Applying LSTM Networks to Predict Multi-drug Resistance Using Binary Multivariate Clinical Sequences

S. Martínez-Agüero
Signal Theory Dept.
Univ Rey Juan Carlos
sergio.martinez@urjc.es

I. Mora-Jiménez
Signal Theory Dept.
Univ Rey Juan Carlos
inmaculada.mora@urjc.es

J. Álvarez-Rodríguez
Intensive Care Dept.
Hospital Univ Fuenlabrada
joaquin.alvarez@urjc.es

A. García-Marqués
Signal Theory Dept.
Univ Rey Juan Carlos
antonio.marques@urjc.es

C. Soguero-Ruiz
Signal Theory Dept.
Univ Rey Juan Carlos
cristina.soguero@urjc.es

Abstract

Longitudinal multivariate clinical sequences recorded in the Electronic Health Records (EHR) provide valuable information about patient health care encounters. Mining effectively these data is challenging because these sequences are characterized by irregular sampling and missing values. Long-Short Term Memory (LSTM) Networks have been recently used to evaluate the potential predictive usefulness in longitudinal EHR data. In this paper, we apply LSTM to predict multi-drug resistance (MDR) in the Intensive Care Unit (ICU) of the University Hospital of Fuenlabrada (Madrid, Spain). Towards that end, we analyze multivariate clinical sequences available in the EHR. Specifically, we consider irregularly sampled clinical measurements about mechanical ventilation and different families of antibiotics taken by the patients. Since the number of patients with MDR is much lower than that with non-MDR, we apply different loss functions and figures of merits to evaluate the LSTM network performance. We conclude that multivariate clinical sequences provide powerful insights to predict MDR in ICU.

1 Introduction

Time series sequences are present in many different fields such as finance [Tay08] or medicine [CBA15]. In the last decade, there has been a growing interest and need in analyzing data collected over time from a diverse range of sources, known as multivariate time series (MTS). For example, in medicine, we can find clinical temporal sequences associated with blood test results or vital signs. In this context, several approaches have been proposed and used to predict and classify MTS, ranging from feature-based methods [WYL18], which provide a new set of features representing the statistical properties of the time series at hand, to deep learning networks, which are able to model and characterize directly MTS [YLH15].

In: Rudolph S., Marreiros G. (eds.): Proceedings of STAIRS, Santiago de Compostela, Spain, august 2020, published at http://ceur-ws.org

Among deep learning approaches, the Long Short-Term Memory (LSTM) networks have been observed as one of the most effective solutions to deal with MTS [HS97a]. LSTMs present several advantages over conventional feed-forward and recurrent neural networks. These advantages include the capacity of LSTM to grasp the long-term dependence of time sequences, which provides them with great versatility. LSTMs have been traditionally used in tasks related to language modeling, image captioning and translation, among others [YPL+17, SSN12]. The use of LSTM in areas such as medicine has significantly increased during the last years. For example, Lipton et al. applied a LSTM network to classify diagnoses based on the temporal data recorded in the Electronic Health Record (EHR) of a pediatric Intensive Care Unit (ICU) [LKEW15]. Nguyen et al. used a bidirectional LSTM model with an attention mechanism to predict ICU mortality outcomes, showing competitive results on the 2012 Physionet data set [NTV17]. Data analyzed in both studies are characterized by MTS reflecting the evolution of the patient statuses and their clinical situations. Despite the clinical insights that MTS may provide, mining effectively these data can be challenging for several reasons: irregular sampling, high dimensionality or imbalanced classification.

In this paper, we deal with these challenges by exploring the potential of the LSTM networks to predict multi-drug resistance (MDR) in the ICU. MDR can be defined as the capacity of bacteria to withstand the effects of a variety of harmful chemical agents designed to damage it [MDHL14]. That means that diseases easily treatable today will require much more complex and expensive treatments than the current ones. Eventually, they could even suddenly become lethal. Even more everyday situations such as minor injuries or cuts could pose a very high risk [MDHL14]. To know the in vitro activity of an antibiotic against a given bacterium (previously isolated in the culture), the results of the antibiogram are analyzed. These results usually take between 24/48 h. Accurate and faster identification of the MDR may enable to implement preventive steps, patient isolation, and therefore a reduction in the rising rates of MDR.

