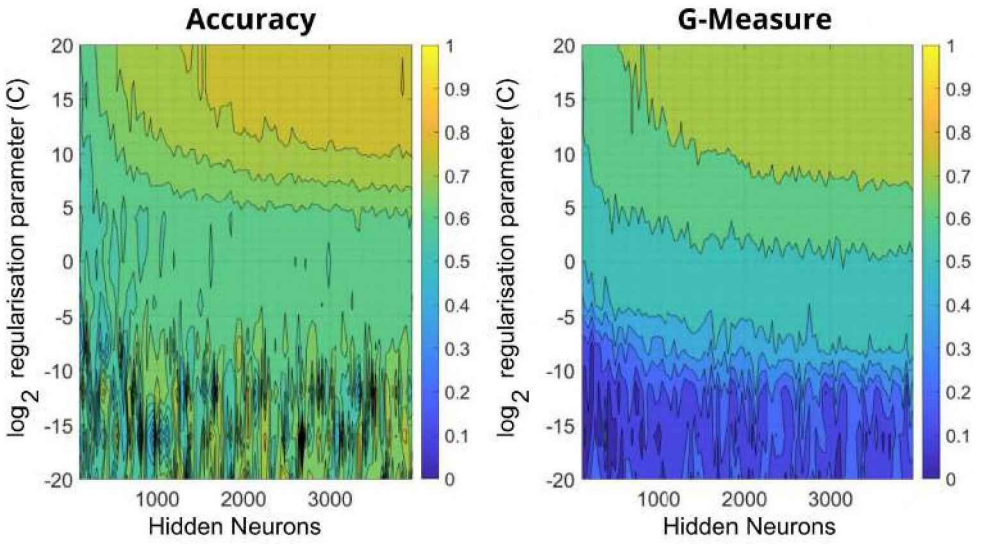

Figure 2: Contour Plot - Unbalanced ELM 1 for the Prediction of Ceftriaxone Resistance in Bacterial *E. coli*.

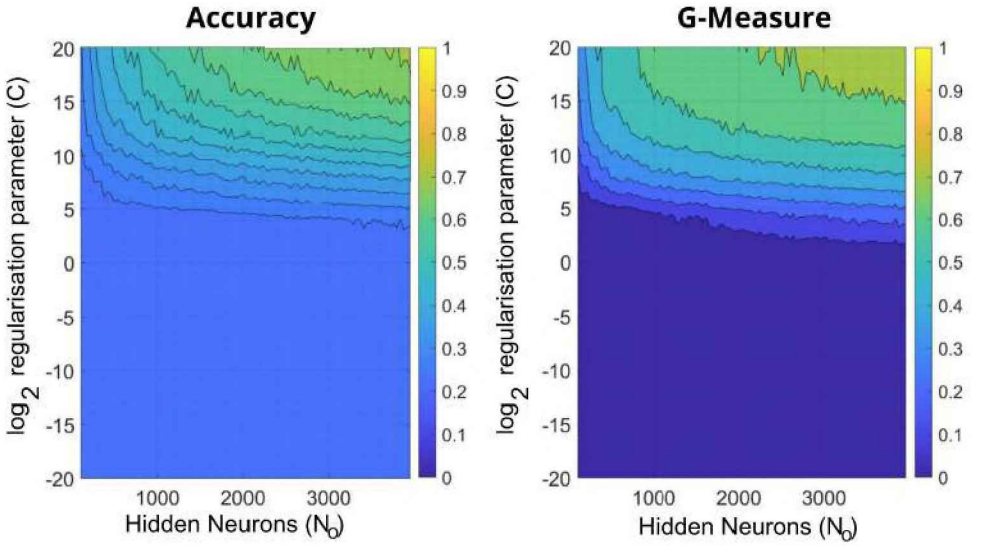

Figure 3: Contour Plot - Unbalanced ELM 2 for the Prediction of Ceftriaxone Resistance in Bacterial *E. coli*.

icantly outperforms its unbalanced version, with all metrics converging at 0.86 for both geometric mean and accuracy. Note that for the application of the different algorithms, the following characteristics of the equipment used should be taken into account: Matlab 2023a software, desktop computer with Windows 10 operating system, Intel(R) Core(TM) i7-2600 CPU @ 3.40GHz processor, 16 GB RAM, and AMD Radeon HD 6700 Series graphics card.

