

Towards an Explainable Model for Sepsis Detection Based on Sensitivity Analysis

M. Chen, Alfredo I. Hernández

▶ **To cite this version:**

M. Chen, Alfredo I. Hernández. Towards an Explainable Model for Sepsis Detection Based on Sensitivity Analysis. Innovation and Research in BioMedical engineering, 2022, 43 (1), pp.75-86. 10.1016/j.irbm.2021.05.006 . hal-03520717

HAL Id: hal-03520717

https://univ-rennes.hal.science/hal-03520717

Submitted on 28 Mar 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial| 4.0 International License

Journal Pre-proof

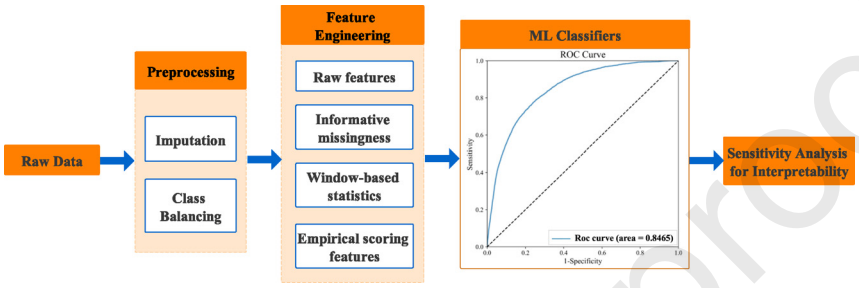
Graphical abstract

IRBM ●●●●, ●●●, ●●●

Towards an explainable model for Sepsis detection based on sensitivity analysis

M. Chen[a], A. Hernández[b,*]

[a] *The Laboratory of Medical Imaging and Artificial Intelligence, School of Biomedical Science and Medical Engineering, Southeast University, Nanjing, China*
[b] *Université de Rennes, Inserm, LTSI - UMR 1099, Rennes, France*



Journal Pre-proof

Journal Pre-proof

Highlights

- Optimized machine learning pipelines for Sepsis early prediction are proposed.
- The models are fitted on the 2019 PhysioNet/Computing in Cardiology Challenge dataset.
- The models achieved comparable performance to the best methods of the challenge.
- A new method for increasing the interpretability of machine learning models is proposed.
- Interpretable information was extracted from the proposed model.

Journal Pre-proof

Journal Pre-proof

Towards an explainable model for Sepsis detection based on sensitivity analysis

M. Chen[a], A. Hernández[b,*]

[a] *The Laboratory of Medical Imaging and Artificial Intelligence, School of Biomedical Science and Medical Engineering, Southeast University, Nanjing, China*
[b] *Université de Rennes, Inserm, LTSI - UMR 1099, Rennes, France*

Abstract

Objectives: Sepsis is a life-threatening condition which is responsible for a high proportion of intra-hospital deaths and related healthcare costs each year. Early detection and treatment of sepsis episodes is critical, since an early treatment may highly improve prognosis. This study proposed an original method to increase the interpretability of a set of machine learning models for the early detection of sepsis onset.

Material and methods: Open data from the electronic medical records of 40,336 patients monitored in intensive care units (ICU), provided by the PhysioNet/Computing in Cardiology Challenge 2019 is used in this paper. We proposed a method including data preprocessing, feature engineering, model construction and tuning, as well as an original interpretability analysis method for the final stage.

Results: A total of 24 models were developed and analyzed. The best model, based on 142 features achieved a 0.4274 utility score. The best compact model integrates only 20 selected features, and provided a utility score of 0.3862. Meanwhile, the proposed sensitivity analysis method allows for the identification of the most relevant markers to early detect the onset of sepsis, as well as their interdependence and relative importance on the final decision.

Conclusion: A set of optimized machine-learning models were proposed for predicting sepsis early in a real-time way with high performance, and interpretable information including the most significant biomarkers were an-

*Corresponding author.
Email address: alfredo.hernandez@inserm.fr (A. Hernández)

Journal Pre-proof

alyzed through novel interpretability method.

Keywords: intensive care unit (ICU), random forest, explainable AI, sensitivity analysis

1. Introduction

Sepsis is a life-threatening condition that is caused by a dysregulated hosts' immune response to infection when the body becomes seriously infected, leading to tissue damage, organ failure, or death [1–3]. According to the recent report of the Center for Disease Control (CDC), more than one-third of patients who die in hospitals in the USA suffer from sepsis [4]. Sepsis is also regarded as a costly syndrome because the cost of its management accounts for more than 13% of the U.S. healthcare expenses per year [5]. Overall, sepsis is a major public health concern that leads to high morbidity, high mortality, and expensive healthcare costs [6].

The syndromic nature of sepsis often complicates identification and diagnosis, which can further contribute to delays in treatment. Two recent studies have emphasized the importance of early detection and treatment of sepsis and have shown that delayed antibiotic treatment of sepsis leads to increased mortality [7, 8]. The effect is more significant for patients with septic shock, which indicates that the delay per hour is associated with an increase of $3.6 - 9.9\%$ per hour in mortality [9].

Traditionally, many general-purpose illness severity scoring systems such as Acute Physiology and Chronic Health Assessment (APACHE II), Simplified Scoring of Acute Physiology (SAPS II), and the Sequential Rating of Organic Failure (SOFA) have been widely used to identify sepsis. These scoring systems monitor laboratory values and vital signs such as heart rate (HR), the fraction of inspired oxygen (FiO2), etc., and give overall scores with threshold-based methods. However, sepsis is a dynamic condition, and these standard scores usually cannot meet the needs of emergencies, that is, to detect sepsis as early as possible to obtain effective treatment, as well as to distinguish patients with specific diseases with high sensitivity and specificity.

Recently, massive data from electronic medical records (EMR) acquired from "real-life" intra-hospital information systems have become publicly available, to support research in this field. The MIMIC (Multiparameter Intelligent Monitoring in Intensive Care)–II Clinical Database has been widely used for developing predictive models. Katharine et al. [10] analyzed routinely

Journal Pre-proof

available data and developed "TREWScore" by training a Cox proportional hazards model with lasso regularization. TREWScore achieved a AUROC of 0.83 for identifying patients who will develop septic shock before the clinical onset time. Calvert et al. [11] developed and applied a simplified sepsis early warning algorithm, InSight, to identify patients with sepsis, severe sepsis and septic shock by implementing a gradient boosting tree algorithm. Insight was trained on MIMIC-II dataset and part of UCSF dataset following the international consensus definitions for sepsis in 2001. This work reported predictions of sepsis 4 hours before the onset of the event, with an AUROC of 0.92 on the remaining UCSF test data, exceeding or rivaling that of existing biomarker detection methods. With the further understanding of sepsis in the medical and scientific communities, the definition and diagnostic criteria of sepsis have been further upgraded to Sepsis-3 [1], and the clinical demand for early prediction of sepsis has become more urgent. In response to this call, Shamim et al. [12] used data available in the ICU in real-time from two Emory University hospitals and passed variables to the Artificial Intelligence Sepsis Expert (AISE) algorithm. Using a modified Weilbull-Cox proportional hazards model, AISE can accurately predict the onset of sepsis in an ICU patient 4 to 12 hours prior to clinical recognition on validation cohort (MIMIC-III ICU dataset [13]) with AUROC in the range of 0.83–0.85. A recent study [14] for early detection of sepsis was implemented on a seven-year-period data from multiple Danish hospitals. They presented a deep learning system combining a convolutional neural network and a long short-term memory network with performance ranging from AUROC 0.758-0.856 (24 to 3 hours before sepsis onset).

However, even though a number of interesting rule-based machine learning and deep learning models have been proposed, the relative strengths and weaknesses of algorithmic approaches are unclear for some reasons [6]. On one hand, researches developed and validated their algorithms with different patient cohorts with different clinical variables and labels arising from different clinical criteria for sepsis. On the other hand, different studies often employed different evaluation metrics that are not explicitly designed for sepsis prediction task. To solve this problem, the PhysioNet/Computing in Cardiology Challenge 2019 provided a common problem statement using the same clinical variables and sepsis criteria, and devised a novel evaluation metric that addresses these issues and could be generally applicable to predicting infrequent events in time series data [6].

