Machine Learning for clinical course analysis in septic patients

Pedro Miguel Ferreira dos Santos
pedro.m.f.santos@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

November 2021

Abstract

According to its latest definition, sepsis is an organic dysfunction caused by a dysregulated host response to infection and if not well managed can further develop into septic shock. These two conditions are one of the leading causes mortality in ICU and their early detection is one of the factors that influence patients' outcome [18]. Therefore, the aim of this work is to predict septic shock onset in ICU. To do so, supervised and unsupervised machine learning techniques were developed and tested with data from Hospital São Francisco Xavier and MIMIC-III. For the unsupervised approach, following an anomaly detection framework, three VAEs were developed and the identification of shock patients was performed using clustering algorithms and anomaly scores. The GMM algorithm was the better clustering algorithm achieving an AUC value of 0.7686 for Hospital São Francisco dataset and 0.9576 for MIMIC-III dataset. The density-based anomaly score applied to the encoded data in latent space outperformed every anomaly score tested, achieving an AUC value of 0.8292 for Hospital São Francisco dataset and 0.9498 for MIMIC-III dataset. These results are competitive with the AUC values of 0.8784 and 0.9988 obtained with supervised models and data from Hospital São Francisco Xavier and MIMIC-III datasets, respectively. The benefits of incorporating information regarding the time elapsed between successive observations was also evaluated with the use of T-LSTM layers, however no significant improvements were observed.

Keywords: Septic shock; Supervised ML; Unsupervised ML; Anomaly scores; T-LSTM

1. Introduction

Throughout the years, the definition of sepsis has been improved and updated and, according to the its latest definition, sepsis is defined as "a serious, potentially fatal, organic dysfunction caused by a dysregulated host response to infection" [18]. Furthermore, septic shock designates "a subset of septic patients in which underlying circulatory and cellular/metabolic abnormalities are sufficiently profound to substantially increase mortality" [18]. This two conditions are considered one of the leading causes of mortality in intensive care units. In fact, according to the CDC, in a typical year, 1.7 million adults in America are affected by sepsis which causes the death of 250,000 individuals [4, 9].

Currently there is no pharmacological treatment for sepsis [19]. Guidelines proposed by SSC recommend fluid resuscitation, administration of vasopressors, measurement of serum lactate as an illness severity marker, acquisition of blood cultures and administration of broad-spectrum antibiotics within the first hour [19]. However, the identification of the infection along with appropriate antimicrobial treatment remains the priority in sepsis management. In fact, studies have reported an increase of 4-7% in the relative risk of mortality for every hour of delayed antibiotic initiation [5, 3, 12, 17].

In this work, supervised and unsupervised machine learning techniques were developed to help predict septic shock onset in ICU. In an unsupervised approach, three different VAEs were developed using data from Hospital São Francisco Xavier and MIMIC-III and an anomaly detection framework was applied. To identify shock patients, both the use of clustering algorithms and anomaly scores were explored and their performance was compared with three supervised LSTM classifiers developed. Furthermore, the benefits on incorporating information regarding the time elapsed between successive observations was also explored through the use of T-LSTM layers.

Hence, the contributions of this work are: 1) the development of machine learning techniques for septic shock prediction using portuguese data, 2) the prediction of septic shock through the use of anomaly scores and 3) the evaluation of the use of T-LSTM models in the prediction of septic shock.

2. Related Work

The integration of machine learning approaches in medicine is not new. Several studies have been

published regarding the use of machine learning to improve diagnosis, treatment and outcomes of different illnesses such as, Alzheimer [6, 8], Parkinson [2, 15], Psychosis [11, 1], and more recently, COVID-19 [14, 10]. In these studies, different techniques were employed, both supervised and unsupervised, and the success of each approach depends not only on the complexity of the disease in study but also on the model used.

Regarding the prediction of sepsis or septic shock, since the models resort to the patients' Electronic Health Records (EHRs), which are a type of sequential data, most models developed are RNN-based [7]. In recent years, new and innovative machine learning approaches for sepsis and septic shock prediction that go beyond the use of RNNs have been developed. One of such examples is the model proposed by Lin et al. in [13], where the model was composed not only by LSTM layers but also a Convolutional Neural Network (CNN) and a Fully Connected Neural Network (FC) with the goal to predict septic shock. CNN was introduced before LSTM with the intention to extract local and time-invariant characteristics from EHR, while the fully connected network was implemented to deal with static data. Two different methods to incorporate static data were explored, one where the static data is incorporated in every step of the LSTM and one where the static data is only incorporated in the last timestep of the LSTM. The first method was denoted Static-repeat, while the other was called Static-last.

This study achieved a highest AUC value of 0.9408 with the model LSTM+CNN+Static-last and demonstrated that models using LSTM can be further improved by incorporating other techniques. Supervised models such as the ones developed in this work require labeled data, which is hard to obtain, specially in the medicine field where most data is unlabeled. This limitation can be overcome with the use of unsupervised models.