Specifically, we analyze MTS measurements of mechanical ventilation as well as the use of different families of antibiotics by the patients at the ICU of the University Hospital of Fuenlabrada (UHF) in Madrid (Spain). Taking into account these series, we propose to predict the MDR imminence by applying a LSTM network.

The rest of the manuscript is organized as follows. Section 2 presents the notation and the statistical approaches used in this paper. In Section 3 we describe the data set, while experiments and results are shown in Section 4. Finally, main conclusions and discussion are drawn in Section 5.

2 Methods

Notation

In this paper, the learning algorithms are performed on a set of D time series per patient, with one temporal series per variable. Each time-series is composed by T time slots. Thus, data associated to the i-th patient can be arranged in the matrix $\mathbf{X}_i = [\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^T] \in \mathbb{R}^{D\times T}$. Column vector $\mathbf{x}_i^t$ contains the D variables associated to the t-th timestlot, i.e., $\mathbf{x}_i^t = (x_i^{(t,1)}, x_i^{(t,2)}, \cdots, x_i^{(t,D)})$. Thus, $x_i^{(t,d)}$ represents the value of the d-th variable in the t-th times slots for the i-th patient. Since we are interested in a binary classification problem, we have considered the label '1' to identify patients with MDR, and the label '0' to identify patients with non-MDR. Thus, we represent the label for the i-th patient by y_i, and the output provided by the network with $\hat{y}_i$.

2.1 Conventional Learning Methods

In this paper we will explore two conventional methods: logistic regression (LR) and voting k-nearest neighbours (k-nn). On the one hand, Logistic Regression (LR) is a parametric approach based on a logistic function to estimate a binary variable as a linear combination of d independent features [DGK02]. The coefficients of the linear model are found by optimizing a regularized cost function, which depends on a penalty coefficient C, to prevent overfitting (Ridge regularization [Tik63]).

On the other hand, k-nearest neighbour (k-nn) is a non-parametric and non-linear classifier well-known in the literature. This classifier depends on a distance measure to determine the similarity between samples. In this paper we have considered the Jaccard distance, which is widely used in the literature for binary variables [Ouy16]. The voting k-nn assigns a class to an unclassified sample taking into account the majority class among the k closest neighbours of the sample in question. Note that the appropriate choice of k conditions the behaviour of the classifier. It is highly dependent on the dimensionality and the training set size, and should be chosen following a cross-validation strategy.

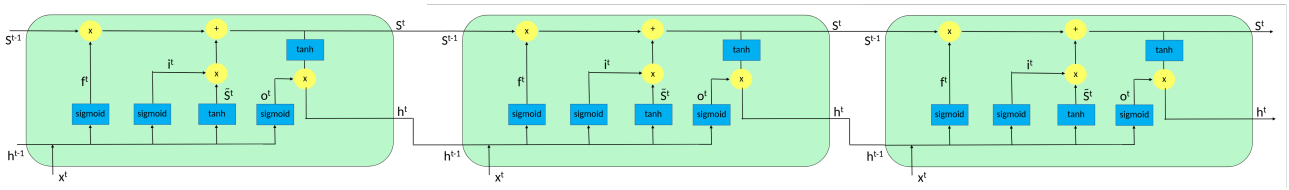


Figure 1: Visual representation of a LSTM cell.

LSTM Networks for Classification of Multivariate Series

The LSTM architecture is a special kind of artificial neural network (NN) composed by a set of recurrently connected subnets, knowns as memory blocks [HS97b]. The main differences between the architecture of standard NN and LSTM are that, on the one hand, conventional NNs can only map from input to output vectors and, on the other hand, conventional NNs do not allow cyclical connections between neurons. As a consequence, standard NNs are not capable of mapping sequential inputs.