Many excellent algorithms that offered valuable references for our work

Journal Pre-proof

have been proposed in the framework of this Challenge. Morrill and colleagues [15] extracted signature-based features from the time series as inputs to represent the longitudinal effects of sepsis and train the early prediction model. Their proposed algorithm won the highest official score on the utility function. Yang et al. [16] implemented a series of feature extraction strategies to obtain a total of 168 variables as input for machine learning phase. Their XGBoost classification model ranking the first in the Challenge (unofficial score), then, was developed and was further improved by a Bayesian optimizer and an ensemble learning framework. Zabihi's team [17] also utilized XGBoost as their base learner. They ensembled five XGBoost models to further improve the prediction results after feature engineering. Lyra et al. [18] developed a the optimized Random Forest model to make the prediction results better. In addition, deep learning algorithms were also widely used in the Challenge, because they are feature extraction-free and some algorithms include memory cells, which are also very suitable for time-series related tasks. The TASP algorithm proposed by Li et al. [18] employed the recurrent neural network (RNN) to capture the long-term dependence factors in the patients' time series data, and finally won the fourth place in the Challenge. Besides, there are other many studies using LSTM, GRU or auto encoder as prediction models to address early prediction task of sepsis [19–21]. After the Challenge, a Smart Sepsis Predictor (SSP) system was trained, validated and tested on the shared data from Challenge 2019 in order to assess whether the patient is suspected of sepsis or not. The SSP was a deep network inserted LSTM, convolutional, and fully connected layers. The performance in Mode 1 (only demographics and vital signs) and Mode 2 (full features) achieved an AUROC of 0.89 and 0.92 for 4 h before sepsis onset, respectively.

Nevertheless, the fundamental need for early and reliable identification of sepsis remains unmet [22]. Moreover, the interpretability of the proposed models should be improved in order to increase the clinical adoption rate of these methods by the medical community and to ease audit and regulatory activities. In this work, we use this set of open data to propose a machine learning pipeline, based on a Random Forest classifier, for early predicting the onset of sepsis, followed by an original approach to increase the explainability of the underlying model.

Journal Pre-proof

2. Material and methods

2.1. Dataset Description and Analysis

Publicly available data and annotations for a total of 40,336 patients (2,932 septic patients and 37,404 non-septic patients) were collected from two independent U.S. hospital systems with distinct Electronic Medical Record (EMR) systems: Beth Israel Deaconess Medical Center (hospital system A) and Emory University Hospital (hospital system B). These data were collected over the past decade with approval from the appropriate institutional review boards [6]. Each patient has 8 hours to 2 weeks' hourly time window of records. Altogether, this dataset included over 1.55 million hourly time windows and 18 million data points.

The Challenge data consist of 40 heterogeneous clinical variables, including 8 vital sign summaries, 26 laboratory values, and 6 demographic descriptions (Appendix: A). As for the annotations, the data are labeled according to Sepsis-3 [1] clinical criteria, based on an acute increase in total SOFA score, consequent to the infection. The definition of the clinical onset time of sepsis has been taken as proposed for the Challenge, and is described in detail in [6]. Furthermore, as introduced in [6], assuming that t^* is the clinical onset time of sepsis, then, for septic patients, the labels are set to '1' if $t \geq t^* - 6$, where the time of $(t^* - 6)$ is defined as the optimal prediction time. The labels are set to '0' otherwise. Also, the labels are set to '0' for non-septic patients at all the time points (Figure 1).

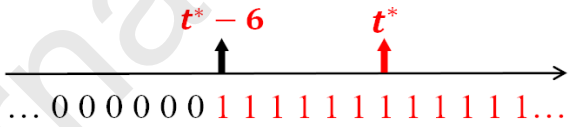


Figure 1: Clinical expert annotations for septic patients. t^* represents the clinical onset time of sepsis.

Table 1 shows the summary of the dataset, where the unbalancing distributions between sepsis and non-sepsis are evident (7.26% patient-wise sepsis prevalence and 1.8% sample-wise sepsis prevalence) and it contains a significant amount of missing values (∼70% missing).

Figure 2 shows histograms of each original variable in the dataset, in which the subtitles of subgraphs indicate the record densities. For each feature, red bins stand for patients with sepsis while green bins represent

Journal Pre-proof

Table 1: Summary of the public dataset.

Hospital system	A	B	Overall
Number of patients	20,336	20,000	40,336
Number of septic patients	1,790	1,142	2,932
Sepsis prevalence	8.8%	5.7%	7.3%
Number of rows	790,215	761,995	1,552,210
Number of entries	9,505,801	9,070,444	18,576,245
Density of entries	30.1%	29.8%	29.9%

non-septic patients. Specifically, among the recorded variables, most vital signs were updated on an hourly basis in most patient records and have approximately more than 85% value density, except for Temperature (Temp) of 34%, Diastolic blood pressure (DBP) of 69%, and End-tidal carbon dioxide (EtCO2) of 3.7%. For laboratory values, except for Glucose missing value of 83%, the other variables were collected with less than 10% density. Both of the two administrative identifiers for ICU unit (Unit 1 and Unit 2) miss nearly 39% of the values, while the remaining four variables (Age, Gender, HospAdmTime and ICULOS) are fixed variables for the demographics.

Serious data imbalance and missing data are two characteristics of this "real-life" quality dataset. The third characteristic is that the sampling time of each patient is inconsistent and random. More specifically, the longest hourly time windows of data are 336 recordings, and the shortest period is 8 hours. Overall, these three characteristics of this database pose great difficulty to the prediction of sepsis by using naïve models.

2.2. Data preprocessing

According to the characteristics of the database listed above, we implemented two data preprocessing strategies to improve the data quality: missing value imputation and class balancing.

Imputation methods. Inspired by related works [15–17, 23, 24] and considering the fact that our machine learning model cannot handle missing values, we compared three imputation strategies.

1) A combination of forward filling and fixed-value imputation methods, where the last valid observation is first used to fill the missing values, and then the remaining missing values (if exist) are replaced with fixed integers (such as 0) to act as placeholders without actual clinical significance.

Journal Pre-proof

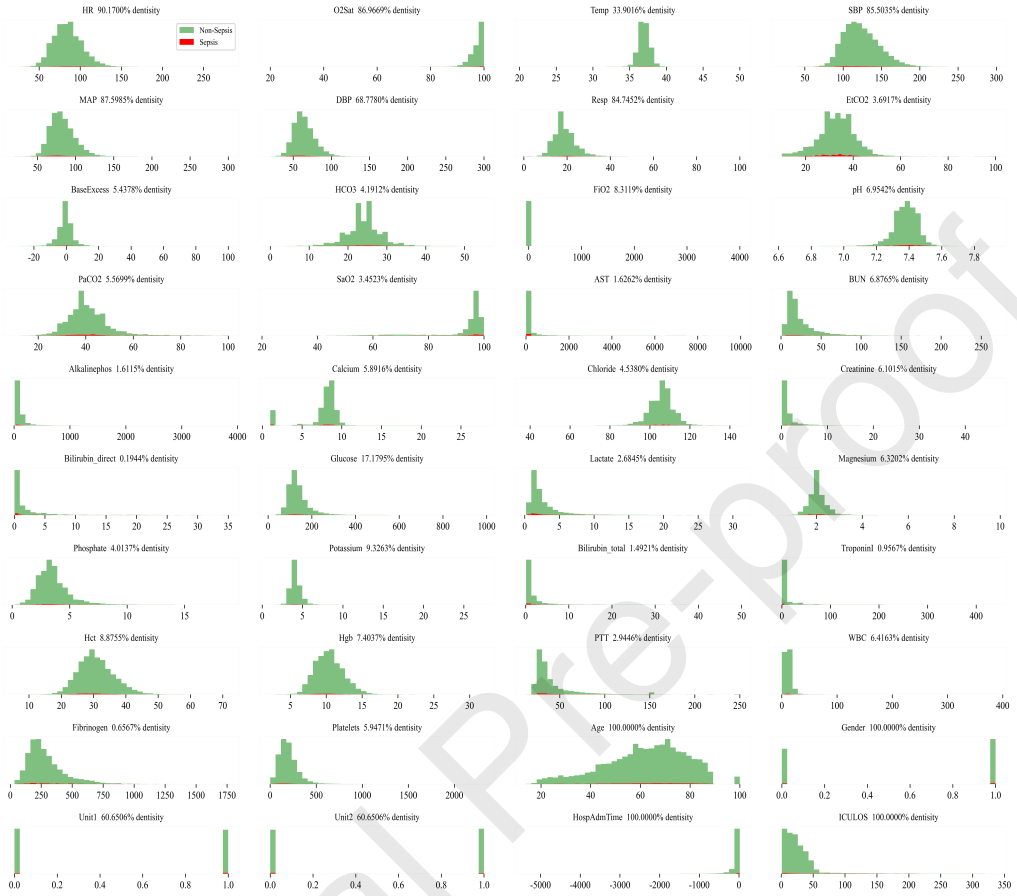


Figure 2: Histogram of 40 raw variables on the whole database.

