Applied Intelligence (2023) 53:30691–30705
https://doi.org/10.1007/s10489-023-05124-z

Comparing artificial intelligence strategies for early sepsis detection in the ICU: an experimental study

Javier Solís-García[1] · Belén Vega-Márquez[1] · Juan A. Nepomuceno[1] · José C. Riquelme-Santos[1] · Isabel A. Nepomuceno-Chamorro[1]

Accepted: 21 October 2023 / Published online: 23 November 2023

Abstract

Sepsis is a life-threatening condition whose early recognition is key to improving outcomes for patients in intensive care units (ICUs). Artificial intelligence can play a crucial role in mining and exploiting health data for sepsis prediction. However, progress in this field has been impeded by a lack of comparability across studies. Some studies do not provide code, and each study independently processes a dataset with large numbers of missing values. Here, we present a comparative analysis of early sepsis prediction in the ICU by using machine learning (ML) algorithms and provide open-source code to the community to support future work. We reviewed the literature and conducted two phases of experiments. In the first phase, we analyzed five imputation strategies for handling missing data in a clinical dataset (which is often sampled irregularly and requires hand-crafted preprocessing steps). We used the MIMIC-III dataset, which includes more than 5,800 ICU hospital admissions from 2001 to 2012. In the second phase, we conducted an extensive experimental study using five ML methods and five popular deep learning models. We evaluated the performance of the methods by using the area under the precision-recall curve, a standard metric for clinical contexts. The deep learning methods (TCN and LSTM) outperformed the other methods, particularly in early detection tasks more than 4 hours before sepsis onset. The motivation for this work was to provide a benchmark framework for future research, thus enabling advancements in this field.

Keywords Sepsis · Early prediction · Onset · Machine learning · Deep learning · Comparative study

1 Introduction

Sepsis is caused by an immunological response to infection. Although bacterial infections are the primary cause of the clinical signs of sepsis, viruses, fungi, or parasites can also result in sepsis [1]. Regardless of the cause of infection, sepsis is a life-threatening condition that causes organ dysfunction due to a dysregulated host response to infection and constitutes a major global health problem [2]. For instance, in the U.S., sepsis affects more than 1.5 million people each year and leads to the death of 250,000 people [3]. Sepsis-related deaths account for an estimated 11 million of all global deaths [2]. In addition to causing alarmingly high mortality and morbidity rates, sepsis poses a substantial burden on the healthcare system's finances [2]; sepsis was responsible for 13% of total U.S. hospital costs but only 3.6% of hospital stays in 2013 [7]. Therefore, a systems medicine approach is necessary for early recognition of sepsis, to treat and predict the prognosis of the condition. With the introduction of electronic health records, genomic and molecular profiling technologies, novel machine learning (ML) algorithms, and modern drug discovery, a wide range of powerful tools is newly unavailable for implementing data-driven strategies [1].

Data-driven biomarker discovery has gained popularity over the past few decades and has the potential to overcome current obstacles. The goal of this approach is to mine and exploit health data by using artificial intelligence, because large amounts of clinical data (vital signs, medications, laboratory measurements, etc.) and health history are available in databases for patients with sepsis [8]. The variability in the data enables data-driven strategies to be implemented in searching for biomarkers in a holistic rather than a reduc-

✉ Javier Solís-García
jsolisg@us.es

Isabel A. Nepomuceno-Chamorro
inepomuceno@us.es

[1] Dpto. Lenguajes y Sistemas Informáticos, University of Seville, Av. Reina Mercedes sn, Sevilla 41012, Spain

tionist manner, focusing on only individual markers, such as hematological markers. ML algorithms can be leveraged to extract patterns from this large volume of available data, thereby addressing one of the primary challenges in sepsis disease: early recognition of sepsis. Timely antibiotic intervention is crucial, because mortality rates increase with every hour of delay, as highlighted in previous studies [9–11].

In recent years, artificial intelligence-based approaches have proven successful in predicting sepsis at the earliest possible time point in intensive care units (ICUs). However, to achieve this prediction, researchers have relied on either private databases provided by health institutions or public databases, such as Multiparameter Intelligent Monitoring in Intensive Care (MIMIC)-III [8] or eICU [12]. These public databases are freely accessible after their terms and conditions are accepted. A major challenges in using ICU health data is that sampling often is irregular, with different variables extracted from patients at varying time points. Therefore, researchers must preprocess the database, considering decisions such as how to label the data, patient inclusion criteria, and case control-onset matching. Several definitions of sepsis are used in the literature, and the selection of the definition used for labeling the data is crucial and affects the analysis in terms of the prevalence[1] of patients with sepsis and the level of difficulty of the prediction task, owing to earlier or later timing of sepsis [13]. On the one hand, more inclusive sepsis labeling indicates higher prevalence, e.g., in a class balance dataset, thus aiding in the training stability of ML models. On the other hand, adding more patients with sepsis could result in a cohort with less severe sepsis, comprising patients difficult to distinguish from non-septic ICU patients. In the literature, the label of severe sepsis or septic shock has rarely been used since a 2016 consensus statement. Instead, internationally acknowledged guidelines, called Sepsis-II [14] and Sepsis-III [15] are used. However, the lack of standardized reporting in the literature has led to low comparability and reproducibility, partly because of the lack of standardization in the use of metrics across studies. Recently published systematic reviews, such as [13] and [1], have reported existing approaches but have not provided experimental comparisons among them. In our work, we address these challenges by using a well-known database, MIMIC-III; the most recent definition of sepsis, Sepsis-III; and the same patient inclusion criteria and case-control alignment as [16]. To ensure comparability and reproducibility, we used several imputation strategies before training to enable a fair comparison. A limitation of this work is that the MIMIC-III database was recorded in one hospital in an 11 year period, thereby introducing potential bias, given its predominantly specific national and regional participant composition. However, the MIMIC database is the most frequently used data source to develop computational models for early prediction of sepsis in the ICU [13].

Importantly, the search for, and acquisition of, other sepsis patient databases is a highly complex process. To enrich our analysis, we used the freely accessible dataset published by PhysioNet in the 2019 challenge [17]. The description of the dataset and the results are reported in Appendix B. This dataset was provided after preprocessing of the data from 40,336 patients in two hospitals.

In summary, the novelty of our work has emerged from the need to establish a consistent experimental benchmark framework for early sepsis prediction to enable future advancements. Challenges include the lack of reproducibility of the results, the comparability of the results (variability in the metrics used), and the lack of transparency in the use of the data sets. To address this issue, this work included the following:

- An updated comparative analysis of the most relevant techniques for sepsis recognition, according to recent studies.
- Performance evaluation of several ML approaches, along with multiple imputation strategies before training, for early sepsis recognition by using the most frequently used data source in the sepsis literature, MIMIC-III. Furthermore, in Appendix B, we report the same analysis using another dataset published by Physionet for a challenge in 2019
- Performance analysis primarily evaluating models by using the area under the precision-recall curve (AUPRC), the best standard metric for clinical contexts.
- Open-source code available to support reproducibility of experiments at the following GitHub links: https://github.com/javiersgjavi/sepsis-review for experiments, and https://github.com/javiersgjavi/tabular-mimic-iii for database cleaning, mapping, and preprocessing. Our motivation was to provide a benchmark framework to assist researchers in developing new approaches for early sepsis prediction in the ICU.

The rest of the article is structured as follows: Section 2 describes the state of the art of ML applications for the early prediction of sepsis in the ICU; Section 3 describes the materials used and the methods proposed for the experimental study; Section 4 reports and discusses the results obtained; and Section 5 presents the conclusions and potential future work.

[1] Prevalence is a measure of the frequency of a disease or health condition in a population for a given point in time.