In [16], Ramos used VAE models and clustering algorithms to predict septic shock onset in a fully unsupervised approach by applying an anomaly detection framework along with clustering algorithms and using data from MIMIC-III. The clustering algorithms tested were K-means, Spectral Clustering and Gaussian Mixture Models (GMM) and the latter registered the best performance with an AUC of 0.8184, achieving a competitive performance when compared with a supervised LSTM model used as baseline. Furthermore, Ramos also verified that by applying an anomaly detection framework, the VAE reconstruction error was higher for shock patients than for non shock patients.

These results mean that the reconstruction error might be a valid anomaly score to use for the prediction of septic shock. More anomaly scores were proposed by Vasilev et al. in [20]. The authors divided the anomaly scores in reconstruction-based, distance-based and density-based scores. When used for the detection of lesions in MRI scans, all these anomaly scores demonstrated a good performance, achieving AUC values above 0.85. These results are very encouraging and the application of these scores for septic shock prediction should be explored.

Most models previously mentioned are LSTM-based. Despite being able to handle sequential data, LSTM layers assume regular time intervals between observations and thus cannot handle sequences with variations in the time elapsed between successive observations. The time intervals between observations in EHRs, either between hospital visits or even between events, such as clinical tests and interventions in one visit, are usually very irregular and these intervals may contain important information. With this in mind, Baytas et al. [2] developed a time-aware LSTM (T-LSTM) layer which incorporates the information regarding the time intervals in the LSTM structure. First, the LSTM memory is decomposed in short term and long term memories. Then the short term memory in each timestep is adjusted according to the time elapsed between observations, which is converted into appropriate weights with the help of a non-increasing function. The mechanism behind T-LSTM layers is expressed by the following equations:

$$i_t = sigmoid\left(W_i \cdot [h_{t-1}, X_t] + b_i\right) \tag{1}$$

$$f_t = sigmoid\left(W_f \cdot [h_{t-1}, X_t] + b_f\right) \tag{2}$$

$$\widetilde{c} = sigmoid\left(W_c \cdot [h_{t-1}, X_t] + b_c\right) \tag{3}$$

$$o_t = sigmoid\left(W_o \cdot [h_{t-1}, X_t] + b_o\right) \tag{4}$$

$$c^s_{t-1} = tanh\left(W_d \cdot c_{t-1} + b_d\right) \tag{5}$$

$$c^l_{t-1} = c_{t-1} - c^s_{t-1} \tag{6}$$

$$\hat{c}^s_{t-1} = c^s_{t-1} * g\left(\Delta t\right) \tag{7}$$

$$c^*_{t-1} = c^l_{t-1} + \hat{c}^s_{t-1} \tag{8}$$

$$c_t = f_t * c^*_{t-1} + i_t * \widetilde{c} \tag{9}$$

$$h_t = o_t * tanh\left(c_t\right) \tag{10}$$

where $W_{[i,f,c,o]}$ are the weight matrices, $b_{[i,f,c,o]}$ are the bias vectors and $\widetilde{c}$ is the candidate cell state. Furthermore, c_{t-1} and c_t are the previous and current cell states, c^s_{t-1} and c^l_{t-1} are the short term and long term components of the previous memory, $\hat{c}^s_{t-1}$ is the discounted short term memory and c^*_{t-1} is the adjusted previous memory. Besides, W_d and b_d are the weight matrix and bias vector of the decomposition network, respectively. The function $g\left(\cdot\right)$ is a non increasing function applied to the time intervals, Δt.

This T-LSTM model was tested in a target sequence prediction and in a clustering task using data from Parkinson's Progression Markers Initiative (PPMI) dataset in order to predict the target sequence of each patient and identify Parkinson's patients, respectively. In both tasks, the T-LSTM model outperformed the vanilla LSTM. The success of the T-LSTM encourages its application in the prediction of septic shock onset.

3. Data Extraction and Preprocessing

Throughout this project, two different datasets were used. The first dataset was provided by Hospital São Francisco Xavier in Lisbon, Portugal. This dataset includes information regarding age, gender, vital signs, clinical tests, procedures/interventions performed and diagnosis from patients admitted since 2015. From the set of features contained in the dataset, 40 variables were selected. Afterwards, some restrictions were implemented in order to select a population of interest for this study. Therefore, from all patients in the dataset, the cohort of interest only included patients over 18 years with ICU stays longer than 24 hours. Furthermore, only the first ICU stay per patient was included. For patients which developed septic shock, only the data until the first septic shock episode and only the data of patients with septic shock onset after the first 13 hours were included.

The models were first developed and tested using the data from this dataset. In a later stage, the models were also trained and evaluated with data from the MIMIC-III dataset. The same criteria for the definition of the cohort of interest were applied.

For the classification of sepsis, the Sepsis-3 criteria were applied, according to which, sepsis is defined by the presence of an infection suspicion along with a change of 2 or more points in the SOFA score. Regarding the classification of septic shock, a patient was considered to have developed septic shock if it had sepsis along with at least one of the following criteria: 1) Score of 3 or more in the hemodynamic component of the SOFA score; 2) Administration of vasopressors or 3) Lactate values superior to 2 mmol/L.