Table 1: Hyperparameters used in the unbalanced ELM model 1 with unbalanced databases.

	E. coli	*K. pneumoniae*	*S. aureus*
Antibiotic	Oxacillin	Ceftriaxone	Ceftriaxone
Hidden Neurons	2500	700	2500
Parameter C	2^{20}	2^{15}	2^{16}
Accuracy	0.78 ± 0.086	0.70 ± 0.0237	0.77 ± 0.010
G-Measure	0.75 ± 0.017	0.68 ± 0.024	0.77 ± 0.170
Training Time (s)	14.71 ± 0.69	1.4 ± 0.17	9.66 ± 0.31

3.2 Model Evaluation

After identifying the studied algorithm's suboptimal hyperparameters, we conducted tests using the values specified in Tables 1, 2, and 3. We present the results of these evaluations in Table 4. The results have been reported using the following metrics: accuracy, geometric mean, sensitivity, and specificity.

Table 4 presents the evaluation results for three variants of the ELM model: Imbalanced ELM 1, Imbalanced ELM 2, and Imbalanced ELM 1 with the application of the SMOTE oversampling method. In particular, the Imbalanced ELM 1 variant performs better in the *S. aureus* dataset and is superior in all four performance measures: accuracy, geometric mean, sen-

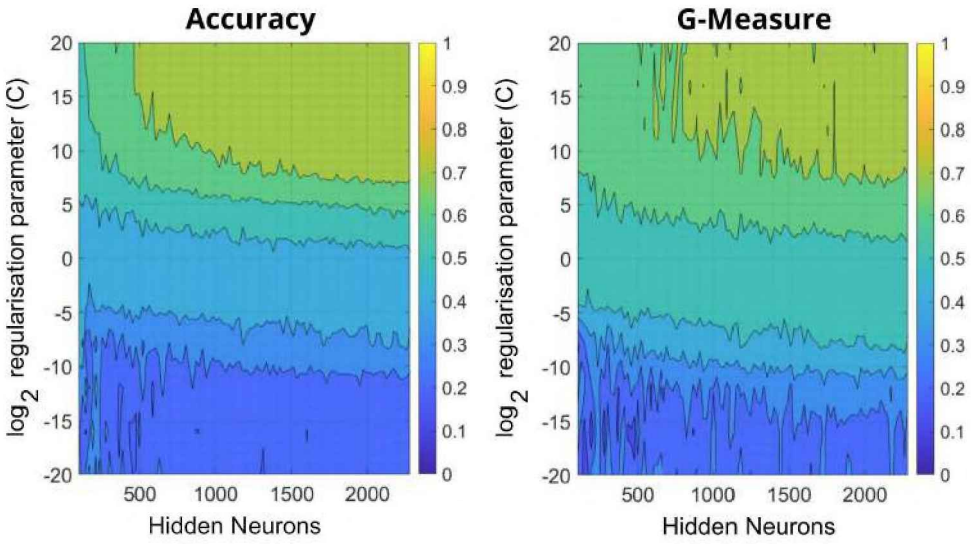

Figure 4: Contour Plot - Unbalanced ELM 1 for the Prediction of Ceftriaxone Resistance in Bacteria *K. pneumoniae*.