2 Related work

Artificial intelligence has been widely used to develop predictive models for the early detection of sepsis. Several studies have used ML-based algorithms as boosted tree models, including random forest (RF) or extreme gradient boosting (XGBoost). On the one hand, RF is a meta-classifier that fits several decision tree classifiers on different subsamples of the dataset, and averages the result to improve predictive performance and control overfitting [18]. RF has been used to predict sepsis in the ICU in [19] and [20]. In these studies, Sepsis-II was used to label the dataset obtained from the ML Healthcare System. However, to increase the training stability of the models, both studies balanced the dataset. In [21], the class imbalance problem was addressed via cost-sensitive learning, which assigned greater weight to the minority class data points according to their proportion. On the other hand, the XGBoost algorithm builds an additive model in a forward stage-wise fashion. This algorithm fits several regression trees to the negative gradient of the loss function and generates only one regression tree in a binary classification scenario [22]. Barton et al. [23] have used XGBoost to predict Sepsis-III as many as 48 hours ahead by using six vital signs, on the basis of data from both a private database, such as the UCSF Medical Center, and a public database, such as MIMIC-III.

In [24], ML algorithms were trained and tested to predict whether a patient would develop sepsis within the next 4 hours, on the basis of recordings from the prior 8 hours. The authors used the Sepsis-II definition and a private database from the Rabin Medical Center. Furthermore, they used support vector machines (SVMs) with linear, radial, and polynomial kernels, neural networks, and logistic regression. Logistic regression, a well-known technique in medical data analysis, has been used for predicting mortality or morbidity. SVM, which offers flexibility in the choice of penalties and loss functions, and can be scaled to large numbers of samples [25], has also been used in [26], with a fixed 4 hour horizon. In that study, the authors used SVM with a linear kernel and a private database from the Emory Healthcare System, and the Sepsis-II definition. In the clinical field, another fundamental aspect that must be considered is the interpretability of the obtained results. In that context, in [27], 24 ML models were used for early detection of sepsis. In addition, a final phase was incorporated to facilitate the interpretation of the results. Finally, a well-known algorithm called AdaBoost has been used in [28], with a private database from Emory University Hospital and a public database such as MIMIC-III.

Recently, deep learning techniques have been found to outperform traditional ML models in many fields, including early recognition of sepsis. However, few articles have provided a rigorous comparison between deep learning and other ML techniques in the literature. One exception is [29], in which feedforward neural networks using long short-term memory (LSTM) has been found to have better performance than a regression model with hand-crafted features [30]. Both studies used the Sepsis-II definition and the public MIMIC-III and MIMIC-II databases, respectively. In [31], the authors also used an LSTM with a recurrent neural network, and the public MIMIC-III database and Sepsis-II definition; this study is one of the few that have made code available.

As mentioned above, comparing approaches is challenging, because scenarios vary among articles, thus preventing meaningful conclusions from being drawn. For instance, in [32] and [33], a classical multilayer perceptron (MLP) and a convolutional neural network (CNN) are used, respectively, but the numbers of variables included in the models is not reported. Another issue is the lack of code availability to reproduce the results, although this aspect is essential for research transparency and reproducibility. In deep learning approaches, only a few studies, such as [16] and [31], have provided the code for data cleaning and analysis, according to our review of the literature [13]. Furthermore, these studies have used different sepsis definitions and databases, thus making comparison of their results challenging.

Finally, forecasting can be performed through either an online or offline training scenario. In the former, the model performs predictions at regular intervals during an ICU stay, by using all the data gathered until the time point of prediction (i.e., predictions in n-hour intervals). To stimulate early prediction, the onset of sepsis can be shifted to the past, as in the PhysioNet challenge. This challenge considers sepsis onset (timepoint labels) to occur 6 hours before the real onset [34]. In contrast, the latter training scenario enables the model to access the entire feature window of patient data. In the literature, most articles have used offline training. However, some exceptions exist, such as [28] and [35], in which online training was used. Furthermore, we found that most articles investigating sepsis onset prediction in the ICU have been based on data from the same center. However, several studies have used multi-center and multi-national data sources, such as [36], although the data sources are not available. An exception is the dataset published by PhysioNet for the 2019 challenge [17], wherein data from two hospitals were made public, along with additional data from another hospital used for validation of the proposed results, which have already been used in scientific publications such as [27].

3 Materials and methods

In this section, we describe how the dataset was created for our experimental study by using the public MIMIC-III database. We begin by explaining the labeling process, which includes defining the labels, and determining the patient inclusion criteria and the case-control onset matching. We also discuss our imputation strategies, and provide an overview of the supervised learning algorithms used in our analysis. A description of the prediction task is provided, including the hyperparameter configurations for training models, how the horizon analysis was conducted, and the performance measures used.

3.1 Data

3.1.1 Data description

In this study, we used version 1.4 of the MIMIC-III database [8], which contains more than 58,000 ICU hospital admissions involving more than 45,000 patients between June 2001 and October 2012. Access to this public database is granted after acceptance of the terms and conditions, which include completion of a recognized course in protecting human research participants and signing a data use agreement.

From the database, we selected 44 laboratory and vital parameters that are irregularly sampled, following the approach in [16]. The list of variables is reported in Section Supplementary Material Table A1.

3.1.2 Data preprocessing

The first step is deciding how to label the patients, to determine the patient inclusion criteria in the dataset, and to perform case-control onset matching. These aspects are explained in the following subsections. The irregularity of the time series and missing values were addressed by implementation of several imputation strategies. Finally, the dataset was normalized with minmax normalization. The data preprocessing steps are summarized in Fig. 1.

Sepsis-III definition We chose the most recent definition of sepsis, Sepsis-III [15], which requires the simultaneous presence of suspected infection and organ dysfunction. The suspected infection cohort was constructed according to the guidelines in [37]. We also followed the organ dysfunction definition given by [15], wherein an increase of at least 2 points in the SOFA score [38] indicates organ dysfunction. To determine the SOFA, we used a time window of 48 to 24 hours around a suspected infection. We selected the most recent definition of sepsis, because most studies have not provided their code. We used the code available in [16], in which the queries provided by [39] were considerably refined and extended to determine the Sepsis-III label hourly.

Patient inclusion criteria Following the recommendations of [40] and [41], we excluded negative laboratory measurements provided by the CareVue system and patients under the age of 15 years, i.e., pediatric cases. Sepsis cases were defined as those in patients with sepsis onset during the ICU stay, whereas case-controls included patients without sepsis onset, but with suspected infection or organ dysfunction. To ensure the validity of case-controls, we included only patients not labeled with any sepsis-related code from the ICD-9 diagnosis code of sepsis as controls.

After applying the above inclusion criteria, we identified 1,797 sepsis cases and 17,276 controls as an initial dataset. Patients who developed sepsis within the first 7 hours of the ICU stay were excluded. After case-control matching, we applied data filtering (described in the next subsection). Finally, we identified 570 sepsis cases and 5,618 controls with a mean time to sepsis onset in the ICU of 16.7 hours. Furthermore, seven patients with fewer than 10 observed measurements were removed.

Case-control onset matching Case-control alignment in a matching procedure is necessary to circumvent a trivial classification task. In [42], the performance of the method has been found to vary depending on the comparison between the window before sepsis onset and either the last window before discharge. A poorly aligned case-control pairing can result in a trivial classification problem.