After extracting data from both datasets, some preprocessing steps were required before feeding this data to the models developed. One of the first steps consisted on the removal of existing outliers, specially in the data from Hospital São Francisco Xavier. With this goal in mind and with the help of an intensive care physician, ranges of acceptable values were defined for each variable. All data outside these ranges were removed and considered missing data.

Since machine learning models cannot deal with missing data, a data imputation method was required. In general, when applying machine learning techniques to health-related data, the forward-filling imputation method is preferred. This method works on the assumption that clinicians in ICU make measurements of certain features when they believe a change in the previous value might have occurred. Therefore, it is safe to assume that at times where there is missing data, no change in that variable has occurred since the last observation. With this in mind, the forward-filling based imputation method proposed in [21] was adopted.

Since the forward filling strategy uses the previous observations to impute missing data, in features where the initial observations are missing this method cannot impute data until an observation is registered. The approach proposed by Wang et all. in [21] tries to overcome this limitation by imputing missing data with the individual-specific mean if there are no previous values or with the global mean in the cases where there is no observations for that variable. Finally all data were normalized with the *MinMaxScaler* tool from *sklearn.preprocessing* python package, which scales data between 0 and 1 by following equation 11.

$$z = \frac{x - min}{(max - min)} \tag{11}$$

After preprocessing, data were split in train, validation and test datasets. This step was performed differently for supervised and unsupervised models. For supervised models, since the classifier needs to learn the patterns of both shock and non-shock patients, data were divided in a stratified fashion using shock as the class label, in order to guarantee that all datasets include patients from both classes. The training, validation and test datasets contained 61.25%, 8.75% and 30% of the total data, respectively. On the other hand, for unsupervised models, since the anomaly detection framework was adopted, the training and validation sets could not include shock patients.

4. Models proposed

4.1. Supervised approach

Three different LSTM model architectures were proposed during the realization of this work. The first model, from here on called *LSTM-All* is composed by a single LSTM layer with 70 units followed by a dense layer with 20 units. For the classification of septic shock, a final dense layer with sigmoid activation is applied. Data is then classified as septic shock if the model output is greater or equal to 0.5 and non-shock if the model output is lesser than 0.5.

In the remaining two models, the variables were first divided in five groups, called *vitals*, *ABG variables*, *clinical tests*, *daily variables* and *static variables*. Then the groups of variables *vitals*, *ABG*

variables, *clinical tests* and *daily variables* are fed to different LSTMs, with 15, 20, 20 and 15 units, respectively. Afterwards, the output of every LSTM layer is concatenated into a single vector, which is then fed to a dense layer with 20 units. The final classification layer is the same as the *LSTM-All* model. The difference between these two models is how they handle the static data.

Following the same approach as in [13], in one of the models, from here on called *LSTM-static-last*, the static data is only incorporated in the last timestep of the LSTMs, i.e. static vector is directly incorporated during the concatenation of the four LSTMs outputs. In the other model, called *LSTM-static-repeat*, the group of static variables is included in the group of daily variables before entering the LSTM layer.

Finally, the benefits of T-LSTM were also analysed by replacing the LSTM layers by T-LSTM layers in every model.

4.2. Unsupervised approach

For the unsupervised approach, three different variational autoencoders were proposed. The difference between them consisted on the encoder used. For each of the VAEs, the encoder was based on one of the models used in the supervised approach, creating the models *VAE-LSTM-All*, *VAE-LSTM-static-last* and *VAE-LSTM-static-repeat*, respectively. Figure 1 demonstrates the encoder structure from model *VAE-LSTM-All*

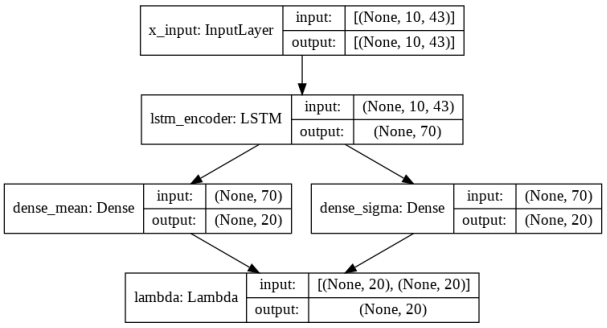

Figure 1: Overview of the encoder from *VAE-LSTM-All*.

For the decoder component, a stochastic approach was followed. To do so, the decoder is composed by a repeat layer followed by a LSTM layer and two dense layers, from which the mean and variance parameters of the data distribution in the feature space were returned. Thereafter, by sampling from this distribution, the reconstruction error can be determined.

For each model, the output from the encoder was visualized with the help of PCA and t-SNE tools, which reduced the encoded data from 20 to 2 dimensions, and the clustering algorithms used in [16] were applied to the encoded data, in order to verify if clusters capable of differentiating shock and non-shock patients were formed.