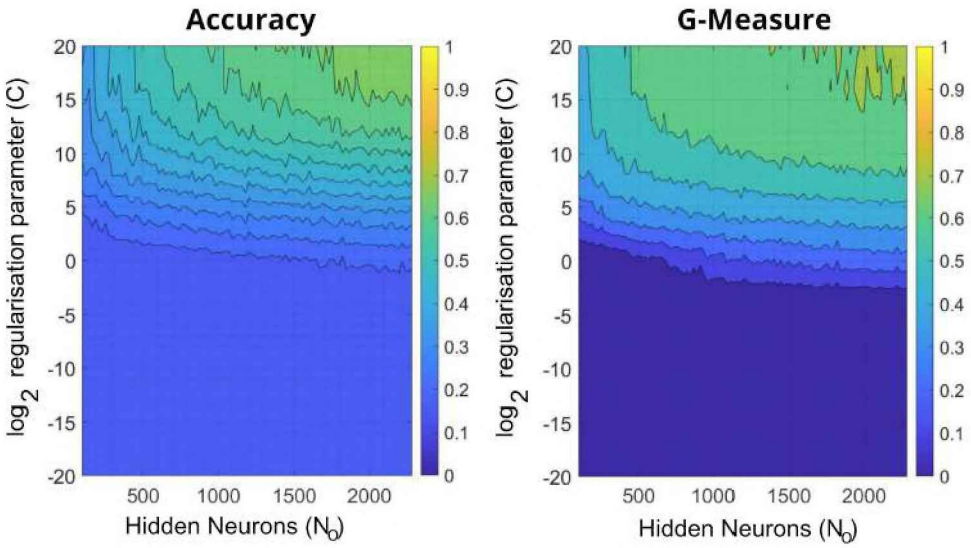

Figure 5: Contour Plot - Unbalanced ELM 2 for the Prediction of Ceftriaxone Resistance in Bacteria *K. pneumoniae*.

Table 2: Hyperparameters used in unbalanced ELM 2 with unbalanced database.

	E. coli	*K. pneumoniae*	*S. aureus*
Antibiotic	Oxacillin	Ceftriaxone	Ceftriaxone
Hidden Neurons	3600	2200	2800
Parameter C	2^{20}	2^{20}	2^{20}
Accuracy	0.70 ± 0.012	0.63 ± 0.0128	0.73 ± 0.0092
G-Measure	0.73 ± 0.018	0.69 ± 0.0085	0.69 ± 0.0091
Training Time (s)	13.95 ± 0.61	5.65 ± 0.19	11.79 ± 0.48

Table 3: Hyperparameters used in unbalanced ELM 1 when applying SMOTE.

	E. coli	*K. pneumoniae*	*S. aureus*
Antibiotic	Oxacillin	Ceftriaxone	Ceftriaxone
Hidden Neurons	3000	1000	2000
Parameter C	2^{16}	2^{16}	2^{16}
Accuracy	0.86 ± 0.0068	0.87 ± 0.0071	0.86 ± 0.0091
G-Measure	0.86 ± 0.0072	0.85 ± 0.0062	0.86 ± 0.01
Training Time (s)	35.13 ± 1.16	3.27 ± 0.13	5.64 ± 0.15

sitivity, and specificity.

On the other hand, Imbalanced ELM 2 performs worse in most metrics; however, its sensitivity is better, implying that this method can better detect the positive class, i.e., resistant bacteria. This result is particularly striking given that positive samples constitute the minority class in the datasets.

Applying the SMOTE technique to the datasets and reevaluating Imbalanced ELM 1 significantly improved accuracy and geometric mean, confirming the trend in Table 3. Although sensitivity did not improve, specificity increased, indicating a better balance in detecting both classes of both resistant and susceptible bacteria. The above reinforces that using the SMOTE

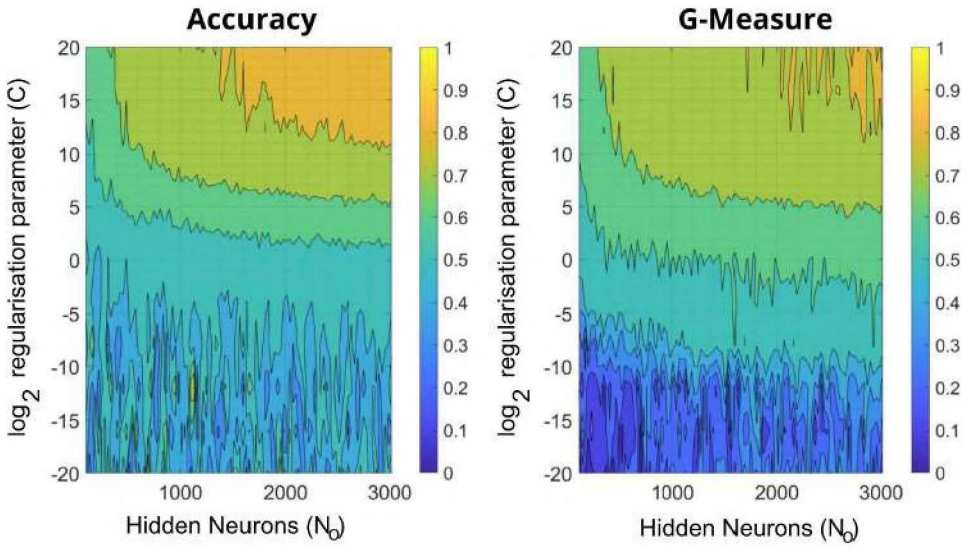

Figure 6: Contour Plot - Unbalanced ELM 1 for the Prediction of Oxacillin Resistance in *S. aureus* Bacteria.