When NNs consider cycles between neurons, the architecture is named as Recurrent Neural Network (RNN). Note that cycles between neurons allow preserving as an 'artificial memory' the information of previous inputs in the internal state of the RNN. Therefore, RNNs can be considered as NNs specialized in mapping sequential inputs [Gra12]. Unfortunately, it is not convenient to work with standard RNN architectures when considering very long MTS. The problem occurs due to the influence of past inputs on the hidden layers because of the cycles between the neurons and the artificial memory commented previously. If the artificial memory makes the gradient either decays or blows up exponentially, the RNN does not work properly. This effect is referred to as the vanishing gradient problem [Gra12]. The LSTM invention was motivated for avoiding the gradient's problems of the RNNs.

The key element of LSTM networks is the cell state (see Fig. 1), which considers information of previous inputs [Gra12]. A simple LSTM cell consists of three kind of gates: forget, input and output.

The forget gate layer ($\mathbf{f}^t$) decides what information of the past states is going to be deleted. It uses $\mathbf{h}^{t-1}$ (the output of the previous state) and $\mathbf{x}^t$, then the result passes through the sigmoid activation function. This process creates a mask that multiplies the previous state $\mathbf{S}^{t-1}$.

The input gate layer ($\mathbf{S}^t$) determines what new information (from $\mathbf{x}^t$) is going to be stored in the cell state. This layer is composed of two mechanism, the first one creates a vector of new candidate values, $\tilde{\mathbf{S}}^t$ and the second one, $\mathbf{i}^t$ decides by how much we decided to update each state value. Next, $\mathbf{i}^t$ and $\tilde{\mathbf{S}}^t$ are multiplied and the result is added to the previous $\mathbf{S}^{t-1} * \mathbf{f}^{t-1}$.

The output gate layer ($\mathbf{o}^t$) decides the exact output, which will be based on our cell state but will be a filtered version. It is composed of a first *sigmoid* sublayer which decides what parts of the cell state will be output. This creates a vector known as $\mathbf{o}^t$. Then the cell state S^t pass through a *tanh* function (to push the values in the interval [-1, 1]) and multiplies $\mathbf{o}^t$ giving as a result $\mathbf{h}^t$.

2.2 Strategies to Handle Imbalanced Data Sets

As it usually happens in the medical context, the classification task is highly imbalanced, with a lower number of patients with MDR regarding those with non-MDR. Dealing with imbalanced classes can lead to a poor model performance since the model can be biased to the majority class [WLW+16]. Since the LSTM network presented above does not have any mechanism for dealing with imbalanced data sets, we have considered two different strategies to train the model: (1) Balancing classes in training, following an undersampling approach to reduce the number of patients (samples) associated to the majority class until both classes have a similar number of patients; (2) Considering cost-sensitive learning, i.e., penalizing errors on the minority class by a cost proportional to how under-represented this class is. In the following, we assume that the network can provide estimates $\hat{y}$ of the posterior probability for class '1'. Thus, assuming that the size of the training set is N, the cost functions considered in this paper are presented below:

- The Binary Cross-Entropy (BCE) function is one of the most used for in binary classification problems. For this reason, we have considered it as a baseline. The BCE function increases as $\hat{y}$ diverges from the actual label y as follows:

$$-\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log\left(\hat{y}_i\right) + \left(1-y_i\right)\log\left(1-\hat{y}_i\right)\right) \tag{1}$$

- The Balanced Binary Cross-Entropy (BBCE) function is an extension of BCE to include a class-balancing weight $\beta \in (0, 1)$. The idea is to penalize more the errors on the less frequent class, as follows:

$$-\frac{1}{N}\sum_{i=1}^{N}\left(\beta y_i \log\left(\hat{y}_i\right) + (1-\beta)\left(1-y_i\right)\log\left(1-\hat{y}_i\right)\right) \tag{2}$$

  The weight β may be set either taking into account the class frequency or by validation. In this paper, we consider a validation strategy.