Furthermore, since a framework based on anomaly detection was followed, the predictive power of some of the anomaly scores proposed in [20] were explored. Since each patient is represented as a Gaussian distribution in the latent space, the distribution of the whole normal population used during the training phase can be estimated as an average of these Gaussians [20], described by

$$q_X(z) = \frac{1}{|X|} \sum_{x \in X} q_\theta\left(z|x\right) \tag{12}$$

The determination of this distribution is fundamental for some of the anomaly scores considered. The following four anomaly scores were analysed:

- Reconstruction error - A higher reconstruction error is expected for shock patients, since the model learnt from data that did not include these type of patients;
- Density-based score - Knowing the distribution of the normal dataset, the probability density function is calculated for the data point sampled from the distribution returned by the VAE;
- Bhattacharyya distance score, which allows to calculate the distance between two distributions. This score calculates the Bhattacharyya distance between the distribution returned from the VAE and the distribution of the normal dataset.
- Mahalanobis distance score, which allows to calculate the distance between a data point and a distribution. This score calculates the Mahalanobis distance between a datapoint sampled from the distribution returned by the VAE and the distribution of the normal dataset.

These scores were determined not only in the latent space but also in the feature space. Afterwards, a threshold was applied to all results in order to separate shock patients from non-shock patients. To choose an appropriate threshold, it was treated as a hyperparameter. This means that several threshold values were tested with part of the test data and then evaluated on the remaining data from the test dataset. Once again, all these results were compared with models using T-LSTM instead of LSTM.

5. Results

5.1. Supervised Approach

As mentioned in the previous section, the dynamic features used in this work can be divided into four

different groups, called *vitals*, *ABG variables*, *Clinical tests* and *Daily variables*. In order to understand the importance of each group of variables and using data from Hospital São Francisco Xavier, a set of experiments was performed to compare how the performance of the model is affected when only one of the groups of variables is included or excluded from the data used as input. The results can be observed in Table 1.

Table 1: Performance of *LSTM-All* model with different inputs using data from Hospital São Francisco Xavier.

Input Data	F1-score	AUC
All data	**0,7207**	**0,8349**
Only *Vitals*	0,5778	0,7207
Only *ABG variables*	0,6531	0,7892
Only *Clinical tests*	0,6552	0,7647
Only *Daily variables*	0,5445	0,6941
All data excluding *vitals*	0,7149	0,8337
All data excluding *ABG variables*	0,6734	0,7977
All data excluding *clinical tests*	0,6762	0,8132
All data excluding *daily variables*	0,585	0,7342

According to Table 1, the worst performance was registered when the model's input included only the group *daily variables*, reaching an AUC value of 0.6941. On the other hand, when only the *daily variables* group was excluded, the model performance suffered the most impact. These results demonstrate that the *daily variables* group contain features fundamental for the prediction of septic shock, despite not being capable to identify this condition by themselves.

Since the different variable groups showed different importance for the prediction of septic shock, the models *LSTM-static-repeat* and *LSTM-static-last* were proposed, in which an independent LSTM for each variable group is included.

The models *LSTM-static-repeat* and *LSTM-static-last* outperformed the model *LSTM-All* in all evaluation metrics used (Table 2), which signifies that dividing the variables into different groups and using an independent LSTM for each group improves the classifier performance. Furthermore, similarly to the results obtained in [13], the model *LSTM-static-last* outperformed *LSTM-static-repeat*.

One limitation of the dataset provided by Hospital São Francisco Xavier, is its reduced population size. To overcome this problem, the three models were trained and tested using data from MIMIC-III database (Table 2). All three models showed significant improvements in their performance, achieving values above 0.9 in all evaluation metrics. Once again, the model *LSTM-static-last* outperformed the remaining models.

Table 2: Performance of models in both datasets. HSFX ≡Hospital São Francisco Xavier.

Dataset	Model	F1-score	AUC
HSFX	LSTM-All	0,7207	0,8349
	LSTM-static-repeat	0,7454	0,8461
	LSTM-static-last	**0,7987**	**0,8784**
MIMIC III	LSTM-All	0,9611	0,9209
	LSTM-static-repeat	0,9862	0,9929
	LSTM-static-last	**0,9874**	**0,9988**

After the previous experiments, all LSTM layers of the three models were replaced by T-LSTM layers, in order to verify whether the incorporation of the information regarding time intervals between observations can improve the performance of the classifiers. The results obtained using T-LSTM models are presented in Table 3.

Table 3: Performance of T-LSTM models in both datasets.

Dataset	Model	F1-score	AUC
HSFX	TLSTM-All	0,7371	0,8487
	TLSTM-static-repeat	0,7202	0,8204
	TLSTM-static-last	**0,7486**	**0,8298**
MIMIC III	TLSTM-All	0,9606	0,9196
	TLSTM-static-repeat	0,975	0,991
	TLSTM-static-last	**0,9893**	**0,9984**

As it can be observed, the incorporation of T-LSTM layers in the models did not bring any significant improvements to the models. These results contradict the findings of Baytas et al. in [2]. One possible justification for this contradiction is the difference in the task for which the models are developed and the data used. In [2], the aim was to identify Parkinson's patients, which is a chronic condition with a slow progress. Therefore, the data used for this task consisted of sequences of patients' hospital visits where the time between each visit can vary from months to years. On the other hand, in this work, the goal was to predict a fast progression condition like septic shock and the data used consisted on sequences of events during a single ICU stay where time intervals vary between a few hours to a few days.

