Temporal Feature Selection for Characterizing Antimicrobial Multidrug Resistance in the Intensive Care Unit

Óscar Escudero-Arnanz[1] and **Inmaculada Mora-Jiménez**[2] and **Sergio Martínez-Agüero**[3] and **Joaquín Álvarez-Rodríguez**[4] and **Cristina Soguero-Ruiz**[5]

Abstract. The emergence and increase of antimicrobial multidrug resistance (AMR) is a demographic and economic problem for current health systems. AMR is particularly problematic in clinical units such as the intensive care unit (ICU), where the risk of infection is high, principally due to the extensive use of antimicrobials and invasive devices. In this work, we propose the use of different temporal feature selection and classification approaches to ascertain the most informative features and extract knowledge for characterizing AMR in the ICU. For this purpose, a set of demographic and temporal features such as antibiotics taken daily by the patient and the use of mechanical ventilation are considered. According to the results obtained in this work, it could be concluded that temporal features such as mecanic ventilation provide powerful insights to predict AMR in ICU.

1 INTRODUCTION

The discovery of antibiotics and their subsequent use in the clinical practice represented a great scientific advance, improving the treatment of infectious diseases and thus saving millions of lives [10]. However, the excessive and incorrect use of antibiotics is contributing a downturn in their effectiveness against bacterial infections, caused by mutations and the acquisition of genetic information from other germs [18]. This fact makes infection control difficult and increases the morbidity and mortality of previously treatable infectious diseases such as malaria or acute respiratory diseases [1].

The impact of antimicrobial multidrug resistance (AMR) can cause an economic burden in hospitals and in the healthcare systems, whose real outcomes still remain unknown. Following the report by the World Health Organisation (WHO), it is estimated an increase in deaths by 2050 caused by antimicrobial resistance, mainly affecting countries such as Africa and Asia [1]. This is a growing problem which needs to be alleviated to avoid the consequences that this could cause. In addition to the demographic effects, the increase in antimicrobial multidrug resistance, this is the resistance of a single bacterium to more than one antibiotic, has a major economic impact, resulting in loss to the world economy of approximately 7% of the Gross Domestic Product by 2050 [13]. From an economic viewpoint, patients infected with antimicrobial resistant bacteria present a higher cost for the healthcare system in comparison to patients who are susceptible to microbial infection [7]. This is caused by the increasing difficulty in treating resistant organisms, making it necessary the breakthrough of new strategies to combat antibiotic resistance. Previous studies have proposed initial analysis based on machine learning models to determine the result (susceptible/resistance) of the antibiogram (a test to measure the in vitro activity of an antibiotic against a given bacterium, which is previously isolated in the culture [12]) or to predict the probability of acquiring a hospital-acquired infection (nosocomial infection), specifically in the ICU [15].

Focusing on a hospital environment, antimicrobial resistance can be acquired by any hospitalised patient, increasing the probability of acquisition for patients admitted to the Intensive Care Unit (ICU). The main reasons are the use of invasive devices, the intensity of treatment and its duration, the high risk of transmission and exposure to antibiotics. The ICU can be considered as the epicenter of development of antimicrobial resistance due to the high rate of nosocomial infections (20-30% of all ICU admissions) [4]. However, the period just before the patient is admitted to the ICU is beginning to take great importance, caused by the increase in the number of patients arriving in the ICU infected by multi-resistant microorganisms [19]. A culture is usually performed to assess bacteria susceptibility/resistance to series of antibiotics. Firstly, an organic sample from the patient (blood or urine samples, among others) is obtained which allows the study of the microorganisms present in their system. Then, the antibiogram is carried out. The result of the antibiogram represents the pair antibiotic/sensibility. Therefore, based on this results, we consider that patients did not acquired the multiresistant bacteria in the ICU if the culture's result is positive within the first 48 hours of the patient's admission, otherwise, the AMR occur during the ICU stay.

The excessive use of antimicrobials during the stay of patients in the ICU (some studies corroborate that more than 60% of patients

[1] Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, Madrid 28943, Spain, o.escudero.2016@alumnos.urjc.es
[2] Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, Madrid 28943, Spain, inmaculada.mora@urjc.es
[3] Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, Madrid 28943, Spain, sergio.martinez@urjc.es
[4] Intensive Care Department, University Hospital of Fuenlabrada, Madrid 28942, Spain, joaquin.alvarez@salud.madrid.org
[5] Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, Madrid 28943, Spain, cristina.soguero@urjc.es

take antibiotics during their ICU stay [4]) along with other factors discussed above, facilitate the emergence of AMR, making this problem the target to be treated. We will study the daily use of antibiotics and mechanical ventilation (MV) in the ICU at University Hospital of Fuenlabrada, Madrid, Spain. The final aim consists in determining the risk factors that best characterize the evolution of critical patients as well as the relevance to identify patients with AMR. To this end, we apply hypothesis tests, linear and non-linear learning algorithms.