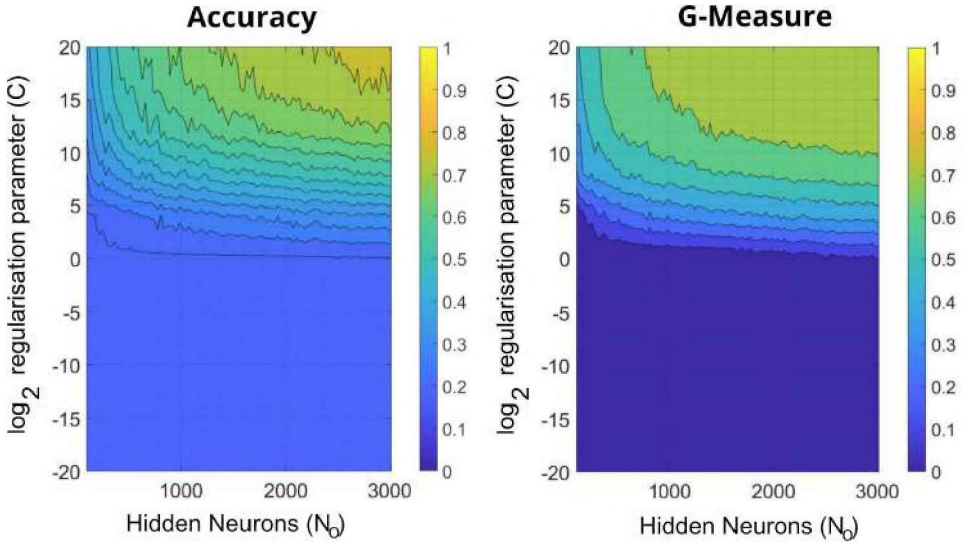

Figure 7: Contour Plot - Unbalanced ELM 2 for the Prediction of Oxacillin Resistance in *S. aureus* Bacteria.

Table 4: Model evaluation results: Unbalanced ELM 1, Unbalanced ELM 2, Unbalanced ELM 1 when applying SMOTE technique.

		E. coli	*K. pneumoniae*	*S. aureus*
ELM unbalanced 1	Accuracy	0,69	0,69	0,82
	G-Measure	0,71	0,71	0,82
	Sensitivity	0,73	0,73	0,79
	Specificity	0,68	0,68	0,83
ELM unbalanced 2	Accuracy	0,64	0,64	0,71
	G-Measure	0,70	0,70	0,74
	Sensitivity	0,83	0,83	0,81
	Specificity	0,59	0,59	0,68
ELM unbalanced 1	Accuracy	0,86	0,86	0,85
When applying SMOTE	G-Measure	0,86	0,86	0,85
	Sensitivity	0,73	0,76	0,77
	Specificity	1,00	0,96	0,93