5.2. Unsupervised Approach

As mentioned in the previous chapter, for the unsupervised approach three different VAEs were developed. The encoder of each of the VAEs developed has a similar structure to the supervised models

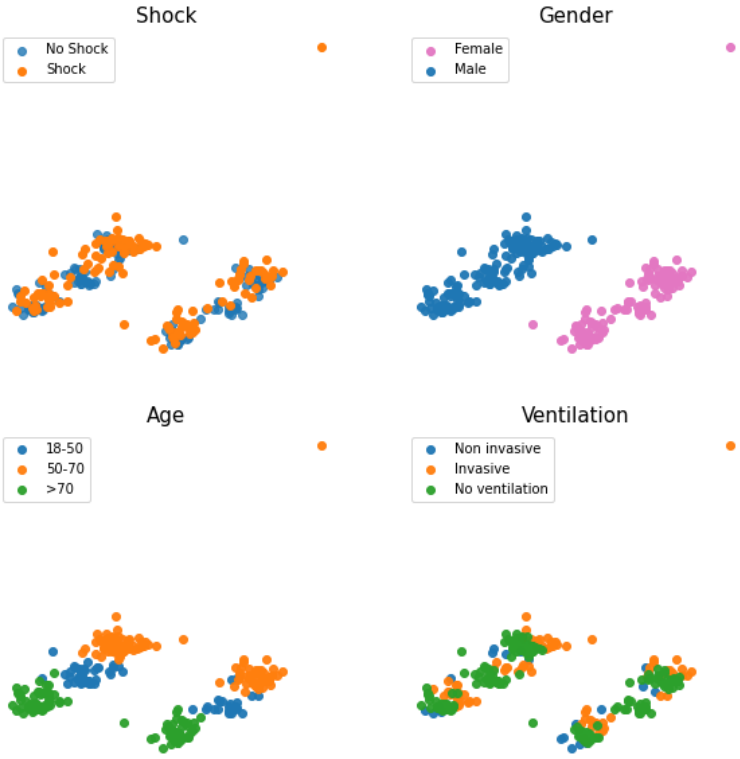

Figure 2: Visualization via PCA of data from Hospital São Francisco Xavier encoded by *VAE-LSTM-All* along with the identification of gender, age and ventilation categories.

proposed and it encodes the original data to distributions in the latent space, which has a size of 20 dimensions. One of the hyperparameters that had to be adjusted was the β weight of the VAE loss function.

For high β values, the data is expected to be concentrated around the origin, since the Kullback-Leibler divergence term dominates the loss function. As the value of this hyperparameter starts to decrease, the encoded data should start to spread out and, since the model is not trained with shock patients, the formation of two clusters would be expected, one for non-shock patients around the origin and one for shock patients. However, according to the results in Figure 2, although some clustering of data can be observed, these clusters do not represent shock and non-shock patients.

Figure 2 represents the results obtained using the model *VAE-LSTM-All*, however the results from the remaining two VAEs are very similar. Observing the results obtained, the patients' age and gender are the variables according to which the clusters are formed. Since the effects of age and gender were so dominant, a new set of experiments was performed in which the data used as input did not include gender and age variables. Since the static variables were not included then the models *VAE-LSTM-static-last* and *VAE-LSTM-static-repeat* were replaced by a new VAE, from here on called *VAE-LSTM-grouped*, in which the dynamic features are also grouped in their four categories (vitals, ABG variables, clinical tests and daily variables) but no static data is included. The PCA of the encoded data obtained in this experiment can be observed in Figure 3.

As it can be observed, although there are regions with higher concentrations of shock patients, there are no clear and well distinguished clusters of patients. However, PCA is only a tool to help in the visualisation of the encoded data in a 2D space. Therefore, despite no clear clusters have been found in the PCA, it might be possible to distinguish shock from non-shock patients resorting to clustering algorithms applied directly to the encoded data.

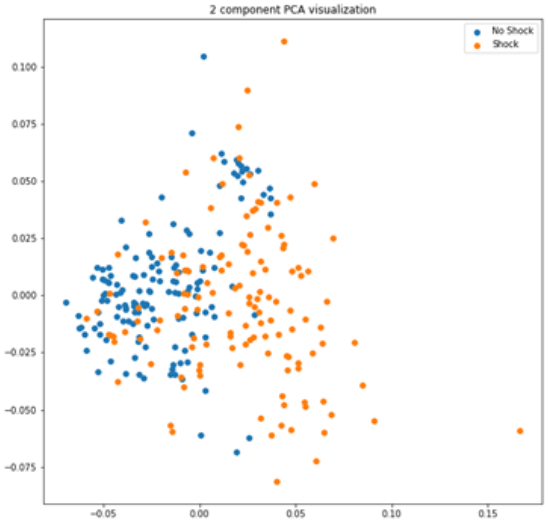

Figure 3: Visualization data without static features from Hospital São Francisco Xavier encoded by *VAE-LSTM-All* via PCA.