2) Same as 1) but replacing the integer placeholder by the mean values of each variable from the training data.
3) Perform linear interpolation first and then apply forward filling strategy.

All these approaches can ensure that the static features (Age, Gender, Unit1, and Unit2) are kept constant for a given patient, along the whole acquisition time frame.

Data split. A total of $40,336$ patients' clinical records were divided into development set (85%) and test set (15%), where the class ratio is consistent with the distribution of the original dataset. During the classification model building process, four-fifths of the development set were used for training while the remaining one fifth was used to validate performance and optimize

Journal Pre-proof

Table 2: Dataset partition (patient-wise).

		Septic	Non-septic	Total
Development set (85%)	Train set (4/5)	1,993	25,434	27,427
	Validation set (1/5)	499	6,359	6,858
Test set (15%)		440	5,611	6,051
Total		2,932	37,404	40,336

the hyperparameters of the models. Table 2 shows the details of the dataset partitions used in this work.

Down-sampling. As mentioned above, the available dataset is seriously imbalanced over two classes. In order to minimize the impact of class imbalance on performance, we down-sampled the excessive instances of the majority class with random seeds before constructing prediction models.

2.3. Feature Engineering

Apart from the original features after imputation, we designed three types of features to explore more information in raw data, including 102 informative features originated from missing patterns of raw variables, 30 time-series features, and 8 empirical scores based on known rule-based disease severity scoring systems.

Raw features. Among 40 original features provided by the Challenge, 8 vital signs and 26 laboratory measurements are time-varying variables, and 6 demographic variables are static and available for each record in the dataset (except for administrative identifiers of Unit1/Unit2). For these original features, multiple imputation strategies are adopted.

Informative missingness. Little[25] and Rubin [26] highlighted the difference between three types of missing data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR). And in [27], Lin et al. proved that introducing explicit representations of data absence is useful for boosting model performance when data is not missing at random. This implies that data missingness at a given time does contain information value. For instance, a patient on which nurses apply regular interventions may be associated with a mean poorer signal quality, due to the patient movements evoked by the interventions. We hypothesize that this information content on missingness is significant in our database. Thus, we compute three missing data indicator sequences to extract the hidden information beneath the measurement pattern. This

Journal Pre-proof

process is performed on 34 time-varying variables (40 raw variables except for six demographics).

1) Binary measurement mask: Indicate whether the measurement is available ('1') or missing ('0').
2) Measurement frequency vector: Record the number of available measurements until the current time point.
3) Measurement time interval vector: Record the time intervals from the previous measurement to the current time point when the current value is missing. We set '-1' when there is no preceding measurement or the current value exists.

Window-based statistics. In order to better inspect the dynamic changes in data, we propose window-based statistics calculated from time series data. Most vital signs variables have the lowest missing rate and the highest quality, and we choose HR, O2Sat, SBP, MAP and Resp for developing new features. According to [16], firstly, a six-hour sliding window is applied to segment each record with a step of one hour. Then, six classical statistics (maximum, minimum, mean, median, standard deviation, and differential standard deviation) are calculated for each recording, in each time window for the selected five variables.

Empirical scoring features. Some commonly accepted scoring systems for quantifying health conditions and identifying sepsis patients are included in this work to take clinically empirical criteria into consideration. Under the rules of (Sepsis-related Organ Failure Assessment) SOFA [28] score, Platelets, Bilirubin, Mean Arterial Pressure (MAP) and Creatinine represent the dysfunction of coagulation, liver, cardiovascular and renal systems, respectively. When respiratory rate (Resp) $\geq$ 22/min and Systolic blood pressure (SBP) $\leq$ 100 mm Hg, the Quick SOFA (qSOFA) score is evaluated as 1, otherwise 0. Besides, Heart rate (HR), Respiratory rate (Resp) and Temperature (Temp) are scored according to the National Early Warning Score (NEWS) [29] score. Therefore, a total of 8 scores are constructed from the variables that are present both in the empirical scores and the available dataset.

2.4. Model Development

Random Forest. In this work, we choose a powerful ensemble supervised machine learning algorithm, Random Forest (RF) [30], which integrates multiple decision trees and uses the Gini coefficient as a measure of purity

Journal Pre-proof

to control the growth and division of the trees. Its advantage over a single decision tree and other ML algorithms is the introduction of two randomnesses: the sample set used to train each subtree is randomly selected from all samples using the bootstrap resampling method; the set of split attributes for each decision node in the subtrees is also randomly selected. This gives RF excellent generalization performance. The output of the RF comes from the majority vote of each subtree (for binary classification).

Model Training. Before building formal classification models, we first used the grid search to find the best parameters greedily for RF by setting $1 - Accuracy$ on the validation set as cost function. The main hyperparameters include the number of CART trees, maximum features for each subtree, maximum depth of subtrees, minimum samples on a node and a leaf, and they are supposed to control subtree growth and pruning. Afterward, we set the hyperparameters as the best ones obtained in grid search and re-train classifiers, and we named these classifiers constructed with unfiltered features as 'Full' models.

Meanwhile, we trained several classifiers on distinct subsets of features (see Section 2.3) in order to explore the predictive ability of the additionally extracted features and find out how newly introduced features influence the performance.

Feature selection. During the construction of each subtree in RF, how much each feature decrease the Gini impurity are measured, and the feature with the highest decrease is selected for internal node. Meanwhile, for each feature, we collected how on average it decreases the Gini impurity, and the mean decrease over all trees in the whole forest is the measure of the variable importance (VI), which can be used to estimate feature impact and select the most significant variables [31]. In this work, this approach was applied in order to i) sort VI generated from the full models in descending order of significance, ii) determine the deletion ratio and iii) re-train new 'Compact' models integrating only the most relevant features.

Model evaluation. The Challenge proposed a specific metric to evaluate the performance of early prediction models. It is the Utility score that penalizes both late or missed identification of sepsis in septic patients as well as wrong sepsis predictions in non-septic patients, and it rewards early positive predictions in septic patients [6]. Specifically, for a true sepsis patient, the prediction model would be rewarded 1 if it can accurately predict sepsis at the optimal prediction time. Both too early or too late prediction would be rewarded less than 1 or even slightly penalized, and the utility

Journal Pre-proof

score would decline as the prediction time moves further from the optimal detection time. For a non-sepsis patient, wrong positive predictions would be slightly penalized considering that this would result in alarm fatigue and poor allocation of healthcare resources and attention. Likewise, non-sepsis predictions in non-septic patients were neither rewarded nor penalized.

The utility score curve can be expressed as follows (Eq. (1)–(4)), in which t^* is the onset time of sepsis as defined in section 2.1. Then the utility score is added up and normalized across all hourly time windows for one specific patient, and it is called the overall utility score.

$$U_{TP}(t) = -0.05\mathbb{I}\left[t < t^* - 6\right] + \frac{(t - t^* + 6)}{6}\mathbb{I}\left[t^* - 6 \leq t \leq t^*\right] + \frac{9 - t + t^*}{9}\mathbb{I}\left[t^* < t \leq t^* + 9\right] \tag{1}$$

$$U_{FN}(t) = 0\mathbb{I}\left[t < t^*\right] - \frac{2\,(t - t^*)}{9}\mathbb{I}\left[t^* \leq t \leq t^* + 9\right] \tag{2}$$

$$U_{FP}(t) = -0.05 \tag{3}$$

$$U_{TN}(t) = 0 \tag{4}$$

Universal metrics of confusion matrix were also calculated for evaluating all the hourly predictions across all patients. The confusion matrix is a matrix that provides visualization of the performance of a classifier by showing distributions of classification results clearly. A series of indexes can be derived from the confusion matrix, including accuracy (ACC) and F1-measure. Both the area under the receiver operating characteristic curve (AUROC) and the precision-recall curve (AUPRC) were used to further reveal the sensitivity and specificity of the model. Here, we specified the patients as positive cases if they were reported developing sepsis during their ICU stay and negative cases if not [16].

2.5. Sensitivity Analysis

Sensitivity analysis is a powerful technique applied to multi-parametric mathematical models in order to estimate the relative influence of each model parameter or model input, and their inter-relation, on the output of the model [32]. In this work, we propose to apply the same theory and methods beneath sensitivity analysis for the estimation of the relative importance of each input

Journal Pre-proof

feature on the output of a machine learning model. We hypothesize that such a powerful analysis may be particularly useful to increase the explainability of black-box machine learning methods.