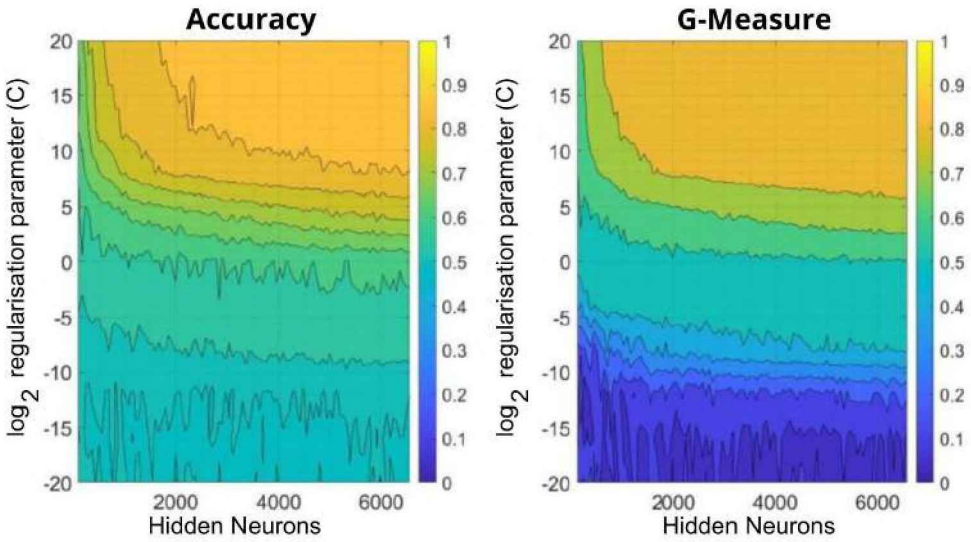

Figure 8: Contour Plot - Unbalanced ELM 1 for the Prediction of Ceftriaxone Resistance in Bacterial *E. coli* with SMOTE technique applied.

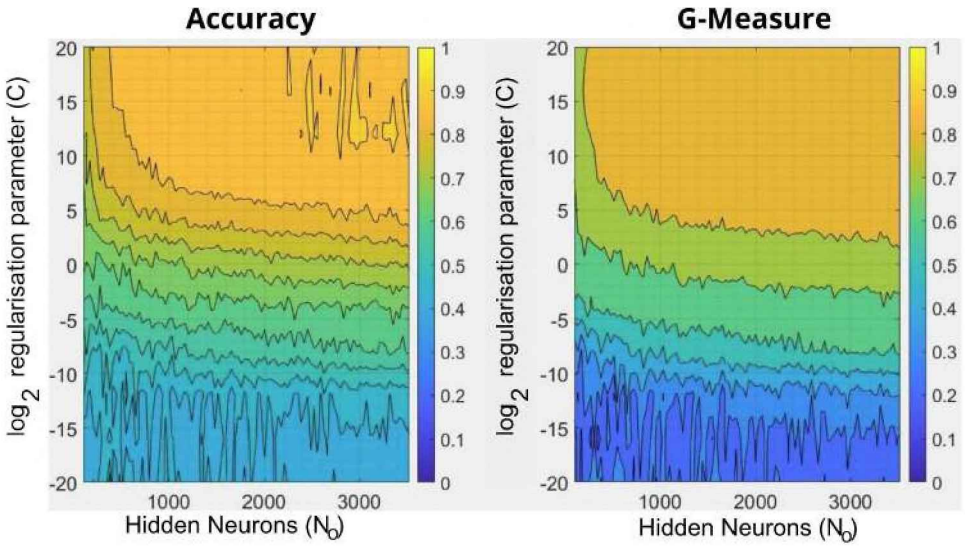

Figure 9: Contour Plot - Unbalanced ELM 1 for the Prediction of Ceftriaxone Resistance in Bacterial *K. pneumoniae* with applied SMOTE technique

oversampling method improves results compared to imbalanced datasets.

4 Conclusions

MALDI-TOF MS combined with ELM and the SMOTE has demonstrated significant potential for predicting bacterial antibiotic resistance in *E. coli*, *K. pneumoniae* and *S. aureus*. Despite initial challenges with data imbalance and model accuracy, implementing weighted ELMs and applying SMOTE significantly enhanced the predictive performance, achieving up to 85% accuracy and a similar geometric mean. As future work, we propose the implementation of deep learning in order to explore a multilabel approach.

Competing interests

The authors have declared that no competing interests exist.

Authors' contribution

V.M.M.: Data curation, Software, Formal Analysis, Writing – original draftT; D.Z.B.: Conceptualization, Methodology, Software, Writing – original draft, Writing – review editing, Supervision; X.A.L.C.: Conceptualization, Methodology, Software, Writing – original draft, Writing – review editing, Supervision; F.T.: Conceptualization, Methodology, Writing-Original draft, Writing – review editing; J.M.M.T.:Data curation, Validation, Formal Analysis, Writing – original draft; R.A.G.: Conceptualization, Methodology, Writing-Original draft, Writing – review editing.All authors reviewed the results and approved the final Version of the manuscript.