- The Focal Loss (FL) function reshapes the BBCE function with a tunable focusing parameter $\gamma \geq 0$ (BBCE is obtained with $\gamma = 0$). As indicated in [LGG+17], the idea is to down-weight samples easy to classify and thus to focus training on those samples of the less frequent class which are hard to classify. The FL function is defined as

$$-\frac{1}{N}\sum_{i=1}^{N}\left(\beta y_i \left(1-\hat{y}_i\right)^{\gamma}\log\left(\hat{y}_i\right) + (1-\beta)\left(1-y_i\right)\hat{y}_i^{\gamma}\log\left(1-\hat{y}_i\right)\right) \tag{3}$$

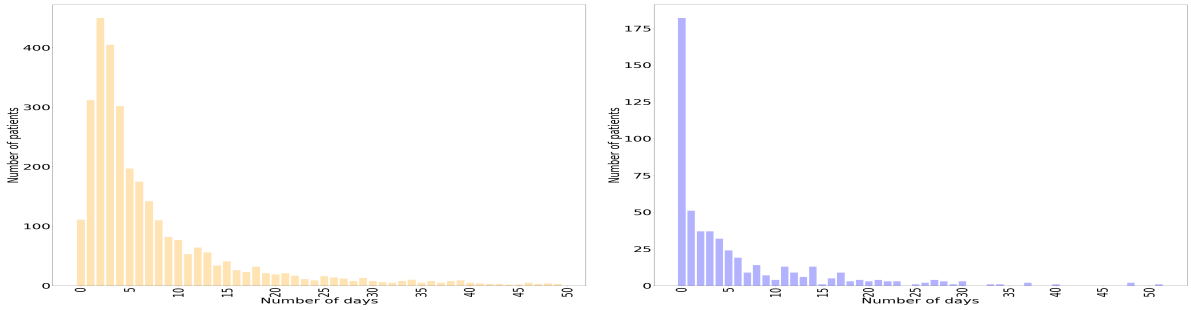
Regarding the figures of merit considered to evaluate the model performance, we have considered the following ones: Classification Accuracy, Specificity, Sensitivity and F1-score. The Accuracy is the ratio of correctly classified samples over all samples used for evaluation. It can provide misleading conclusions because results can overestimate the model performance, e.g. high accuracy rate can be achieved while most samples for the minority class are incorrectly classified. To avoid the potential bias of the Accuracy towards the majority class, Specificity and Sensitivity get the scores for each class. In particular, Sensitivity (also called Recall) has been computed as the proportion of patients with MDR who are correctly identified by the model as having the condition; Specificity refers to the proportion of non-MDR patients who are correctly identified by the model as not having the condition. With regard to the F1-score, it is the weighted average of Sensitivity and Precision (number of correctly predicted MDR-patients over the total number of patients labelled by the model as MDR-patients). The use of Accuracy is more appropriate when false positives and false negatives have similar "cost", otherwise, the use of F1-score is more convenient. At this work, we have stopped the LSTM network training according to an early-stopping approach based on the F1-score.

3 Database Description and Pre-processing

Data analyzed in this work were collected from the EHR of UHF for a period of 13 years, running from 2004 to 2016. A total of 2,540 patients were admitted to the ICU during this period, and 449 of them had MDR bacteria. A patient is identified as MDR if a culture is flagged as positive, indicating that MDR bacteria was found. Since several cultures of the same patient can be flagged with MDR bacteria, we have limited our research to the prediction of the first of these cultures. As a consequence, we have not considered the patient's history after the time step when this culture is performed.