With this in mind, the same three algorithms used in [16], were applied to the encoded data (Table 4). Similarly to the results obtained in [16], the GMM was the better clustering algorithm, however the performance of all clustering algorithms was subpar when compared with the results reported by Ramos.

The disparity in performance might be due to the limiting size of the dataset provided by Hospital São Francisco Xavier. Therefore, the same experiments were performed using data from MIMIC-III. According to Figure 4, two clear clusters which represent shock and non-shock patients can be defined. Regarding the performance of the clustering algorithms, once again the GMM algorithm outperformed the remaining ones, reaching values above 0.9 across all evaluation metrics. However, the other two clustering algorithms showed a decrease on the model performance.

Table 5: Performance of clustering algorithms using T-LSTM models and data from Hospital São Francisco Xavier.

Model	**Clustering algorithm**	**F1-score**	**AUC**
VAE-TLSTM All	K-means	0,616	0,6496
	Spectral Clustering	0,6129	0,6493
	GMM	0,6507	0,7276
VAE--TLSTM grouped	K-means	0,7297	0,7765
	Spectral Clustering	**0,733**	**0,78**
	GMM	0,5434	0,6844

Table 4: Performance of clustering algorithms using data from Hospital São Francisco Xavier and MIMIC-III.

Dataset	**Model**	**Clustering algorithm**	**F1-score**	**AUC**
HSFX	VAE-LSTM All	K-means	0,5842	0,6759
		Spectral Clustering	0,5935	0,6786
		GMM	0,6931	0,7538
	VAE-LSTM grouped	K-means	0,6068	0,6652
		Spectral Clustering	0,6006	0,6605
		GMM	**0,7253**	**0,7686**
MIMIC III	VAE-LSTM All	K-means	0,5068	0,67
		Spectral Clustering	0,5124	0,6726
		GMM	**0,923**	**0,9576**
	VAE-LSTM grouped	K-means	0,5287	0,6803
		Spectral Clustering	0,53	0,6809
		GMM	0,802	0,8554

These models were then compared with T-LSTM models and, as in the supervised approach, no significant improvements were registered (Table 5). In fact, in some clustering algorithms the performance decreased considerably.

Besides clustering algorithms, the performance of anomaly scores was also explored. By defining an appropriate threshold for each anomaly score, the distinction between shock and non-shock patients can be established reasonably well. The results obtained can be observed in Table 6.

As it can be observed in Table 6, the distance-based scores outperformed the others and all anomaly scores performed better in the latent space than in the feature space. This result was expected since VAE imposes a restriction in the latent space, not in the feature space. Although these anomaly scores could not outperform the clustering algorithms when using data from MIMIC-III, they showed a more consistent performance overall, achieving a relatively good performance even with the small dataset from Hospital São Francisco Xavier. One downside of this approach is that, for the decision of the threshold value, data labels were required and therefore this approach is not fully unsupervised, unlike clustering.

Regarding the use of T-LSTM layers, once again no significant improvement was observed.

6. Conclusions

6.1. Main conclusions

In this work, supervised and unsupervised machine learning approaches for the prediction of septic shock were explored. These models were

Table 6: AUC values of anomaly scores using data from Hospital São Francisco Xavier and MIMIC-III.

Dataset	Model	Space	Error	Density score	Bhattacharyya	Mahalanobis
Hospital São Francisco Xavier	VAE-LSTM-All	Latent	-	**0,8292**	0,8258	0,8222
		Feature	0,6791	0,779	0,7632	0,7845
	VAE-LSTM-grouped	Latent	-	0,7683	0,7782	0,7803
		Feature	0,7657	0,6037	0,6241	0,7644
MIMIC-III	VAE-LSTM-All	Latent	-	**0,9498**	0,9454	0,9364
		Feature	0,6877	0,8929	0,91	0,8995
	VAE-LSTM-grouped	Latent	-	0,9179	0,9262	0,9223
		Feature	0,6721	0,8218	0,8181	0,7984

Table 7: AUC values of anomaly scores using T-LSTM models and data from Hospital São Francisco Xavier.