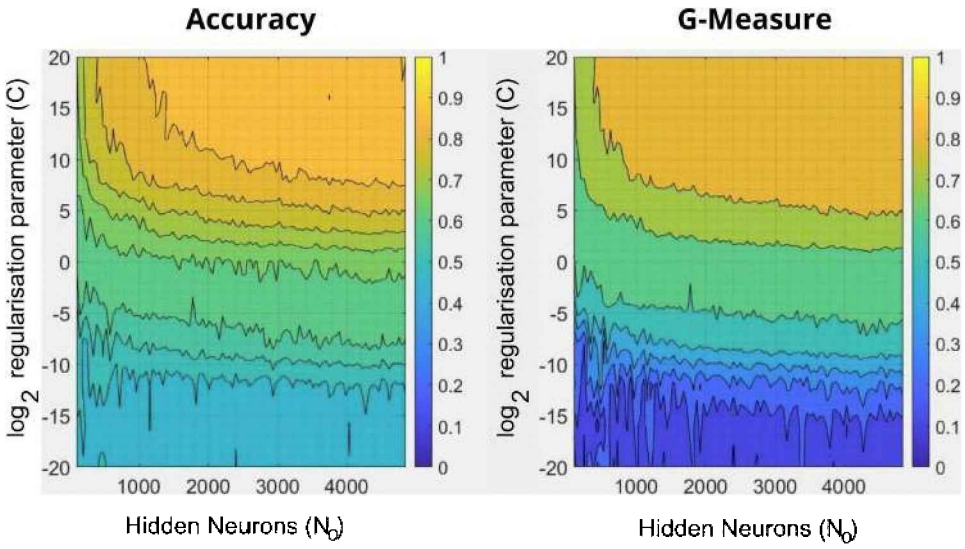

Figure 10: Contour Plot - Unbalanced ELM 1 for the Prediction of Ceftriaxone Resistance in *S. aureus* Bacteria with SMOTE technique applied.

Acknowledgements

X.A.L.-C. thanks to financial support from Research Project ANID FONDECYT Iniciación en Investigación N. 11220897. The author R.A-G., a Ph.D. student in Engineering at UCM, acknowledges the funding from ANID-Subdirección de Capital Humano/Doctorado Nacional/2024-21241043.

References

[1] J. O'Neill, "Tackling drug-resistant infections globally: final report and recommendations." 2016.

[2] G. Mancuso, A. Midiri, E. Gerace, and C. Biondo, "Bacterial antibiotic resistance: The most critical pathogens," *Pathogens*, vol. 10, no. 10, p. 1310, 2021.

[3] M. F. Varela, J. Stephen, M. Lekshmi, M. Ojha, N. Wenzel, L. M. Sanford, A. J. Hernandez, A. Parvathi, and S. H. Kumar, "Bacterial resistance to antimicrobial agents," *Antibiotics*, vol. 10, no. 5, p. 593, 2021.

[4] X. A. López-Cortés, J. M. Manríquez-Troncoso, R. Hernández-García, and D. Peralta, "Msdeepamr: antimicrobial resistance prediction based on deep neural networks and transfer learning," *Frontiers in Microbiology*, vol. 15, p. 1361795, 2024.

[5] 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.

[6] X. A. López-Cortés, F. Matamala, B. Venegas, and C. Rivera, "Machine-learning applications in oral cancer: A systematic review," *Applied Sciences 2022, Vol. 12, Page 5715*, vol. 12, p. 5715, 6 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/11/5715/htmhttps://www.mdpi.com/2076-3417/12/11/5715

[7] A. Tapia-Castillo, C. A. Carvajal, X. López-Cortés, A. Vecchiola, and C. E. Fardella, "Novel metabolomic profile of subjects with non-classic apparent mineralocorticoid excess," *Scientific Reports 2021 11:1*, vol. 11, pp. 1–12, 8 2021. [Online]. Available: https://www.nature.com/articles/s41598-021-96628-6

[8] V. R. Olate-Olave, L. Guzmán, X. A. López-Cortés, R. Cornejo, F. M. Nachtigall, M. Doorn, L. S. Santos, and A. Bejarano, "Comparison of chilean honeys through maldi-tof-ms profiling and evaluation of their antioxidant and antibacterial potential," *Annals of Agricultural Sciences*, vol. 66, pp. 152–161, 12 2021.