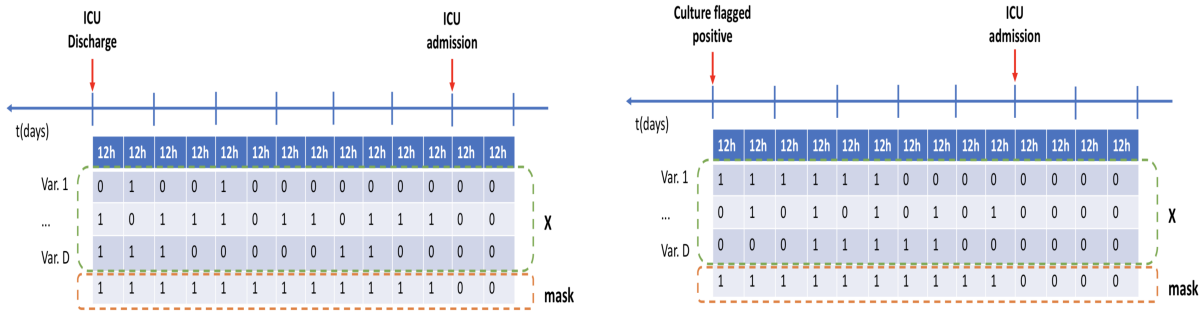
To have an idea of the length of the series, we show in Figure 2 the distribution of these time series both for MDR and non-MDR patients. Note that the time series for MDR patients are considered just until the day any culture is flagged as positive. Figure 2 (a) shows the histogram of the elapsed time from the ICU admission to the ICU discharge for patients with no culture flagged as positive along with the whole ICU stay. For patients with any culture flagged as positive, the distribution of the elapsed time from the ICU admission to the time the first MR culture flagged as positive is shown in Figure 2 (b). We have tacked the patient characterization by a MTS registered every 12 hours during a 7-day window. For MDR patients, we consider the first time slot (t=1 in $\mathbf{X}$, the first 12 hours) as the day the culture that will be flagged as positive was performed and then continue completing the series in reverse time order. For non-MR patients, we also consider a reverse time order in matrix $\mathbf{X}$, but the first time slot corresponds to the day the patient was admitted in the ICU. Figure 3 shows a sketch illustrating it.

In this work, the MTS are represented by the family of antibiotics taken by the patient, as well as the time the patient has been assisted with mechanical ventilation. Regarding the first group of variables, a total of 15 antibiotic families were extracted: Aminoglycosides, antifungals, carbapenems, third generation cephalosporins, fourth generation cephalosporins, glycylcyclines, lincosamides, nitroimidazoles, oxazolidinones, broad-spectrum penicilins, penicilins, polymyxins, quinolones, sulfonamides and non grouped antibiotics. Associated to each family of antibiotics, a binary variable was created to indicate whether the patient has taken (or not) that family



(a) (b)

Figure 2: Histogram of the elapsed time from the ICU admission to: (a) the ICU discharge for non-MDR patients; (b) the first culture flagged as positive for MDR patients.



(a) (b)

Figure 3: Structure of matrix **X** and binary **mask** for two arbitrary patients: (a) Non-MDR, and (b) MDR.

of antibiotics during a period of 12 hours. Regarding mechanical ventilation, a binary variable indicates whether the patient has been connected (or not) to a breathing machine for a temporary period of 12 hours.

Apart from the clinical registered variables, it has been reported that missing values are usually informative missingness [CPC+18], i.e., missing values can provide rich information about target labels in supervised learning tasks (e.g, time series classification). To effectively exploit the information of missingness patterns, we will use a mask (an additional binary variable) indicating for each time slot whether the patient was in the ICU during the period of 12 hours under study or not. As it is shown in Figure 3, values in the mask are zero whenever the patient is not in the ICU.

4 Experiments and Results

The goal of the experimental work in this paper is to design a data-driven model enabling the prediction of the imminent onset of MRD for the ICU patients. Since we have information about the antibiotics and the use of mechanical ventilation along with their stay in ICU, we propose to design a LSTM network by considering MTS. Taking into account that the number of MDR and non-MDR patients are imbalanced, we consider different strategies to train the model and evaluate its performance. Apart from that, we also analyze the knowledge provided by mechanical ventilation when considered in conjunction with the antibiotics.

