Machine Learning for clinical course analysis in septic patients

Pedro Miguel Ferreira dos Santos

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Prof. Maria Margarida Campos da Silveira
Prof. Luís Miguel da Cruz Coelho

Examination Committee

Chairperson: Prof. João Orlando Marques Gameiro Folgado
Supervisor: Prof. Maria Margarida Campos da Silveira
Member of the Committee: Prof. Susana Margarida da Silva Vieira

November 2021

Declaration
I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Preface

The work presented in this thesis was performed at Instituto Superior Técnico (Lisbon, Portugal) in collaboration with Hospital São Francisco Xavier, during the period February-November 2021, under the supervision of Prof. Margarida Silveira and Dr. Luís Coelho.

Acknowledgments

First of all, I would like to thank my family for all the support, encouragement and strength they gave me throughout these years, specially this last semester, during most of which I was locked at home working.

I would also like to express my gratitude to Professor Margarida Silveira for all the insights, suggestions and feedback throughout this semester. Without the supervision of Professor Margarida Silveira this thesis would not be possible. I'm also grateful to Doctor Luís Coelho who was always available to clarify any clinical related questions that emerged during the development of this work.

Last but not least, to all my friends and colleagues that I could always count on and were always there for me during the good and bad times in my life and enriched my university life. Thank you!

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 [1]. 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.

Resumo

De acordo com a sua definição mais recente, sepsis consiste numa disfunção orgânica provocada por uma resposta desregulada a uma infeção e caso esta resposta não seja controlada, o paciente pode entrar em choque séptico. Estas duas condições são as principais causas de mortalidade em unidades de tratamento intensivo e a sua deteção precoce é um dos principais fatores que influenciam o desfecho dos pacientes [1]. Deste modo, o objetivo deste trabalho consiste na predição de choque séptico em unidades de tratamento intensivo. Com este objetivo em mente, técnicas supervisionadas e não supervisionadas de aprendizagem automática foram desenvolvidas e testadas com dados do Hospital São Francisco Xavier e da MIMIC-III. Numa abordagem não supervisionada, e aplicando um processo de deteção de anomalias, três VAEs foram desenvolvidos e a identificação de choque séptico foi realizada recorrendo a algoritmos de agrupamento e pontuações de anomalias.

O algoritmo GMM foi o melhor algoritmo de agrupamento do qual resultou uma AUC de 0.7686 para os dados do Hospital São Francisco Xavier e 0.9576 para os dados da MIMIC-III. A pontuação de anomalias baseada na densidade probabilística superou o desempenho das restantes pontuações testadas, alcançando uma AUC de 0.8292 para os dados do Hospital São Francisco Xavier e 0.9498 para os dados da MIMIC-III. Estes resultados são competitivos com a AUC de 0.8784 e 0.9988 obtidos com modelos supervisionados, usando esses dois conjuntos de dados. Os benefícios de incorporar informação sobre o intervalo de tempo entre observações sucessivas também foram avaliados através do uso de T-LSTM, no entanto não se observou melhoramentos significativos.

Palavras Chave

Choque séptico; Aprendizagem supervisionada; Aprendizagem não supervisionada; Deteção de anomalias; T-LSTM.

Contents

List of Figures

List of Tables

Acronyms

ABG	Arterial Blood Gas
AUC	Area Under the Receiver Operating Characteristic Curve
BCA	Balanced Class Accuracy
CARS	Compensatory Anti-inflammatory Response Syndrome
CDC	Centers for Disease Control and Prevention
CNN	Convolutional Neural Network
EHRs	Electronic Health Records
ELBO	Evidence Lower Bound
FN	False Negative
FP	False Positive
GMM	Gaussian Mixture Models
GRU	Gated Recurrent Units
HBO	Hours Before Onset
ICU	Intensive care unit
KL	Kullback-Leibler
KPNC	Kaiser Permanente Northern California
LOS	Length of stay
LR	Logistic Regression
LSS	Linear State Space

LSTM	Long-Short Term Memory
mAUC	Multiclass Area Under the Operating Curve
MLP	Multilayer Perceptron
MNIST	Modified National Institute of Standards and Technology
MSE	Mean Square Error
NPV	Negative Predictive Value
PCA	Principal Component Analysis
PICS	Persistent Inflammation-Immunosuppression Catabolism Syndrome
PPMI	Parkinson’s Progression Markers Initiative
PPV	Positive Predictive Value
qSOFA	quick SOFA
RNNs	Recurrent Neural Networks
SCCM	Society of Critical Care Medicine
SIRS	Systemic Inflammatory Response Syndrome
SOFA	Sequential Organ Failure Assessment
SSC	Surviving Sepsis campaign
TN	True Negative
TP	True Positive
TR	Sequential Target Replication
T-LSTM	Time aware LSTM
t-SNE	t-Distributed Stochastic Neighbor Embedding
VAE	Variational Autoencoder
WHO	World Health Organization

1

Introduction

Contents

1.1 Sepsis: Definitions throughout the years

Sepsis is considered as one of the oldest illnesses described in mankind history and it was described by Hippocrates as the processes that turn infected wounds purulent, around 400 BCE. Almost 2000 years later, a new definition was proposed by Hugo Schottmüller in 1914, according to which "Sepsis is present if a focus has developed from which pathogenic bacteria, constantly or periodically, invade the blood stream in such a way that this causes subjective and objective symptoms" [8]. One common point between these two definitions is that they consider that the development from infection to sepsis is caused by the pathogens themselves.

In 1992, the American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference was held. In this conference a new definition of sepsis was created (Sepsis-1), which related the development of sepsis with the immune response of the host and not with pathogens [9,10]. The goal of this conference was to propose new criteria standards for the sake of an early detection and diagnosis of this illness. From this consensus, the concept of Systemic Inflammatory Response Syndrome (SIRS) arose. SIRS corresponds to "an exaggerated defense response of the body to a noxious stressor" and it is defined by four criteria, which are easily accessible in any hospital setting [9, 11]:

- Body temperature over 38 or under 36 degrees Celsius
- Heart rate greater than 90 beats/minute
- Respiratory rate greater than 20 breaths/minute or partial pressure of CO2 less than 32 mmHg
- Leucocyte count greater than 12000 or less than 4000 /microliters or over 10% immature forms or bands

According to the new definition, the diagnosis of sepsis requires a suspicion of infection along with 2 or more of the SIRS criteria [9]. It is important to take into consideration that a patient can present SIRS without an infection and in these cases there shouldn't be a sepsis diagnosis.

Besides this new sepsis definition, new concepts were created: severe sepsis and septic shock, which corresponds to sepsis with acute organ dysfunction and sepsis with hyperlactataemia or hypotension refractory to fluid resuscitation, respectively [12]. These new definitions were kept unchanged even after a new conference held in 2003 [13]. This conference gathered north american and european entities, during which the PIRO model (later known as Sepsis-2) was proposed. This model took into consideration the predisposition, pathogen, host response and organ dysfunction, allowing the stratification of sepsis patients [8, 13]. However, a lack of sufficient evidence to support a change of definitions was reported. Furthermore, a few limitations in the use of the SIRS criteria were highlighted:

- Despite presenting a good sensitivity, SIRS criteria for sepsis has a poor specificity

- SIRS cannot distinguish between a normal and a pathological immune response
- These criteria cannot predict organ dysfunction
- There might be some criteria with a more important role, but the same weight is allotted to each SIRS criteria

In [14], using a sample of over 130000 septic patients, it was reported that one in eight sepsis patients did not fulfilled SIRS criteria. Furthermore, a correlation between SIRS criteria and organ dysfunction or death was not observed. Apart from the limitations already mentioned, these definitions of sepsis, severe sepsis and septic shock lead one to wrongly believe that they correspond to a continuous progress of this illness [1].

To tackle these problems, a task force composed by the European Society of Intensive Care Medicine and the Society of Critical Care Medicine (SCCM) was created which lead to new definitions of sepsis and septic shock [1]. According to this task force, sepsis is defined as “a serious, potentially fatal, organic dysfunction caused by a dysregulated host response to infection”, septic shock designates “a subset of septic patients in which underlying circulatory and cellular/metabolic abnormalities are sufficiently profound to substantially increase mortality” and the term of severe sepsis was removed. This definitions are known as Sepsis-3.

The use of SIRS as a criteria for sepsis was also abandoned and the Sequential Organ Failure Assessment (SOFA) was adopted. SOFA score is composed by six different scores that evaluate the respiratory, cardiovascular, hepatic, coagulation, renal and neurological systems allotting a value from 0 to 4, as the organ dysfunction of the corresponding system worsens. According to [15], it was possible to establish a correlation between patients’ SOFA score and mortality. Thus, this scoring system allows to quantify the number and severity of organs in dysfunction and provides an objective tool of organ dysfunction and death risk stratification.

According to these new definitions, a patient is diagnosed with sepsis if there is a suspicion of infection conjugated with an acute change of 2 or more points on the SOFA score. The presence of sepsis with hypotension requiring vasopressor therapy to maintain a mean arterial blood pressure greater or equal to 65 mmHg and a serum lactate level greater than 2 mmol/L after adequate fluid resuscitation leads to a septic shock diagnosis [1].

Moreover, the task force also reported three useful criteria to identify a higher probability to have a longer stay in hospital and higher risk of mortality in patients with infection, leading to the development of a new scoring system called quick SOFA (qSOFA). The three criteria are:

- Alteration in mental status, indicated by a Glasgow Coma Scale score of 13 or less
- Systolic blood pressure of less than 100 mm Hg,

- Respiratory rate of more than 22 breaths per minute.

This qSOFA score was reported to perform better in identifying high risk patients than the original SIRS criteria [11, 12], providing a simple and relatively fast risk stratification tool that can be used to identify patients at risk of sepsis. However, according to [16] referred in [12], this score should not be used to exclude high risk patients.

Along with the definition of sepsis, the understanding of the mechanisms behind this illness has also been improved along the years. The pathogenesis of sepsis is a complex process where multiple aspects of the interaction between host and pathogen play a roll in this process [17], where there are two main mechanisms involved: inflammatory response and anti-inflammatory response.

In sepsis, at the time of infection, there is an early exaggerated proinflamatory response to an infection, characterized by SIRS. In order to regulate this response, a compensatory anti-inflammatory response syndrome CARS is triggered [9]. If both responses are balanced the infection is under control. However, if there is a predominance of one of the responses serious problems start to arise. While a predominance of the inflammatory response can cause organ dysfunction and death, the predominance of the anti-inflamatory response leads to the persistence of the infection or even the development of new infections [17].

Initially, it was assumed that these two responses were sequential, first occurring an exaggerated inflammatory response that later evolves into a phase of immunosuppression [9,12]. Later, this hypothesis was discarded and replaced by a theory that suggests that both inflammatory and immunosupression phases actually occur simultaneously with one response prevailing over the other, and the intensity of each response depends on several factors related not only to the pathogen but also the host [12,17,18]. After both responses, the patient may recover but can become chronically ill, developing a syndrome called Persistent Inflammation-Immunosuppression Catabolism Syndrome (PICS). This syndrome is defined as ”ongoing inflammation, manageable organ failure, ongoing protein catabolism and poor nutrition leading to cachexia, poor wound healing, and immunosuppression with increased susceptibility to secondary infections” [17, 18].

1.2 Motivation: Sepsis management

Despite the advances made regarding the understanding of sepsis and its mechanisms, one fact still remains. Currently there is no pharmacological treatment for sepsis [19]. In 2002, an initiative called Surviving Sepsis campaign (SSC) was created with the goal of reducing sepsis mortality. This initiative created the SSC Sepsis Bundles, composed by several measurable interventions. Over a period of 5 years, these bundles were tested and a sepsis mortality decrease was reported [8, 20].

These guidelines proposed by SSC have been updated throughout the years and in 2018 the “Hour-1

Bundle”, which recommends 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, was created [19]. However, the identification of the infection along with appropriate antimicrobial treatment remains the priority in sepsis management. Some of these interventions are controversial. According to [21] referred in [8], the early administration of fluids might not be favorable and may even be harmful to some sepsis patients. Moreover, the administration of antibiotics must proceed with care. While an increase of 4-7% in the relative risk of mortality for every hour of delayed antibiotic initiation has been reported, the early administration of antibiotics, if not done appropriately, might lead to an increase of antibiotic resistance and sepsis incidence [8, 12, 22, 23]. These guidelines have been updated in [24] published this year.

Since a delay in the administration of appropriate antibiotics comes with an increase of mortality, the early identification of sepsis along with a stratification of patients regarding the risk of mortality is crucial for an early treatment initiation [12]. Although most common infections that evolve to sepsis are respiratory, abdominal, bloodstream and renal infections, virtually any infecting agent can lead to sepsis [12, 25–27]. Thus, this illness presents a wide variety of signs and symptoms which hinders its diagnosis.

Nowadays studies regarding sepsis focus not only in finding new biomarkers [9, 28, 29] but also in the incorporation of machine learning techniques able to provide support in the diagnosis process. The work performed in this thesis concerns the latter area of research.

1.3 Relevance: Incidence and mortality of sepsis

Sepsis is a syndrome that affects millions of individuals per year and along with septic shock is one of the leading causes of mortality in intensive care units, being recognized by the World Health Organization WHO as a global health priority. Besides, not only a decreased health-related quality of life, substantial cognitive impairment and functional disability is observed in patients who survive this illness, these patients also present an increased risk of death in the year following hospital discharge [8, 19, 30].

Several studies report an increase of incidence of sepsis but a decrease of its mortality over the years [1, 31]. Although the mortality of sepsis has been decreasing, its numbers remain unacceptably high. According to [32] referred in [12], in 2001, more than 750 000 cases of sepsis were reported in the USA. More recently, the Centers for Disease Control and Prevention (CDC) stated that, in a typical year, 1.7 million adults in America are affected by sepsis which causes the death of 250,000 individuals and is responsible for one out of three deaths in hospital [33, 34]. In a review study, performed in 2020 using data from the previous decade, the average 30-day sepsis mortality and the average 30-day septic shock mortality in North America, Europe and Australia was 24.4% and 34.7%, respectively.

These values varied between the three regions. Furthermore, it also reported a statistically significant decrease of 30-day septic shock mortality rate between 2009 and 2011, but not after 2011 [35].

These data represent the incidence and mortality of sepsis in high-resource countries, however this illness has a greater impact in low-resource countries, where the true values of incidence and mortality are difficult to estimate [12]. These high mortality rates in sepsis patients once again reinforce the need of a fast and accurate diagnosis that machine learning techniques might help to provide.

1.4 Organization of the document

This thesis is organized as follows. Chapter 1 is an introduction to the topic presenting the definition of important concepts, the motivation and the relevance of this work. A review on some of the studies performed regarding the incorporation of machine learning techniques in the health sector is performed in Chapter 2. Chapter 3 details the methodology followed for the realization of this work while Chapter 4 includes the results obtained and the discussion of said results. Finally, the main conclusions and limitations of the work along with perspectives for future work is presented in Chapter 5.

2

State of the Art

Contents

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 [3, 36–38], Parkinson [39, 40], Psychosis [41, 42], and more recently, Covid-19 [43–48]. 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.