Morris method [33] is a particularly interesting, "one-step-at-a-time" (OAT) sensitivity analysis method. Morris method generates several random trajectories in the parameter (input features) space. During this process, it does not return to the original base point after perturbation but continues perturbing another dimension from the perturbed point, leading to an efficient exploration for parameter space in given levels. The method is based on the estimation of "Elementary Effects" (EE) that are associated with a given trajectory in the parameter space. EE is a measure of the changes in the output of the model $Y(\vec{x})$ related to a given trajectory, when perturbing one parameter x_k at a time. The elementary effect for parameter k (EE_k) is defined as follows:

$$EE_k = \frac{1}{\Delta}\left[Y\left(x_1, x_2, \ldots, x_k + \Delta, \ldots, x_K\right) - Y\left(x_1, x_2, \ldots, x_K\right)\right] \tag{5}$$

where Δ is a predefined variation. For each parameter (input feature) x_k, two essential statistical indicators are obtained from EE_k. The first is the absolute mean [34] that represents the linear effect of the variable x_k on the model output Y, defined as

$$\mu_k^* = \frac{1}{n_R}\sum_{r=1}^{n_R} |EE_k^r| \tag{6}$$

And the second is the standard deviation of the EE_k from all the trajectories and it indicates the presence of nonlinearity and/or interactions between variables.

$$\sigma_k = \sqrt{\frac{1}{n_R}\sum_{r=1}^{n_R}\left(EE_k^r - \mu_k\right)^2} \tag{7}$$

By examining the μ^* of EE_k, we can estimate a relative importance ranking among parameters and, in turn, screen out non-influential parameters or make a deeper exploration of each parameter's influence on the model prediction outcomes. Moreover, the analysis of σ_k gives an indication of feature interactions, which cannot be correctly addressed with the VI method.

Journal Pre-proof

The range of possible values for each variable that we are interested in was derived from the development data. For each variable, we calculated its first quartile $Q1$ and the third quartile $Q3$ respectively and then calculated the maximum and minimum values by ignoring data points outside the range of $Q1 - 3.0 \times (Q3 - Q1)$ to $Q3 + 3.0 \times (Q3 - Q1)$. To supplement, some manual adjustments for the maximum and the minimum were conducted to frame a more reasonable variable space.

3. Results

3.1. The sepsis early prediction model

The proposed models predict whether a patient in ICU will present sepsis on an hourly basis. To compare the efficiency of 3 imputation strategies for dealing with missing values (details in Section 2.3), we composed 3 sets of development data. For each set of data, there are 8 groups of features to be used for fitting RF classification models. Before constructing the model, the negative class (non-sepsis) is in the absolute majority, so it was down-sampled to balance with the positive class (sepsis), and there were a total of $\sim 37,000$ rows of hourly data for supervised learning and $\sim 9,500$ rows for validation.

Table 3: Utility Scores of all models.

Feature Set	R Only		+IM		+SW		+ES	
Development Set	F (40)	C (15)	F (142)	C (20)	F (172)	C (25)	F (180)	C (25)
ffill+0	0.38	0.37	0.42	0.37	0.40	0.37	0.40	0.36
ffill+mean	0.38	0.36	0.42	0.36	0.41	0.36	0.41	0.36
linear+mean	0.39	0.38	**0.43**	**0.39**	0.41	0.38	0.42	0.38

Feature set: R—Raw features, IM—Informative missingness features, SW—Sliding window-based statistics, ES—Empirical scoring system based features. **Imputation strategies:** 'ffill+0'—Forward filling & zero imputation, 'ffill+mean'—forward filling & mean values imputation, 'linear+mean': linear interpolation & mean values imputation. **F:** Full, models built with unselected variables. **C:** Compact, models built with selected variables based on corresponding full models' VI ranking. The numbers below 'F' and 'C' in the second row are the numbers of features included in the corresponding model.

Hyperparameters in classifiers were tuned by grid search on the training set and validation set. Overall utility scores on our test set of a total of 24

Journal Pre-proof

Table 4: The results of the best models on the test set.

	AUROC	AUPRC	Accuracy	F1-measure	Utility Score
Full	**0.8465**	0.1030	**0.8188**	**0.1227**	**0.4274**
Compact	0.8310	0.1131	0.7862	0.1080	0.3862

models are summarized in Table 3. As shown in the table, the "best Full model", which is comprised of raw features and informative missingness features and that is trained with 'linear+mean' strategy-imputed development data, achieved the highest utility score of 0.4272 on the test set, as well as it realized the highest AUROC, accuracy, and F1-measure among the 24 models that we developed.

Figure 3 shows the ROC curve, PRC curve and the confusion matrix for the "best Full model". The variable importance was calculated during the construction process of full models, which offers a reference to conduct feature selection.

Among all compact models trained with top features, the "best Compact model" is the one that trained with the top 20 features (accounting for 62.46% of the VI) defined by VI of the best full model, and it achieved the utility score of 0.3862 on the local test set (see Table 4). It is expected that such compact models provide lower performance than their corresponding full models, since a significantly smaller variable set is included. In our results, the compact models' utility scores are 0.01 to 0.06 lower than that of the full models', depending on different imputation approaches and feature sets used. Moreover, results show that the best efficient imputation strategy is the 'linear+mean' approach for fixing missing values.

The results listed in the first two columns of Table 3 titled "R Only" belong to models constructed with raw variables from the available database solely. With the introduction of informative missingness features, the utility score of the full models increased significantly by ~ 0.4. However, sliding-window based statistics and empirical scoring systems based features did not further improve the performance. On the other hand, the inclusion of new variables has basically no effect on the performance of the compact models, this may due to the limited and similar number of input feature dimensions.

Concerning patient-wise statistics, the "best Full model" correctly detected 397 septic patients (sensitivity = 0.9023) while wrongly identifying $2,613$ non-septic patients as positive ones (specificity = 0.5343).

Figure 4 illustrates the VI ranking of the best compact model, in which

Journal Pre-proof

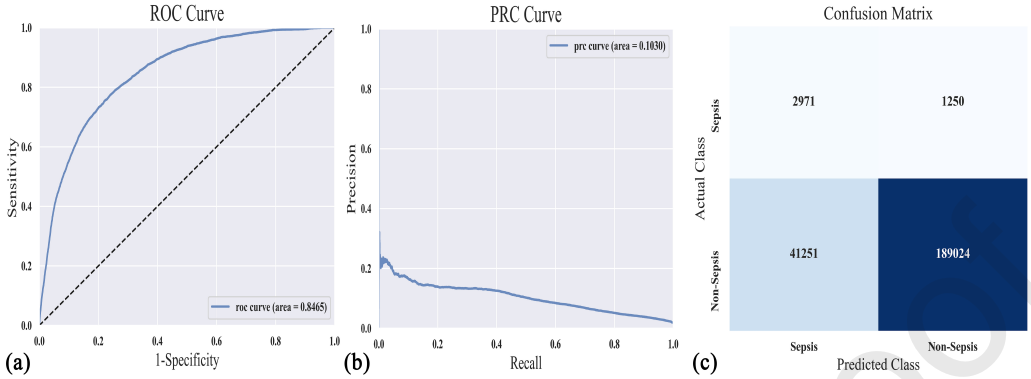


Figure 3: Performance of the Best Full Model.
(a), ROC curve with a 0.8465 area under the curve score (the AUROC baseline is 0.5). **(b),** PRC curve with a 0.1030 area under the curve (the AUPRC baseline is 0.018). **(c),** Confusion matrix, the number annotated on each quadrant is the rows of samples.