The rest of the paper is organized as follows. Section 2 introduces the methods used for the temporal patient characterization. In Section 3, a brief description of the data set is presented, while in Section 4 the experimental work and prediction results are shown. Finally, discussion and conclusions are presented in Section 5.

2 METHODS

Notation

In this paper, each sample is a patient represented by a set of D features, being each feature composed by a time series of T consecutive time slots. Therefore, the data associated to the i-th patient can be arranged in a feature matrix $\mathbf{X}_i = [\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^T] \in \mathbb{R}^{D \times T}$. Where the column vector $\mathbf{x}_i^t$ contains the D features of the i-th patient in the time slot t. Thus, $\mathbf{x}_i^t$ can be represented as the column vector $\mathbf{x}_i^t = [x_{i,1}^t, x_{i,2}^t, \cdots, x_{i,D}^t]^T$, where $[.]^T$ denotes the transpose operator and $x_{i,d}^t$ shows the value of the d-th feature associated to the i-th patient in the t-th time slot. Since we are tackling with a binary classification task, we have considered the label '1' to identify patients with AMR, and the label '0' to identify patients with non-AMR. Therefore, the label (desired output) for the i-th patient is defined by y_i, whereas the output provided by the model is represented as $\hat{y}_i$.

2.1 Feature Selection

There are different methods for feature selection in the literature. The goal is to eliminate features that may be noisy, irrelevant or redundant when building a data-driven model [17]. Also, selecting the most important features can increase the knowledge and the model interpretability. In this work, we want to select features based on hypothesis tests. For each feature, our null hypothesis is that there is no difference between the two populations (AMR patients and non-AMR patients). If there is no evidence to rule out the null hypothesis, then the tested feature is not selected. Since we are dealing with binary and numerical features, we evaluate a test of proportions for the first kind of features, and a two-sample Kolmogorov-Smirnov test for the latter.

Two-proportion z-test. This hypothesis test evaluates whether the presence on a single feature differs in two populations [16]. The null hypothesis states that there is no evidence of difference in the proportion between both populations, whereas the opposite applies for the alternative hypothesis.

Two-sample Kolmogorov-Smirnov test. It is a hypothesis test based on the empirical distribution function and used to estimate whether values of the same feature in two populations are from the same continuous distribution [2]. An advantage of this test over parametric test is the independence of the statistic from the expected frequency distribution, depending only on the sample size.

2.2 Imbalanced sampling

In healthcare-related data sets, it is very common to deal with imbalanced data [8], i. e., one class predominates over the other. This imbalance is a challenge for designing data-driven models, since conventional approaches will mostly learn from the majority class and lead to biased models, reducing the performance for the minority class. Data-driven approaches tend to learn better the mapping of patients belonging to the majority class (far more numerous) than that of the minority class. To tackle this challenge, several strategies could be followed [8]. In this work, we followed a random undersampling strategy [20] with no replacement for the majority class. The final sample size is such that the class frequency is similar. Thus, the number of patients of the majority class is matched before training the model according to the number of patients of the minority class. The undersampling process and subsequent model training is repeated several times not to be conditioned to a particular subsampling, providing statistics on the performance. We benchmark the results obtained with random undersmpling with a synthetic minority oversampling technique (SMOTE), which consists of oversampling the examples in the minority class [6].

2.3 Classification Approaches

Classification approaches encompasses statistical techniques to build models based on the underlying relationships among data. The set of N available samples is split into two independent subsets, named training set and test set. The former is used to create the classifier following a learning process, whereas the latter is used to evaluate the performance of the built model. Normally, the 70% of samples are randomly assigned to the training set and the rest to the test [5].

2.3.1 Logistic Regression

The model provided by Logistic Regression (LR) is a linear combination of the different features. Despite its name, it is a classification approach since the result of the linear combination is the input to a logistic function. To carry out the linear combination of the features, a set of coefficients $w_{i\,i=1}^{d}$ should be found by optimizing a binary cross-entropy cost function. In this work, we considered a regularized term in the cost function, in particular the Ridge regularization [9] for preventing the model from overfitting. To find an appropriate value for the hyperparameter weighting the penalization term in the cost function, named penalty coefficient $C > 0$, we followed a 5 fold cross-validation approach on the training set.