Model	Space	Error	Density score	Bhattacharyya	Mahalanobis
VAE-LSTM-All	Latent	-	**0,7926**	0,7874	0,7917
	Feature	0,766	0,756	0,7675	0,7271
VAE-LSTM-grouped	Latent	-	0,7623	**0,7945**	0,7875
	Feature	0,7423	0,5987	0,6169	0,6047

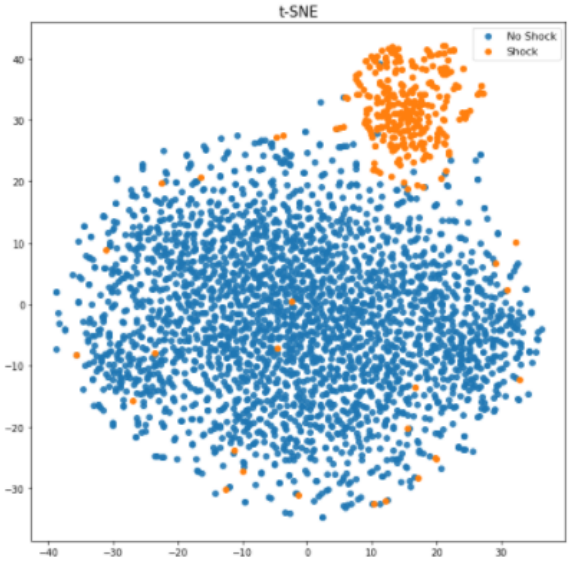

Figure 4: Visualization data without static features from MIMIC-III encoded by *VAE-LSTM-All* via t-SNE.

then trained and tested with data from two different datasets: Hospital São Francisco Xavier and MIMIC-III. For the supervised approach, three different classifiers were proposed. In both datasets considered, *LSTM-static-last* was the best model, reaching an AUC value of 0.8784, in Hospital São Francisco Xavier dataset, and 0.9968, in MIMIC-III dataset. Moreover, an importance analysis of the variable groups was also conducted which revealed that the group *daily variables* include fundamental features for the prediction of septic shock, while the features included in the group *vitals* are too broad and not specific enough for the prediction of this condition.

In the unsupervised approach, three different models were also developed. All three models are variational autoencoders trained only with non-shock patients. The encoded data was then clustered with the help of three clustering algorithms, K-means, Spectral Clustering and GMM, and the latter performed the best. In this approach, the reduced size of the Hospital São Francisco Xavier dataset had a significant impact in the clustering task. In fact, the best algorithm only reached an AUC value of 0.7686 with data from Hospital São Francisco Xavier but using data from MIMIC-III the best model achieved an AUC value of 0.9576.

Besides clustering algorithms, the prediction of septic shock through the use of anomaly scores was also evaluated. The anomaly scores considered were VAE reconstruction error, Density-based score, Bhattacharyya distance score and Mahalanobis distance score, however none of them could outperform the GMM clustering algorithm when using MIMIC-III data. However, when the models were trained and tested with data from Hospital São Francisco Xavier, their evaluation metrics remained elevated, unlike what happened with clustering algorithms. Therefore, these anomalies scores might be a better criteria to predict sepsit shock since their performance is not impacted by the size of the dataset as much as with clustering algorithms. All these results demonstrated that unsupervised techniques can be a competitive approach to supervised models in the prediction of septic shock.

Finally, models with T-LSTM to account for the irregularity in the time intervals between successive observations in the patients' timeseries were also evaluated. However, no significant improvements could be registered both in supervised and unsu-

pervised approaches.

6.2. Limitations and Future Work

Importance analysis of the groups of variables only indicated the groups which had the greatest impact in the model performance, but it could not identify which variables affect and how they influence the classifier decision. The identification of the variables that influenced the model decision and explanation of why the model classified data the way it did holds great interest, specially in the field of medicine where each choice has a great impact on patients' lives. The use of attention mechanisms might be one of the focus for future investigations with the goal to help enlighten how each variable influences the classifier decision.

In this work, the T-LSTM models registered contradicting results when compared with the results obtained in [2]. As previously mentioned, this contradiction might be due to the different health conditions considered and data used. Therefore, further investigation should be performed regarding the benefits brought by the incorporation of the time intervals information into the LSTM structure.

Another limitation of this work is related with the anomaly scores. The anomaly scores analysed required the use of labeled data for the determination of an appropriate threshold, therefore this method could not be considered fully unsupervised. Since one of the goals in using unsupervised machine learning techniques is to eliminate the need of labeled data, new alternatives to predict septic shock with anomaly scores without the use of labeled data should be explored.

Acknowledgements

This document was written and made publically available as an institutional academic requirement and as a part of the evaluation of the MSc thesis in Biomedical Engineering of the author at Instituto Superior Técnico. The work described herein was performed at Instituto Superior Técnico (Lisbon, Portugal) in collaboration with Hospital São Francisco Xavier, during the period February-October 2021, under the supervision of Prof. Margarida Silveira and Dr. Luís Coelho.

References

[1] S. Amoretti, N. Verdolini, G. Mezquida, F. D. R. da Ponte, M. J. Cuesta, L. Pina-Camacho, M. Gomez-Ramiro, C. D. la Cámara, A. González-Pinto, C. M. Díaz-Caneja, I. Corripio, E. Vieta, E. de la Serna, A. Mané, B. Solé, A. F. Carvalho, M. Serra, and M. Bernardo. Identifying clinical clusters with distinct trajectories in first-episode psychosis through an unsupervised machine learning technique. *European Neuropsychopharmacology*, 47:112–129, 2021.

[2] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou. Patient subtyping via time-aware LSTM networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Part F1296:65–74, 2017.