One of the most important factors that allowed the increasing and successful use of machine learning in medicine was the implementation and adoption of Electronic Health Records (EHRs) [49]. EHRs consist on a longitudinal collection of patients' healthcare data. They comprise observations of different variables, which can either be static, such as age and gender, or dynamic, such as vital signs.

Although EHRs allow an easier data extraction, they also present some disadvantages. First, since these records contain all healthcare data of a patient, EHR data is inherently heterogeneous comprising of different types of features. Besides, this type of data also suffers from sparsity and noise due to several factors, such as, irregular intervals between visits and even tests, misdiagnosis and incomplete or erroneously recording of data [49–51]. Despite all these barriers, machine learning techniques using EHR data have been demonstrated to be a useful tool in predicting, modeling and monitoring different diseases.

2.1 Supervised ML approaches in medicine

Machine learning approaches can be broadly divided into three main categories: supervised, unsupervised or reinforcement learning. In the supervised approach, the model is given a set of data and its corresponding output and the goal is to find a map between them. Generally this approach wields better results than unsupervised methods. For example, [52] and [2] are two studies where this approach was used for a multilabel classification of diagnosis and outperformed the baselines considered for comparison.

In [52], the main goal was to predict the diagnosis and medication order of the next patients visit. In addition, the ability of the model to predict the time to the patients next visit was also examined. To do so, the authors developed a model called Doctor AI, which resorted to a RNN with Gated Recurrent Units (GRU) and to the electronic health records of over 200.000 patients from the Sutter Health Palo Alto Medical Foundation. In order to evaluate the results obtained, two different evaluation metrics were used. For the multilabel classification problem it was used the Top-k recall, which, according to the authors, "mimics the behavior of doctors conducting differential diagnosis, where doctors list most probable diagnoses and treat patients accordingly". This metric is calculated by dividing the number of true positives in the top *k* predictions by the total number of true positives. The coefficient of determination (R^2)

was the metric used for the prediction of the time to the patients' next visit. As baselines for comparison, the authors used frequency baselines, where they use a patient's diagnosis and treatment of the last visit as prediction for the current one, Logistic Regression (LR) and a Multilayer Perceptron (MLP) using information from the last five visits. Four different configurations of the Doctor AI were trained and tested. The model that performed best was a model composed by a RNN with two hidden layers initialized by a embedding matrix with the Skip-gram vectors trained on the entire dataset. This model achieved a top-30 recall of 0.7248 and a R^2 value of 0.2534, outperforming all baselines considered. For example, the MLP model only achieved a top-30 recall of 55.74 and a R^2 value of 0.1221.

Lipton et al. reported similar results in [2]. The goal of this study was to classify 128 diagnoses using thirteen variables extracted from the EHR system at Children's Hospital LA. This dataset originally contains 429 distinct diagnosis but only the most common 128 were considered. To do so, the proposed model used RNNs with Long-Short Term Memory (LSTM) and the results were compared against logistic regression and a MLP as baselines. Moreover, two different techniques to improve the learning task were implemented and tested: Sequential Target Replication (TR) and Auxiliary Output Training. With sequential target training, an output is generated at each sequence step, in order to provide a local error signal (Figure 2.1), while auxiliary output training resorts to the unused 301 labels as auxiliary targets, serving as regularizers.

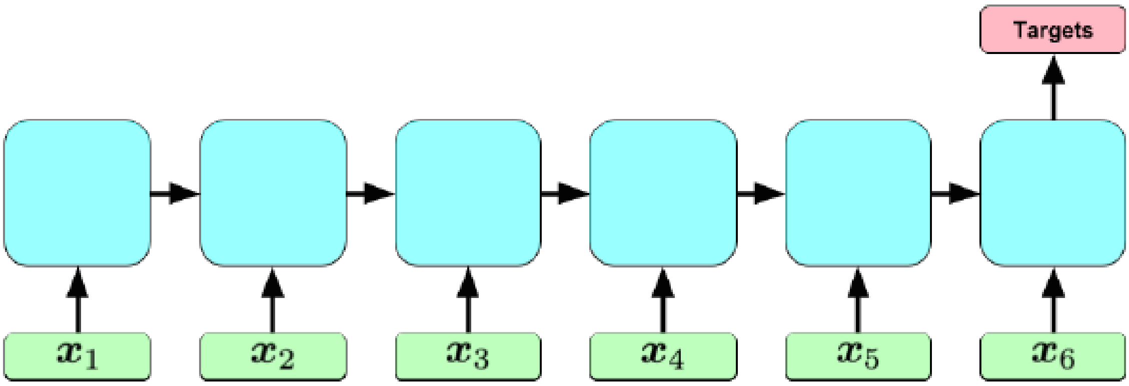


Figure 2.1: Representation of a RNN model with target replication. Source: [2]

Once again, the RNN models outperformed the baselines. Furthermore, among the RNN models, the ones which resorted to target replication and auxiliary output training performed the best, reporting micro and macro AUCs of over 0.84 and 0.78, respectively. Although the auxiliary outputs improved the performance, this improvement came at the cost of slower training.

Both studies demonstrated the superiority of RNNs when it comes to deal with sequential data such as EHR, and its usefulness for predicting future diagnoses. They have also shown that the number of visits and the rarity of the labels influence the performance of the model. However in both studies some limitations were revealed. One of the limitations is related to the noisy and sparse characteristic of electronic records and the irregular intervals between data (visits and medical tests). According to [2], imputation methods may remove useful and important information from clinical time series.

Aware of this limitation, Nguyen et al. included in their study regarding the prediction of Alzheimer's disease progression [3], an analysis of the impact different imputation strategies had on the results. In this study, the authors used an RNN architecture called minimalRNN, developed in [53], along with other models such as LSTM and Linear State Space (LSS) as baselines, and explored three different imputation methods: forward-filling, linear-filling and model filling. With forward-filling, missing data is imputed using the last time point with observed data. On the other hand, in linear filling approaches, data is imputed according to a linear interpolation between the previous and next time points. Finally, model filling strategy, unlike the other two, is an integrative approach. Resorting to the previous time points, the model itself is used to predict the observations of the next time point and then use this prediction to fill in missing data. A representation of these three imputation methods can be observed in Figure 2.2.

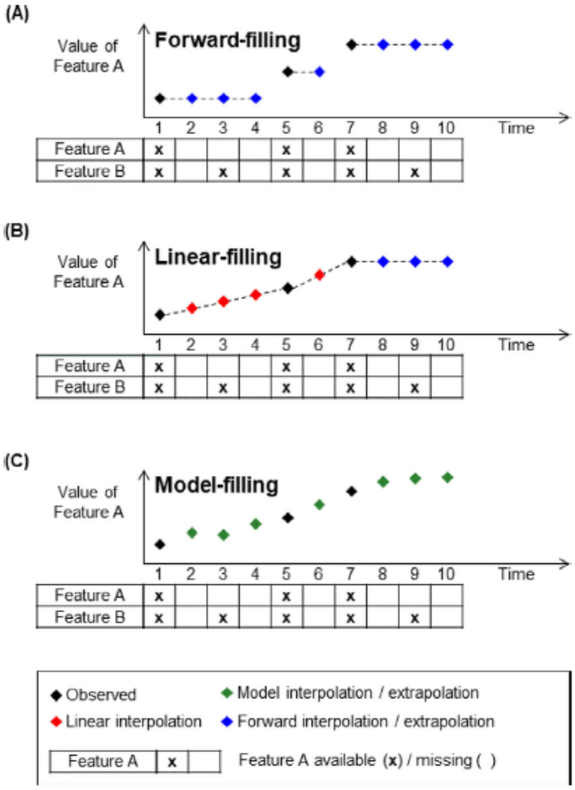


Figure 2.2: Representation of three different imputation methods: (A) Forward-filling; (B) Linear-filling; (C) Model-filling. Source: [3]

In this study, the best model was the minimalRNN with model filling imputation, achieving a Multiclass Area Under the Operating Curve (mAUC) of 0.944 and a Balanced Class Accuracy (BCA) of 0.887, which led the authors to conclude that this imputation approach was better than the remaining two. However, after a careful inspection of the results, the model filling strategy wasn't always the best approach. In fact, both in LSTM and LSS model, the model filling method did not perform better than the forward or the linear filling. Therefore, while model filling demonstrated to be the better imputation strategy when using minimalRNN, it may not be the case if other models of machine learning are used.

As a matter of fact, the success of machine learning in predicting disease progression depends not only on the models and imputation strategy used but also on the disease itself. For instance, although the prediction of Alzheimer and the prediction of sepsis or septic shock are conceptually the same (prediction of disease onset), these are two illnesses with completely different characteristics. Sepsis is characterized by its wide variety of symptoms and its fast progression [12]. For this reason, along the years several models for the early detection of sepsis have been developed, such as "InSight" [54] and "SepLSTM" [55].

In [56], Fagerström et al. developed the LiSep LSTM, which consisted on a LSTM neural network designed for the early identification of septic shock and compared it with five state of the art machine learning algorithms for early detection of sepsis, four of which are based on RNN models and the remaining one resorted to a Cox proportional hazards model. The data used in this study originated from MIMIC-III. MIMIC-III is a clinical database containing information related to patients admitted to the Beth Israel Deaconess Medical Center in Boston between 2001 and 2012 [57]. These data include information regarding vital signs, medications, demographics, laboratory measurements, imaging reports and clinicians notes. For this study, the data used as input include patient biometrics, vital parameters, and laboratory test results. The model was evaluated resorting to the Area Under the Receiver Operating Characteristic Curve (AUC) and Hours Before Onset (HBO), which corresponds to the number of hours by which the model can anticipate the onset of septic shock.

The LiSep LSTM performed worse or on par with the existing state of the art models when considering the AUC metric, achieving a value of 0.83 while the best models achieved 0.93. However, regarding the HBO metric, the LiSep achieved a value of 48h, surpassing the five remaining models by over 20 hours. Besides, according to the results obtained, the predictions are more reliable the closer they are to the onset of septic shock. One of the main differences between the different LSTM models compared in this study concerns to the features considered as input. According to the authors, the significant variation of results between the considered LSTM models shows there may be a set of features better suited for early prediction of septic shock. One limitation of this study lies on the classification of sepsis and septic shock, which was based on the outdated SIRS criteria.

Most models already mentioned are trained with a high number of features. In order to improve real-world practicability, Wernly et al., in [58], restricted the variables used to only features included in Arterial Blood Gas (ABG) tests, since they are globally standardized on ICU and collected at a relatively high frequency and the goal was to predict mortality of septic patients in the first 96 hours after admission.

This model was composed by a single hidden LSTM layer with 140 units and it used ABG values from the first 48 hours in order to predict mortality in the next 48 hours. The model achieved an AUC value of 0.88, a Positive Predictive Value (PPV) of 0.60 and a Negative Predictive Value (NPV) of 0.90.

The setup of this study was meant to mimic the evaluation of a patient after an ICU trial, which usually consists on 48 hours of ICU treatment followed by a re-triage. The high predictive power for mortality reported in this study, demonstrates the usefulness of this model as a support tool for mortality risk stratification in septic patients during re-triage.

In recent years, new and innovative machine learning approaches for sepsis and septic shock prediction have been developed that go beyond the use of RNNs. One of such examples is the model proposed by Lin et al. in [4], where the model was composed not only by LSTM layers but also a Convolutional Neural Network (CNN) and a Fully Connected Neural Network. CNN was introduced before LSTM with the intention to extract local and time-invariant characteristics from EHR. On the other hand, the fully connected network was implemented to deal with static data. Doing so, this model can handle both dynamic and static information in order to predict septic shock.

In this work, two different approaches for septic shock prediction were used: the visit level early diagnosis (also known as left aligned) and the event level early diagnosis (also known as right aligned). In visit level approach, the model uses the first *k* hours after admission to predict whether the patient will develop septic shock anytime during the ICU stay (Figure 2.3 A). On the other hand, the goal of the event level early diagnosis is to determine if a patient will develop septic shock *n* hours later, so the patients are aligned by their end point (Figure 2.3 B). This n-hour window is called the hold-off window.

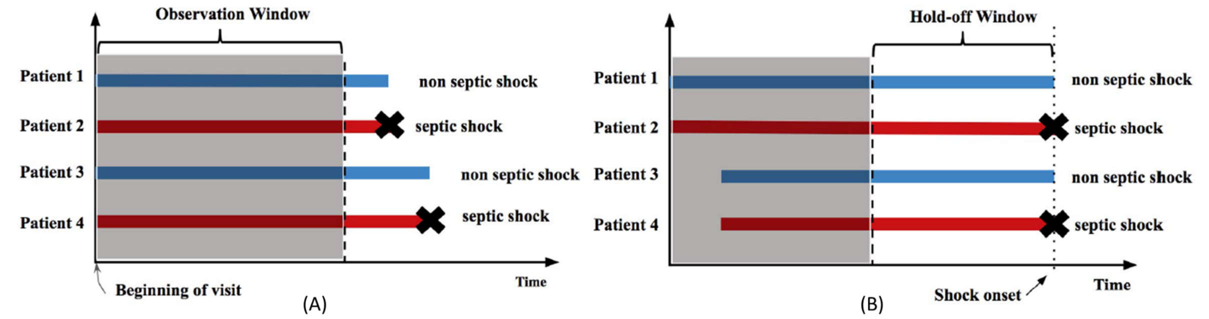


Figure 2.3: Schematic of sequence alignment: (A) Left align (B) Right align. Source: [4]

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.

Moreover, the advantages each additional component (CNN and fully connected network) brings to the model were also analysed. Thus, six different configurations were compared:

- Model using only LSTM (LSTM-Origin)

- Models using LSTM along with fully connected network for static data (LSTM+Static-repeat and LSTM+Static-last)

- Model using LSTM along with CNN (LSTM+CNN)

- Models using LSTM along with CNN and fully connected network for static data (LSTM+CNN+Static-repeat and LSTM+CNN+Static-last)

Regarding the left aligned task, the length of the observation window was varied between 3h to 24h and the LSTM+CNN+Static-last model achieved the best AUC with a value of 0.9408 and the best F1 Score with a value of 0.8579. It was also reported that the longer the observation window, the better the performance of the models, which was expected since the models are provided with more data to learn from. As for the impact of each additional component, the authors observed that while adding both the CNN and the fully connected network to the LSTM brings better results, the model performance does not improve when only one of these components is incorporated in the model. These models also outperformed six classical machine learning techniques used as baselines.

Regarding the right aligned task, the authors varied the hold-off window size between 2h and 24h. Two different behaviours emerged for hold-off windows shorter than 5 hours and hold-off windows longer or equal to 5 hours. When the hold-off window was shorter than 5 hours, LSTM+CNN+Static-last achieved the best AUC (0.8517) and F1 Score (76.74) and incorporating CNN or the static information alone did not help the model. On the other hand, when the hold-off window was longer than 5 hours, LSTM+CNN achieved the best metrics (AUC of 0.7717 and F1 score of 0.6909) and adding CNN alone benefited LSTM greatly, whereas incorporating static information (alone or along with CNN) did not help. These models were compared with the same six baselines used in the left aligned task and once again outperformed them in every metric used.

