Antibiotic Resistance Mechanisms in Staphylococcus aureus based on MALDI-TOF

Guillermo Grande Santi, Jesús Martín Trilla, Lucía Ferrer Duaso

Exploratory data analysis

As the data was previously preprocessed and binned, any missing values or issues in data collection were already handled, so no additional methods were applied to address them.

The data is continuous in a stepwise manner, meaning there are no categorical features. Since the step sizes (bins) are quite small, the data was treated as continuous values.

The imbalance of classes corresponds to the proportions of (00:0.641, 10:0.254, 01:0.053, 11:0.050)[1]. The knowledge that they are imbalanced implies that classic metrics like accuracy will not be used, as well as the need to use special objective functions to train the models, and/or a data augmentation (either positive or negative).

To mitigate class imbalance, various resampling techniques from the imbalanced-learn library were employed or even specific models. Additionally, when training Neural Networks, methods like Balanced Batch Generator were utilized. These approaches helped rebalance the dataset and enhance model performance.

Preprocessing

Looking at the magnitude of the values, with up to 5 decimal places, the proposed scaling methods include **standardization** (transforming to a distribution with mean 0 and standard deviation 1), **normalization** (using a min-max scaler to map values into the range $[0, 1]$), and **logarithmic scaling**.

Due to the significant imbalance in the ratio of the number of features to the number of samples ($\approx 6 : 1$), feature selection was necessary. The analyzed dimensionality reduction methods included **PCA**, either retaining components that explain 95% of the variance or selecting the **elbow point** at the 324^{th} principal component where the variance explained is 65%; **LASSO**-based feature selection; **autoencoder**-based embedded space representation with a latent dimension of N; and Minimum Redundancy Maximum Relevance (mRMR) for approximated feature selection[2]. Additionally, **oversampling** and **undersampling** techniques were explored, including traditional undersampling methods like **random undersampling**, and oversampling methods such as **SMOTE**, which, however, did not yield satisfactory results due to the high dimensionality and interpolation issues.

All preprocessing steps were applied only to the training subset, while the test data was transformed without refitting the scalers to prevent data leakage and ensure fidelity in metric evaluations.

The t-SNE visualization 1 of the dataset highlights its highly non-linear structure. This observation suggests that the failure of previous preprocessing techniques could be attributed to their reliance on linear operations (e.g., PCA). Since the data distribution is inherently non-linear, traditional scaling and dimensionality reduction methods may struggle to effectively represent the underlying structure, leading to suboptimal model performance.

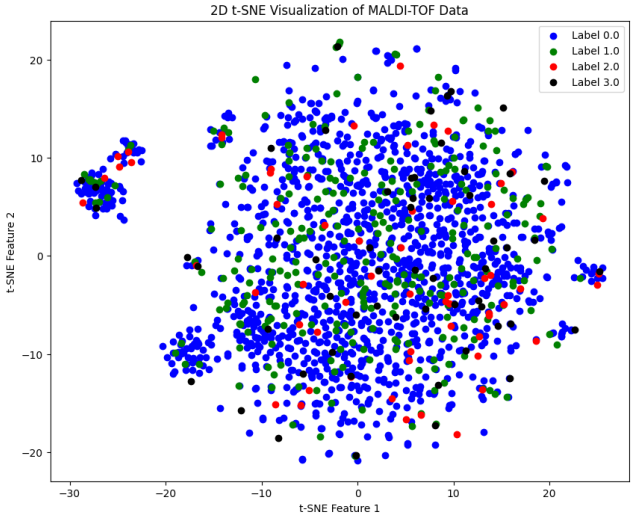

Figure 1: 2D Multilabel Visualization t-SNE

Models

Random Forest Variations

For the **Erythromycin** classifier, the highest balanced accuracy of **0.63** was achieved using a Balanced Random Forest

[1]Notation $ij : perct$, with $i : bool|$ resistant to Erythromycin; $j : bool|$resistant to Ciprofloxacin

[2]Due to heavy computation, it was not applied directly on all features, but on a lax lasso, and no significant improvement was seen

model with StandardScaler and Lasso regularization. Other preprocessing strategies, such as StandardScaler with PCA, Random Under Sampling, and SMOTE, scored 0.59, while MinMaxScaler combinations showed slightly lower performance around 0.56.

For the **Ciprofloxacin** classifier, the best result was also **0.70** with a Balanced Random Forest model using StandardScaler and Lasso. Other combinations, including MinMaxScaler with PCA, Random Under Sampling, and SMOTE, yielded scores between 0.52 and 0.61, with the lowest performance occurring when using SMOTETomek with MinMaxScaler and PCA.

In general, the best method for balancing was the use of the **Balanced Random Forest** approach. **StandardScaler** performed better than **MinMaxScaler**, and regularization aided in the processing time. However, it did not result in a significant improvement in performance, likely due to the linear nature of **Lasso** and **PCA**.

Dense Neural Network

Although **autoencoders** were not successful in dimensionality reduction due to the small amount of data and the difficulty in capturing a useful latent space, other neural networks were tested to predict bacterial resistance.