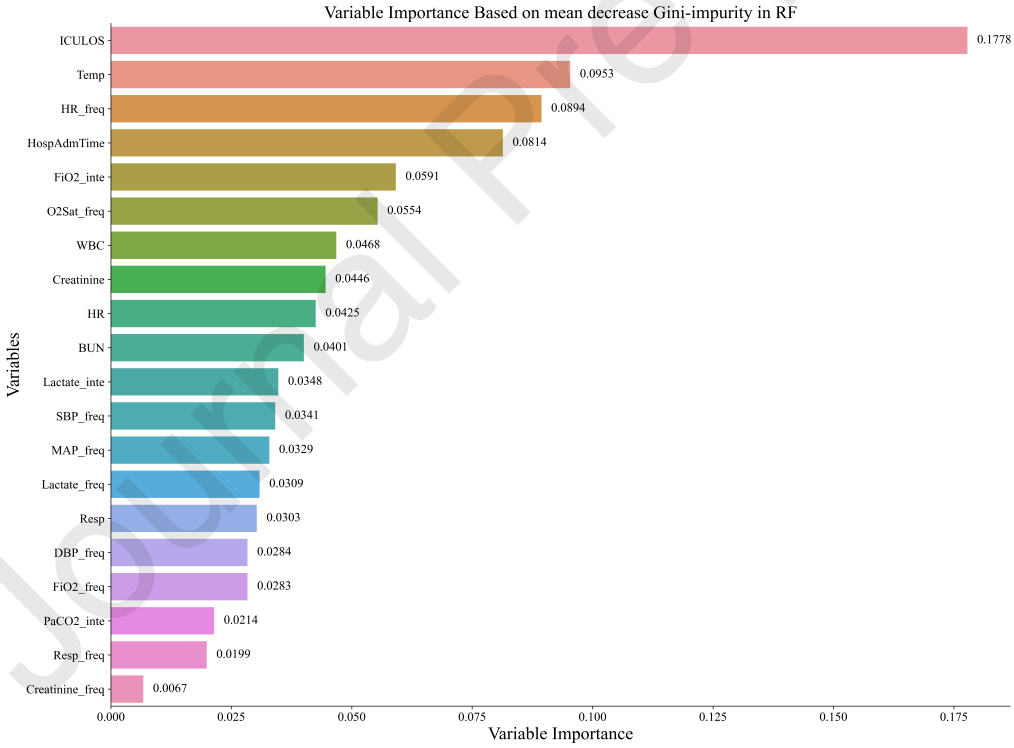


Figure 4: Variable Importance ranking for the best compact RF model.

Journal Pre-proof

ICULOS is the most significant variable, and 8 of those selected features are raw features while 12 of them are informative missingness features. The VI ranking of the best full model given in Supplementary Figure S1, where ICULOS (length of stay in ICU) ranks first and is much greater than the other features. Next comes the HR_freq (measurement frequency of HR values) and Temp (temperature). Using only the top 13 features are able to represent more than half of the overall VI, while the last 47 features merely take 10% of the VI altogether, which means that a high proportion of features do not contribute to the prediction task and they are redundant.

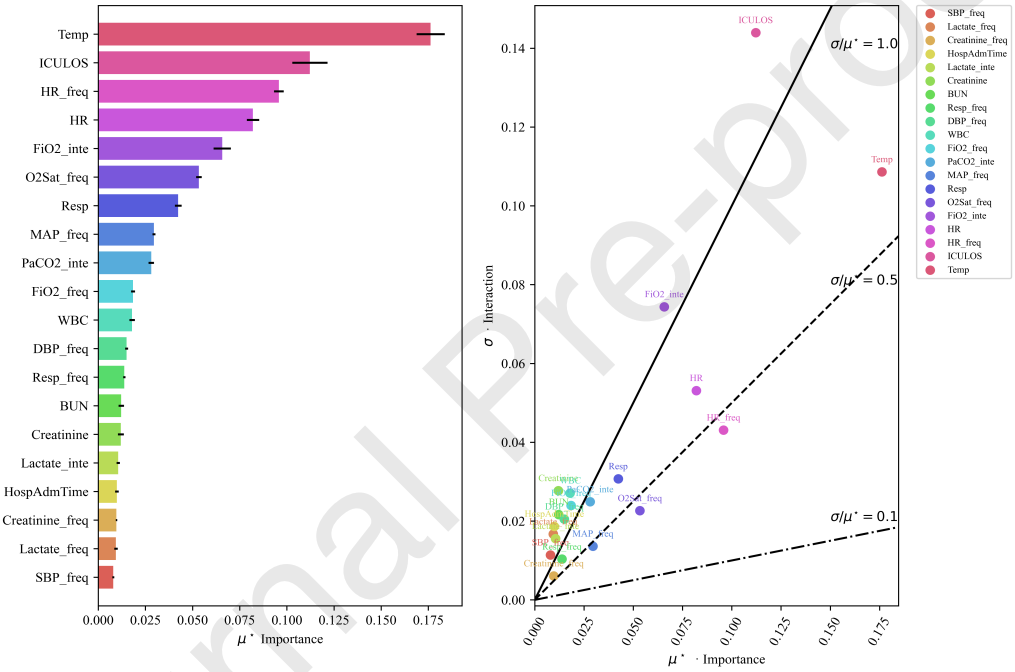


Figure 5: Morris Sensitivity analysis results of the Best Compact Model. The histogram shows the variable global sensitivity with error bars representing 0.95 confidence interval level of μ^*; and the scatters in the right display the distributions of variables on the importance μ^* versus interaction σ plane. The three lines represent different ratios of σ and μ^*. The solid line is for the ratio equals 1.0, the dashed line has a ratio of 0.5, and the dotted line has a ratio of 0.1.

3.2. Sensitivity analysis

For sensitivity analysis, we performed the Morris method on each compact model. We obtained the global sensitivity and interaction factors of every

Journal Pre-proof

variable. Results of the sensitivity analysis for the "best Compact model" are displayed in Figure 5. Similar to the VI results, Temp, ICULOS and HR_freq occupy the top three places in terms of Morris analysis and their sensitivity is higher on the classification output than for other variables. Furthermore, Temp gains the largest importance value followed by ICULOS. ICULOS and Temp are two features that have the highest interaction impacts, which implies that there is an obvious interactional impact between these two features and the remaining features. The top 20 most influential features for sepsis early prediction computed by the Morris method are summarized in Table 5. 8 Raw features contributed to 56% of the overall impact (μ^*) and 60% of the interaction (σ) while 12 IM features devoted to 44% and 40% of μ^* and σ, respectively. In addition, the mean and standard deviation of top features are shown in Table 5. Naturally, for both sepsis and non-sepsis groups, there is no significant difference in the mean and standard deviation of discussed features between the development set and test set, because the data set was split randomly. While the mean value of each top feature is quite discriminating between the sepsis and non-sepsis samples, e.g. the average Temp and ICULOS values of sepsis data are markably higher than that of non-sepsis data both in training set and test set. Besides, the standard deviation reflects the dispersion of each feature's distribution, the vast majority of the top features' standard deviation are showing obvious differences between positive and negative samples.

4. Discussion

This paper provides two main contributions: the first one on the proposal of optimized machine learning methods for early detection of sepsis and the second one on novel methods to increase the interpretability of these methods.

Concerning the first point, a set of data processing and machine learning methods for the early detection of sepsis, based on Random Forests, have been proposed, evaluated and optimized. 24 model variants have been explored, while optimizing the feature sets and the handling of missing values. Among these, the best model, based on 142 features including 40 raw features and 102 informative missingness features, shows the highest utility score as well as the highest AUROC and recall/sensitivity. Only a small partition of events is missed. In contrast, the specificity of the best model is relatively low. Since the prevalence of events is low and the dataset is inclined to non-septic patients, it is difficult to achieve a low false-alarm rate.

Journal Pre-proof

Table 5: Top 20 features for sepsis prediction based on Morris analysis. The values in the table are standing for **mean(std)**. Detailed descriptions and units of the top features are listed in Appendix B.

Rank	Features	Development Set		Test Set	
		Sepsis	Non-sepsis	Sepsis	Non-sepsis
1	Temp	37.15(0.92)	36.88(0.65)	37.17(0.95)	36.87(0.69)
2	ICULOS	55.27(58.64)	26.46(27.17)	57.88(62.28)	26.74(29.09)
3	HR_freq	50.86(55.56)	23.51(25.69)	52.90(57.98)	23.54(26.61)
4	HR	91.25(18.90)	84.39(17.28)	88.16(19.12)	83.96(17.37)
5	FiO2_inte	7.10(23.26)	3.82(12.38)	6.62(21.42)	3.47(11.85)
6	O2Sat_freq	49.68(54.55)	22.80(25.22)	52.04(57.48)	22.93(26.36)
7	Resp	20.27(6.18)	18.64(4.93)	19.95(6.51)	18.58(5.03)
8	MAP_freq	50.01(54.96)	22.97(25.35)	52.09(57.21)	23.03(26.40)
9	PaCO2_inte	9.48(24.10)	4.52(12.42)	8.38(22.99)	4.46(13.33)
10	WBC	12.74(7.32)	11.28(6.48)	13.32(9.30)	11.28(8.29)
11	FiO2_freq	9.12(14.27)	2.72(6.06)	9.58(16.06)	3.01(7.32)
12	DBP_freq	41.43(53.08)	18.64(25.00)	40.71(52.51)	18.37(24.90)
13	Resp_freq	47.69(53.33)	22.02(24.37)	49.45(56.23)	22.08(25.80)
14	BUN	28.20(20.59)	23.26(16.96)	30.47(22.16)	22.67(18.53)
15	Creatinine	1.65(1.64)	1.46(1.54)	1.63(1.58)	1.43(1.72)
16	Lactate_inte	13.76(31.51)	4.33(15.03)	12.90(31.70)	4.27(16.14)
17	Lactate_freq	3.00(5.86)	0.97(2.83)	2.37(4.35)	0.88(2.37)
18	Creatinine_freq	3.71(4.11)	1.71(1.92)	3.72(3.80)	1.74(2.58)
19	HospAdmTime	-79.93(209.56)	-58.77(187.95)	-73.58(176.25)	-52.12(125.09)
20	SBP_freq	47.30(54.47)	22.44(25.12)	48.37(55.27)	22.31(25.73)

Journal Pre-proof

Journal Pre-proof

Moreover, the utility score defined by the challenge is rewarding true positive predictions way more than true negatives, and it is penalizing missing events heavily compared to false alarms. This fact of encouraging to seek as many as possible septic patients could explain why the models with high sensitivity provide the best utility scores. It is interesting to note that this score has been criticized since its initial proposal [21], but remains useful for comparison purposes with past works performed with this same database. Although a direct comparison of our proposed model with respect to those proposed in the Challenge is not possible, due to the fact that final evaluation dataset is not public, the performance provided by the model proposed in this paper is comparable to that published by the best teams of the challenge during the learning phase (Table 6).