4.1 Experimental Setup

The methodology performed to train and evaluate the classifiers is the following. First, two-thirds of samples (patients) were randomly assigned to the training set, whereas the rest were assigned to the test set. A 5-fold validation strategy was carried out just on the training set to find the most appropriate value for different hyperparameters. In particular, for LR we chose the penalty coefficient $C \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\%\}$, whereas for voting k-nn we found the number of neighbours k. Regarding hyperparameters associated to the

LSTM network architecture, we considered the number of neurons in the hidden layer (with {5, 7, 10} as explored values) and the dropout rate ({0.2, 0.35, 0.5%}), used as a regularization technique to reduce overfitting on the training data. Regarding hyperparameters associated to the LSTM network trained according to the BBCE function, the considered β = {0.5, 0.66, 0.75, 0.83 0.875, 0.91}. For the LSTM networks optimizing the FL function, different values of β {0.2, 0.4, 0.6, 0.8} and γ {0.25, 0.50, 1, 2} were considered. To provide statistics about the model performance, we repeat the training/test partitioning process 20 times. Specifically, the mean and standard deviation of the following figures of merit in the test set was provided: accuracy, specificity, sensitivity and F1-score.

4.2 Prediction Results

We present in Table 1 the mean and the standard deviation of the performance provided by the LSTM networks in terms of Accuracy, Specificity, Sensitivity and F1-score. Several conclusions can be drawn. On the one hand, when considering only the use of antibiotics in the ICU in a 7-day window, the use of the FL cost function as a strategy to handle imbalanced data sets provides the highest averaged performance in terms of accuracy (66.83), specificity (68.67) and F1-score (37.99). The LSTM network used to obtain these results was trained using 5 neurons (with tanh activation function) in the hidden layer, dropout of 0.2, β=0.8, and γ=1. It is important to remark that the highest values of sensitivity are obtained when applying the BBCE function. On the other hand, when analyzing both the use of antibiotics and mechanical ventilation jointly, the predictive results improved with respect to the use of MTS just related antibiotics. In this case, the highest values of Accuracy and Specificity are obtained also when the FL function is evaluated (84.34 and 94.05, respectively). The best sensitivity value (62.33) is provided when undersampling the training data set.

Respect to the LSTM network on the hand one, when considering only the use of antibiotics in the ICU in a 7-day window, the use of the FL cost function as a strategy to handle imbalanced data sets provides the highest averaged performance in terms of accuracy (66.83), specificity (68.67) and F1-score (37.99). The LSTM network used to obtain these results was trained using 5 neurons (with tanh activation function) in the hidden layer, dropout of 0.2, β=0.8, and γ=1. It is important to remark that the highest values of sensitivity are obtained when applying the BBCE function. On the other hand, when analyzing both the use of antibiotics and mechanical ventilation jointly, the predictive results improved with respect to the use of MTS just related antibiotics. In this case, the highest values of Accuracy and Specificity are obtained also when the FL function is evaluated (84.34 and 94.05, respectively). The best Sensitivity value (62.33) is provided when undersampling the training data set. Training the LSTM with the BBCE function, 10 neurons in the hidden layer, dropout of 0.2, and β=0.66, provides the highest F1-score (50.31). Since F1-score is the harmonic mean of precision and recall, it takes both false positives and false negatives into account. Therefore, the higher the F1-score, the better the classification results.