With this study, the authors not only confirmed the superiority of LSTM compared to other classical machine learning approaches when dealing with sequential data, but also demonstrated that models using LSTM can be further improved by incorporating other techniques. Besides, this work is also very relevant since it was possible to achieve great results using short observation windows, making possible to assess patient risk shortly after being admitted in an ICU.

Despite achieving good results, all models exposed until this point suffer from a common limitation: lack of interpretability. Kaji et al. tried to overcome this problem in [59], where attention mechanisms were incorporated to a LSTM model. In this framework, an attention vector learns weights corresponding to each feature. Then, the input time series are weighted by this learned attention vector before being made available to the LSTM as input, allowing the model to focus on specific features.

This work used data from MIMIC-III and achieved an AUC of 0.952 for predicting sepsis on the same day and an AUC of 0.876 in the task of next-day sepsis prediction. Furthermore, resorting to the learned attention weights, attention maps were constructed, allowing to determine not only the variables that had the most influence but also which time steps were predictive for the sepsis diagnosis. However, according to the authors, although these attention maps can indicate if a certain variable is important, they cannot determine whether or not that variable increases or decreases the probability of sepsis.

Nevertheless, the results from this study demonstrate that attention mechanisms are a promising approach for increasing the interpretability of machine learning models, increasing clinicians' reliance and trust on these models as support tools in the diagnostic process.

One downside of this study is the fact that no comparison was made between model performance with and without attention mechanisms. However, other studies have been published that reported an increase of performance when using attention mechanisms [49, 50, 60]. In fact, attention mechanisms were developed in order to improve model performance. The increase of interpretability was a secondary benefit.

2.2 Unsupervised ML approaches in medicine

One of the major limitations of supervised learning approaches is the need of good quality labeled data, however most EHR datasets are labeled according to the International Classification of Diseases (ICD9) code, which is manually inserted by clinicians. This process is not only time consuming but also highly susceptible to errors. Unsupervised learning may provide an alternative method to overcome this problem. In this paradigm, the models are developed in order to identify patterns in the dataset without resorting to labeled data. Therefore, these algorithms can be very useful for tasks such as feature extraction and clustering. However, the use of unsupervised approaches in medicine has not been so extensively explored as supervised methods.

In [61], Mayhew et al. try to identify latent phenotypes associated with a a higher risk of mortality in septic patients resorting to an unsupervised method called composite mixture models (CMM). According to the authors, CMM is described as “a flexible joint probability model for multi-typed, multivariate data” and it is based on two assumptions. First, the population in study is heterogeneous and can be decomposed into subgroups or clusters. Second, it is possible to determine the full joint distribution of a multi-typed observation vector by specifying appropriate univariate, exponential distributions for each feature of said vector.

This study included data from Kaiser Permanente Northern California (KPNC) dataset which contains both dynamic and static features of patients with sepsis diagnosis admitted in KPNC medical centers between 2009 and 2013. Furthermore, the maximum, minimum, median, and standard deviation of patient vital signs over three different post-admission periods (3, 6, or 12 h) were determined and then concatenated into a single dataset, which the authors called combined dataset. This combined dataset was used for mortality enrichment and cluster trajectory analysis, from which 20 final clusters were obtained. These clusters were shown to represent different phenotypes and clinical course trajectories, for example while one cluster represented respiratory failure in chronic disease patients, other represented patients with moderate hemodynamic compromise. These clusters were also shown to be associated with different risks of mortality. Therefore, the CMM model was shown to be a useful tool for risk stratification of patients, since it was able to identify clusters strongly associated with a higher risk of mortality.

This study supports the applicability of unsupervised machine learning techniques in the medicine field. However, some limitations with this model were noted. For example, to fit this model, first it is required to specify the univariate distributions for each data feature.

In a completely different approach, Yao et all. resorted to autoencoders to predict sepsis in [5]. Autoencoders are a machine learning technique composed by two components: encoder and decoder. The encoder receives an input sequence and returns a compact vector, which can then be used for clustering, representation learning or as input in other supervised methods. The decoder receives the output from the encoder and reconstructs the original sequence by minimizing the reconstruction error.

In this work, four different autoencoders were developed and tested:

- An autoencoder to extract temporal features composed by a LSTM, denoted as TAE
- An autoencoder to extract spatial features composed by a multi-layer neural network, denoted as SAE
- Two autoencoders to extract both spatial and temporal features by stacking the two previous autoencoders, denoted as TSAE, when the SAE is attached after the TAE, and STAE, when the TAE is attached after the SAE (Figure 2.4)

The data encoded by these four autoencoders were then used in three different classifiers (decision tree, random forest and logistic regression).The best performance was registered when using the hybrid autoencoder TSAE along with logistic regression with an AUC of 0.566 and a F1-score of 0.313. Besides, TAE performed better than SAE, which demonstrates the importance of learning temporal patterns in patients data in order to predict sepsis onset. Furthermore, given that TSAE consistently outperformed STAE, the order of the stacked autoencoders is shown to be important, as it influences the performance of the models.

Finally, since both TSAE and STAE outperformed the temporal autoencoder, this work proves that learning spatial and time-invariant patterns present in the dataset used can improve the performance of an LSTM model, also demonstrated by Lin et al. in [4].

Another interesting study using autoencoders is the one developed in [39] by Baytas et al., in which the authors tried to overcome the irregular intervals present in EHR data using a Time aware LSTM (T-LSTM) when subtyping patients. Normally, LSTM models assume regular time intervals between

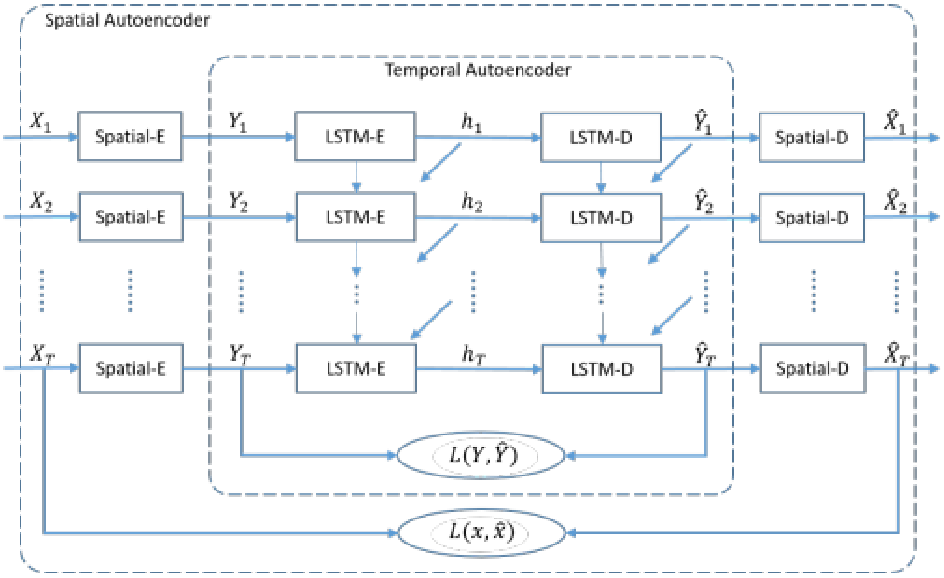


Figure 2.4: STAE model architecture. Source: [5]

each timestep of the input sequence. However, since the time intervals between EHR data are highly irregular, there was a need to include information regarding the time lapse between successive EHR observations in the machine learning model. T-LSTM is a technique derived from the LSTM architecture where, in each timestep, the memory of the previous time step is decomposed into short and long term memories. The short term memory is then adjusted according to the time interval between two successive timesteps in order not to lose the global profile of a patient. To do so, a non-increasing function is used, converting the time vector into appropriate weights, later applied to the short term memory component. This architecture will be further explored in the following chapters.

This new architecture was experimented with two different datasets. First, the authors used a synthetic dataset, which simulated EHR records of up to 100.000 patients, with lab results, diagnoses, and start and end dates of the admissions in order to predict Diabetes Mellitus. The performance of T-LSTM was analysed using supervised and unsupervised approaches. In the supervised approach, the performance of T-LSTM is compared against a classic LSTM and a traditional Logistic Regression through their AUC values. In the unsupervised approach, the T-LSTM was incorporated with an autoencoder, employed to extract expressive features from the raw data, later used to predict Diabetes Mellitus by clustering with *k*-means. The representative power of these clusters was expressed through the Rand index and compared with the results obtained using an autoencoder composed by a classic LSTM. In both approaches, the T-LSTM outperform every baseline used, achieving an AUC value 0.91 and a Rand index of 0.96 for the supervised and unsupervised approaches, respectively.

Afterwards, the Parkinson's Progression Markers Initiative (PPMI) dataset was used in two different tasks: target sequence prediction and patient subtyping. This dataset includes clinical and behavioral assessments, imaging data, and biospecimens information of Parkinson's patients. These data can be divided in two categories called features and targets. While features are related to patients characteristics, such as motor symptomss and cognitive functioning, targets correspond to variables related with the progression of Parkinson's disease.

In the target sequence prediction, as the name implies, T-LSTM was used to predict the target sequence of each patient. The model was evaluated using the Mean Square Error (MSE) and compared against a model composed by a classical LSTM. In patient subtyping task, a T-LSTM autoencoder was used to obtain clusters and identify different subtypes of patients. In this task, since the ground truth was unknown, the clusters were statistically analysed and compared with clusters obtained using an LSTM autoencoder. Once again, the T-LSTM models outperformed classic LSTM models, reporting a lower MSE, in the target sequence prediction, and better expressive clusters in the patient subtyping task.

The consistent better performance of T-LSTM models across all experiments performed in this study, suggests that information regarding the time interval between successive time steps in sequential data like EHR, can be an asset capable of improving model performance.

In the field of machine vision, anomaly detection is a common unsupervised framework used. According to this approach, anomalies rarely occur in the data and their features are significantly different from normal data. In [6], Krissaane et al. adopted this framework for the early detection of sepsis. An autoencoder was trained only with normal data (non-septic patients' data) and was tested in a group composed by septic and non-septic patients. As it would be expected, on average the reconstruction error of patients that developed sepsis was higher (Figure 2.5 A). A precision/recall curve was drawn by applying different thresholds to the reconstruction error (Figure 2.5 B). These results demonstrate the applicability of this framework for sepsis prediction.

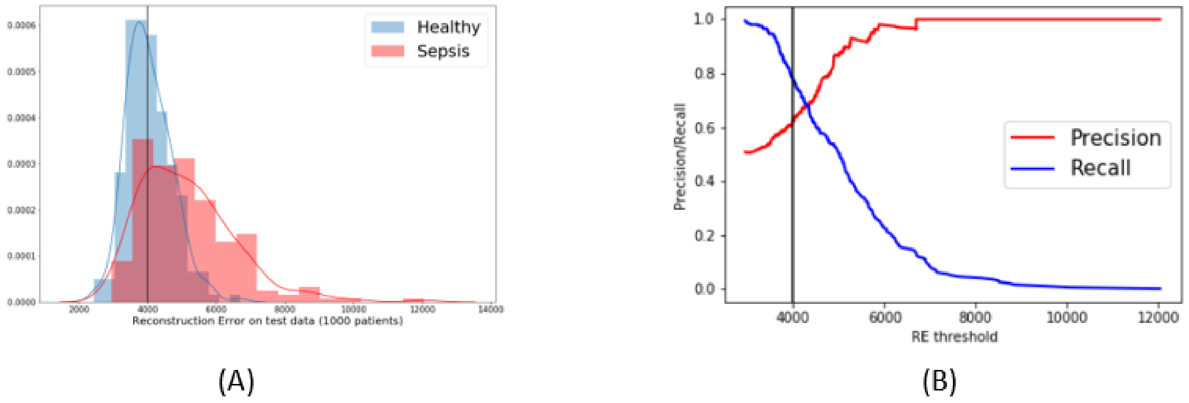


Figure 2.5: Results obtained by Krissaane et al. (A) Reconstruction error for healthy and sepsis groups; (B) Precision/recall curve for different reconstruction error thresholds. Source: [6]

New improvements have been developed in the field of anomaly detection with the surge of the Variational Autoencoder (VAE). VAE works similarly to an autoencoder, however instead of encoding to a single point in the latent space, the input data is encoded as a distribution over the latent space. Besides, a regularization term is added to the loss function in order to obtain a better organisation of the latent space, generally by forcing the model to map the inputs closely to the unit Gaussian distribution $\mathcal{N}(0, I)$ in this space. The VAE architecture will be further explained in Chapter 3.

Applying a anomaly detection framework, Ramos developed a VAE model with the aim to predict septic shock onset in [62]. The VAE model was trained using only septic patients which did not developed septic shock from MIMIC-III. In order to identify shock patients, three different clustering algorithms were applied to the data encoded by VAE. The algorithms used were K-means, Spectral Clustering and Gaussian Mixture Models (GMM). The latter technique showed a more consistent performance and achieved an AUC value of 0.8184. The performance of this model along with the clustering algorithms showed to be very competitive when compared to the performance of a supervised model used as baseline. Furthermore, similarly to [6], higher reconstruction error was verified in shock patients.

In [63], Vasilev et al. resorted to a VAE model to propose new metrics to detect anomalies based on the distributions learned by VAE, which were denoted anomaly scores. These anomaly scores proposed could be divided in three categories: VAE reconstruction-based, distance-based or density-based. The authors tested the anomaly detection power of these scores in two different datasets.

The Modified National Institute of Standards and Technology (MNIST) dataset was one of the datasets used in this study, and this dataset contains 60,000 small images of handwritten single digits between 0 and 9. In this task, one of the digits was considered an anomaly and VAE model was trained with the remaining digits. Since there are 10 digits, ten different anomaly detection experiments were performed and the performance of the anomaly scores proposed were compared with the performance of linear PCA, PCA with a Gaussian kernel (kPCA) and a classic VAE-based approach. For every digit, the proposed anomaly scores outperformed the baseline methods, presenting a higher AUC value. For example, when the digit 7 was considered an anomaly, while the three methods used as baseline achieved AUC scores between 0.5 and 0.6, some of the new proposed anomaly scores registered AUC values between 0.7 and 0.8.

The second dataset used was composed by diffusion MRI scans and the goal was to detect multiple sclerosis lesions. The dataset contained 26 diffusion MRI scans of healthy volunteers, 20 of which were used for training, and 3 diffusion MRI scans of multiple sclerosis patients, used for testing along with the remaining six scans of healthy patients. In this approach every voxel is a sample, rather than every scan. In this experiment, the anomaly scores proposed were compared with a score based on the Euclidean distance between the test datapoint and its nearest neighbor from the reference dataset in the feature space. Most proposed anomaly scores showed a good performance, achieving AUC values above 0.8

in most cases, with at least one of the new proposed scores outperforming the baseline. In this study, the proposed anomaly scores showed promising results and the application of these scores for early prediction of sepsis might be an interesting focus to be further explored.

Table 2.1 and Table 2.2 include a summary of the unsupervised and supervised machine approaches mentioned along this chapter, respectively, and their main results.