To ensure a realistic class balance of no more than 10% [43], we assigned 10 unassigned random controls for each case. The simulated onset of control was defined as the hour after admission when the matched case met the sep-

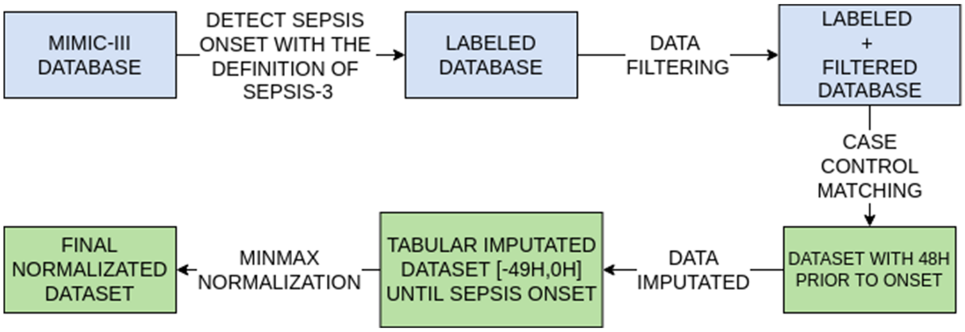
MIMIC-III DATABASE
DETECT SEPSIS ONSET WITH THE DEFINITION OF SEPSIS-3
LABELED DATABASE
DATA FILTERING
LABELED + FILTERED DATABASE
CASE CONTROL MATCHING
DATASET WITH 48H PRIOR TO ONSET
DATA IMPUTATED
TABULAR IMPUTATED DATASET [-49H,0H] UNTIL SEPSIS ONSET
MINMAX NORMALIZATION
FINAL NORMALIZATED DATASET

Fig. 1 Flowchart of the data prepocessing

sis requirements. Finally, 48 h of input data were extracted for each case and its corresponding matched controls before sepsis onset [16]. Further details regarding the constructed dataset are shown in Table 1.

3.2 Imputation strategies

To achieve optimal training of classification algorithms, converting real-world temporal data into temporal data with a uniform structure is crucial. This process is referred to as an imputation problem. We used various imputation techniques when applying signatures to irregular time series data.

According to [44], the following imputation techniques are available:

- **Linear interpolation** Imputation of a specific value is performed through linear interpolation. This interpolation is performed by using the value before and after the point of interest.
- **Forward filling** The value of the point of interest is imputed by forward propagating the value of the last observed point. If a missing value is present before any real value is sampled, the imputed value is 0.
- **Carry forward** This method is similar to forward filling; the main difference is that if a missing value appears before any actual value, it is set as the mean value of the measure for the patient.
- **Indicator imputation** Missing values in measurements are set to 0 in the time series and 1 in a binary indicator to mark the presence of a missing value. If the value is not missing, the value is 0 in the binary indicator.
- **Zero imputation** Missing values in measurements are set to 0.
- **Gaussian process adapter** This imputation method can connect irregularly sampled time series data to any black-box classifier that can be learned by using gradient descent.

We utilized all the above imputation techniques except the Gaussian process adapter. In a prior study [45], the Gaussian process adapter has been found to perform better when optimized jointly with the classifier rather than separately. Of note, the aim of our work was to provide a fair comparison among several models. Therefore, we selected imputation strategies ensuring that all models were evaluated by using the same four datasets (one for each imputation method).

Table 1 Characteristics of the dataset

Parameters	Values
Source	MIMIC-III database
Number of patients	6181
Clinical variables	44
Negative cases	5618
Positive cases	570
Prevalence	9.22%

3.3 Overview of supervised learning models

To experimentally compare the most commonly used methods for early recognition of sepsis during ICU stays, we divided the methods from Section 2 into ML models and deep learning models (neural network-based methods).

3.3.1 Machine learning models

- **AdaBoost classifier** This model belongs to the family of boosting algorithms framed within the ensemble models. As in any ensemble model, the objective is to obtain a higher accuracy rate through the combination of several classifiers [46].
- **Linear SVC** This model is a faster implementation of support vector classification for the case of a linear kernel. SVM models are learning tools that recognize patterns and decision boundaries within a dataset [47] and have been widely used in healthcare fields [48]. The main objective of this type of model is to build hyper-planes in a multidimensional space that are able to separate different classes of the dataset.
- **Logistic regression** Logistic regression is a statistical regression model used to determine the relationship between one or more independent variables and a binary outcome feature [49]. This method has many similarities with linear regression, but logistic regression uses an iterative maximum likelihood method to fit the final model.
- **Random forest classifier** this model belongs to the family of bagging algorithms framed within the ensemble models. One of the main advantages of using trees in classification tasks is that they are models that are easily understood by humans; however, the hit rates are lower than those of other models. To overcome this disadvantage, several trees are usually aggregated to achieve a higher hit ratio.
- **XGBoost** XGBoost is an updated version of the gradient boosting decision tree (GBDT) algorithm [50]. This algorithm consists of a sequential assembly of decision trees. These trees are added sequentially to learn from the results of previous trees and correct the errors produced, until the errors can no longer be corrected (gradient descent).

3.3.2 Deep learning models

- **MLP** Multilayer perceptron is a fully connected feedforward type of neural network; i.e., the backpropagation technique is used for its training process. This type of network consists of three main layers: an input layer that acts as the initial data entry point, several intermediate layers that are responsible for training the network, and finally an output layer with a single neuron that is responsible for providing the final prediction [51].
- **CNN** CNNs were first introduced in [52], in which Fukushima described the concept of convolutional layers. This type of layer applies a convolution operation to merge two sets of information and mimics the behavior of the human eye toward any stimulus [53].
- **LSTM** Long short-term memory (LSTM) networks are a special type of recurrent neural network that can model long-term temporal dependencies without neglecting temporal dependencies that may exist in the short term [54]. LSTM was introduced in [55] and is able to model temporal dependencies on large horizons.
- **TCN** Temporal CNNs are a special type of CNN adapted to work with temporal data. TCN was introduced in [56] as a CNN with special characteristics: convolutions are causal to prevent information loss, and the architecture can process a sequence of any length and map it to an output of the same length [57].
- **GRU** Gated recurrent units are a type of deep neural network for solving the problems of long-term memory and gradient in backpropagation by following the principle of nonlinear function approximation [58]. This type of network emerged as a simplification of the already implemented LSTM networks to avoid overfitting. The novelty of these networks is in the neurons of the hidden layers, in this case combining the forget gate and input gate into a single update gate.

3.4 Prediction task

3.4.1 Model validation and parameterization

According to Moor et al. in [16], we performed random splitting, using 80% of the samples for training, 10% for testing, and 10% for validation, as shown at the top of Fig. 2. For hyperparameter tuning, we used the parameter sampler provided by the Scikit-learn model selection library, which generated parameters sampled from specified distributions [59].

To define the search space for the model parameters in this study, we relied on the parameterization presented in a well-known review [60], in which various deep learning models applied to time series data were explored. To fine-tune the parameters, a grid search approach was used on the hyperparameters of each model. Subsequently, the configuration that yielded the best results was selected. The values tested for each hyperparameter in each model, along with the optimal values discovered, are shown in Table 2 for ML models and in Table 3 for deep learning models.

To ensure a fair comparison, we aimed to explore as many parameter combinations as possible for each model while still maintaining a reasonable total training time. To achieve this, we used a random search approach to sample from the parameter space, conducting a total of 25 iterations per classification and imputation model.

3.4.2 Settings of prediction task

We used offline training combined with a horizon evaluation, because this is the most frequently used setting [13], and early detection of sepsis is crucial.

In the offline scenario, the feature window has two cases. In the case of patients with sepsis, the feature window extends from hospital admission to the onset of sepsis. In the case of controls, the window endpoint corresponds to a matched onset, achieved through the case-control matching process described in Section 3.1.2.