To show the potential of LSTM in this scenario, conventional approaches as LR and voting k-nn were also considered with the undersampling approach. We checked that performance provided by LR and k-nn worsen to those obtained with LSTM in terms of balancing between specificity and sensitivity. Specifically, when considering all features we obtained Sensitivity values of 31.51 $\pm$ 5.93 and 30.16 $\pm$ 8.57 for LR and k-nn, respectively; whereas the Specificity values were 96.46 $\pm$ 1.51 and 93.52 $\pm$ 3.72. Performance of conventional approaches even worsen when considering just antibiotics (Sensitivity values, 22.35 $\pm$ 7.32 and 26.48 $\pm$ 9.04 for LR and k-nn, respectively; Specificity values, 96.36 $\pm$ 1.60 and 91.47 $\pm$ 4.98 for LR and k-nn, respectively.

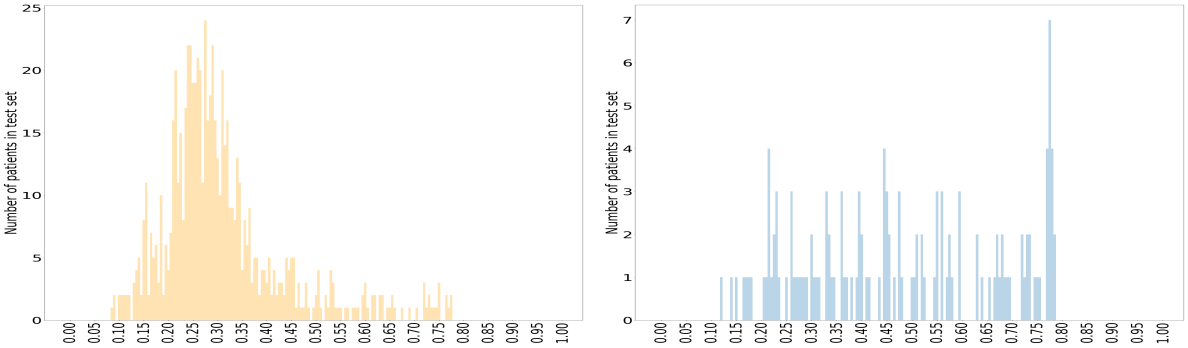
Furthermore, since the sigmoid activation function is used at the output node of the LSTM network, the output can be considered as an estimation of the posterior probability for the positive class. Fig. 4 shows the output distribution for one of the test partitions, both for non-MDR (a) and MDR patients (b). For non-MDR patients, note that the output distribution is clearly shifted towards low probability values, as would be desirable. However, the lower number of MDR patients makes it difficult to characterize the output distribution for these patients. In particular, the mode of the distribution does not seem to be as clearly shifted to the right as would be desirable. A more detailed analysis, considering more MDR patients, could help to identify the reasons for this behaviour and therefore to propose strategies to modify the design of the classifier.

5 Discussion and Conclusions

Multi-drug resistance is a growing and worldwide problem in current societies, directly associated with the use of antimicrobial drugs. The large-scale and sometimes inappropriate use of antibiotics in hospitals in general, and ICUs in particular, has accelerated the emergence of resistant bacteria. This means that bacteria that were

Model	Data source	Training Strat.	Accuracy	Specificity	Sensitivity	F1-score
LSTM	**All features**	Original data	85.12 ± 1.05	96.99 ± 1.18	29.05 ± 5.67	40.19 ± 5.32
		Undersampling	67.08 ± 4.47	68.05 ± 5.94	**62.33 ± 4.28**	39.93 ± 2.37
		BBCE	84.19 ± 1.21	92.23 ± 1.99	46.15 ± 5.42	**50.31 ± 2.99**
		FL	**84.34** ± 1.48	**94.05 ± 1.53**	38.44 ± 4.26	46.08 ± 3.50
	Antibiotics	Original data	82.23 ± 1.08	99.21 ± 0.56	2.26 ± 1.07	4.21 ± 1.92
		Undersampling	57.40 ± 5.91	56.56 ± 10.88	62.04 ± 18.10	32.69 ± 5.62
		BBCE	61.07 ± 2.37	59.90 ± 3.64	**66.89 ± 5.15**	37.45 ± 1.98
		FL	**66.83 ± 2.41**	**68.67 ± 3.85**	58.36 ± 5.25	**37.99 ± 2.00**

Table 1: Mean ± standard deviation performance on 20 test sets in terms of Accuracy, Specificity, Sensitivity and F1-score when training LSTM with different data sources (second column) and training strategies (third column). Highest performance for each data source is in bold.