Table 2.1: Review of the unsupervised ML models mentioned along the chapter and their performance.

Article	Task	Model	Results	
			AUC	Other
Baytas et al., 2017 [39]	Diabetes Mellitus Prediction	T-LSTM	-	RI: 0.96
		LSTM	-	RI: 0.91
Krissaane et al., 2019 [6]	Sepsis Prediction	AENN + RF + LR	0.614	-
Vasilev et al., 2018 [63]	Multiple sclerosis lesion detection	VAE+baseline	0.7218	-
		VAE+reconstruction based score	0.923	-
		VAE+distance based score	0.91	-
		VAE+density based score	0.944	-
Yao et al., 2019 [5]	Sepsis Prediction	DT	0.529	-
		RF	0.531	-
		LR	0.511	-
		SAE+DT	0.515	-
		SAE+RF	0.522	-
		SAE+LR	0.541	-
		TAE+DT	0.541	-
		TAE+RF	0.511	-
		TAE+LR	0.534	-
		STAE+DT	0.527	-
		STAE+RF	0.509	-
		STAE+LR	0.544	-
		TSAE+DT	0.525	-
		TSAE+RF	0.533	-
		TSAE+LR	0.566	-
Ramos, 2021 [62]	Septic Shock Prediction	VAE+k-means	0.9114	-
		VAE+SC	0.7376	-
		VAE+GMM	0.8184	-
		LSTM	0.8038	-

Table 2.2: Review of the supervised ML models mentioned along the chapter and their performance.

Article	Task	Model	Results	
			AUC	Other
Choi et al., 2016 [52]	Diagnosis and Medication Prediction	GRU	-	Recall@30: 0.7248
		Most Frequent Label	-	Recall@30: 0.66
		LR	-	Recall@30: 0.5253
		MLP	-	Recall@30: 0.5574
Lipton et al., 2016 [2]	Diagnosis Prediction	LR	0.7218	-
		MLP	0.777	-
		LSTM	0.7625	-
		LSTM+TR+AuxOut	0.7926	-
Nguyen et al., 2020 [3]	Alzheimer Prediction	LR	0.7218	-
		RNN+FF	0.923	-
		RNN+LF	0.91	-
		RNN+MF	0.944	-
		LSS+MF	0.926	-
		LSTM+MF	0.925	-
Fagerstrom et al., 2019 [56]	Septic Shock Prediction	LiSep	0.7218	-
		TREWScore	0.83	-
		InSight	0.83	-
		Multitask LSTM	0.85	-
		SepLSTM	0.93	-
Wernly et al., 2021 [58]	Mortality Prediction in Septic Patients	LSTM	0.88	-
		LR	0.82	-
Lin et al., 2018 [4]	Septic Shock Prediction (Left Align)	LSTM-Origin	0.9168	-
		LSTM+CNN+Static-last	0.9411	-
		LR	0.7985	-
		NB	0.7147	-
		SVM	0.7707	-
		DT	0.6462	-
		RF	0.7977	-
		MLP	0.4896	-
	Septic Shock Prediction (Right Align)	LSTM-Origin	0.841	-
		LSTM+CNN+Static-last	0.865	-
		LR	0.742	-
		NB	0.687	-
		SVM	0.705	-
		DT	0.597	-
		RF	0.740	-
		MLP	0.744	-
Kaji et al., 2019 [59]	Septis Prediction (Same day)	LSTM+Attention	0.952	-
	Septis Prediction (Next day)	LSTM+Attention	0.876	-
Baytas et al., 2017 [39]	Diabetes Mellitus Prediction	T-LSTM	0.91	-
		LSTM	0.85	-
		LR	0.56	-
	Target Sequence Prediction	T-LSTM	-	MSE: 0.50
		LSTM	-	MSE: 0.51

3

Methodology

Contents

As already mentioned in previous chapters, the aim of this work is to develop a machine learning model capable of predicting septic shock onset on ICU patients, and this prediction task can be performed on a visit-level (left aligned) or on a event-level (right aligned). The models and results presented in this thesis are based on the later approach, since this approach was believed to provide more relevant results.

Patient in-hospital stays can last from a few days to months. Since left-aligned approaches only analyse the first *k* hours after admission, for cases where the shock onset occurred in the late stages of patients' ICU stay, the model might not be able to identify them. Furthermore, some sepsis and septic shock cases might develop from infections acquired during the ICU stay. Therefore, while left aligned models can be useful for risk stratification shortly after a patient admission in ICU, since right aligned models resort to observations from the previous *n* hours, this might be a better approach for patient real time monitoring. With this approach, an analysis of different patterns between patients who develop septic shock and those who do not is also possible.

Models from both supervised and unsupervised approaches were developed and the results compared. Although supervised models are expected to perform better since they are provided information regarding patient classification (septic shock or non-septic shock) during the training phase, the aim of this comparison is to determine if an unsupervised model can achieve the same degree of success as a supervised one. Every model proposed, either supervised or unsupervised, resorts to an observation window of 10 hours and a 3 hour hold-off window. This means that, at a given point in time, the model predicts whether a patient will develop septic shock three hours later, according to the observations made in the previous ten hours. Moreover, in consideration of the results and conclusions presented in the previous chapter regarding the performance of machine learning techniques when handling sequential data, all models proposed are RNN-based.

In this chapter, the methodology followed for the realization of this work is explained in detail, including data extraction, data preprocessing, the models proposed along with their background, and the evaluation metrics used.

3.1 Data Extraction

Throughout this project, two different datasets were used. The first dataset was provided by Hospital São Francisco Xavier in Lisbon, Portugal. This dataset included information regarding age, gender, vital signs, clinical tests, procedures/interventions performed and diagnosis from patients admitted since 2015. From the set of variables contained in the dataset, 40 variables were selected (Table A.1). These data were distributed among several *Excel* files, where patients and ICU stays were identified by unique numbers, called *ProcessID* and *EpisodeID*. Therefore, the first required step was to merge the data from all files according to these IDs. Afterwards, some restrictions were implemented in order to select a population of interest for this study. The criteria used were:

- **Only include patients with ICU stays longer than 24 hours**. Doing so, patients which are already severely ill at the time of admission are excluded.
- **Only include patients over 18 years**, since the development of sepsis is influenced by age. Besides, patients under 18 years old are a whole different population with significantly different behaviours and patterns when compared with adult patients
- **Only include the first ICU stay per patient**. Some repeat ICU stays might be due to complications from previous visits, therefore, for the sake of consistency, only the first ICU stay per patient is considered.
- For patients which developed septic shock, **only include data until the first septic shock episode**. In some cases, one patient might suffer from more than one septic shock in only one ICU stay, thus, for patients which developed septic shock, only information preceding this episode is considered.
- For patients which developed septic shock, **only include patients with septic shock onset after the first 13 hours**, since the models proposed require a ten hour observation window and a three hour hold-off window.

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. To extract data from MIMIC-III, an open source pipeline called MIMIC Extract [7] was followed, according to which, all data was first required to be uploaded to a local PostgreSQL database. Afterwards, a python script, provided by the authors of [7], and last updated in December 2020, was executed. This script allows the customization of the cohort selection and extraction criteria, after which a file is created with the static information of the patients included in the cohort.

Regarding information related to the clinical tests and vital labs of the patients, this tool performs an outlier removal and unit conversion step in order to guarantee consistent units for all values in each variable. Afterwards, all data is aggregated in hourly-bins and the mean and standard deviation are calculated and included in the patients timeseries. Furthermore, a file for clinical interventions is also created, where these features are extracted with the use of binary indicators. An overview of this pipeline can be observed in Figure 3.1.

The final step for the data extraction is the classification of sepsis and septic shock. For the classification of sepsis, the Sepsis-3 criteria were applied, according to which, sepsis is defined by the presence of a infection suspicion along with a change of 2 or more points in the SOFA score. The first dataset contains information of patients diagnosed with any type of infection and the diagnosis are timestamped.

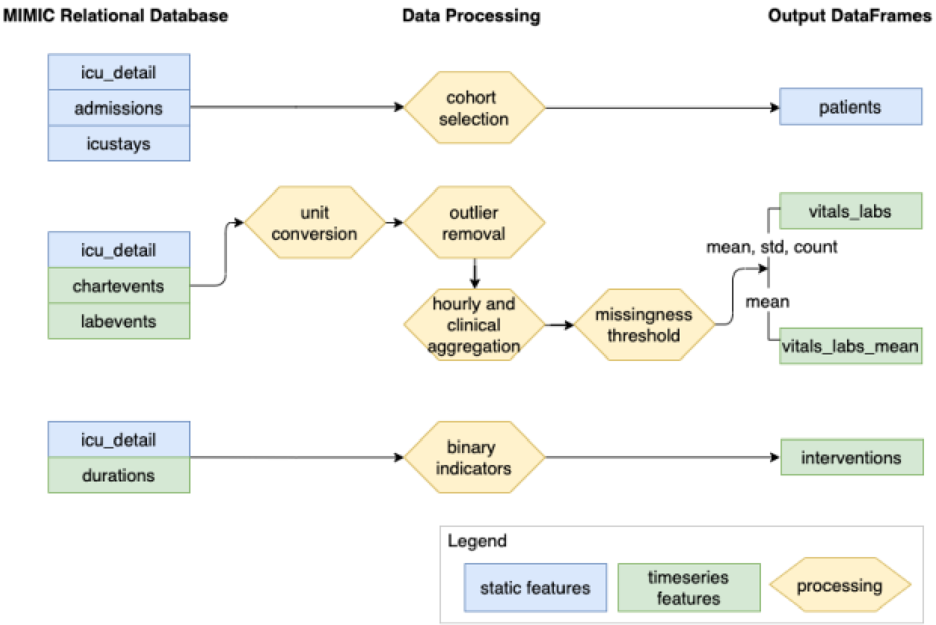


Figure 3.1: Overview of the MIMIC Extract pipeline. Source: [7]

Therefore, if a patient from this dataset is diagnosed with an infection and has a change of 2 points or more in the SOFA score, which is also one of the features provided, the patient can be considered to have sepsis. On the other hand, the MIMIC-III dataset does not possess timestamped information regarding infection diagnosis, so the identification of septic patients is more complicated. Considering this limitation, for this database, there is a suspicion of infection if an antibiotic is administered within 72 hours after a collection of body fluid culture or if a culture is ordered within 24 hours after administration of antibiotic, and the onset time is considered to be either the culture time or the antibiotic time, depending on which was ordered first.

Regarding the classification of septic shock, some changes suggested by an intensive care physician from Hospital São Francisco Xavier were made to the Sepsis-3 definition. While Sepsis-3 criteria defines septic shock as a subset of sepsis with hypotension requiring vasopressor therapy to maintain a mean arterial blood pressure greater or equal to 65 mmHg and a serum lactate level greater than 2 mmol/L after adequate fluid resuscitation [1], according to the medical experience of the physician, some cases of septic shock do not present both criteria. Therefore, for this work, a patient is considered to have developed septic shock if it has sepsis along with at least one of the following criteria:

- Score of 3 or more in the hemodynamic component of the SOFA score
- Administration of vasopressors
- Lactate values superior to 2 mmol/L

Table 3.1: Detailed information of the population of interest from both datasets considered (Hospital São Francisco Xavier and MIMIC-III). $HSFX \equiv$ Hospital São Francisco Xavier, LOS≡ Length of stay.

Dataset	Group	Number of Patients	Gender		Age		LOS (hours)	
			M	F	mean	std	mean	std
HSFX	Healthy	406	195	211	63.04	19.18	128.96	105.7
	Sepsis	438	273	165	63.97	15.86	359.83	267.31
	Shock	133	56	76	66.18	13.79	418.21	408.51
MIMIC	Healthy	11112	6424	4688	72.4	52.04	189.61	190.06
	Sepsis	3888	2118	1770	79.6	60.29	273.41	253.75
	Shock	1343	765	578	83.36	61.76	276.63	274.03

According to these classification criteria, the final cohort of the dataset from Hospital São Francisco Xavier is composed by a 844 patients, 133 of which developed septic shock during their ICU stay, while the final cohort from the MIMIC-III database is composed by 15000 patients and 1343 of them developed septic shock. A more detailed overview of the final cohort is described in Table 3.1.

3.2 Data Preprocessing

After extracting data from both datasets, some preprocessing steps were required before feeding these data to the models developed. One of the first steps consisted on the removal of existing outliers, specially in data from Hospital São Francisco Xavier. With this goal in mind and with the help of the intensive care physician, ranges of acceptable values were defined for each variable. All data outside these ranges were removed and considered missing data. The defined ranges can be observed in Table A.1 in Appendix A. Regarding MIMIC-III data, although the extraction tool used already includes an outlier removal step, in order to guarantee consistency in the data of Hospital São Franciso Xavier and MIMIC-III dataset, the same ranges were applied to the latter.

As already mentioned, electronic health records are inherently sparse, and observations are irregularly sampled. Besides, different variables are sampled at different rates, for example, vital signs are collected almost continuously while clinical lab results are collected occasionally. These factors contribute for high missing data rates, and since machine learning models cannot deal with missing data, a data imputation method was required. As already mentioned in [3], the method chosen for data imputation is one of the factors that influence model performance, therefore it was important to choose an appropriate data imputation approach. 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 consideration, the forward-filling based imputation method proposed in [7] 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 [7] 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 (3.1).

$$z = \frac{x - min}{(max - min)} \tag{3.1}$$

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.

3.3 Supervised Models

3.3.1 LSTM

One of the main characteristics of feedforward networks is their lack of memory. Since these networks do not possess memory, when dealing with sequential data, the entire sequence must be provided to the model at once, losing the temporal information contained in the sequence. Therefore, recurrent neural networks were developed [64]. These networks process sequences by iterating through the sequence elements allowing to process information incrementally. During these iteration throughout the sequence, the network maintains a state containing information relative to previous observations, which is updated as new information is provided. This means that, the output at a certain timestep is influenced by the state of the previous timestep. This relation can be described by the equation

$$h_t = f\,(W \cdot x_t \; + \; U \cdot h_{t-1}) \tag{3.2}$$

where h_t is the state at the timestep *t*, h_{t-1} is the state at the timestep *t-1*, x_t is the element of the sequence at timestep *t*, W and U correspond to weight matrices and f is the activation function.

However, simple RNNs cannot deal with long sequences due to a limitation called vanishing gradient problem, according to which, as the size of the sequence increases the amount of information required to be memorized increases as well and at a certain point the network becomes untrainable, since the value of gradient becomes too small and no learning is done [64, 65].