2.3.2 Decision Trees

Decision trees (DT) are non-parametric classifiers which can be graphically represented in a tree shape as a hierarchical structure starting from a root node [14]. For building the tree, a recursive splitting process is carried out dividing the decision space into subspaces based on a criteria related to entropy or Gini index. In this work, we have chosen the Gini criterion to make the splitting process [11]. When a node is created, a region in the feature space is splitted in two parts. A label is assigned to each partition according to the majority class among the training samples in that particular partition. One advantage of DT is the model interpretability, that partly relies on the fact that the most discriminative features are closest to the root node, what implicitly could be considered as a feature selection process.

In this work we considered DT built following the classification and regression tree algorithm named CART [3], since it has been

extensively used in the literature when dealing with heterogeneous features (numerical and categorical).

3 DATASET DESCRIPTION AND TEMPORAL FEATURES

In this work, an anonymized dataset provided by the University Hospital of Fuenlabrada (UHF) in Madrid (Spain) has been analysed. This dataset contains demographic and clinical features of 2889 patients admitted in the ICU of the UHF during a period of 13 years, from 2004 to 2016. The goal is leverage these data to characterize AMR in the ICU. From a clinical viewpoint, clinicians at UHF considered that patients with a positive culture (presence of multiresistant germs) in the first 48 hours, had acquired the AMR before their ICU admission. On the contrary, we considered that patients with a positive culture after the early 48 hours of their admission, had acquired the AMR during their ICU stay. Therefore, 507 of the total number of patients acquired antimicrobial resistance, of which 171 (33.73%) acquired AMR before their ICU admission and 336 (66.27%) during their ICU stay. The average age of AMR patients is 62.39 years, and 59.29 for non-AMR patients. In both cases, the standard deviation is high (13.00 and 16.02, respectively). Regarding gender, the percentage of men is higher for both AMR and non-AMR patients (63.71% and 61.13%, respectively).

The dataset has been preprocessed to characterize the evolution of the patient's health status by a set of features suitable to feed the predictive model inputs. Thus, the d-th temporal feature corresponding to the i-th patient is represented by a a row vector associated to a T-days time window, and it is given by: $\mathbf{x}_{i,d} = [x^1_{i,d}, x^2_{i,d}, \cdots, x^T_{i,d}]$, with $d = 1, \cdots, D$. In this work, we have considered $T = 7$ time slots, i.e, the temporal characterization of a patient has been done in a 7-days time window, with t_0 the first 24 hours from the ICU admission for the non-AMR patients. Regarding AMR patients, the time slot t_0 represents the time slot furthest from the first positive culture, and therefore, closest to the ICU admission. Since the length of the ICU stay can be shorter than 7 days for some patients, we created a new binary feature, called mask, which takes a value of '1' if the patient was in the ICU at this time slot, or '0' otherwise. The upper panel in Fig. 1 illustrates ficticious values for the mask and the D features associated to one AMR patient. In this example, since the culture flagged as positive the fifth day since the patient's ICU admission, all features assigned to t_0 and t_1 have null values. The bottom panel in Fig. 1 represents the hypothetical values for the mask and features associated to a potential non-AMR patient with a stay of at least 7 days, being t_0 the time slot nearest to the patient's ICU admission.

The features represented as $\mathbf{x}_{i,d}$ in Fig. 1 are associated to the family of antibiotics taken by the patient (23 features), as well as to the mechanical ventilation (MV), to the result of the albumin blood test and to the number of times this blood test was required. The families of the antibiotics the patient can take are the following: Aminoglycosides (AMG), Antifungals (ATF), Carbapenemes (CAR), 1st generation Cephalosporins (CF1), 2nd generation Cephalosporins (CF2), 3rd generation Cephalosporins (CF3), 4th generation Cephalosporins (CF4), unclassified antibiotics (Others), Glycyclines (GCC),Glycopeptides (GLI), Lincosamides (LIN), Lipopeptides (LIP), Macrolides (MAC), Monobactamas (MON), Nitroimidazolics (NTI), Miscellaneous (OTR), Oxazolidinones (OXA), Broad-Spectrum Penicillins (PAP), Penicillins (PEN), Polypeptides (POL), Quinolones (QUI), Sulfamides (SUL) and Tetracyclines (TTC). Regarding the feature associated to MV, for each time slot we have considered the number of hours the patient was assisted with mechanical ventilation. The use of these features is supported by the fact that the incorrect and excessive use of antibiotics or external devices are one of the main causes for the AMR onset. In addition, two demographic features (not time-dependent), the age and the gender of the patient, have been used as input of the models.