Table 6: The results of the best teams' models of the Challenge on the shared dataset.

Rank in the Challenge	ACC	AUROC	AUPRC	F1-measure	U-Score
Unofficial 1 [16]	0.818	0.847	-	-	0.425
Official 1 [15]	-	-	-	-	0.430
Official 2 [24]	0.858	0.834	0.111	0.133	0.400
Official 3 [17]	0.844	0.8333	-	-	0.4281
Official 4 [23]	-	-	-	-	0.4149
Our best Full model	0.8188	0.8465	0.1030	0.1227	0.4274

We have also proposed a compact model, which is based on 20 features from the most important variables in the best Full model. This compact model provided a lower utility score but higher specificity and AUPRC. These results were derived from an independent patient group and demonstrated comparable performance as in the development data. This model is particularly interesting, in our sense, due to its compactness and has been retained for the sensitivity analysis phase.

A final interesting result in the model construction phase of this work is related to the informative function embedded into missing values (or low-quality data segments). Indeed, results show that the models integrating informative missingness features provide better performance than the other approaches. Specifically, two of the informative missingness features: the measurement frequency features and the measurement time interval features, when sum up, are of great impact to discriminate the spetic patients, while the binary masks indicating the absence of measurements provide trivial information. In the best full model (include raw features and informative

Journal Pre-proof

missingness features), the "frequency" IM features account for nearly 40% of the variable importance based on RF and the "interval" features take up about 14.5%, and most of them rank high among all features. When adding the other two features—window-based statistics features and empirical scoring features, informative missingness features are still the dominant features of the three sets of newly introduced variables, even if the total variable importance of informative missingness features decrease marginally ($\sim$ 5% of decline). In the model with "R+IM+WS" features, window-based statistics make up for approximately 16% of variable importance, which is only one third of that of informative missingness's. Similarly, with the introduction of empirical scoring features, the proportion between informative missingness features' variable importance and window-based statistics features' variable importance keep almost the same, while empirical scoring sets of features have no apparent contribution to the prediction (only $\sim 0.27\%$).

The second contribution of this paper is the thorough study of feature importance and sensitivity analysis in order to give insight on the interpretability of the model. Indeed, reliable onset time prediction of sepsis using AI models with interpretable and quantitative clinical biomarkers remains a priority for both clinicians and healthcare decision support [35].

From a methodological standpoint, we have shown that the classical RF-based variable importance and the proposed sensitivity analysis methods provide complementary information. In the first case, VI is established through calculating the normalized decrease Gini impurity for a given variable in the learning phase. In the proposed sensitivity analysis approach, the significance of a variable is estimated on the impact of a modification of its value on the classification output. This impact is estimated both on the mean direct effect of a given variable and from the interactions of this variable with the others. Within the top 10 variables, Temp, ICULOS, HR_freq, HR, FiO2_inte, O2Sat_freq and WBC are common between these two approaches, but their order is not the same. Our analysis shows, for instance, that variations on the observed temperature (Temp) are by far the most sensitive source of information for the proposed model for the estimation of Sepsis. The role of temperature monitoring is widely known in this application field. Our method highlights the importance of monitoring temperature variations. More interestingly, two other physiological parameters are highlighted not only for their direct impact, but also for their interactions: HR and Respiration. Although the role of HR is known from the literature on sepsis detection, the interest in respiratory analysis and, more importantly, on the

Journal Pre-proof

interactions between HR and respiration is more recent in the literature. Advanced analysis of HR and respiratory time-series are of major importance in this field. However, this requires a much higher observability rate, in order to obtain meaningful HR and respiratory time-series, as well as an increased data quality. Variables regarding the frequency and the intervals of observation of physiological variables are more related to the observation interest of a given individual and provide significant information on the attention given to a patient, in a similar manner than the informative missingness variables. It is particularly interesting to note the importance of these variables on the proposed model. The proposed sensitivity analysis provides quantitative information on the extent on which these physiological or non-physiological variables are exploited by the model, giving thus a further step towards the interpretability of the proposed classifier.

The main limitations of our study are related to alert fatigue, because of the high sensitivity but the quite low specificity. This lack of specificity is mainly due to the cost function that we have used. We used simple classification error rate in this work, while in other works in this same database have taken the utility score into consideration when designing the cost function [24, 36], which is an important and possible way to improve the results.

Another major limitation is the poor quality of the dataset, which is noisy, heterogeneous, and randomly sampled. This data quality problem, characteristic of such "real-life" datasets, are one of the major challenges for the development of machine learning systems in this field. In order to overcome this difficulty, we compared different strategies for missing value handling, giving interesting results. Moreover, the proposed sensitivity analysis gives also interesting information for the identification of which variables should be acquired with the highest quality level. Finally, we should notice that this dataset was collected over the past decade [6] and that current monitoring methods have allowed for an increased observability. In particular, it is now possible to observe directly the continuous signals acquired by the monitor, and this approach allows for the extraction of many more informative features, such as those from heart rate variability analysis. Our future work is directed towards the integration of such high-resolution data into our processing pipeline.

Journal Pre-proof

5. Conclusions

Early detection of sepsis events is still a major challenge on the handling of patients in intensive care units. In this work, we proposed optimized machine-learning methods for early detection of sepsis, integrating a data preprocessing step with different imputation stategies and class balancing, to deal with the low quality and limited observability on the dataset. Feature engineering was conducted by creating additional features and selecting the best feature subsets. The optimal hyperparameters for our machine learning models were greedily sought with the grid search strategy. A total of 24 models were developed, and the best one based on 142 features achieved a 0.4274 utility score, and the best compact model trained with 20 selected features obtained 0.3862. The second contribution of this work is the proposal of novel methods to increase the interpretability of the proposed models. The most important digital markers for the early detection of sepsis included Temp, HR series, FiO2 series, Resp series variables, etc. Future works are directed towards the use of high-resolution cardiorespiratory data, to increase the performance of the proposed models.

Journal Pre-proof

Journal Pre-proof

Appendix A: Descriptions of 40 features in the dataset.

Vital signs (columns 1-8)	
HR	Heart rate(beats per minute)
O2Sat	Pulse oximetry (%)
Temp	Temperature (Deg C)
SBP	Systolic BP (mm Hg)
MAP	Mean arterial pressure (mm Hg)
DBP	Diastolic BP (mm Hg)
Resp	Respiration rate (breaths per minute)
EtCO2	End tidal carbon dioxide (mm Hg)
Laboratory values (columns 9-34)	
BaseExcess	Measure of excess bicarbonate (mmol/L)
HCO3	Bicarbonate (mmol/L)
FiO2	Fraction of inspired oxygen (%)
pH	N/A
PaCO2	Partial pressure of carbon dioxide from arterial blood (mm Hg)
SaO2	Oxygen saturation from arterial blood (%)
AST	Aspartate transaminase (IU/L)
BUN	Blood urea nitrogen (mg/dL)
Alkalinephos	Alkaline phosphatase (IU/L)
Calcium	(mg/dL)
Chloride	(mmol/L)
Creatinine	(mg/dL)
Bilirubin_direct	Bilirubin direct (mg/dL)
Glucose	Serum glucose (mg/dL)
Lactate	Lactic acid (mg/dL)
Magnesium	(mmol/dL)
Phosphate	(mg/dL)
Potassium	(mmol/L)
Bilirubin_total	Total bilirubin (mg/dL)
TroponinI	Troponin I (ng/mL)
Hct	Hematocrit (%)
Hgb	Hemoglobin (g/dL)
PTT	partial thromboplastin time (seconds)
WBC	Leukocyte count (count*10^3/µL)
Fibrinogen	(mg/dL)
Platelets	(count*10^3/µL)
Demographics (columns 35-40)	
Age	Years (100 for patients 90 or above)
Gender	Female (0) or Male (1)
Unit1	Administrative identifier for ICU unit (MICU)
Unit2	Administrative identifier for ICU unit (SICU)
HospAdmTime	Hours between hospital admit and ICU admit
ICULOS	ICU length-of-stay (hours since ICU admit)

Journal Pre-proof

Appendix B: Descriptions of top 20 features based on Morris analysis.