To overcome this limitation, several RNN-based models were developed, including LSTM. LSTM architecture solves this problem through the implementation of three gates in its structure, called input gate (i_t), forget gate (f_t) and output gate (o_t), and an additional data flow responsible to carry information across timesteps, called the cell state (c_t). The interaction of these three gates controls the flow of information in a LSTM model. As the names imply, the forget gate is responsible for defining which information carried by the cell state should be removed, the input gate is responsible for adding new information to the cell state and the output gate filters the information in the cell state relevant for the task at hand to be returned by the model [4, 64]. These interactions are described by the following equations:

$$i_t = sigmoid\,(W_i \cdot [h_{t-1}, X_t] + b_i) \tag{3.3}$$

$$f_t = sigmoid\,(W_f \cdot [h_{t-1}, X_t] + b_f) \tag{3.4}$$

$$\widetilde{c} = sigmoid\,(W_c \cdot [h_{t-1}, X_t] + b_c) \tag{3.5}$$

$$o_t = sigmoid\,(W_o \cdot [h_{t-1}, X_t] + b_o) \tag{3.6}$$

$$c_t = c_{t-1} \cdot f_t + \widetilde{c} \cdot i_t \tag{3.7}$$

$$h_t = o_t * tanh\,(c_t) \tag{3.8}$$

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. Figure 3.2 demonstrates the mechanisms behind an LSTM unit.

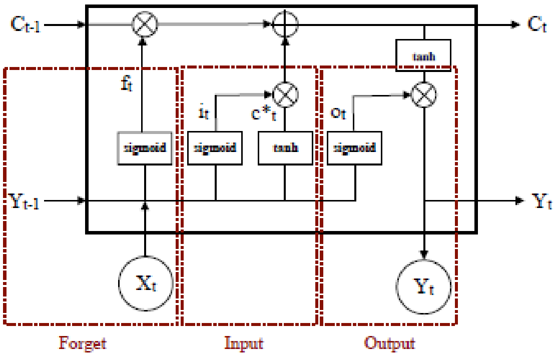


Figure 3.2: Overview of a LSTM structure. Source: [4]

3.3.2 T-LSTM

Although LSTMs brought improvements to RNN models some limitations remain. As mentioned in the previous chapter, LSTM models assume that elements in a sequence are sampled at regular intervals, however the distribution of observations in a temporal patient health record are non-uniform. The time elapsed between two consecutive visits can vary from days to years and even in the same ICU stay the interval between two measurements of the same variable is highly irregular. Furthermore, regarding the processing of EHR data, these varying intervals might provide useful information. Therefore, a variation of LSTM capable of integrating these irregular time intervals in its structure was developed, called T-LSTM [39].

The integration of the time intervals in T-LSTM structure is achieved through the following equations:

$$c_{t-1}^{s} = tanh\left(W_d \cdot c_{t-1} + b_d\right) \tag{3.9}$$

$$c_{t-1}^{l} = c_{t-1} - c_{t-1}^{s} \tag{3.10}$$

$$\hat{c}_{t-1}^{s} = c_{t-1}^{s} * g\left(\Delta t\right) \tag{3.11}$$

$$c_{t-1}^{*} = c_{t-1}^{l} + \hat{c}_{t-1}^{s} \tag{3.12}$$

$$c_t = f_t * c_{t-1}^{*} + i_t * \widetilde{c} \tag{3.13}$$

$$h_t = o_t * tanh\left(c_t\right) \tag{3.14}$$

where c_{t-1} and c_t are the previous and current cell states, c_{t-1}^{s} and c_{t-1}^{l} are the short term and long term components of the previous memory, $\hat{c}_{t-1}^{s}$ 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 memory decomposition network, respectively. The function $g\left(\cdot\right)$ is a non increasing function applied to the time intervals, Δt. Finally, i_t, f_t, o_t and $\widetilde{c}$ are the input, forget, and output gates and the candidate cell memory determined by following the equations 3.2-3.5.

As previously mentioned, according to [39], one of the main contributions of T-LSTM is the subspace decomposition of its memory in short-term and long-term memories (Equations 3.9 and 3.10). This decomposition is data driven, and the parameters of the decomposition network, W_d and b_d, are learned simultaneously with the rest of network parameters by back-propagation during the training phase. Regarding the activation function of the decomposition network, Baytas et al. [39] experimented different function types and although *tanh* function performed slightly better, no significant differences were observed.

According to [39], during the training phase while the long term memory should not be entirely discarded, the short term memory should be adjusted depending on time interval between the current observation and the previous one. The adjustment of this memory component is accomplished by weighting it with the elapsed time (Equation 3.11), but first the latter must be converted into appropriate weights. This conversion is performed resorting to non-increasing continuous function. The type of function chosen is dependent on the task at hand, however, as guidelines, the authors recommended to use the function $g(\Delta t) = 1/\Delta t$ for datasets with small time intervals and $g(\Delta t) = 1/log(e + \Delta t)$ for datasets with large time intervals.

After obtaining the adjusted short-term memory, it is combined with the long term component to obtain the adjusted previous cell state (Equation 3.12). From this point forwards, T-LSTM operates in the same manner that vanilla LSTM, with input, forget and output gates.

3.3.3 Proposed architectures

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*, according to Table A.1. 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 [4], 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.

All three models were developed using *Keras* and trained with 500 epochs using a batch size of 50, a binary cross-entropy loss function, an *Adam optimizer*, and an early stop criteria with 25 epochs of patience, in order to avoid overfitting. Furthermore, the impact of the five different groups of variables on the performance of the model was also analysed. To do so, for each variable group, the performance was evaluated for when only that variable group was included and for when only that variable group was excluded. This will allow to understand whether some of the variable groups are more important when classifying sepsis.

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

3.4 Unsupervised Models

3.4.1 VAE

Autoencoders are one common approach for unsupervised machine learning. They are specially usefull for feature extraction and data compression, or dimensionality reduction. This technique is composed by two components: an encoder, which takes the input and encodes it to a latent space, usually with a lower dimension, and a decoder, which decodes the original input from the encoded data by minimizing the reconstruction error. This structure allows for autoencoders to learn only the main structured part of the information. However, this learning scheme has some limitations, such as the lack of regularity of the latent space [66]. In fact, since autoencoders are only trained to encode and decode the original data with as low reconstruction error as possible, how the latent space is organised is not taken into consideration. Therefore, it is hard to guarantee a good organization of the latent space and, consequently, some data points in the latent space, once decoded, are meaningless.

VAE are developed models capable of overcoming this limitation by incorporating variational inference in the autoencoder. Unlike classic autoencoders, instead of encoding a single data point, VAE encode the original data into a statistical distribution in the latent space. In other words, the encoded data is forced to obey some type of prior probability distribution $p_\theta(z)$. Usually, the prior probability distribution used is the Gaussian distribution $\mathcal{N}(0, I)$ [63, 66, 67]. Therefore, the output of a VAE encoder are the parameters of a distribution. From this distribution, a data point is sampled and provided to the decoder which reconstructs the original data.

Since the network learns by error backpropagation, a reparametrisation trick is required during the sampling step, in order to make the error backpropagation possible despite the random sampling. According to this reparametrisation trick [66], a random datapoint is sampled following equation 3.16, where ϵ is noise that follows a normal distribution $\mathcal{N}(0, I)$ and $\odot$ corresponds to element wise multiplication. Sampling using this expression allows the separation of a deterministic and a stochastic components.

$$z = \mu_z + \sigma_z \odot \epsilon \tag{3.15}$$

Considering the original data as x and the encoded data as z, by assuming that x is generated from z, the encoder of a VAE can be defined by $p_\theta(z|x)$ and the decoder by $p_\theta(x|z)$. According to the Bayes Theorem the relation between $p_\theta(z|x)$ and $p_\theta(x|z)$ can be established by:

$$p\left(z|x\right) = \frac{p\left(x|z\right)p\left(z\right)}{p\left(x\right)} = \frac{p\left(x|z\right)p\left(z\right)}{\int p\left(x|u\right)p\left(u\right)du} \tag{3.16}$$

Therefore, theoretically, knowing the distributions $p_\theta(z)$ and $p_\theta(x|z)$, it is possible to determine $p_\theta(z|x)$. However, the true posterior distribution $p_\theta(z|x)$ is usually intractable [68]. Therefore, VAE resort to variational inference to find a deterministic approximation, $q_\theta(z|x)$, of the intractable true posterior. Hence, the goal of the VAE is to find the function $q_\theta(z|x)$ that better approximates $p_\theta(z|x)$, and the quality of this approximation is evaluated with the Kullback-Leibler divergence, $D_{KL}\left(q_\phi\left(z|x\right)\ ||\ p_\theta\left(z|x\right)\right)$. This term usually cannot be calculated directly, however it is possible to minimize it by maximizing the Evidence Lower Bound (ELBO) [63]. Therefore, the loss function of a VAE is described by the expression:

$$\mathcal{L}\left(\theta, \phi, x\right) = \mathbb{E}_{q_\theta(z|x)}\left[log\ p_\theta\left(x|z\right)\right] - \beta\ \mathcal{D}_{KL}\left(q_\theta\left(z|x\right)\ ||\ p_\theta\left(z\right)\right) \tag{3.17}$$

The first term of the loss function corresponds to the reconstruction error and the second term corresponds to the Kullback-Leibler (KL) divergence and it is used as a regularization term [63]. The trade-off between the two loss terms is assured by the weighting parameter, β. Considering the prior distribution, $p_\theta(z)$, a Gaussian function $\mathcal{N}(0, I)$, the Kullback-Leibler divergence term can be simplified using equation 3.18 [68]

$$\mathcal{D}_{KL}\left(q_\theta\left(z|x\right)\ ||\ p_\theta\left(z\right)\right) = -\frac{1}{2}\left[1 + log\left(\sigma^2\right) - \sigma^2 - \mu^2\right] \tag{3.18}$$

Using this approach, VAE guarantees a certain regularity in the latent space, such as continuity. For instance, from two close points in the latent space, two similar data will be decoded.

3.4.2 Anomaly Detection and scores

For the unsupervised approach, three different variational autoencoders were proposed. The difference between them consisted on the encoder used. For each of the VAEs one of the models used in the supervised approach was used as the encoder, creating the models *VAE-LSTM-All*, *VAE-LSTM-static-last* and *VAE-LSTM-static-repeat*, respectively. Figure 3.3 demonstrates the encoder structure from model *VAE-LSTM-All*

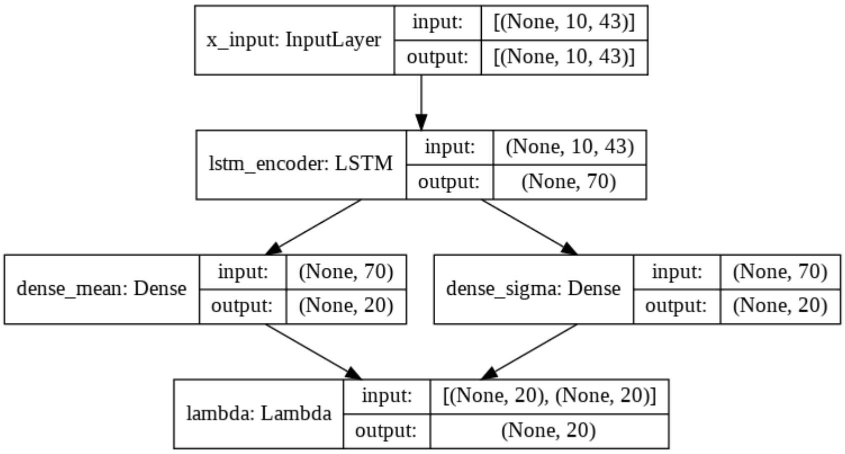


Figure 3.3: Overview of the encoder from *VAE-LSTM-All*.

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. Figure 3.4 demonstrates the structure of the decoder.

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 [62] 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 [63] 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 [63], described by

$$q_X(z) = \frac{1}{|X|} \sum_{x \in X} q_\theta\left(z|x\right) \tag{3.19}$$

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

- Reconstruction error - This score was also explored in [6] and 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 dataset 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.

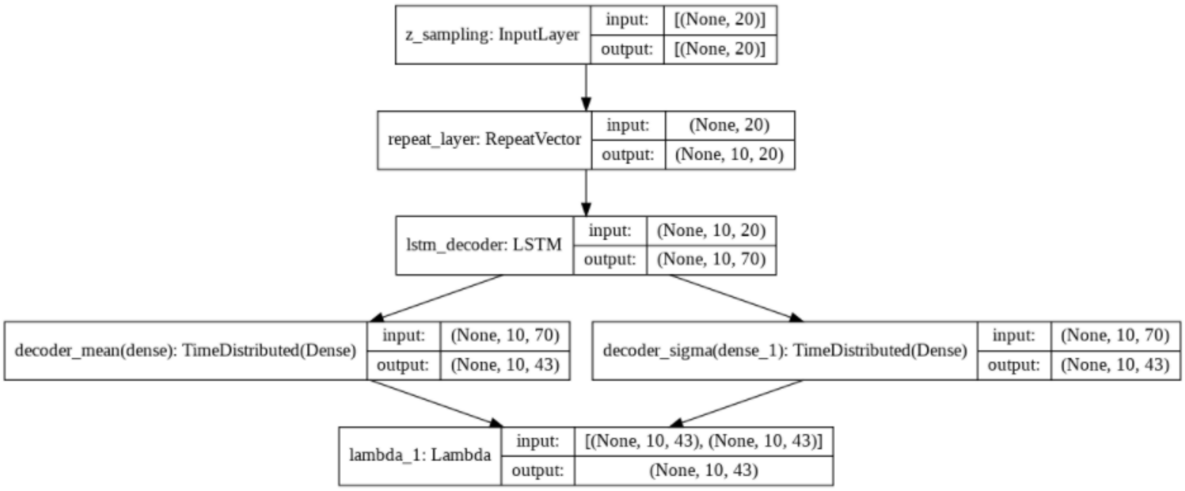


Figure 3.4: Overview of the decoder from VAEs proposed.

3.5 Evaluation metrics

In order to evaluate and compare the performance of all models proposed, both supervised and unsupervised, some evaluation metrics were determined. The metrics used were recall, precision, f1-score and AUC. Considering the following confusion matrix, where TN and TP correspond to true negatives and true positives respectively, and false negatives and false positives are represented by FN and FP respectively.

Table 3.2: Confusion Matrix for a binary classification problem.

Predicted **Actual**	**Negative**	**Positive**
Negative	TN	FP
Positive	FN	TP

Recall can be calculated by the Equation (3.20) and it represents the proportion of correct diagnosis amongst shock patients. This evaluation metric is specially important, since septic shock is a severe condition and therefore it is important for the model to not exclude shock patients.

$$Recall = \frac{TP}{TP + FN} \tag{3.20}$$

On the other hand, precision defines the proportion of shock predictions that actually correspond to shock patients and it is determined by Equation (3.21).

$$Precision = \frac{TP}{TP + FP} \tag{3.21}$$

Sometimes an improvement in precision comes at the cost of a decrease of recall and vice versa. Therefore, there needs to be a balance between these two metrics. With this in mind, the metric f1-score was also determined. This score takes into consideration both the recall and precision of a model and is a better evaluation metric in cases where there is an uneven class distribution, such as in the case of septic shock.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.22}$$

Finally, the last score evaluated was the Area Under the Curve (AUC) score, which measures model performance at different threshold settings.

4

Results and Discussion

Contents

4.1 Supervised Models

4.1.1 Comparison between architectures

As a first step, a model composed only by a LSTM layer and a Dense layer followed by the final classification layer (*LSTM-All*) was developed and the hyperparameters of the network were chosen. A size of 70 and 20 units was established for the LSTM and Dense layers, respectively. As mentioned in the previous chapter, 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 4.1.