(a) (b)

Figure 4: Histogram of the posterior probabilities in the test set for: (a) non-MDR patients; (b) MDR patients.

sensitive to different antibiotics become resistant and, as a result, the length of stay, morbidity, mortality rates and healthcare costs may increase considerably. It has been estimated that antibiotic resistance would affect to 10 million people by 2050, with a cost up to 100 trillion USD [oAR16].

These alarming figures have resulted in big efforts against bacterial pathogens. Recently, several studies have analyzed the DNA-sequencing using machine learning techniques to predict antimicrobial resistance [NLM+19]. Although the whole genome sequencing provides the potential to predict antimicrobial susceptibility, it is not available in all hospitals nowadays. To overcome this drawback, we analyze MTS recorded at the EHR of the ICU, specifically, we focus on the use of antibiotics and mechanical ventilation.

Leveraging EHR data, the goal of this paper is to find underlying and complex relationships between the use of antibiotics, mechanical ventilation and MDR by using LSTM networks. We evaluate the performance when different training strategies are considered to deal with imbalanced classes. The models presented in this work could speed up the workflow of the ICU, helping to identify and isolate patients potentially at risk of MDR.

This study was conducted using only two sources of MTS. To generalize the obtained conclusions, different MTS should be included, as well as clinical and demographic data such as age, gender or diagnoses. As future research, we plan to explore interpretable methods to find the risk factors associated to multi-drug resistance.

5.0.1 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 under the research project TEC2016-75361-R, 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

[CBA15] K. Caballero Barajas and R. Akella. Dynamically modeling patient's health state from electronic medical records: A time series approach. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 69–78, 2015.

[CPC+18] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

[DGK02] M. Klein D. G. Kleinbaum. *Logistic regression*. Springer, 2002.

[Gra12] A. Graves. *Supervised sequence labelling with recurrent neural networks*. Springer, 2012.

[HS97a] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 1997.

[HS97b] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[LGG+17] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[LKEW15] Z. Lipton, D. Kale, C. Elkan, and R. Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[MDHL14] C.A. Michael, D. Dominey-Howes, and M. Labbate. The antimicrobial resistance crisis: causes, consequences, and management. *Frontiers in public health*, 2:145, 2014.

[NLM+19] M. Nguyen, S. W. Long, P. F. McDermott, R. Olsen, Rick L. Olson, R. S., G. H. Tyson, S. Zhao, and J. J. Davis. Using machine learning to predict antimicrobial mics and associated genomic features for nontyphoidal salmonella. *Journal of clinical microbiology*, 57(2):e01260–18, 2019.

[NTV17] P. Nguyen, T. Tran, and S. Venkatesh. Deep learning to attend to risk in ICU. *CoRR*, abs/1707.05010, 2017.

[oAR16] Review on Antimicrobial Resistance. *Tackling drug-resistant infections globally: final report and recommendations*. Review on antimicrobial resistance, 2016.

[Ouy16] Ming Ouyang. Knn in the jaccard space. In *2016 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–7. IEEE, 2016.

[SSN12] M. Sundermeyer, R. Schlüter, and H. Ney. Lstm neural networks for language modeling. *Interspeech 2012*, 2012.

[Tay08] S. J. Taylor. *Modelling financial time series*. world scientific, 2008.

[Tik63] Alexander N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. 1963.

[WLW+16] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy. Training deep neural networks on imbalanced data sets. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 4368–4374, 2016.

[WYL18] J. Wu, L. Yao, and B. Liu. An overview on feature-based classification algorithms for multivariate time series. In *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pages 32–38. IEEE, 2018.

[YLH15] Y. Bengio Y. LeCun and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[YPL+17] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. 2017.