In a horizon analysis, all input data gathered until n hours before onset are provided to the model to evaluate its predictive performance at the n hours horizon. Instead of optimizing separate models for each individual horizon, which would result in n different models, we optimized a single model using all available training data. We then challenged this model by gradually restricting access to the data closer to the sepsis onset, as illustrated at the bottom of Fig. 2. Ultimately, we used a horizon of 7 hours before sepsis onset, following the approach outlined in [41], to assess the model's ability to recognize sepsis at an early stage.

3.4.3 Performance measures

The early prediction of sepsis presents a challenge, because it is an imbalanced classification problem, in which the distribution of samples across the known classes is biased or skewed. Simple metrics such as accuracy do not effectively capture the performance of such problems [61]. Previous studies on sepsis prediction have often reported the area under the receiver operating characteristic curve (AUROC or AUC). However, this metric may be influenced by the prevalence of the disease and is less informative in highly imbalanced datasets [62]. In line with [63], we opted to use the precision-recall curve (AUPRC), because it provides more informative results for binary classifiers in unbalanced datasets. Moreover, AUPRC is recommended for conditions with low prevalence [64] and is considered the preferred met-

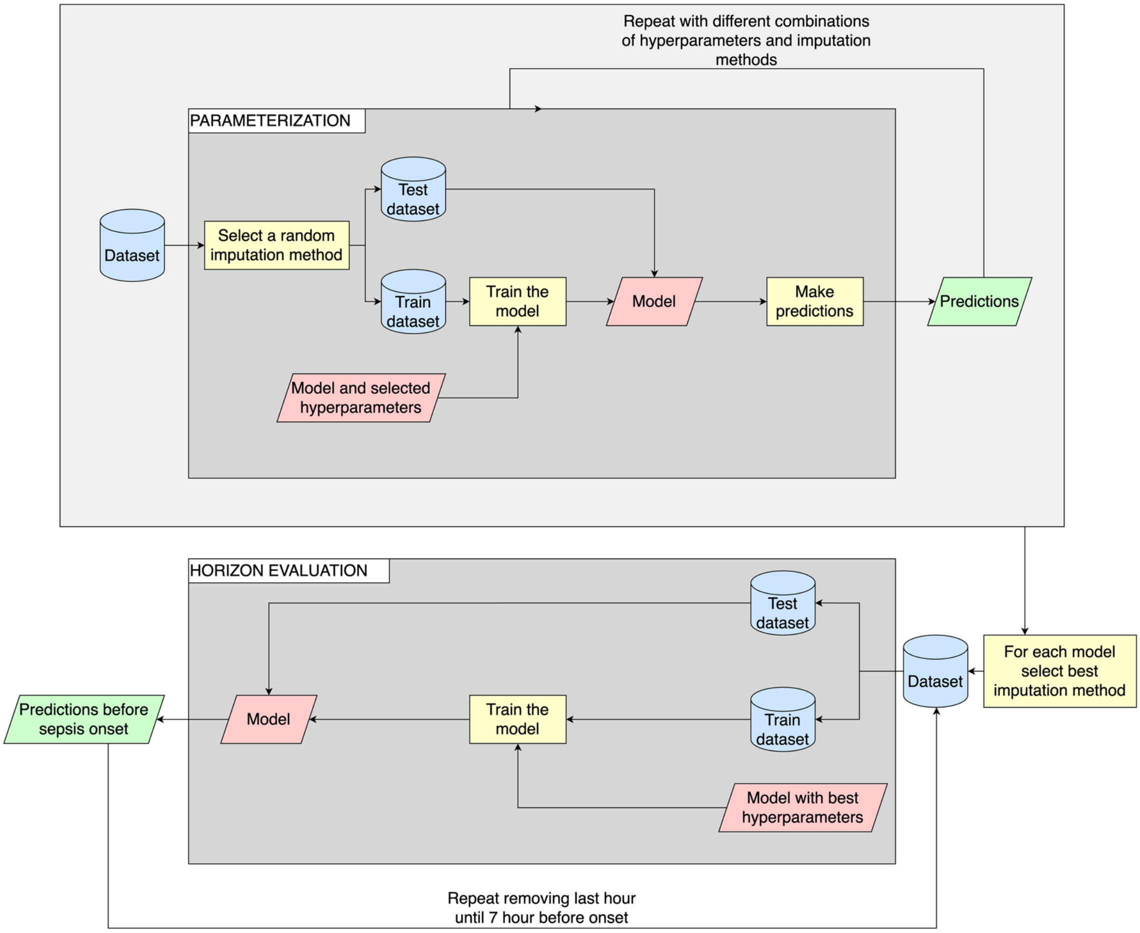
Repeat with different combinations of hyperparameters and imputation methods
PARAMETERIZATION
Dataset
Select a random imputation method
Test dataset
Train dataset
Train the model
Model
Make predictions
Predictions
Model and selected hyperparameters
HORIZON EVALUATION
Test dataset
Dataset
For each model select best imputation method
Predictions before sepsis onset
Model
Train the model
Train dataset
Model with best hyperparameters
Repeat removing last hour until 7 hour before onset

Fig. 2 Search for best hyperparameters and model imputation method

Table 2 Parameter grid used for ML models

Model	Parameters	Values	Best value
LinearSVC	n max iter	500, 1000, 2000	2000
	loss	hinge, squared_hinge	squared_hinge
	penalty	l2	12
XGBClassifier	eta	0.1, 0.3, 0.6	0.3
	max depth	2, 4, 6, 8, 10	10
	grow policy	depthwise, lossguide	depthwise
	sampling method	uniform, gradient-based	gradient based
LogisticRegression	penalty	l2, none	l2
	solver	lbfgs, sag, saga	sag
AdaBoostClassifier	n estimators	25, 50, 100	100
	learning rate	0.8, 0.9, 1.0	0.9
RandomForestClassifier	n estimators	100, 300, 600	100
	max depth	2, 4, 6, 8, 10	8
	min samples split	2, 4, 6, 8	6
	min samples leaf	1, 3, 5, 7	7

The last column specifies the values for the optimal model. For each model, parameters were randomly selected with 25 iterations

Table 3 Parameter grid used for deep learning models

Model	Parameters	Values	Best value
MLP		10^{-3}, 10^{-2}	
		32, 64	
	learning rate	(8), (8, 16), (16, 8), (8, 16, 32),	10^{-3}
	batch size	(32, 16, 8), (8, 16, 32, 16, 8),	64
	hidden layers	(32, 64), (64, 32), (32, 64, 128),	(32, 64, 128, 64, 32)
		(128, 64, 32), (32, 64, 128, 64, 32)	
TCN	learning rate	10^{-3}, 10^{-2}	10^{-3}
	batch size	32, 64	64
	nb stacks	1, 3	1
	nb filters	32, 64	64
	dilations	(1, 2, 4, 8), (1, 2, 4, 8, 16)	(1, 2, 4, 8)
	kernel size	3, 6	6s
	return sequences	True, False	True
	tcn dropout	0, 0.2	0
	dense dropout	0, 0.2	0
	dense layers	MLP hidden layers values	(128, 64, 32)
CNN	learning rate	10^{-3}, 10^{-2}	10^{-3}
	batch size	32, 64	32
	layers	1, 2, 3	3
	units	16, 32, 64	16
	pool size	0, 2	0
	dense dropout	0, 0.2, 0.4	0
	dense layers	MLP hidden layers values	(128, 64, 32)
GRU	learning rate	10^{-3}, 10^{-2}	10^{-3}
	batch size	32, 64	64
	layers	1, 2, 4	2
	units	32, 64, 128	32
	return sequence	True, False	True
	recurrent dropout	0, 0.2, 0.4	0.2
	dense dropout	0, 0.2, 0.4	0.2
	dense layers	MLP hidden layers values	32, 64, 128
LSTM	learning rate	10^{-3}, 10^{-2}	10^{-3}
	batch size	32, 64	64
	layers	1, 2, 4	1
	units	32, 64, 128	128
	return sequence	True, False	True
	recurrent dropout	0, 0.2	0
	dense dropout	0, 0.2	0
	dense layers	MLP hidden layers values	(32, 16, 8)