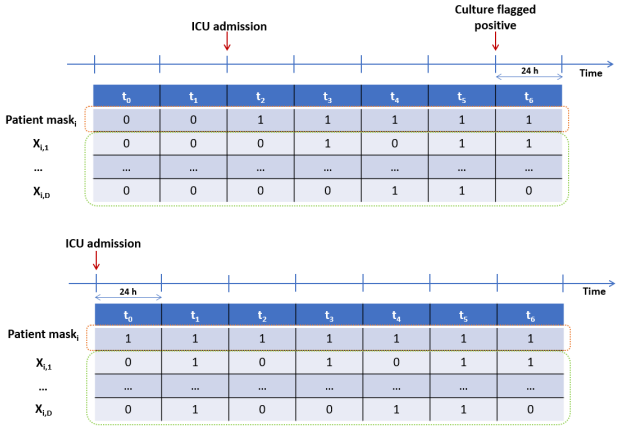

Figure 1. Temporal feature matrix construction with a time window of 7 consecutive slots of 24 hours: AMR patient (upper panel) and non-AMR patient (bottom panel). For the AMR patient, t_6 represents the time slot closest to the date the positive culture is performed. For the non-AMR patient, t_0 represents the time slot closest to the patient's ICU admission.

We present in Fig. 2 the percentage of AMR and non-AMR patients who take each family of antibiotics. Note that this percentage is similar for some families of antibiotics such as Broad-Spectrum Penicillins, Quinolones and Lipopeptides. However, the percentage of Antifungals, Glycopeptides and Carbapenemes is higher for AMR patients, while non-AMR patients present a higher percentage of Penicillins, among others.

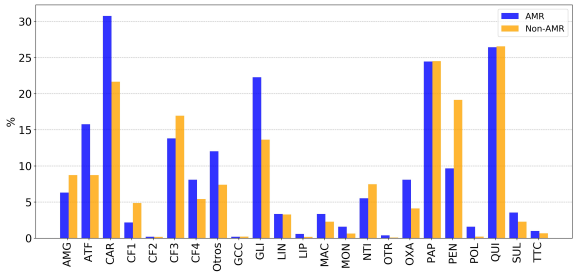

Figure 2. Percentage of patients with respect to the total of each population (AMR and non-AMR) taking a particular family of antibiotics.

4 EXPERIMENTS AND RESULTS

The goal of this work was twofold. On the one hand, a feature selection strategy was applied to find the most relevant features to discriminate between AMR and non-AMR patients. On the other hand, the chosen features were considered to evaluate the potential of different prediction models when classifying AMR and non-AMR patients. Towards that end, we start this section by discussing the ex-

perimental set-up, then we present the feature selection process and the prediction results.

4.1 Experimental Setup

The methodology to select relevant features and train different classifiers is as follows. First, patients in the dataset were randomly separated, assigning the 70% to the train set and the 30% to the test set [5]. In order to reduce the potential bias in the results produced by good or bad partitions, we repeat this process 1000 times. The metrics used for measuring the performance of the classifiers are the mean and the standard deviation of the Accuracy, Specificity, Sensitivity, F1-score and the Area Under the Curve (AUC).

For tuning hyperparameters, a 5-fold cross-validation strategy was considered in the training set. For the LR models the hyperparameter used was the penalty coefficient $C \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 0.75, 1.0\}$. The hyperparameters associated to the decision tree were the depth of the tree (ranging from 4 to 22) and the minimum of samples per leaf (between 6 and 15).

4.2 Temporal Feature Selection

We performed a hypothesis test for each time slot and for all features described in Section 3, except for demographic features due to the non-dependence in time of this kind of features. For both imbalanced and balanced data, we considered the p-value provided by the two-proportion z-test for antibiotics and by the two-sample Kolmogorov-Smirnov test for MV and albumin (see Table 1). In the case of imbalanced data, we determined as significant features those with a p-value < 0.1. When considering balanced datasets, we perform $N = 1000$ subsamplings of the majority class and obtain the median of the p-values, selecting those features such that the median of the p-values is lower than 0.1.