[3] D. Berg and H. Gerlach. Recent advances in understanding and managing sepsis [version 1; peer review: 3 approved]. *F1000Research*, 7(0):1–8, 2018.

[4] CDC. What is sepsis? `https://www.cdc.gov/sepsis/what-is-sepsis.html#anchor_1547214418`, Updated : 2021-10-07.

[5] M. Cecconi, L. Evans, M. Levy, and A. Rhodes. Sepsis and septic shock. *The Lancet*, 392(10141):75–87, 2018.

[6] S. El-Sappagh, T. Abuhmed, S. M. R. Islam, and K. S. Kwak. Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. *Neurocomputing*, 412:197–215, 2020.

[7] J. Fagerström, M. Bång, D. Wilhelms, and M. S. Chew. LiSep LSTM : A Machine Learning Algorithm for Early Detection of Septic Shock. pages 1–8, 2019.

[8] C. K. Fisher, A. M. Smith, J. R. Walsh, A. J. Simon, C. Edgar, C. R. Jack, D. Holtzman, D. Russell, D. Hill, D. Grosset, F. Wood, H. Vanderstichele, J. Morris, K. Blennow, K. Marek, L. M. Shaw, M. Albert, M. Weiner, N. Fox, P. Aisen, P. E. Cole, R. Petersen, T. Sherer, and W. Kubick. Machine learning for comprehensive forecasting of Alzheimer's disease progression. *Scientific Reports*, 9:1–14, 2019.

[9] J. Hajj, N. Blaine, J. Salavaci, and D. Jacoby. The "centrality of sepsis": A review on incidence, mortality, and cost of care. *Healthcare (Switzerland)*, 6, 2018.

[10] M. M. Khodeir, H. A. Shabana, A. S. Alkhamiss, Z. Rasheed, M. Alsoghair, S. A. Alsagaby, M. I. Khan, N. Fernández, and W. A. Abdulmonem. Early prediction keys for COVID-19 cases progression: A meta-analysis. *Journal of Infection and Public Health*, 14:561–569, 2021.

[11] N. Koutsouleris, L. Kambeitz-Ilankovic, S. Ruhrmann, M. Rosen, A. Ruef, D. B.

Dwyer, M. Paolini, K. Chisholm, J. Kambeitz, T. Haidl, A. Schmidt, J. Gillam, F. Schultze-Lutter, P. Falkai, M. Reiser, A. Riecher-Rössler, R. Upthegrove, J. Hietala, R. K. Salokangas, C. Pantelis, E. Meisenzahl, S. J. Wood, D. Beque, P. Brambilla, and S. Borgwardt. Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: A multimodal, multisite machine learning analysis. *JAMA Psychiatry*, 75:1156–1172, 2018.

[12] A. Kumar, D. Roberts, K. E. Wood, B. Light, J. E. Parrillo, S. Sharma, R. Suppes, D. Feinstein, S. Zanotti, L. Taiberg, D. Gurka, A. Kumar, and M. Cheang. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical Care Medicine*, 34, 2006.

[13] C. Lin, Y. Zhangy, J. Ivy, M. Capan, R. Arnold, J. M. Huddleston, and M. Chi. Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-LSTM. *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, pages 219–228, 2018.

[14] M. S. Mottaqi, F. Mohammadipanah, and H. Sajedi. Contribution of machine learning approaches in response to SARS-CoV-2 infection. *Informatics in Medicine Unlocked*, 23:100526, 2021.

[15] M. Nilashi, O. Ibrahim, H. Ahmadi, L. Shahmoradi, and M. Farahmand. A hybrid intelligent system for the prediction of Parkinson's disease progression using machine learning techniques. *Biocybernetics and Biomedical Engineering*, 38:1–15, 2018.

[16] G. B. Ramos. Unsupervised learning approach for understanding critical infectious disease progression in ICU patients. Master's thesis, Instituto Superior Técnico, January 2021.

[17] M. Singer. Antibiotics for sepsis: Does each hour really count, or is it incestuous amplification? *American Journal of Respiratory and Critical Care Medicine*, 196, 2017.

[18] M. Singer, C. S. Deutschman, C. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J. D. Chiche, C. M. Coopersmith, R. S. Hotchkiss, M. M. Levy, J. C. Marshall, G. S. Martin, S. M. Opal, G. D. Rubenfeld, T. D. Poll, J. L. Vincent, and D. C. Angus. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA - Journal of the American Medical Association*, 315(8):801–810, 2016.

[19] K. Thompson, B. Venkatesh, and S. Finfer. Sepsis and septic shock: current approaches to management: Sepsis and septic shock. *Internal Medicine Journal*, 49:160–170, 02 2019.

[20] A. Vasilev, V. Golkov, M. Meissner, I. Lipp, E. Sgarlata, V. Tomassini, D. K. Jones, and D. Cremers. q-Space Novelty Detection with Variational Autoencoders. pages 1–11.

[21] S. Wang, M. B. A. Mcdermott, M. C. Hughes, and T. Naumann. MIMIC-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III. 2020.