Rank	**Features**	**Description**
1	Temp	Temperature (°C)
2	ICULOS	ICU length of stay (hr)
3	HR_freq	Measurement frequency of heart rate
4	HR	Heart rate (beats/min)
5	FiO2_inte	Measurement interval of fraction of inspired oxygen
6	O2Sat_freq	Measurement interval of pulse oximetry
7	Resp	Respiration rate (breaths/minute)
8	MAP_freq	Measurement frequency of mean arterial pressure
9	PaCO2_inte	Measurement interval of partial pressure of carbon dioxide from arterial blood
10	WBC	Leukocyte count (count/L)
11	FiO2_freq	Measurement frequency of FiO2
12	DBP_freq	Measurement frequency of diastolic BP
13	Resp_freq	Measurement frequency of Resp
14	BUN	Blood urea nitrogen (mg/dL)
15	Creatinine	Creatinine (mg/dL)
16	Lactate_inte	Measurement interval of lactic acid
17	Lactate_freq	Measurement frequency of Lactate
18	Creatinine_freq	Measurement frequency of Creatinine
19	HospAdmTime	Time between hospital and ICU admission (hr)
20	SBP_freq	Measurement frequency of systolic BP

Journal Pre-proof

Journal Pre-proof

References

[1] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, et al., The third international consensus definitions for sepsis and septic shock (sepsis-3), Jama 315 (8) (2016) 801–810.

[2] C. W. Seymour, V. X. Liu, T. J. Iwashyna, F. M. Brunkhorst, T. D. Rea, A. Scherag, G. Rubenfeld, J. M. Kahn, M. Shankar-Hari, M. Singer, et al., Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3), Jama 315 (8) (2016) 762–774.

[3] M. Shankar-Hari, G. S. Phillips, M. L. Levy, C. W. Seymour, V. X. Liu, C. S. Deutschman, D. C. Angus, G. D. Rubenfeld, M. Singer, Developing a new definition and assessing new clinical criteria for septic shock: for the third international consensus definitions for sepsis and septic shock (sepsis-3), Jama 315 (8) (2016) 775–787.

[4] Centers for disease control and prevention: Sepsis., [EB/OL], `https://www.cdc.gov/sepsis/datareports/index.html,`. Accessed February 1 2019.

[5] C. J. Paoli, M. A. Reynolds, M. Sinha, M. Gitlin, E. Crouser, Epidemiology and costs of sepsis in the united states—an analysis based on timing of diagnosis and severity level, Critical care medicine 46 (12) (2018) 1889.

[6] M. A. Reyna, C. Josef, S. Seyedi, R. Jeter, S. P. Shashikumar, M. B. Westover, A. Sharma, S. Nemati, G. D. Clifford, Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019, in: 2019 Computing in Cardiology (CinC), IEEE, 2019, pp. Page–1.

[7] C. W. Seymour, F. Gesten, H. C. Prescott, M. E. Friedrich, T. J. Iwashyna, G. S. Phillips, S. Lemeshow, T. Osborn, K. M. Terry, M. M. Levy, Time to treatment and mortality during mandated emergency care for sepsis, New England Journal of Medicine 376 (23) (2017) 2235–2244.

[8] V. X. Liu, V. Fielding-Singh, J. D. Greene, J. M. Baker, T. J. Iwashyna, J. Bhattacharya, G. J. Escobar, The timing of early antibiotics and

Journal Pre-proof

Journal Pre-proof

hospital mortality in sepsis, American journal of respiratory and critical care medicine 196 (7) (2017) 856–863.

[9] A. Kumar, D. Roberts, K. E. Wood, B. Light, J. E. Parrillo, S. Sharma, R. Suppes, D. Feinstein, S. Zanotti, L. Taiberg, et al., Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock, Critical care medicine 34 (6) (2006) 1589–1596.

[10] K. E. Henry, D. N. Hager, P. J. Pronovost, S. Saria, A targeted real-time early warning score (trewscore) for septic shock, Science translational medicine 7 (299) (2015) 299ra122–299ra122.

[11] J. S. Calvert, D. A. Price, U. K. Chettipally, C. W. Barton, M. D. Feldman, J. L. Hoffman, M. Jay, R. Das, A computational approach to early sepsis detection, Computers in biology and medicine 74 (2016) 69–73.

[12] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, T. G. Buchman, An interpretable machine learning model for accurate prediction of sepsis in the icu, Critical care medicine 46 (4) (2018) 547.

[13] A. Johnson, T. Pollard, L. Shen, L. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, R. Mark III, Mimic-iii, a freely accessible critical care database. sci data. 2016; 3: 160035 (2016).

[14] S. M. Lauritsen, M. E. Kalør, E. L. Kongsgaard, K. M. Lauritsen, M. J. Jørgensen, J. Lange, B. Thiesson, Early detection of sepsis utilizing deep learning on electronic health record event sequences, Artificial Intelligence in Medicine 104 (2020) 101820.

[15] J. Morrill, A. Kormilitzin, A. Nevado-Holgado, S. Swaminathan, S. Howison, T. Lyons, The signature-based model for early detection of sepsis from electronic health records in the intensive care unit, in: 2019 Computing in Cardiology (CinC), IEEE, 2019, pp. Page–1.

[16] M. Yang, C. Liu, X. Wang, Y. Li, H. Gao, X. Liu, J. Li, An explainable artificial intelligence predictor for early detection of sepsis, Critical Care Medicine 48 (11) (2020) e1091–e1096.

Journal Pre-proof

[17] M. Zabihi, S. Kiranyaz, M. Gabbouj, Sepsis prediction in intensive care unit using ensemble of xgboost models, in: 2019 Computing in Cardiology (CinC), IEEE, 2019, pp. Page–1.

[18] S. Lyra, S. Leonhardt, C. H. Antink, Early prediction of sepsis using random forest classification for imbalanced clinical data, in: 2019 Computing in Cardiology (CinC), IEEE, 2019, pp. 1–4.

[19] L. Tran, M. Nguyen, C. Shahabi, Representation learning for early sepsis prediction, in: 2019 Computing in Cardiology (CinC), IEEE, 2019, pp. 1–4.

[20] E. Macias, G. Boquet, J. Serrano, J. Vicario, J. Ibeas, A. Morel, Novel imputing method and deep learning techniques for early prediction of sepsis in intensive care units, in: 2019 Computing in Cardiology (CinC), IEEE, 2019, pp. 1–4.

[21] B. Roussel, J. Behar, J. Oster, A recurrent neural network for the prediction of vital sign evolution and sepsis in icu, in: 2019 Computing in Cardiology (CinC), IEEE, 2019, pp. Page–1.

[22] H. C. Prescott, T. J. Iwashyna, Improving sepsis treatment by embracing diagnostic uncertainty, Annals of the American Thoracic Society 16 (4) (2019) 426–429.

[23] X. Li, Y. Kang, X. Jia, J. Wang, G. Xie, Tasp: A time-phased model for sepsis prediction, in: 2019 Computing in Cardiology (CinC), IEEE, 2019, pp. Page–1.

[24] J. A. Du, N. Sadr, P. de Chazal, Automated prediction of sepsis onset using gradient boosted decision trees, in: 2019 Computing in Cardiology (CinC), IEEE, 2019, pp. Page–1.

[25] R. J. Little, D. B. Rubin, Statistical analysis with missing data, Vol. 793, John Wiley & Sons, 2019.

[26] D. B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.

[27] J.-H. Lin, P. J. Haug, Exploiting missing clinical data in bayesian network modeling for predicting medical problems, Journal of biomedical informatics 41 (1) (2008) 1–14.

Journal Pre-proof

[28] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. D. Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, L. G. Thijs, The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure, Intensive Care Medicine (1996).

[29] G. B. Smith, D. R. Prytherch, P. Meredith, P. E. Schmidt, P. I. Featherstone, The ability of the national early warning score (news) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death, Resuscitation 84 (4) (2013) 465–470.