These models employed **two dense layers** with **dropout** and **regularization** techniques to mitigate overfitting caused by data scarcity. Results were comparable both with and without dimensionality reduction methods, as well as with different oversampling and undersampling techniques and **BCE** as the selected objective function. The highest balanced accuracy achieved was **0.61** for Erythromycin and **0.66** for Ciprofloxacin using cross-validation.

CNN Classifier with probabilities calibrated with One Class Detection

The classifier model is trained with the manual augmented data[3]. The actual architecture selected is a convolutional architecture with **two conv1d layers**, **batch normalization**, and **relu** activations. The selected objective function was **Focal BCE**, and then probabilities for classification were calibrated with one class detection. The ensemble proposed was first a CNN, then a **one-class model** trained where the anomalies (positive class) are the resistant cells. The use of this model on top of the classifier while lowering the probabilities of the normal classes, it did not improve the recall, plus the convolutions did not seem to provide any improvement from previous tabular methods. The balanced accuracy reported is $0.601 \pm .05$ on resistance to erythromycin and $0.58 \pm .05$ on resistance to Ciprofloxacin.

Method Performance Analysis Looking at the methods not explained and explained in this segment that were mainly with accuracy around 0.58, we assume that no linear representation can explain the data distribution, thus the best performing methods are those that are not linear and with dimensionality reduction, either recursive selection methods or deep learning embed space. Final selected to model to interpret Model C.

[3]The combination of 3 different spectrums techniques: intensity modification, shift, and peak dropout

Explainability

The Explainability selected was SHAPley Values over the Balanced Random Forest and Deep SHAPley Values over the Neural Networks Classifiers. As the results provided in the top contributing mostly to the same proteins, we will only be showing one image, the range (proteins Da) will be extracted ($3 \times feature\_i$+2000).

For Erythromycin, **proteins within 4508-4513** Da **and proteins within 5438-5446** Da **are highly important**, where the bigger the intensity Da the more it contributes to increment the likelihood to be resistant, seen in figure 2. We observe that the third range has an **inverse relationship**, where lower values correspond to higher probabilities.

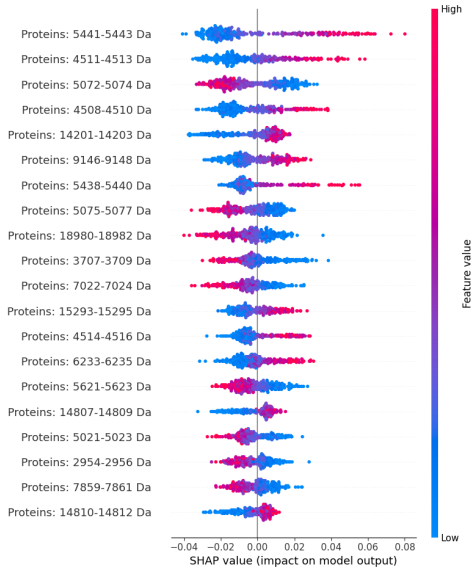

Figure 2: SHAP Explainability for Erythromycin

For Ciprofloxacin, proteins in the range of **6554-6556** Da contribute positively with low values, while the likelihood decreases. Meanwhile, the remaining bins have a positive effect. The fourth bin has the **highest contribution** overall, providing significant explainability if it appears within those values. Additionally, the values in the **4997-5010** Da range also seem to be important. This information is extracted from figure 3.

Conclusions

Our study explored various machine learning and deep learning techniques to classify bacterial resistance, emphasizing the challenges posed by class imbalance, high dimensionality, and the non-linear nature of the dataset.

Among the evaluated models, Balanced Random Forest and dense neural networks demonstrated the best performance, with accuracies up to 0.70 for Ciprofloxacin resistance prediction. Feature selection, data augmentation, and

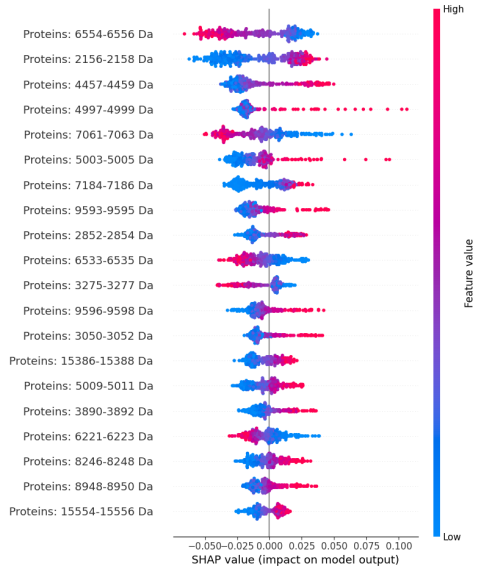

Figure 3: SHAP Explainability for Ciprofloxacin

different scaling methods were applied to improve performance, though their impact was often limited due to the complexity of the data.

The explainability analysis using SHAP values provided us insights into the most influential features, reinforcing the importance of specific protein ranges in determining bacterial resistance.