To perform the experiments, we have used those features selected by the above tests when using balanced subsets, together with the demographic features of the patient. We have selected those features that are statistically significant (p-value ¡ 0.1) during the first 48 hours (t_0 and t_1), from 48 hours (t_2,t_3,t_4,t_5, and t_6) onwards or throughout the time window (from t_0 to t_6). According to these conditions, we have obtained the following features: all time slots of ATF, PEN, OXA, and Albumin (Value), from time slot t_2 to t_6 for Others and MV (hours), and the first two time slots for AMG, CF3, GLI, NTI, QUI and Albumin (Count). Some of these features are clinically relevant. For example, QUI and AMG are antimicrobial families employed to treat the pseudomona aeruginosa infections, OXA family are the main antimicrobial given to tackle the staphylococcus aureus (both pseudomonas aeruginosa and staphylococcus aureus are the most common MDR bacteria). The mechanical ventilation and the level of albumin in the blood are related to the patient's state of health. The p-values associated to these features and time slots are in bold in Table 1.

4.3 Prediction Results

In this subsection, the results of predicting whether a patient will be considered AMR or non-AMR are presented in Table 2. For the prediction, we considered both a linear (LR) and non-linear (DT) models, designed using the features selected in Subsection 4.2.

Several conclusions can be obtained from Table 2, where the mean and standard deviation of several performance measurements on the test subsets of 1000 subsamplings are provided using random undersampling and SMOTE to balance the data. In general, the LR model (a linear model) achieves better results, especially in terms of Sensitivity (69.97 ± 3.68). On the contrary, better results in term of Specificity (86.13 ± 1.87) are obtained when considering DT (non-linear model). The results obtained through the use of SMOTE for LR improve, except Sensitivity. On the other hand, in DT, better results are obtained for Specificity and Accuracy, while the other metrics worsen.

Figure 3 shows the importance of the features provided by 1000 different models when considering LR and DT. To estimate the feature importance in LR, we have considered the absolute values of the weights associated to the features, while we have used the Gini index in DT. The results are presented in box-plots, with features sorted increasingly according to median of the p-values provided by 1000 models. Features with the highest importance are approximately the same in both classifiers, highlighting MV in some time slots, the blood albumin value and the age of the patient.

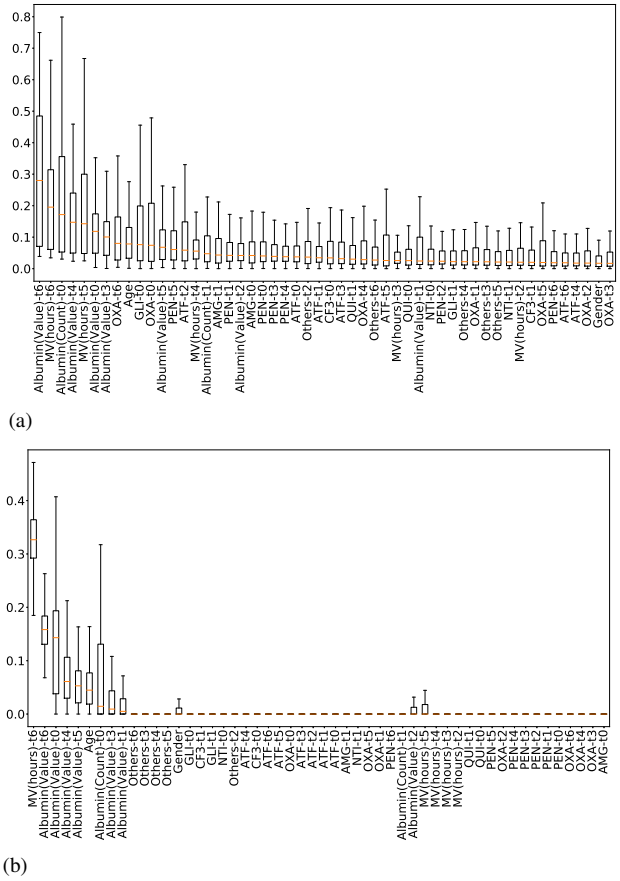

Figure 3. Box-plots of the importance of features provided by 1000 different models: (a) absolute value of the coefficients for the LR models; (b) importance based on gini index.

5 CONCLUSIONS

Nowadays, AMR has become a real and growing problem due to the inappropriate use of antimicrobials. Bacteria that were previously

Table 1. p-value obtained when performing a hypothesis test on a single feature per time slot, associated to AMR and non-AMR populations. First p-value in the cell corresponds to unbalanced datasets, while the second value shows the median of the p-values for balanced dataset on 1000 subsamplings. Bold figures denote those features satisfying the alternative hypothesis (p-value lower than 0.1).