The last column specifies the values for the optimal model. For each model, parameters were randomly selected with 25 iterations

Table 4 Best results by imputation method

Imputation	Model	Time	AUPRC	AUROC	Accuracy
Carry forward	TCN	0.4597	**0.7553**	0.9333	0.9315
Forward filling	LSTM	1.0676	0.7336	**0.9491**	*0.925*
Indicator imputation	GRU	1.276	*0.7089*	*0.9097*	0.9347
Linear interpolation	TCN	**0.3646**	0.7298	0.9472	0.9266
Zero imputation	GRU	*1.5048*	0.7134	0.9272	**0.9396**

The best results were obtained with the methods for each imputation strategy. For each metric, the best result is in bold, and the worst is in italics

ric in clinical contexts [65].

$$p = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{1}$$

$$r = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2}$$

$$\text{AP} = \int_0^1 p(r)dr \tag{3}$$

Specifically, AUPRC measures the maximum value when the model successfully identifies all positive examples without any false positive detection. Because precision and recall do not consider true negatives in their calculations, this metric focuses solely on predicting positive samples. Consequently, it is well-suited for unbalanced datasets and clinical contexts when the aim is correctly identifying positive cases [66]. In summary, the AUPRC is defined by using the precision metric (1) and the recall (2). However, in evaluation of models with this metric, not working with the graphical representations of the curves but instead translating them into a numerical value is recommended. This process is achieved by approximating the AUPRC with the average precision (AP) value, as defined in (3).

Although a baseline AUROC value is typically set to 0.5, interpreting the baseline AUPRC can be misleading, because it is influenced by the ratio of positive examples to the dataset size. In our case, the baseline AUPRC is determined by the prevalence of sepsis cases in our preprocessed MIMIC-III database, corresponding to a value of 0.0922. Any higher value achieved by our models would indicate an improvement over baseline.

4 Results and discussion

This section reports and discusses the performance obtained from the conducted experiments, which involved various ML and deep learning methods, along with the imputation strategies discussed in Section 3. The results are presented in terms of AUPRC together with other metrics used in the sepsis literature, such as AUROC and accuracy, although the other metrics are less informative in the case of unbalanced data.

Importantly, the experiments were performed on a computer system with the following specifications: Intel(R) Core(TM) i7-8700 3.20GH as CPU, NVIDIA GeForce 2080 as GPU 1 and Titan Xp as GPU 2 and 4X16 GB (64 GB) - DDR4 as memory RAM.

4.1 Classification task

Tables 4 and 5 present a comprehensive comparison of the best result achieved for each imputation method and ML

Table 5 Best results obtained by model

Model	Imputation	Time	AUPRC	AUROC	Accuracy
TCN	Carry forward	0.4597	**0.7553**	0.9333	0.9315
CNN	Linear interpolation	0.0849	0.7208	0.9279	0.9331
MLP	Linear interpolation	0.0896	0.636	0.89	*0.9184*
GRU	Carry forward	0.2971	0.7235	0.9147	0.9413
LSTM	Forward filling	*1.0676*	0.7336	**0.9491**	0.925
LinearSVC	Linear interpolation	0.383	0.4707	0.7623	0.9396
XGBClassifier	Linear interpolation	0.2019	0.6067	0.8297	**0.956**
LogisticRegression	Linear interpolation	0.212	0.4953	0.7714	0.9429
AdaBoostClassifier	Linear interpolation	0.7513	0.511	0.8006	0.9429
RandomForestClassifier	Linear interpolation	**0.011**	*0.3942*	*0.6785*	0.9331

For each metric, the best result is in bold, and the worst is in italics method, respectively. The tables report various metrics, with the best result in bold and the worst result in italics.

Table 4 shows the best performance obtained for each imputation strategy. The deep learning methods, particularly the TCN model, consistently yielded the best results for each imputation method. Interestingly, none of the classical ML algorithms achieved the top performance in this comparison. Notably, the carry-forward imputation method produced the highest AUPRC score. A more detailed analysis of the distribution of results obtained by each imputation method can be found in a supplementary figure (Fig. A6). This figure presents violin plots of each imputation method by model, considering the AUPRC. The deep learning models exhibit superior performance, whereas the other algorithms show a lower standard deviation. This disparity is attributable primarily to the sensitivity of deep learning models to training parameters.

Focusing on the methods rather than the imputation strategies, Table 5 reports the best result per method independently of the imputation strategy, i.e., it shows the top-performing result achieved among all experiments conducted for each specific method. In terms of AUPRC, the best result was obtained with TCN, whereas the worst result was obtained with RF, in terms of both AUROC and AUPRC. Thus, RF may not be suitable for addressing this particular problem. Notably, linear interpolation emerged as the preferred imputation method in 7 of the 10 models used, thus indicating its effectiveness. However, Tables 4 and 5 demonstrate the interdependency between imputation strategies and classifier algorithms. For instance, when LI is the chosen imputation strategy, Table 4 suggests that the optimal model is TCN. Conversely, the TCN model is prioritized, Table 5 indicates that the preferred imputation strategy is CF. These results highlight the interconnected nature of imputation strategies and classifier algorithms in achieving optimal performance.

Because the AUPRC metric is widely recommended in clinical settings, a detailed analysis of this metric was conducted in Section Supplementary Material Figs. A1 and A2. These figures illustrate the distribution of AUPRC scores in box plots, differentiating between the imputation method and the model. First, Fig. A1 demonstrates that although the carry forward imputation method yielded the best individual result, on average, linear interpolation had the highest performance. This finding is consistent with the observations in Table 5. Second, as shown in Fig. A2, deep learning models outperformed the others in terms of AUPRC. One general advantage of deep learning methods, in the context of biased databases such as data from real-world healthcare systems, is the possibility of fine-tuning them to slightly different data associated with the same problem [67]. However, the substantial effects of parameter optimization on these models must be noted, as indicated by the wide dispersion of results for each model.

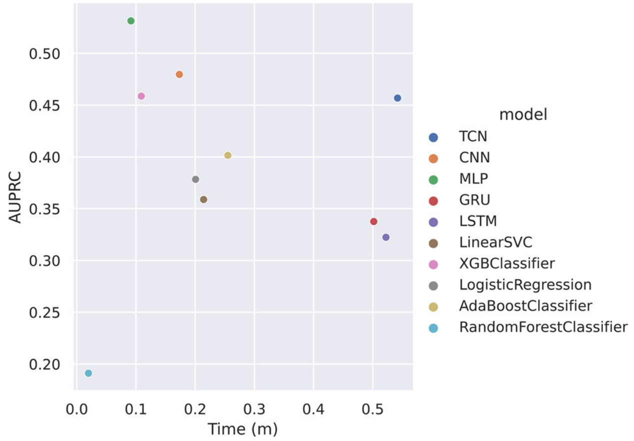

Fig. 3 Dot plot representing the average computation time in minutes per model on the x-axis and the average AUPRC on the y-axis. Each point represents a different classification model. Models located in the upper left corner are best, whereas those located in the lower left corner are worst in terms of computation time and AUPRC

Finally, in the Supplementary Material Tables A2 and A3, the mean, standard deviation, maximum, and minimum values of all conducted experiments provide a comprehensive overview of the results.