Table 4.1: Performance of *LSTM-All* model with different inputs using data from Hospital São Francisco Xavier.

Input Data	Recall	Precision	F1-score	AUC
All data	**0,7378**	0,7133	**0,7207**	**0,8349**
Only *Vitals*	0,4814	0,7223	0,5778	0,7207
Only *ABG variables*	0,6518	0,6673	0,6531	0,7892
Only *Clinical tests*	0,5629	0,7852	0,6552	0,7647
Only *Daily variables*	0,4707	**0,8219**	0,5445	0,6941
All data excluding *vitals*	0,7185	0,7525	0,7149	0,8337
All data excluding *ABG variables*	0,6592	0,698	0,6734	0,7977
All data excluding *clinical tests*	0,7111	0,6537	0,6762	0,8132
All data excluding *daily variables*	0,5259	0,6831	0,585	0,7342

According to Table 4.1, the worst performance was registered when the model's input included only the group *daily variables*, reaching a recall of only 47% and an AUC value of 0.6941. This result was already expected since this group of variables is composed by interventions performed, such as ventilation and administration of vasopressors, and features which are observed in a daily basis on an ICU setting, such as SOFA score, and these variables show the least variation during a 10 hours observation window when compared to the other variables. Since there is so little variation, there is no sufficient information for the model to be able to distinguish between non-shock and shock patients. The model using only vital signs as input did not report a good performance too, achieving only a value of 0.48 of recall. This means that amongst all shock patients, this model only identified correctly 48% of the patients. This leads one to conclude that the variables included in this group are too broad and not specific enough to identify septic shock. In fact, none of the variable groups demonstrated to be able to identify septic shock by themselves.

To understand if all variable groups are required to identify septic shock, a set of experiments was performed, in which from the whole patients timeseries a certain variable group was excluded before using the data as input and the performance of the models were compared. According to the results obtained, the model performance suffered the most impact when the *daily variables* group was excluded. 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. The model performed best when all data was used as input, which means that the four variable groups bring benefits to the model performance

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. Furthermore, the benefits of incorporating masking layers in the models were also explored. These layers allow for the LSTM layers to skip timesteps by masking sequences with a mask value and therefore were implemented with the aim to skip timesteps where no observations were registered. The models *LSTM-static-repeat* and *LSTM-static-last* outperformed the model *LSTM-All* in all evaluation metrics used (Table 4.2), 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 [4], the model *LSTM-static-last* outperformed *LSTM-static-repeat*. Regarding the masking layers, their application did not improve model performance. In fact, in most cases, there was a decrease of performance and therefore these layers should not be included in the model for this task.

Table 4.2: Model performances with and without masking layers using data from Hospital São Francisco Xavier.

Model	Recall	Precision	F1-score	AUC
LSTM-All	0,7378	0,7133	0,7207	0,8349
LSTM-static-repeat	0,7481	0,7429	0,7454	0,8461
LSTM-static-last	**0,8**	**0,8016**	**0,7987**	0,8784
LSTM-All with masking layer	0,68	0,694	0,6756	0,806
LSTM-static-repeat with masking layer	0,7481	0,7511	0,748	0,8469
LSTM-static-last with masking layer	0,7795	0,8124	0,7971	**0,8813**

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 4.3). Since this dataset includes much more patients, there is more information and the models learn to classify better shock and non shock patients. In fact, 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* achieved better values in all metrics except in precision where it was outperformed by the model *LSTM-static-repeat*.

Table 4.3: Models performance using data from MIMIC-III.

Model	**Recall**	**Precision**	**F1-score**	**AUC**
LSTM-All	0,9381	0,9854	0,9611	0,9209
LSTM-static-repeat	0,9862	**0,9862**	0,9862	0,9929
LSTM-static-last	**0,9987**	0,9765	**0,9874**	**0,9988**

4.1.2 LSTM vs TLSTM

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 4.4.

Table 4.4: Performance of T-LSTM models in both datasets.

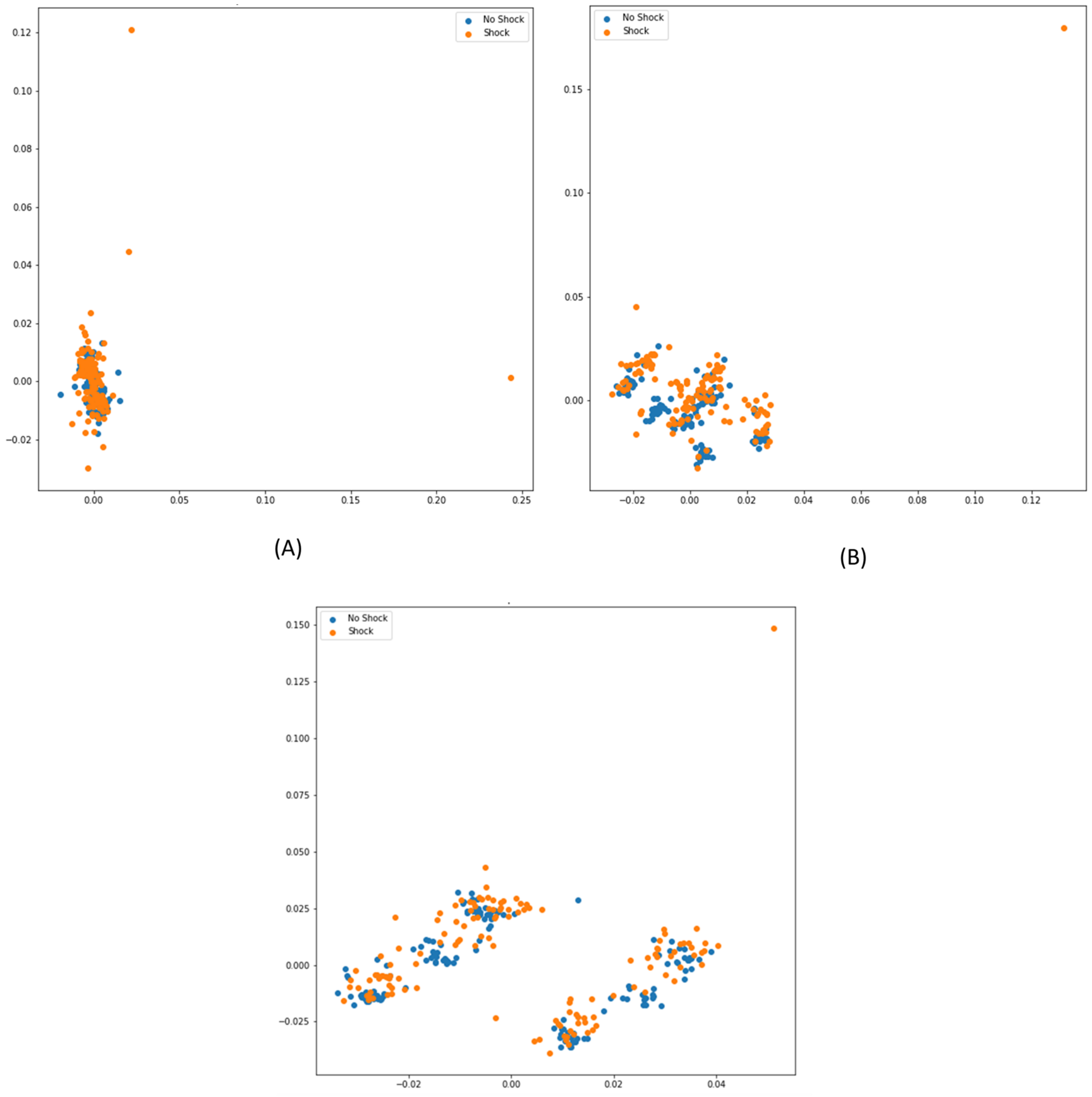
Dataset	**Model**	**Recall**	**Precision**	**F1-score**	**AUC**
Hospital São Francisco Xavier	TLSTM-All	**0,7644**	0,7161	0,7371	0,8487
	TLSTM-static-repeat	0,6889	0,7595	0,7202	0,8204
	TLSTM-static-last	0,6933	**0,825**	**0,7486**	**0,8298**
MIMIC-III	TLSTM-All	0,9421	0,9799	0,9606	0,9196
	TLSTM-static-repeat	0,9628	**0,9877**	0,975	0,991
	TLSTM-static-last	**0,998**	0,9813	**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 [39]. One possible justification for this contradiction is the difference in the task for which the models are developed and the data used. In [39], 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.

4.2 Unsupervised Models

4.2.1 Representation Learning

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 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. As it can be observed in Figure 4.1, resorting to PCA to visualize the encoded data in two dimensions, for a high β value, the data is concentrated around the origin as it would be expected, since the Kullback-Leibler divergence term dominates the loss function.



(A)

(B)

(C)

Figure 4.1: Visualization via PCA of data from Hospital São Francisco Xavier encoded by *VAE-LSTM-All*, using: (A) β =200, (B) β =100, (C) β =50.

As the value of this hyperparameter starts to decrease, the encoded data starts to spread out. Since the model is not trained with shock patients, it would be expected to observe the formation of two clusters, one for non-shock patients around the origin and one for shock patients. However, according to the results in Figure 4.1, although some clustering of data can be observed, these clusters do not seem to represent shock and non-shock patients. Therefore, in order to understand how these data were being clustered, a more detailed analysis was performed (Figure 4.2).

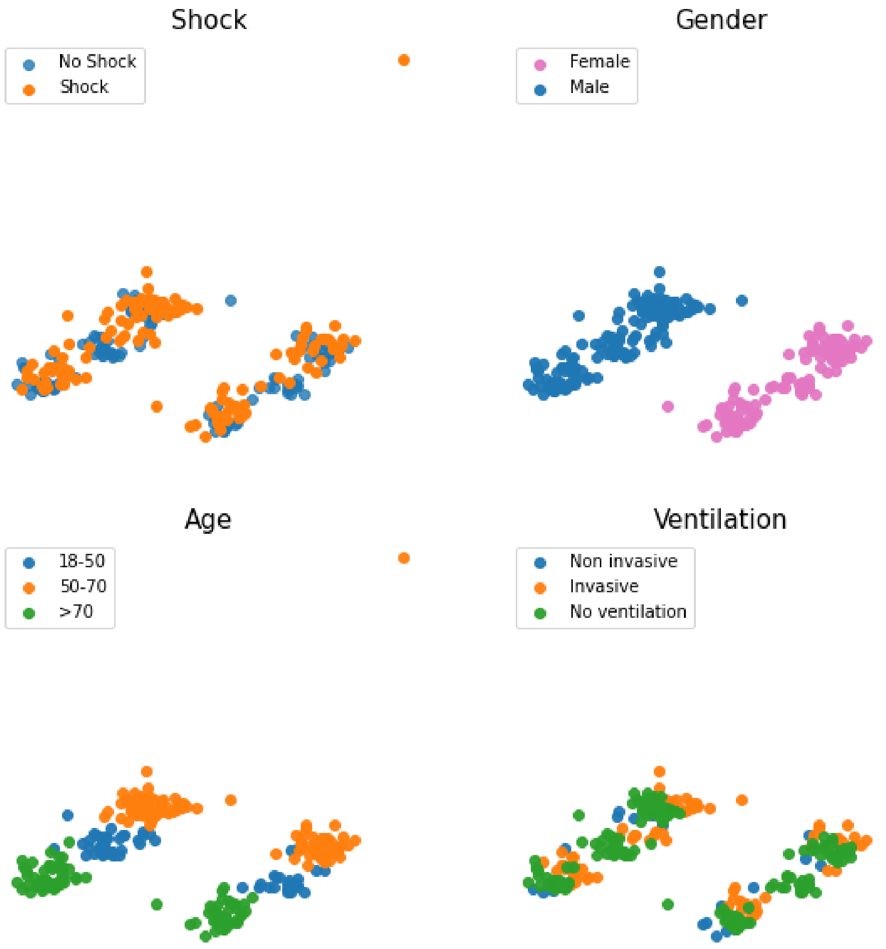
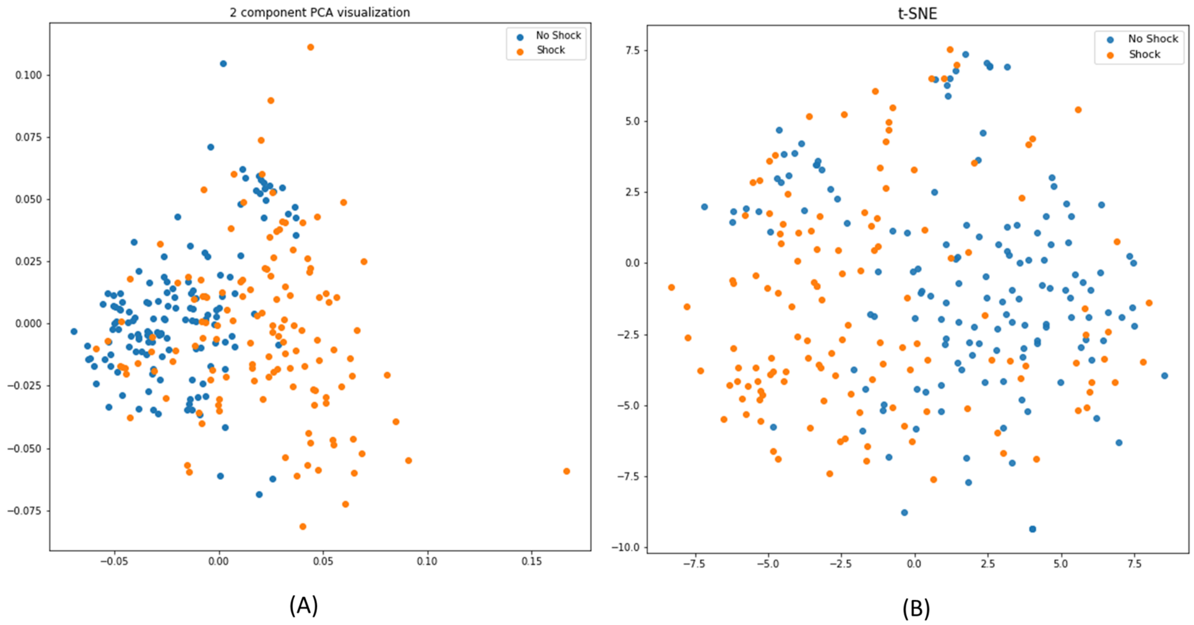


Figure 4.2: Visualization via PCA of data from Hospital São Francisco Xavier encoded by *VAE-LSTM-All* along with the identification of gender, age and ventilation categories.

Figure 4.2 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 and t-SNE of the encoded data obtained in this experiment can be observed in Figure 4.3.



(A)

(B)

Figure 4.3: Visualization data without static features from Hospital São Francisco Xavier encoded by *VAE-LSTM-All* via (A) PCA; (B) t-SNE.