Feature \ Time Slot	t_0	t_1	t_2	t_3	t_4	t_5	t_6
AMG	2.943e-03	6.167e-03	1.606e-02	5.128e-02	1.467e-01	1.929e-01	8.895e-01
	7.440e-03	**1.411e-02**	3.733e-02	9.571e-02	2.465e-01	2.781e-01	7.193e-01
ATF	4.497e-07	1.483e-07	4.722e-09	1.543e-07	8.290e-08	2.623e-06	6.851e-03
	4.948e-04	**6.484e-04**	**1.938e-04**	**5.590e-04**	**3.763e-04**	**1.656e-03**	**4.029e-02**
CAR	1.186e-01	7.480e-02	1.003e-01	2.500e-03	5.866e-03	2.408e-04	1.924e-04
	2.130e-01	1.958e-01	2.171e-01	6.678e-02	5.954e-02	2.019e-02	5.289e-03
CF1	2.808e-02	2.199e-01	2.432e-01	2.430e-01	8.337e-02	4.779e-01	7.706e-01
	3.229e-02	3.127e-01	3.132e-01	2.541e-01	5.744e-02	5.242e-01	7.044e-01
CF2	4.798e-01	1.0e+00	1.0e+00	6.686e-01	1.405e-01	3.281e-01	6.444e-01
	1.0e+00	1.0E+00	1.0e+00	1.0e+00	3.169e-01	3.169e-01	1.0e+00
CF3	1.230e-03	1.608e-02	3.104e-02	9.467e-02	6.654e-01	8.983e-01	1.845e-01
	4.685e-03	**3.716e-02**	5.323e-02	1.321e-01	6.257e-01	6.527e-01	3.128e-01
CF4	2.870e-01	3.282e-01	3.930e-01	3.498e-01	3.258e-02	5.092e-02	2.525e-01
	5.024e-01	5.032e-01	5.669e-01	4.227e-01	1.978e-01	1.386e-01	3.330e-01
Others	6.868e-02	1.405e-01	4.077e-04	3.253e-04	4.630e-05	1.042e-04	2.756e-02
	2.079e-01	2.723e-01	**1.712e-02**	**1.133e-02**	**5.216e-03**	**7.956e-03**	**9.556e-02**
GCC	3.866e-01	5.120e-01	6.722e-01	5.981e-01	5.222e-01	6.900e-01	2.274e-01
	3.167e-01	3.168e-01	3.168e-01	3.169e-01	3.169e-01	3.169e-01	3.170e-01
GLI	1.406e-04	4.152e-03	4.495e-02	2.368e-02	2.255e-04	8.859e-06	8.654e-06
	5.365e-03	**3.866e-02**	1.243e-01	1.008e-01	8.780e-03	3.000e-03	1.531e-03
LIN	1.288e-01	5.057e-01	7.211e-01	8.0236e-01	7.746e-01	6.766e-01	7.989e-01
	2.007e-01	3.989e-01	6.486e-01	6.322e-01	6.326e-01	6.331e-01	6.343e-01
LIP	6.175e-01	2.537e-01	6.501e-01	1.768e-01	3.314e-01	2.781e-03	6.444e-01
	1.0e+00	3.168e-01	1.0e+00	3.169e-01	3.169e-01	8.263e-02	1.0e+00
MAC	5.310e-01	7.344e-01	6.374e-01	8.605e-01	3.339e-01	1.379e-02	9.732e-02
	5.223e-01	7.034e-01	7.036e-01	7.038e-01	5.239e-01	8.075e-02	2.454e-01
MON	1.852e-01	6.557e-01	4.626e-01	5.367e-01	4.164e-01	4.321e-02	1.021e-02
	1.563e-01	5.624e-01	6.532e-01	6.533e-01	6.534e-01	2.545e-01	9.410e-02
NTI	1.642e-02	3.691e-02	3.441e-01	5.944e-01	8.174e-01	1.247e-01	5.465e-01
	4.325e-02	**6.454e-02**	4.214e-01	6.640e-01	7.086e-01	2.253e-01	5.842e-01
OTR	1.0e+00	1.0e+00	6.501e-01	6.686e-01	1.405e-01	3.281e-01	2.524e-02
	1.0e+00	1.0e+00	1.0e+00	1.0e+00	3.169e-01	3.169e-01	1.568e-01
OXA	2.821e-04	7.593e-03	8.882e-03	3.699e-03	1.195e-05	3.403e-04	1.686e-05
	9.830e-03	**6.735e-02**	**7.648e-02**	**4.899e-02**	**6.101e-03**	**2.435e-02**	**3.685e-03**
PAP	1.319e-01	6.248e-02	1.286e-02	1.798e-02	5.152e-02	1.540e-01	8.518e-01
	2.256e-01	1.048e-01	3.750e-02	6.685e-02	1.387e-01	2.650e-01	6.804e-01
PEN	4.679e-06	9.010e-06	2.468e-04	1.372e-04	1.342e-03	3.036e-03	1.492e-02
	1.030e-05	**1.966e-05**	**1.313e-03**	**5.027e-04**	**4.512e-03**	**1.231e-02**	**4.455e-02**
POL	6.175e-01	2.537e-01	2.034e-03	9.855e-04	1.062e-02	7.761e-03	1.985e-07
	1.0e+00	3.168e-01	8.241e-02	8.250e-02	8.256e-02	1.782e-01	4.516e-03
QUI	9.795e-03	1.982e-02	4.590e-03	4.443e-02	7.272e-01	7.191e-01	8.757e-01
	3.028e-02	**5.761e-02**	1.700e-02	7.849e-02	5.774e-01	6.558e-01	6.735e-01
SUL	8.343e-01	6.689e-01	1.280e-01	2.232e-01	4.833e-02	3.959e-03	4.582e-01
	7.032e-01	4.758e-01	2.429e-01	4.003e-01	3.113e-01	8.586e-02	5.904e-01
TTC	4.105e-01	1.686e-01	3.659e-01	8.036e-01	8.793e-01	3.018e-01	4.320e-01
	5.623e-01	3.153e-01	4.762e-01	5.627e-01	5.628e-01	4.123e-01	4.128e-01
MV (hours)	7.451e-01	4.393e-01	1.441e-04	2.180e-12	0.0e+00	1.882e-26	1.332e-15
	7.705e-01	7.423e-01	**3.997e-03**	**1.238e-07**	**9.859e-14**	**5.943e-16**	**1.046e-19**
Albumin (Value)	5.551e-16	4.593e-06	1.646e-04	9.611e-05	1.145e-03	2.981e-01	1.332e-15
	2.016e-19	**6.524e-04**	**5.390e-03**	**4.703e-03**	**2.541e-02**	**5.892e-01**	2.689e-10
Albumin (Count)	5.551e-16	6.335e-06	1.646e-04	3.506e-04	4.085e-02	6.951e-01	1.665e-15
	2.016e-19	**9.377e-04**	7.237e-03	1.065e-02	2.209e-01	9.211e-01	6.278e-10