4.2 Computation time

Given the critical nature of the study data, the computation time is a fundamental aspect to be considered. Figure 3 shows a dot plot in which the x-axis represents the average training time of each classification model, and the y-axis is the average AUPRC. Models located higher on the left would be optimal, because of their very short time and very low prediction error; however, models located at the lower right corner would be the most disadvantageous. Thus, the figure shows that the best model, on the basis of AUPRC and time, is MLP, whereas the worst model, considering these two variables, is LSTM. Of note, the position of RF in the lower left corner indicates that its training time is very short, and the results in terms of prediction error are very high. The results for this model suggest that the problem is that the classifier cannot converge in the training process, and the training process consequently stops before finding an optimal solution.

Finally, the total time spent on the parameter optimization process, i.e., the number of parameter combinations tested, with each model limited to 150 combinations, is illustrated in Fig. A5.

4.3 Classification task before sepsis onset

Figure 4 depicts the predictive performance for different time horizons. The x-axis represents the prediction horizon in hours before sepsis onset, whereas the y-axis indicates

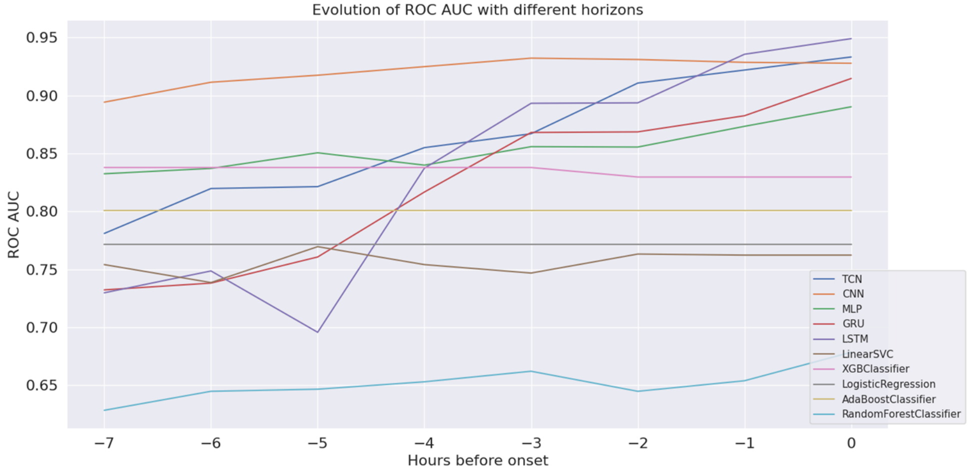

Fig. 4 Predictive performance for the different time horizons. The x-axis indicates the prediction horizon hours before sepsis onset. The y-axis indicates AUPRC

the AUPRC. The CNN model consistently outperformed the other methods, including classical ML approaches, for early sepsis detection. Both TCN and LSTM also exhibited superior performance to the other methods, particularly for tasks involving early detection more than 4 hours before sepsis onset. In contrast, RF had the lowest performance. Additional measures, which provide less informative insights for class imbalance problems, are shown in Supplementary Material Figs. A3 and A4.

5 Conclusions

In this study, we performed a comparative analysis of ML algorithms for the early prediction of sepsis in the ICU because the early diagnosis of sepsis has been shown to decrease treatment delays, increase appropriate care, and decrease mortality. However, in this study, the economic consequences of forecasting the onset of sepsis were not analyzed. Several studies have provided some analysis in this regard, such as [68]. Briefly, we reviewed the literature and selected the MIMIC-III database [8] because of its widespread use. Although the MIMIC-III data agreement guidelines state that code using MIMIC-III must be published, few studies have made their code available. However, making the code publicly accessible does not necessarily make it easily reproducible. In our experience, we had difficulty in reproducing findings from other studies. Therefore, a good experimental comparison of ML algorithms for the early detection of sepsis is mandatory. We implemented our own code, and we encourage others to use container technologies such as Docker to facilitate reproducibility.

To perform the experimental comparison, we made several decisions, including the sepsis definition, data filtering, and imputation strategies for building the dataset. We used offline training combined with a horizon evaluation, which determines how early the model can recognize sepsis, and we used a horizon of 7 hours before sepsis onset. Case-control alignment was used in a matching procedure to avoid a trivial classification task, with 10 random unassigned controls assigned for each case. We also applied random training-test splitting for model validation and used the AUPRC as well as ROC curves and accuracy for evaluation. We conducted an extensive experimental study using five ML methods and five neural network-based models or deep learning architectures. Neural network-based models were found to outperform traditional ML methods for the early prediction of sepsis.

In summary, this study provides valuable insights into the use of ML algorithms for the early prediction of sepsis in ICU patients and highlights the importance of reproducibility in research. The main conclusions of this experimental study can be summarized as follows:

1. To ensure a fair comparison, we used as many parameter combinations as possible for each model without compromising the total training time. We used a random search of parameters with a total of 25 iterations per classification and imputation model.
2. Linear interpolation was found to be the best imputation method in 7 of the 10 models used, thus highlighting its efficiency.
3. The best results were obtained by using TCN followed by LSTM, GRU, and CNN, thus demonstrating that deep learning models outperform other ML models. However, the importance of parameter optimization in these models is crucial, as shown by the wide dispersion of results for each model.
4. The computational time of most traditional classification models showed a clear downward trend, and their dispersion was usually lower than DL methods, thereby indicating that almost all experiments using these types of models had a running time very close to the average.
5. In the horizon prediction, both the TCN and LSTM models significantly outperformed the other methods,

particularly for early detection tasks more than 4 hours before the onset of sepsis. In contrast, the RF model had the worst performance.

In this study, we performed a fair comparison among several models, and used several imputation strategies before training. In future work, a combination of multitask Gaussian processes imputation together with a neural network in one end-to-end trainable framework may be analyzed. This scheme optimizes the imputation and classifier optimization process together. Finally, we suggest applying generative models in imputation schemes, a new trend in this field.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10489-023-05124-z.

Acknowledgements We thank Karsten Borgwardt and Michael Moor for availability to answer our questions.

Author Contributions JSG and BVM carried out the analysis and prepared the scripts. JN and INCH conceived the methods. JRS and INCH led the project. JRS and INCH wrote the manuscript. All authors read and approved the final manuscript. Finally, all authors certify that they have participated sufficiently in the work to take public responsibility for the content.

Funding Funding for open access publishing: Universidad de Sevilla/CBUA. This research was supported partly by the FEDER Andalucía US-1263341 and by the Spanish Ministry of Science and Innovation under grant PID2020-117954RB-C22.

Data availability and access The data set used in our work can be obtained from version 1.4 of the MIMIC-III database [8]. Reproducibility of the experiments can be performed with the code from our comparative analysis, available at https://github.com/javiersgjavi/sepsis-review. Preprocessing steps to handle the data set are provided at https://github.com/javiersgjavi/tabular-mimic-iii. Of note, the dataset is not stored in the GitHub link and must be downloaded from the MIMIC-III database. However, the data used in Appendix B can be obtained without a need for registration on the PhysioNet website: https://physionet.org/content/challenge-2019/1.0.0/

Declarations

Ethical and informed consent for data used The MIMIC-III database is freely accessible after acceptance of the terms and conditions, and completion of a recognized course in protecting human research participants.

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

Credits The logos incorporated in this article were acquired from the Flaticon website and designed by Freepik.