[9] C. González, C. A. Astudillo, X. A. López-Cortés, and S. Maldonado, "Semi-supervised learning for maldi–tof mass spectrometry data classification: An application in the salmon industry," *Neural Computing and Applications*, vol. 35, no. 13, pp. 9381–9391, 2023.

[10] X. A. López-Cortés, F. M. Nachtigall, V. R. Olate, M. Araya, S. Oyanedel, V. Diaz, E. Jakob, M. Ríos-Momberg, and L. S. Santos, "Fast detection of pathogens in salmon farming industry," *Aquaculture*, vol. 470, pp. 17–24, 3 2017.

[11] X. A. Lopez-Cortes, F. Avila-Salas, C. Orellana, and L. S. Santos, "Strategy based on data mining and maldi-mass spectrometry for control disease of srs in salmo salar," *IEEE ICA-ACCA 2018 - IEEE International Conference on Automation/23rd Congress of the Chilean Association of Automatic Control: Towards an Industry 4.0 - Proceedings*, 1 2019.

[12] Y. Charretier and J. Schrenzel, "Mass spectrometry methods for predicting antibiotic resistance," *PROTEOMICS–Clinical Applications*, vol. 10, no. 9-10, pp. 964–981, 2016.

[13] C. Weis, A. Cuénod, B. Rieck, O. Dubuis, S. Graf, C. Lang, M. Oberle, M. Brackmann, K. K. Søgaard, M. Osthoff *et al.*, "Direct antimicrobial resistance prediction from clinical maldi-tof mass spectra using machine learning," *Nature Medicine*, vol. 28, no. 1, pp. 164–174, 2022.

[14] H. Cruz-Reyes, A. Reyes-Nava, E. R. Lara, and R. Alejo, "Estudio del desbalance de clases en bases de datos de microarrays de expresión genética mediante técnicas de deep learning." *Res. Comput. Sci.*, vol. 147, no. 5, pp. 197–207, 2018.

[15] W. Zong, G.-B. Huang, and Y. Chen, "Weighted extreme learning machine for imbalance learning," *Neurocomputing*, vol. 101, pp. 229–242, 2013.

[16] S. R. Shrivastava, P. S. Shrivastava, and J. Ramasamy, "World health organization releases global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics," *Journal of Medical Society*, vol. 32, no. 1, pp. 76–77, 2018.

[17] D. Zabala-Blanco, M. Mora, R. J. Barrientos, R. Hernández-García, and J. Naranjo-Torres, "Fingerprint classification through standard and weighted extreme learning machines," *applied sciences*, vol. 10, no. 12, p. 4125, 2020.

[18] R. Penrose, "A generalized inverse for matrices," in *Mathematical proceedings of the Cambridge philosophical society*, vol. 51. Cambridge University Press, 1955, pp. 406–413.

[19] W. Deng, Q. Zheng, and L. Chen, "Regularized extreme learning machine," in *2009 IEEE symposium on computational intelligence and data mining*. IEEE, 2009, pp. 389–395.

[20] M. S. Shelke, P. R. Deshmukh, and V. K. Shandilya, "A review on imbalanced data handling using undersampling and oversampling technique," *Int. J. Recent Trends Eng. Res*, vol. 3, no. 4, pp. 444–449, 2017.

[21] J. García Abad *et al.*, "Comparativa de técnicas de balanceo de datos. aplicación a un caso real para la predicción de fuga de clientes," 2021.

[22] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of artificial intelligence research*, vol. 61, pp. 863–905, 2018.

[23] A. D. Rasamoelina, F. Adjailia, and P. Sinčák, "A review of activation function for artificial neural network," in *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE, 2020, pp. 281–286.

Citation: V. Macaya Mejías, D. Zabala-Blanco, X. A. López-Cortés, F. Tirado, J. M. Manríquez-Troncoso and R. Ahumada-García. *Predicting Bacterial Antibiotic Resistance using MALDI-TOF Mass Spectrome-try Databases with ELM Applications*. Journal of Computer Science & Technology, vol. 24, no. 2, pp. 88–98, 2024. **DOI:** 10.24215/16666038.24.e08.
Received: April 15, 2024 **Accepted:** August 15, 2024.