Table 2. Mean $\pm$ standard deviation of several performance measurements (Specificity, Sensitivity, Accuracy, F1-score and AUC) on 1000 test sets when designing a lineal model (LR) and a non linear model (DT).

Training Strat.	Model	Specificity	Sensitivity	Accuracy	F1-score	AUC
Random Under.	**LR**	73.45 ± 2.09	69.97 ± 3.68	72.83 ± 1.64	47.79 ± 2.54	71.71 ± 1.82
	DT	78.52 ± 3.77	60.45 ± 4.38	75.23 ± 2.84	47.08 ± 3.25	69.48 ± 2.19
SMOTE	**LR**	77.52 ± 1.78	66.6 ± 3.91	75.58 ± 1.44	49.2 ± 2.62	72.06 ± 1.9
	DT	86.34 ± 1.87	48.78 ± 4.63	79.65 ± 1.56	45.97 ± 3.5	67.56 ± 2.25

easily treatable have now become an issue difficult to deal with, especially in the ICUs. In these units, AMR has created a great impact on morbidity, hospital costs, and sometimes patient survival.

It is necessary to be aware of the growing problem caused by the expansion of AMR, for which new research, efforts, and approaches are needed to prevent further spread of AMR. The use of automatic learning methods is a very useful tool to solve problems related to the clinical environment following a data-driven strategy. These methods allows us to reduce the time of detection of infectious diseases, resulting in a reduction in the number of deaths as well as in health economic costs.

In this work, we proposed the use of feature selection and machine learning approaches to extract knowledge and predict the appearance of AMR of patients admitted in the ICU. Features such as the performance provided by LR (71.71% AUC) suggests that the analysis presented in this paper could be a first step to identify the bacteria appearance and isolate the patients at risk of AMR.

As future work, we propose the analysis of more features related to the patients such as blood samples or vital signs, as wells as the use of more advanced machine learning methods, as for example, long short-term memory networks which are capable of learning long-term dependencies.