References

1. Liu AC, Patel K, Vunikili RD, Johnson KW, Abdu F, Belman SK, Glicksberg BS, Tandale P, Fontanez R, Mathew OK, Kasarskis A, Mukherjee P, Subramanian L, Dudley JT, Shameer K (2020) Sepsis in the era of data-driven medicine: personalizing risks, diagnoses, treatments and prognoses. Brief Bioinform 21(4):1182–1195. https://doi.org/10.1093/bib/bbz059
2. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, Colombara DV, Ikuta KS, Kissoon N, Finfer S, Fleischmann-Struzek C, Machado FR, Reinhart KK, Rowan K, Seymour CW, Watson RS, West TE, Marinho F, Hay SI, Lozano R, Lopez AD, Angus DC, Murray CJL, Naghavi M (2020) Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. The Lancet 395(10219):200–211. https://doi.org/10.1016/S0140-6736(19)32989-7
3. Center for Disease Control. CDC Urges Early Recognition, Prompt Treatment of Sepsis (2019)
4. Hotchkiss RS, Moldawer LL, Opal SM, Reinhart K, Turnbull IR, Vincent J-L (2016) Sepsis and septic shock. Nat Rev Dis Primers 2(1):1–21. https://doi.org/10.1038/nrdp.2016.45
5. Kaukonen K-M, Bailey M, Suzuki S, Pilcher D, Bellomo R (2014) Mortality Related to Severe Sepsis and Septic Shock Among Critically Ill Patients in Australia and New Zealand, 2000–2012. JAMA 311(13):1308–1316. https://doi.org/10.1001/jama.2014.2637
6. Torio, C.M., Moore, B.J.: National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2013. In: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs, Rockville (MD) (2006)
7. Paoli CJ, Reynolds MA, Sinha M, Gitlin M, Crouser E (2018) Epidemiology and costs of sepsis in the united states–an analysis based on timing of diagnosis and severity level. Crit Care Med 46(12):1889
8. Johnson AEW, Pollard TJ, Shen L, Lehman L-wH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG (2016) MIMIC-III, a freely accessible critical care database. Scientific Data 3(1):160035. https://doi.org/10.1038/sdata.2016.35
9. Ferrer R, Martin-Loeches I, Phillips G, Osborn TM, Townsend S, Dellinger RP, Artigas A, Schorr C, Levy MM (2014) Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. Crit Care Med 42(8):1749–1755. https://doi.org/10.1097/CCM.0000000000000330
10. Weiss SL, Fitzgerald JC, Balamuth F, Alpern ER, Lavelle J, Chilutti M, Grundmeier R, Nadkarni VM, Thomas NJ (2014) Delayed antimicrobial therapy increases mortality and organ dysfunction duration in pediatric sepsis. Crit Care Med 42(11):2409–2417. https://doi.org/10.1097/CCM.0000000000000509
11. Pruinelli L, Westra BL, Yadav P, Hoff A, Steinbach M, Kumar V, Delaney CW, Simon G (2018) Delay within the 3-hour surviving sepsis campaign guideline on mortality for patients with severe sepsis and septic shock. Crit Care Med 46(4):500–505. https://doi.org/10.1097/CCM.0000000000002949

12. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O (2018) The eICU collaborative research database, a freely available multi-center database for critical care research. Scientific Data 5(1):180178. https://doi.org/10.1038/sdata.2018.178
13. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K (2021) Early prediction of sepsis in the ICU using machine learning: a systematic review. Front Med 8
14. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, Cohen J, Opal SM, Vincent J-L, Ramsay G (2003) For the international sepsis definitions conference: 2001 SCCM/ESICM/ACCP/ATS/SIS international sepsis definitions conference. Intensive Care Med 29(4):530–538. https://doi.org/10.1007/s00134-003-1662-x
15. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche J-D, Coopersmith CM, Hotchkiss RS, Levy MM, Marshall JC, Martin GS, Opal SM, Rubenfeld GD, Poll T, Vincent J-L, Angus DC (2016) The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 315(8):801–810. https://doi.org/10.1001/jama.2016.0287
16. Moor M, Horn M, Rieck B, Roqueiro D, Borgwardt K (2019) Early recognition of sepsis with gaussian process temporal convolutional networks and dynamic time warping. In: Proceedings of the 4th machine learning for healthcare conference, pp 2–26. ISSN: 2640-3498
17. Reyna MA, Josef C, Seyedi S, Jeter R, Shashikumar SP, Westover MB, Sharma A, Nemati S, Clifford GD (2019) Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. In: 2019 computing in cardiology (CinC). IEEE, p 1
18. Ho TK (1995) Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, 1:278–2821. https://doi.org/10.1109/ICDAR.1995.598994
19. Wyk F, Khojandi A, Kamaleswaran R (2019) Improving prediction performance using hierarchical analysis of real-time data: a sepsis case study. IEEE J Biomed Health Inform 23(3):978–986. https://doi.org/10.1109/JBHI.2019.2894570
20. Wyk F, Khojandi A, Mohammed A, Begoli E, Davis RL, Kamaleswaran R (2019) A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier. Int J Med Inform 122:55–62. https://doi.org/10.1016/j.ijmedinf.2018.12.002
21. Mahmud F, Pathan NS, Quamruzzaman M (2020) Early detection of sepsis in critical patients using random forest classifier. In: 2020 IEEE Region 10 Symposium (TENSYMP). IEEE, pp 130–133
22. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. KDD'16, pp. 785–794. https://doi.org/10.1145/2939672.2939785
23. Barton C, Chettipally U, Zhou Y, Jiang Z, Lynn-Palevsky A, Le S, Calvert J, Das R (2019) Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. Comput Biol Med 109:79–84. https://doi.org/10.1016/j.compbiomed.2019.04.027
24. BLoch E, Rotem T, Cohem J, Singer P, Aperstein Y (2019) Machine learning models for analysis of vital signs dynamics: a case for sepsis onset prediction - PMC. J Healthc Eng
25. Scikit-Learn: Linear SVC Documentation. https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html Accessed 12/09/2023
26. Shashikumar SP, Stanley MD, Sadiq I, Li Q, Holder A, Clifford GD, Nemati S (2017) Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. J Electrocardiol 50(6):739–743. https://doi.org/10.1016/j.jelectrocard.2017.08.013
27. Chen M, Hernández A (2022) Towards an explainable model for sepsis detection based on sensitivity analysis. IRBM 43(1):75–86
28. Abromavičius V, Plonis D, Tarasevičius D, Serackis A (2020) Two-stage monitoring of patients in intensive care unit for sepsis prediction using non-overfitted machine learning models. Electronics 9(7):1133. https://doi.org/10.3390/electronics9071133. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute
29. Kam HJ, Kim HY (2017) Learning representations for the early detection of sepsis with deep neural networks. Computers in Biology and Medicine 89:248–255. https://doi.org/10.1016/j.compbiomed.2017.08.015
30. Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, Jay M, Das R (2016) A computational approach to early sepsis detection. Comput Biol Med 74:69–73. https://doi.org/10.1016/j.compbiomed.2016.05.003
31. Kaji DA, Zech JR, Kim JS, Cho SK, Dangayach NS, Costa AB, Oermann EK (2019) An attention based deep learning model of clinical events in the intensive care unit. PLoS ONE 14(2):0211057. https://doi.org/10.1371/journal.pone.0211057
32. Lukaszewski RA, Yates AM, Jackson MC, Swingler K, Scherer JM, Simpson AJ, Sadler P, McQuillan P, Titball RW, Brooks TJG, Pearce MJ (2008) Presymptomatic prediction of sepsis in intensive care unit patients. Clinical and Vaccine Immunology: CVI 15(7):1089–1094. https://doi.org/10.1128/CVI.00486-07
33. Lauritsen SM, Kalør ME, Kongsgaard EL, Lauritsen KM, Jørgensen MJ, Lange J, Thiesson B (2020) Early detection of sepsis utilizing deep learning on electronic health record event sequences. Artif Intell Med 104:101820. https://doi.org/10.1016/j.artmed.2020.101820
34. Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, Clifford GD, Sharma A (2020) Early prediction of sepsis from clinical data: the PhysioNet/computing in cardiology challenge 2019. Crit Care Med 48(2):210–217. https://doi.org/10.1097/CCM.0000000000004145
35. Barton C, Chettipally U, Zhou Y, Jiang Z, Lynn-Palevsky A, Le S, Calvert J, Das R (2019) Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. Comput Biol Med 109:79–84
36. Moor M, Bennett N, Plečko D, Horn M, Rieck B, Meinshausen N, Bühlmann P, Borgwardt K (2023) Predicting sepsis using deep learning across international sites: a retrospective development and validation study. EClinicalMedicine 62
37. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, Rubenfeld G, Kahn JM, Shankar-Hari M, Singer M, Deutschman CS, Escobar GJ, Angus DC (2016) Assessment of clinical criteria for sepsis. JAMA 315(8):762–774. https://doi.org/10.1001/jama.2016.0288
38. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart CK, Suter PM, Thijs LG (1996) The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the European society of intensive care medicine. Intensive Care Med 22(7):707–710. https://doi.org/10.1007/BF01709751
39. Johnson AEW, Stone DJ, Celi LA, Pollard TJ (2018) The MIMIC code repository: enabling reproducibility in critical care research. J Am Med Inform Assoc 25(1):32–39. https://doi.org/10.1093/jamia/ocx084
40. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, Shimabukuro D, Chettipally U, Feldman MD, Barton C, Wales DJ, Das R (2016) Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR Med Inform 4(3):28. https://doi.org/10.2196/medinform.5909
41. Moor M, Horn M, Rieck B, Roqueiro D, Borgwardt K (2021) MGP-TCN for sepsis prediction on the MIMIC Dataset. GitHub