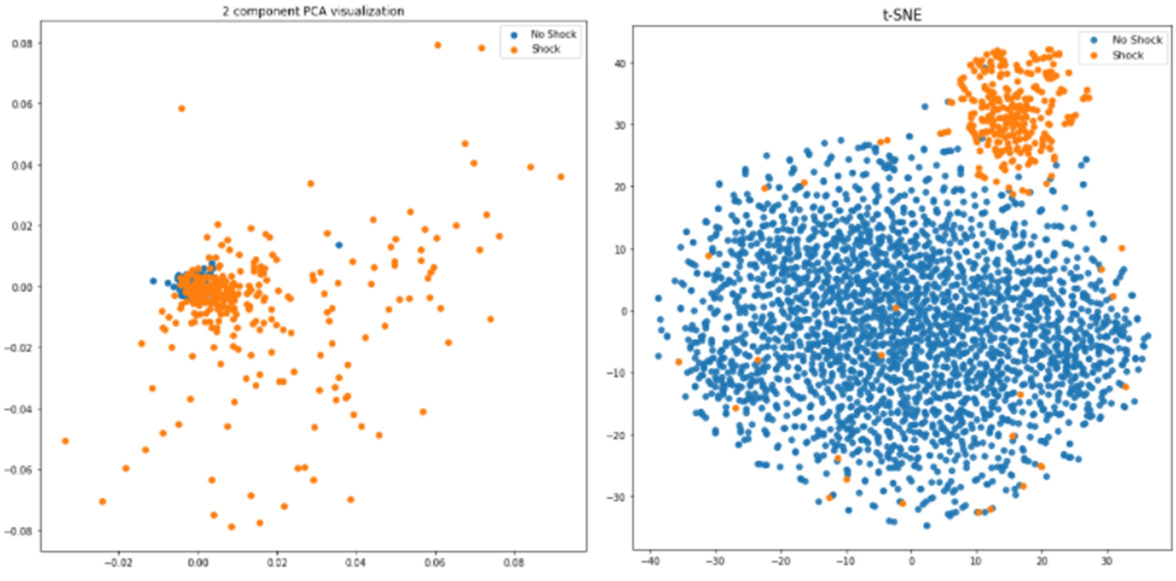
As it can be observed, although there are regions with higher concentrations of shock patients, such as the region where the first PCA component is higher than zero, there are no clear and well distinguished clusters of patients. However, PCA and t-SNE are only tools to help in the visualisation of the encoded data in a 2D space. Therefore, despite no clear clusters have been found in the PCA and t-SNE results, it might be possible to distinguish shock from non-shock patients resorting to clustering algorithms applied directly to the encoded data. With this in mind, the same three algorithms used in [62], were applied to the encoded data (Table 4.5). These algorithms were K-means, Spectral clustering and GMM. Similarly to the results obtained in [62], 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 4.4, although shock and non-shock clusters cannot be observed using PCA, in the results from the t-SNE two clear clusters which represent these populations can be defined. Regarding the performance of the clustering algorithms, once again the GMM algorithm outperformed the remaining ones,

Table 4.5: Performance of clustering algorithms using data from Hospital São Francisco Xavier and MIMIC-III.

Dataset	Model	Clustering algorithm	Recall	Precision	F1-score	AUC
Hospital São Francisco Xavier	VAE-LSTM-All	K-means	0,4661	0,826	0,5842	0,6759
		Spectral Clustering	0,4761	0,8141	0,5935	0,6786
		GMM	0,5613	**0,9067**	0,6931	0,7538
	VAE-LSTM-grouped	K-means	0,5262	0,7199	0,6068	0,6652
		Spectral Clustering	0,5262	0,7019	0,6006	0,6605
		GMM	**0,619**	0,8757	**0,7253**	**0,7686**
MIMIC-III	VAE-LSTM-All	K-means	0,3405	0,991	0,5068	0,67
		Spectral Clustering	0,3456	0,9934	0,5124	0,6726
		GMM	**0,9248**	0,9216	**0,923**	**0,9576**
	VAE-LSTM-grouped	K-means	0,361	0,9914	0,5287	0,6803
		Spectral Clustering	0,362	**0,9946**	0,53	0,6809
		GMM	0,717	0,9362	0,802	0,8554

reaching values above 0.9 across all evaluation metrics. However, the other two clustering algorithms showed a decrease on the model performance. These results indicate that the poor results obtained when using data from Hospital São Francisco Xavier might indeed be caused by the reduced size of the population. When comparing with the results reported in [62], the model *VAE-LSTM-All* achieved higher values in all metrics. This difference might be due to not only differences in the set of variables used but also in the different cohorts of interest. While in [62], the model proposed was trained only with septic patients which did not developed septic shock, in this work the models were trained with all patients which did not entered septic shock, regardless of having developed sepsis or not.



(A)

(B)

Figure 4.4: Visualization data without static features from MIMIC-III encoded by *VAE-LSTM-All* via (A) PCA; (B) t-SNE.

These models were then compared with T-LSTM models and, as in the supervised approach, no significant improvements were registered (Table 4.6). In fact, in some clustering algorithms the performance decreased considerably.

Table 4.6: Performance of clustering algorithms using T-LSTM models and data from Hospital São Francisco Xavier.

Model	Clustering algorithm	Recall	Precision	F1-score	AUC
VAE-TLSTM-All	K-means	0,5789	0,6581	0,616	0,6496
	Spectral Clustering	0,5714	0,6608	0,6129	0,6493
	GMM	0,5112	0,8947	0,6507	0,7276
VAE-TLSTM-grouped	K-means	**0,609**	0,9101	0,7297	0,7765
	Spectral Clustering	**0,609**	0,9204	**0,733**	**0,78**
	GMM	0,3759	**0,9803**	0,5434	0,6844

4.2.2 Anomaly scores

Besides clustering algorithms, the performance of anomaly scores was also explored. As previously mentioned, four different anomaly scores were analyzed: reconstruction error, Density-based score, Bhattacharyya distance score and Mahalanobis distance score. Regarding the reconstruction error, and similarly to the results obtained in [6] and [62], as a general rule shock patients registered higher reconstruction error as it can be observed in Figure 4.5.

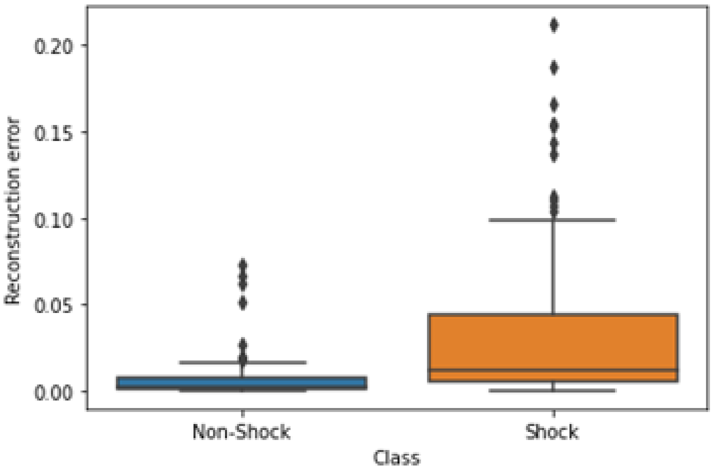


Figure 4.5: Boxplot of the Reconstuction error for shock and non-shock patients using model *VAE-LSTM-grouped*.

By defining an appropriate threshold, the distinction between shock and non-shock patients can be established reasonably well. In fact, similar results were observed for the remaining anomaly scores and the results obtained can be observed in Table 4.7.

Table 4.7: 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

As it can be observed in Table 4.7, 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, for *VAE-TLSTM-grouped* no significant improvement was observed. On the other hand, for the model *VAE-TLSTM-All* some improvements can be observed, specially in the reconstruction error score. Table 4.7 and Table 4.8 only present the AUC values of the anomaly scores analysed. In Appendix B, Table B.1 and Table B.2 present a more detailed report of the results obtained.

Table 4.8: 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

5

Conclusion

Contents

5.1 Main conclusions

In this work, supervised and unsupervised machine learning approaches for the prediction of septic shock were explored. These models were tested with data from two different datasets: Hospital São Francisco Xavier and MIMIC-III. For the supervised approach, three different classifiers were proposed. While one of the models only include a single LSTM layer for all data (*LSTM-All*), in the remaining two models the dynamic variables are first divided into four different groups and each group is handled by an independent LSTM layer. The difference between these latter two models is on the incorporation of the static data. While one model incorporates static data in every timestep of the LSTM layers (*LSTM-static-repeat*), the other only incorporates it in the last timestep (*LSTM-static-last*). 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 by including or excluding only one of the variable groups in the models' input and registering the impact on the performance. This importance analysis 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 encoder of each of the models proposed is similar to one of the supervised models developed. These encoders return the parameters of the data distribution in a latent space, which can then be sampled with a reparameterization trick. 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. This reinforces that the size of the cohort in study is an important factor that affect model performance.

Besides clustering algorithms, the prediction of septic shock through the use of four distinct 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 affected 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. This conclusion is very encouraging, specially in the medicine field where most data is unlabeled, since unsupervised techniques do not require labeled data.

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 unsupervised approaches.

5.2 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. Increasing the interpretability of these models can improve the clinicians' trust and confidence in the results obtained. 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 and help to define a better and more relevant set of variables for the prediction of sepsis.

In this work, the T-LSTM models registered contradicting results when compared with the results obtained in [39]. 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 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

Bibliography

[1] 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*, vol. 315, no. 8, pp. 801–810, 2016.

[2] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, “Learning to diagnose with LSTM recurrent neural networks,” *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2016.

[3] M. Nguyen, T. He, L. An, D. C. Alexander, J. Feng, and B. T. Yeo, “Predicting Alzheimer’s disease progression using deep recurrent neural networks,” *NeuroImage*, vol. 222, p. 117203, 2020. [Online]. Available: https://doi.org/10.1016/j.neuroimage.2020.117203

[4] 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*, pp. 219–228, 2018.

[5] J. Yao, M. L. Ong, K. K. Mun, S. Liu, and M. Motani, “Hybrid feature learning using autoencoders for early prediction of sepsis,” *2019 Computing in Cardiology Conference (CinC)*, vol. 45, 2019.

[6] I. Krissaane, K. Hampton, J. Alshenaifi, and R. Wilkinson, “Anomaly detection semi-supervised framework for sepsis treatment,” *2019 Computing in Cardiology Conference (CinC)*, vol. 45, 2019.

[7] 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.

[8] D. Berg and H. Gerlach, “Recent advances in understanding and managing sepsis [version 1; peer review: 3 approved],” *F1000Research*, vol. 7, no. 0, pp. 1–8, 2018.

[9] J. D. Faix, "Biomarkers of sepsis," *Critical Reviews in Clinical Laboratory Sciences*, vol. 50, no. 1, pp. 23–36, 2013.

[10] J. L. Vincent, S. M. Opal, J. C. Marshall, and K. J. Tracey, "Sepsis definitions: Time for change," *The Lancet*, vol. 381, 2013.

[11] B. B. Chakraborty RK, "Systemic inflammatory response syndrome," 2020, https://www.ncbi.nlm.nih.gov/books/NBK547669/, Updated : 2021-07-28.

[12] M. Cecconi, L. Evans, M. Levy, and A. Rhodes, "Sepsis and septic shock," *The Lancet*, vol. 392, no. 10141, pp. 75–87, 2018. [Online]. Available: http://dx.doi.org/10.1016/S0140-6736(18)30696-2

[13] M. M. Levy, M. P. Fink, J. C. Marshall, E. Abraham, D. Angus, D. Cook, J. Cohen, S. M. Opal, J. L. Vincent, and G. Ramsay, "2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference," *Critical Care Medicine*, vol. 31, 2003.

[14] K.-M. Kaukonen, M. Bailey, D. Pilcher, D. J. Cooper, and R. Bellomo, "Systemic inflammatory response syndrome criteria in defining severe sepsis," *New England Journal of Medicine*, vol. 372, 2015.

[15] A. E. Jones, S. Trzeciak, and J. A. Kline, "The sequential organ failure assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation," *Critical Care Medicine*, vol. 37, 2009.

[16] M. M. Churpek, A. Snyder, X. Han, S. Sokol, N. Pettit, M. D. Howell, and D. P. Edelson, "Quick sepsis-related organ failure assessment, systemic inflammatory response syndrome, and early warning scores for detecting clinical deterioration in infected patients outside theintensive care unit," *American Journal of Respiratory and Critical Care Medicine*, vol. 195, 2017.

[17] R. Salomão, B. Ferreira, M. Salomão, S. Santos, L. C. Azevedo, and M. Brunialti, "Sepsis: evolving concepts and challenges," *Brazilian Journal of Medical and Biological Research*, vol. 52, 04 2019.

[18] J. C. Mira, L. F. Gentile, B. J. Mathias, P. A. Efron, S. C. Brakenridge, A. M. Mohr, F. A. Moore, and L. L. Moldawer, "Sepsis pathophysiology, chronic critical illness and pics," *Critical care medicine*, vol. 45, 2017.

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

[20] "Surviving sepsis campaign," 2020, https://www.sccm.org/SurvivingSepsisCampaign/About-SSC/History, last checked : 2021-10-01.

[21] C. W. Seymour, F. Gesten, H. C. Prescott, M. E. Friedrich, T. J. Iwashyna, G. S. Phillips, S. Lemeshow, T. Osborn, K. M. Terry, and M. M. Levy, “Time to treatment and mortality during mandated emergency care for sepsis,” *New England Journal of Medicine*, vol. 376, 2017.

[22] 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*, vol. 34, 2006.

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

[24] L. Evans, A. Rhodes, W. Alhazzani, M. Antonelli, C. M. Coopersmith, C. French, F. R. Machado, L. Mcintyre, M. Ostermann, H. C. Prescott, C. Schorr, S. Simpson, W. J. Wiersinga, F. Alshamsi, D. C. Angus, and Y. Arabi, “Surviving sepsis campaign : international guidelines for management of sepsis and septic shock 2021,” *Intensive Care Medicine*, 2021. [Online]. Available: https://doi.org/10.1007/s00134-021-06506-y

[25] J. L. Vincent, J. Rello, J. Marshall, E. Silva, A. Anzueto, C. D. Martin, R. Moreno, J. Lipman, C. Gomersall, Y. Sakr, and K. Reinhart, “International study of the prevalence and outcomes of infection in intensive care units,” *JAMA - Journal of the American Medical Association*, vol. 302, 2009.

[26] J. L. Vincent, Y. Sakr, C. L. Sprung, V. M. Ranieri, K. Reinhart, H. Gerlach, R. Moreno, J. Carlet, J. R. L. Gall, and D. Payen, “Sepsis in european intensive care units: Results of the soap study,” *Critical Care Medicine*, vol. 34, 2006.

[27] S. Karlsson, M. Varpula, E. Ruokonen, V. Pettilä, I. Parviainen, T. I. Ala-Kokko, E. Kolho, and E. M. Rintala, “Incidence, treatment, and outcome of severe sepsis in ICU-treated adults in finland: The finnsepsis study,” *Intensive Care Medicine*, vol. 33, 2007.

[28] C. Pierrakos and J. L. Vincent, “Sepsis biomarkers: A review,” *Critical Care*, vol. 14, 2010.