ACKNOWLEDGEMENTS

This work has been partly supported by the Institute of Health Carlos III, Spain (grant DTS 17/00158), by the Spanish Ministry of Economy, Industry and Competitiveness under the Research Project Klinilycs (TEC2016-75361-R), by the Science and Innovation Ministry Grants AAVis-BMR (PID2019-107768RA-I00) and BigTheory (PID2019-106623RB-C41), by Project Ref. F656 financed by Rey Juan Carlos University, by the Young Researchers R&D Project Ref. 2020-661, financed by Rey Juan Carlos University and Community of Madrid (Spain), and by the Youth Employment Initiative (YEI) R&D Project Ref. TIC-11649 financed by the Community of Madrid (Spain).

REFERENCES

[1] Matteo Bassetti, Garyphallia Poulakou, Etienne Ruppe, Emilio Bouza, Sebastian J Van Hal, and Adrian Brink, 'Antimicrobial resistance in the next 30 years, humankind, bugs and drugs: a visionary approach', *Intensive care medicine*, **43**(10), 1464–1475, (2017).

[2] W Baumgartner, P Weiß, and H Schindler, 'A nonparametric test for the general two-sample problem', *Biometrics*, 1129–1135, (1998).

[3] L Breiman, JH Friedman, RA Olshen, and CJ Stone, *Classification and Regression Trees*, Chapman and Hall, 1984.

[4] Nele Brusselaers, Dirk Vogelaers, and Stijn Blot, 'The rising problem of antimicrobial resistance in the intensive care unit', *Annals of intensive care*, **1**(1), 47, (2011).

[5] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad, 'Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission', in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730, (2015).

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer, 'Smote: synthetic minority over-sampling technique', *Journal of artificial intelligence research*, **16**, 321–357, (2002).

[7] Sara E Cosgrove, 'The relationship between antimicrobial resistance and patient outcomes: mortality, length of hospital stay, and health care costs', *Clinical Infectious Diseases*, **42**(Supplement_2), S82–S89, (2006).

[8] José F Díez-Pastor, Juan J Rodríguez, César García-Osorio, and Ludmila I Kuncheva, 'Random balance: ensembles of variable priors classifiers for imbalanced data', *Knowledge-Based Systems*, **85**, 96–111, (2015).

[9] Jianqing Fan and Runze Li, 'Statistical challenges with high dimensionality: Feature selection in knowledge discovery', *arXiv preprint math/0602133*, (2006).

[10] Infectious Diseases Society of America (IDSA), 'Combating antimicrobial resistance: policy recommendations to save lives', *Clinical Infectious Diseases*, **52**(suppl_5), S397–S428, (2011).

[11] G James, D Witten, T Hastie, and R Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2013.

[12] Sergio Martínez-Agüero, Inmaculada Mora-Jiménez, Jon Lérida-García, Joaquín Álvarez-Rodríguez, and Cristina Soguero-Ruiz, 'Machine learning techniques to identify antimicrobial resistance in the intensive care unit', *Entropy*, **21**(6), 603, (2019).

[13] Marc Mendelson and Malebona Precious Matsoso, 'The world health organization global action plan for antimicrobial resistance', *SAMJ: South African Medical Journal*, **105**(5), 325–325, (2015).

[14] J. Ross Quinlan, 'Induction of decision trees', *Machine learning*, **1**(1), 81–106, (1986).

[15] Paz Revuelta-Zamorano, Alberto Sánchez, José Luis Rojo-Álvarez, Joaquín Álvarez-Rodríguez, Javier Ramos-López, and Cristina Soguero-Ruiz, 'Prediction of healthcare associated infections in an intensive care unit using machine learning and big data tools', in *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*, pp. 840–845. Springer, (2016).

[16] Pilar Talón-Ballestero, Lydia González-Serrano, Cristina Soguero-Ruiz, Sergio Muñoz-Romero, and José Luis Rojo-Álvarez, 'Using big data from customer relationship management information systems to determine the client profile in the hotel sector', *Tourism Management*, **68**, 187–197, (2018).

[17] Jiliang Tang, Salem Alelyani, and Huan Liu, 'Feature selection for classification: A review', *Data classification: Algorithms and applications*, 37, (2014).

[18] Magnus Unemo and William M Shafer, 'Antimicrobial resistance in neisseria gonorrhoeae in the 21st century: past, evolution, and future', *Clinical microbiology reviews*, **27**(3), 587–613, (2014).

[19] Dirk Vogelaers, De Bels, et al., 'Patterns of antimicrobial therapy in severe nosocomial infections: empiric choices, proportion of appropriate therapy, and adaptation rates—a multicentre, observational survey in critically ill patients', *International journal of antimicrobial agents*, **35**(4), 375–381, (2010).

[20] Show-Jane Yen and Yue-Shi Lee, 'Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset', in *Intelligent Control and Automation*, 731–740, Springer, (2006).