42. Futoma J, Hariharan S, Heller K, Sendak M, Brajer N, Clement M, Bedoya A, O'brien C (2017) An improved multi-output gaussian process rnn with real-time validation for early sepsis detection. In: Machine learning for healthcare conference. PMLR, pp 243–254
43. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG (2018) An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 46(4):547–553
44. Moor M, Horn M, Bock C, Borgwardt K, Rieck B (2020) Path imputation strategies for signature models. In: ICML Workshop on the art of learning with missing values (artemiss)
45. Li SC-X, Marlin BM (2016) A scalable end-to-end gaussian process adapter for irregularly sampled time series classification. Adv Neural Inf Process Syst 29:1804–1808
46. Vega-Márquez B, Nepomuceno-Chamorro IA, Rubio-Escudero C, Riquelme JC (2021) Ocean: ordinal classification with an ensemble approach. Inf Sci 580:221–242
47. Ghosh S, Dasgupta A, Swetapadma A (2019) A study on support vector machine based linear and non-linear pattern classification. In: 2019 International conference on intelligent sustainable systems (ICISS). IEEE, pp 24–28
48. Gupta V (2023) Application of chaos theory for arrhythmia detection in pathological databases. Int J Med Eng Inform 15(2):191–202
49. Schober P, Vetter TR (2021) Logistic regression in medical research. Anesth Analg 132(2):365
50. Wang Y, Sun S, Chen X, Zeng X, Kong Y, Chen J, Guo Y, Wang T (2021) Short-term load forecasting of industrial customers based on svmd and xgboost. Int J Electric Power Energy Syst 129:106830
51. Desai M, Shah M (2021) An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (mlp) and convolutional neural network (cnn). Clinical eHealth 4:1–11
52. Fukushima K, Miyake S (1982) Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition. In: Competition and cooperation in neural nets, pp 267–285
53. Dhillon A, Verma GK (2020) Convolutional neural network: a review of models, methodologies and applications to object detection. Prog Artif Intell 9(2):85–112
54. Vega-Márquez B, Rubio-Escudero C, Nepomuceno-Chamorro IA, Arcos-Vargas Á (2021) Use of deep learning architectures for day-ahead electricity price forecasting over different time periods in the Spanish electricity market. Appl Sci 11(13):6097
55. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
56. Bai S, Kolter JZ, Koltun V (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. CoRR abs/1803.01271
57. Lara-Benítez P, Carranza-García M, Luna-Romera JM, Riquelme JC (2020) Temporal convolutional networks applied to energy-related time series forecasting. Appl Sci 10(7):2322
58. Zhang Y-G, Tang J, He Z-Y, Tan J, Li C (2021) A novel displacement prediction method using gated recurrent unit model with time series analysis in the Erdaohe landslide. Nat Hazards 105(1):783–813
59. Scikit-Learn:Parameter Sampler Documentation. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.ParameterSampler.html#sklearn.model_selection.ParameterSampler Accessed 12/09/2023
60. Lara-Benítez P, Carranza-García M, Riquelme JC (2021) An experimental review on deep learning architectures for time series forecasting. Int J Neural Syst 31(03):2130001
61. Mohammed R, Rawashdeh J, Abdullah M (2020) Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: 2020 11th international conference on information and communication systems (ICICS), pp 243–248. https://doi.org/10.1109/ICICS49469.2020.239556. ISSN: 2573-3346
62. Lobo JM, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. Global Ecol Biogeogr 17(2):145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x
63. Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 10(3):0118432. https://doi.org/10.1371/journal.pone.0118432 Publisher: Public Library of Science
64. Ozenne B, Subtil F, Maucort-Boulch D (2015) The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol 68(8):855–859. https://doi.org/10.1016/j.jclinepi.2015.02.010 Publisher: Elsevier
65. Pinker E (2018) Reporting accuracy of rare event classifiers. npj Digital Medicine 1(1):1–2. https://doi.org/10.1038/s41746-018-0062-0. Number: 1 Publisher: Nature Publishing Group
66. Draelos R (2019) Measuring performance: AUPRC and average precision
67. Scherpf M, Gräßer F, Malberg H, Zaunseder S (2019) Predicting sepsis with a recurrent neural network using the mimic iii database. Comput Biol Med 113:103395. https://doi.org/10.1016/j.compbiomed.2019.103395
68. Ericson O, Hjelmgren J, Sjövall F, Söderberg J, Persson I (2022) The potential cost and cost-effectiveness impact of using a machine learning algorithm for early detection of sepsis in intensive care units in Sweden. J Health Econ Outcomes Res 9(1):101

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Javier Solís-García earned a B.Sc. degree in Computer Science from the University of Sevilla in 2021 and went on to complete an M.Sc. degree in Artificial Intelligence at the Universidad Politécnica de Madrid in 2022. He is currently pursuing a Ph.D. in Computer Sciences and Artificial Intelligence at the Department of Computing Systems and Languages, where he also serves as a researcher. His current research interests focus on Machine Learning and Deep Learning techniques and their application across various fields.

Belén Vega-Márquez received her Ph.D. in computer science from the University of Sevilla, Spain in 2022. She is an Assistant Professor in the Department of Computing Systems and Languages, University of Sevilla, and a researcher in the Minerva research lab. Her current interests are the use of data science techniques such as artificial intelligence and data mining for its application in health and nutrition.

José C. Riquelme-Santos received an M.Sc. degree in mathematics and a Ph.D. degree in computer science from the University of Seville, Spain. Since 1987, he has been with the Department of Computer Science, University of Seville, where he is currently a full professor. His primary research interests include data mining, machine learning techniques, and evolutionary computation.

Juan A. Nepomuceno has a degree in Mathematics and a PhD in Computer Science from the University of Seville (Spain). He is a tenured Associate Professor in the Department of Computer Systems and Languages, University of Seville. His research topics focus on the analysis of clinical data and the use of machine learning techniques in the context of biological problems.

Isabel A. Nepomuceno-Chamorro received her Ph.D. in applied computer science from the University Pablo de Olavide, Spain, in 2011. She is a tenured Associate Professor in the Department of Computing Systems and Languages at the University of Sevilla. Her primary research interests include machine learning techniques and their application in the area of personalized medicine. She has participated in regional, national and European projects in this topic.