[29] D. G. Goswami, L. F. Garcia, C. Dodoo, A. K. Dwivedi, Y. Zhou, D. Pappas, and W. E. Walker, “Evaluating the Timeliness and Specificity of CD69, CD64, and CD25 as Biomarkers of Sepsis in Mice,” *Shock (Augusta, Ga.)*, vol. 55, 2021.

[30] C. W. Seymour, J. N. Kennedy, S. Wang, C. C. H. Chang, C. F. Elliott, Z. Xu, S. Berry, G. Clermont, G. Cooper, H. Gomez, D. T. Huang, J. A. Kellum, Q. Mi, S. M. Opal, V. Talisa, T. Van Der Poll, S. Visweswaran, Y. Vodovotz, J. C. Weiss, D. M. Yealy, S. Yende, and D. C. Angus, “Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis,” *JAMA - Journal of the American Medical Association*, vol. 321, no. 20, pp. 2003–2017, 2019.

[31] C. Rhee and M. Klompas, “Sepsis trends: Increasing incidence and decreasing mortality, or changing denominator?” *Journal of Thoracic Disease*, vol. 2, 2020.

[32] D. C. Angus, W. T. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, and M. R. Pinsky, “Epidemiology of severe sepsis in the united states: Analysis of incidence, outcome, and associated costs of care,” *Critical Care Medicine*, vol. 29, 2001.

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

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

[35] M. Bauer, H. Gerlach, T. Vogelmann, F. Preissing, J. Stiefel, and D. Adam, “Mortality in sepsis and septic shock in Europe, North America and Australia between 2009 and 2019-results from a systematic review and meta-analysis,” *Critical Care*, vol. 24, 2020.

[36] 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*, vol. 412, pp. 197–215, 2020. [Online]. Available: https://doi.org/10.1016/j.neucom.2020.05.087

[37] 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*, vol. 9, pp. 1–14, 2019.

[38] M. M. Ghazi, M. Nielsen, A. Pai, M. J. Cardoso, M. Modat, S. Ourselin, and L. Sørensen, “Training recurrent neural networks robust to incomplete data: Application to Alzheimer’s disease progression modeling,” *Medical Image Analysis*, vol. 53, pp. 39–46, 2019. [Online]. Available: https://doi.org/10.1016/j.media.2019.01.004

[39] 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*, vol. Part F1296, pp. 65–74, 2017.

[40] 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*, vol. 38, pp. 1–15, 2018.

[41] 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*, vol. 75, pp. 1156–1172, 2018.

[42] 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*, vol. 47, pp. 112–129, 2021. [Online]. Available: https://doi.org/10.1016/j.euroneuro.2021.01.095

[43] A. D. Castelnuovo, M. Bonaccio, S. Costanzo, A. Gialluisi, A. Antinori, N. Berselli, L. Blandi, R. Bruno, R. Cauda, G. Guaraldi, I. My, L. Menicanti, G. Parruti, G. Patti, S. Perlini, F. Santilli, C. Signorelli, G. G. Stefanini, A. Vergori, A. Abdeddaim, W. Ageno, A. Agodi, P. Agostoni, L. Aiello, S. A. Moghazi, F. Aucella, G. Barbieri, A. Bartoloni, C. Bologna, P. Bonfanti, S. Brancati, F. Cacciatore, L. Caiano, F. Cannata, L. Carrozzi, A. Cascio, A. Cingolani, F. Cipollone, C. Colomba, A. Crisetti, F. Crosta, G. B. Danzi, D. D’Ardes, K. de Gaetano Donati, F. D. Gennaro, G. D. Palma, G. D. Tano, M. Fantoni, T. Filippini, P. Fioretto, F. M. Fusco, I. Gentile, L. Grisafi, G. Guarnieri, F. Landi, G. Larizza, A. Leone, G. Maccagni, S. Maccarella, M. Mapelli, R. Maragna, R. Marcucci, G. Maresca, C. Marotta, L. Marra, F. Mastroianni, A. Mengozzi, F. Menichetti, J. Milic, R. Murri, A. Montineri, R. Mussinelli, C. Mussini, M. Musso, A. Odone, M. Olivieri, E. Pasi, F. Petri, B. Pinchera, C. A. Pivato, R. Pizzi, V. Poletti, F. Raffaelli, C. Ravaglia, G. Righetti, A. Rognoni, M. Rossato, M. Rossi, A. Sabena, F. Salinaro, V. Sangiovanni, C. Sanrocco, A. Scarafino, L. Scorzolini, R. Sgariglia, P. G. Simeone, E. Spinoni, C. Torti, E. M. Trecarichi, F. Vezzani, G. Veronesi, R. Vettor, A. Vianello, M. Vinceti, R. D. Caterina, and L. Iacoviello, “Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST Study,” *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 30, pp. 1899–1913, 2020.

[44] V. Arvind, J. S. Kim, B. H. Cho, E. Geng, and S. K. Cho, “Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19,” *Journal of Critical Care*, vol. 62, pp. 25–30, 2021. [Online]. Available: https://doi.org/10.1016/j.jcrc.2020.10.033

[45] A. C. Chang, “Artificial intelligence and COVID-19: Present state and future vision,” *Intelligence-Based Medicine*, vol. 3-4, p. 100012, 2020. [Online]. Available: https://doi.org/10.1016/j.ibmed.2020.100012

[46] M. S. Mottaqi, F. Mohammadipanah, and H. Sajedi, "Contribution of machine learning approaches in response to SARS-CoV-2 infection," *Informatics in Medicine Unlocked*, vol. 23, p. 100526, 2021. [Online]. Available: https://doi.org/10.1016/j.imu.2021.100526

[47] 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*, vol. 14, pp. 561–569, 2021. [Online]. Available: https://doi.org/10.1016/j.jiph.2021.03.001

[48] M. F. Aslan, M. F. Unlersen, K. Sabanci, and A. Durdu, "CNN-based transfer learning–BiLSTM network: A novel approach for COVID-19 infection detection," *Applied Soft Computing*, vol. 98, p. 106912, 2021. [Online]. Available: https://doi.org/10.1016/j.asoc.2020.106912

[49] Y. Zhang, X. Yang, J. Ivy, and M. Chi, "Attain: Attention-based time-aware LSTM networks for disease progression modeling," *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2019-Augus, pp. 4369–4375, 2019.

[50] Q. Suo, F. Ma, G. Canino, J. Gao, A. Zhang, P. Veltri, and G. Agostino, "A multi-task framework for monitoring health conditions via attention-based recurrent neural networks," *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2017, pp. 1665–1674, 2017.

[51] M. B. Mayhew, B. K. Petersen, A. P. Sales, J. D. Greene, V. X. Liu, and T. S. Wasson, "Flexible, cluster-based analysis of the electronic medical record of sepsis with composite mixture models," *Journal of Biomedical Informatics*, vol. 78, no. November 2017, pp. 33–42, 2018. [Online]. Available: https://doi.org/10.1016/j.jbi.2017.11.015

[52] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks." *JMLR workshop and conference proceedings*, vol. 56, pp. 301–318, 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28286600%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5341604

[53] M. Chen, "MinimalRNN: Toward More Interpretable and Trainable Recurrent Neural Networks," 2017. [Online]. Available: http://arxiv.org/abs/1711.06788

[54] J. S. Calvert, D. A. Price, U. K. Chettipally, C. W. Barton, M. D. Feldman, J. L. Hoffman, M. Jay, and R. Das, "A computational approach to early sepsis detection," *Computers in Biology and Medicine*, vol. 74, 2016.

[55] H. J. Kam and H. Y. Kim, "Learning representations for the early detection of sepsis with deep neural networks," *Computers in Biology and Medicine*, vol. 89, 2017.

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

[57] A. E. Johnson, T. J. Pollard, L. Shen, L. W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, “MIMIC-III, a freely accessible critical care database,” *Scientific Data*, vol. 3, 2016.

[58] B. Wernly, B. Mamandipoor, P. Baldia, C. Jung, and V. Osmani, “Machine learning predicts mortality in septic patients using only routinely available ABG variables: a multi-centre evaluation,” *International Journal of Medical Informatics*, vol. 145, p. 104312, 2021. [Online]. Available: https://doi.org/10.1016/j.ijmedinf.2020.104312

[59] D. A. Kaji, J. R. Zech, J. S. Kim, S. K. Cho, N. S. Dangayach, A. B. Costa, and E. K. Oermann, “An attention based deep learning model of clinical events in the intensive care unit,” *PLoS ONE*, vol. 14, no. 2, pp. 1–17, 2019.

[60] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, “RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism,” *Advances in Neural Information Processing Systems*, pp. 3512–3520, 2016.

[61] M. B. Mayhew, B. K. Petersen, A. P. Sales, J. D. Greene, V. X. Liu, and T. S. Wasson, “Flexible, cluster-based analysis of the electronic medical record of sepsis with composite mixture models,” *Journal of Biomedical Informatics*, vol. 78, pp. 33–42, 2018. [Online]. Available: https://doi.org/10.1016/j.jbi.2017.11.015

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

[63] 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,” pp. 1–11, 2018.

[64] N. Ketkar and J. Moolayil, *Deep Learning with Python*, 2021.

[65] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” *IEEE Transactions on Neural Networks*, vol. 5, 1994.

[66] J. Rocca, “Understanding variational autoencoders (VAEs),” 2019, https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73, Last visited : 2021-10-20.

[67] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10265 LNCS, pp. 146–147, 2017.

[68] X. Wang, Y. Du, S. Lin, P. Cui, Y. Shen, and Y. Yang, “adVAE: A self-adversarial variational autoencoder with Gaussian anomaly prior knowledge for anomaly detection,” *Knowledge-Based Systems*, vol. 190, p. 105187, 2020. [Online]. Available: https://doi.org/10.1016/j.knosys.2019.105187

A

Variables

Table A.1: Dynamic variables used in this work along with the 4 groups considered and whether the variable is present in MIMIC-III dataset.

Variable Group	**Variable**	**Range**		**Present in MIMIC**
		min	**max**	
vitals	Temperature	34	42	Yes
	Dyastolic blood pressure	20	160	Yes
	Mean blood pressure	20	160	Yes
	Systolic blood pressure	40	260	Yes
	Heart rate	0	250	Yes
	Respiratory rate	0	150	Yes
	Peripheral capillary oxygen saturation	50	100	**No**
ABG variables	Calcium	0,2	7	Yes
	Chloride	60	170	Yes
	Bicarbonate	0	60	Yes
	Potassium	1,5	10	Yes
	Lactate	0	24	Yes
	Glucose	-	-	Yes
	Hemoglobin	1,5	20	Yes
	Hematocrit	5	70	Yes
	Bilirubins	0,1	50	Yes
	P|F ratio	29	600	**No**
	pH	6,6	7,9	Yes
clinical tests	Leukocytes	0	160	**No**
	Neutrophils	0	100	Yes
	Lymphocytes	0	100	Yes
	Basophils	0	100	Yes
	Monocytes	0	100	Yes
	Eosinophils	0	100	Yes
	Erythrocytes	0,8	7,2	Yes
	Average Globular Volume	-	-	Yes
	Platelets	-	-	Yes
	Prothrombin time	0	23	Yes
	C-Reactive Protein	-	-	**No**
	Magnesium	0	7	Yes
	Albumin	0	7	Yes
daily	Invasive Ventilation	-	-	**No**
	Non-invasive ventilation	-	-	**No**
	Hemodynamic SOFA	0	4	Yes
	Epinephrine	-	-	Yes
	Dopamine	-	-	Yes
	Norepinephrine	-	-	Yes
	SOFA score	0	24	Yes

B

Detailed Results

Table B.1: Detailed performance report of anomaly scores using LSTM models. L ≡ Latent space, F ≡ Feature space.

Dataset	Model	Score	Recall	Precision	F1-score	AUC
Hospital São Francisco Xavier	VAE-LSTM-All	RE	0,9248	0,6103	0,7322	0,6791
		Density (L)	0,8157	0,8228	0,8217	0,8292
		Density (F)	0,7368	0,7943	0,7638	0,779
		Bhattacharyya (L)	0,8195	0,8193	0,8192	0,8258
		Bhattacharyya (F)	0,7781	0,7441	0,7598	0,7632
		Mahalanobis (L)	0,8157	0,8159	0,8155	0,8222
		Mahalanobis (F)	0,7518	0,7937	0,7722	0,7849
	VAE-LSTM-grouped	RE	0,8671	0,7076	0,7784	0,7657
		Density (L)	0,8095	0,7337	0,7696	0,7683
		Density (F)	0,8671	0,5501	0,6731	0,6037
		Bhattacharyya (L)	0,8596	0,73	0,7873	0,7782
		Bhattacharyya (F)	0,8846	0,5895	0,69	0,6241
		Mahalanobis (L)	0,8195	0,7479	0,7814	0,7803
		Mahalanobis (F)	0,842	0,5799	0,6863	0,6354
MIMIC-III	VAE-LSTM-All	RE	0,4135	0,5654	0,4777	0,6877
		Density (L)	0,9074	0,9333	0,9201	0,9498
		Density (F)	0,7932	0,9277	0,8552	0,8929
		Bhattacharyya (L)	0,895	0,9634	0,928	0,9454
		Bhattacharyya (F)	0,8333	0,8823	0,8571	0,91
		Mahalanobis (L)	0,8796	0,9405	0,909	0,9364
		Mahalanobis (F)	0,8117	0,8855	0,847	0,8995
	VAE-LSTM-grouped	RE	0,3919	0,496	0,4379	0,6721
		Density (L)	0,8611	0,804	0,8315	0,9179
		Density (F)	0,6574	0,852	0,7421	0,8218
		Bhattacharyya (L)	0,8734	0,8323	0,8524	0,9262
		Bhattacharyya (F)	0,6481	0,8677	0,742	0,8181
		Mahalanobis (L)	0,8765	0,7675	0,8184	0,9223
		Mahalanobis (F)	0,608	0,8678	0,715	0,7984

Table B.2: Detailed performance report of anomaly scores using T-LSTM models and data from Hospital São Francisco Xavier. L ≡ Latent space, F ≡ Feature space.

Model	Score	Recall	Precision	F1-score	AUC
VAE-TLSTM-All	RE	0,9097	0,6914	0,7857	0,766
	Density (L)	0,7706	0,7946	0,7824	0,7926
	Density (F)	0,7255	0,7608	0,7424	0,756
	Bhattacharyya (L)	0,7856	0,8162	0,778	0,7874
	Bhattacharyya (F)	0,7518	0,7628	0,7568	0,7675
	Mahalanobis (L)	0,787	0,7806	0,785	0,7917
	Mahalanobis (F)	0,748	0,7126	0,7256	0,7271
VAE-TLSTM-grouped	RE	0,8996	0,6699	0,7667	0,7423
	Density (L)	0,837	0,7147	0,7708	0,7623
	Density (F)	0,8921	0,5443	0,676	0,5987
	Bhattacharyya (L)	0,7568	0,8113	0,7783	0,7945
	Bhattacharyya (F)	0,8796	0,5598	0,6831	0,6169
	Mahalanobis (L)	0,8195	0,7593	0,7879	0,7875
	Mahalanobis (F)	0,8971	0,5496	0,681	0,6047