[30] L. Breiman, Random forest, Machine Learning 45 (2001) 5–32.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[32] A. K. Saltelli, K. Chan, E. M. Scott, Sensitivity analysis, 2000.

[33] M. D. Morris, Factorial sampling plans for preliminary computational experiments, Technometrics 33 (2) (1991) 161–174.

[34] F. Campolongo, J. Cariboni, A. Saltelli, An effective screening design for sensitivity analysis of large models, Environmental Modelling & Software 22 (10) (2007) 1509–1518.

[35] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable ai for trees, Nature machine intelligence 2 (1) (2020) 2522–5839.

[36] Y. Chang, J. Rubin, G. Boverman, S. Vij, A. Rahman, A. Natarajan, S. Parvaneh, A multi-task imputation and classification neural architecture for early prediction of sepsis from multivariate clinical time series, in: 2019 Computing in Cardiology (CinC), IEEE, 2019, pp. Page–1.

Journal Pre-proof

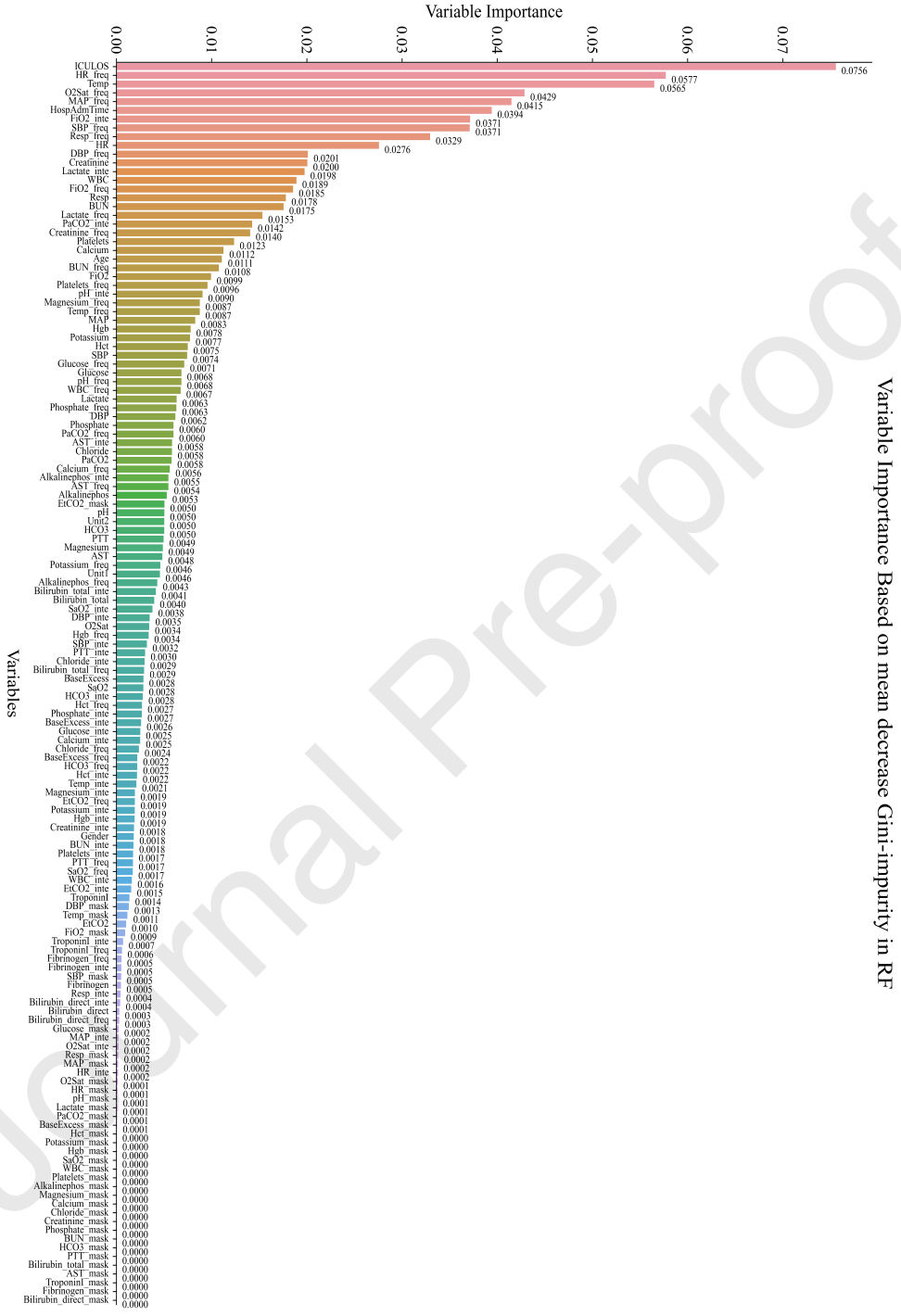
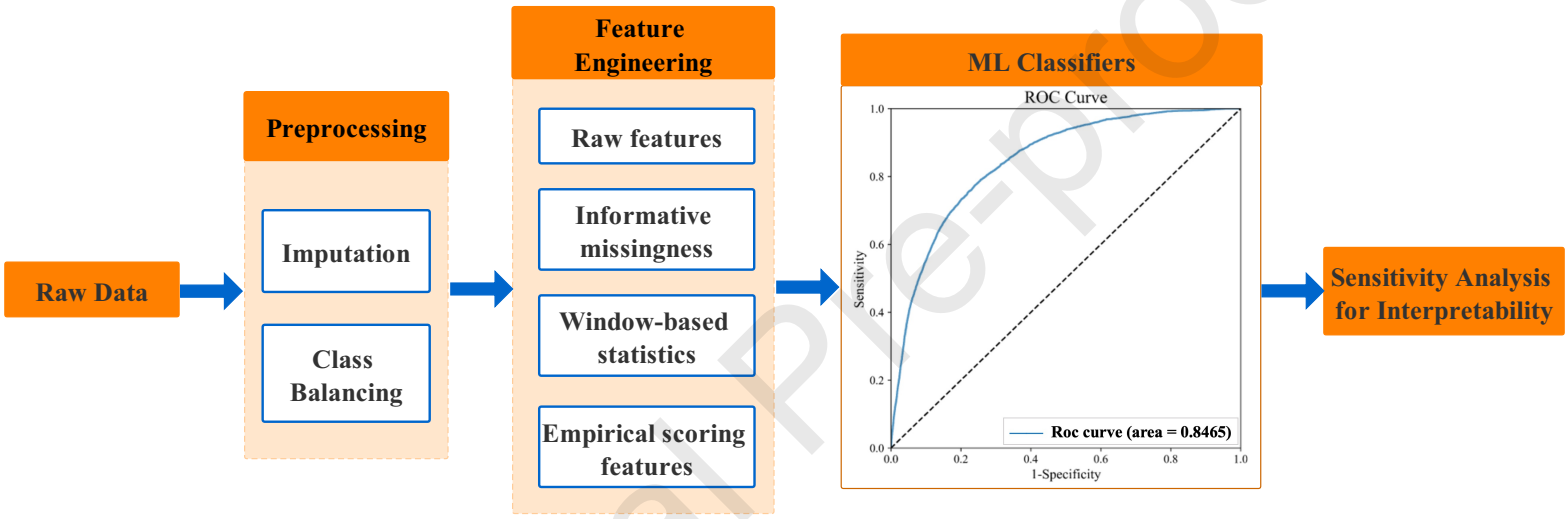


Figure S1: Variable Importance ranking for the best full RF model.



Journal Pre-proof

IRBM-graphical abstract

Journal Pre-proof

Author responsibilities, integrity, ethics

This is an **editable** PDF form. It should **be saved to your computer, then completed** using Adobe reader or equivalent. Please **do NOT substitute** any other document (text file, scanned image, etc.).

Article title : Towards an explainable model for Sepsis detection based on

Human and animal rights

☐ The authors declare that the work described has been carried out in accordance with the Declaration of Helsinki of the World Medical Association revised in 2013 for experiments involving humans as well as in accordance with the EU Directive 2010/63/EU for animal experiments.

☐ The authors declare that the work described has not involved experimentation on humans or animals.

Informed consent and patient details

☑ The authors declare that this report does not contain any personal information that could lead to the identification of the patient(s) and/or volunteers.

☐ The authors declare that they obtained a written informed consent from the patients and/or volunteers included in the article and that this report does not contain any personal information that could lead to their identification.

☐ The authors declare that the work described does not involve patients or volunteers.

Disclosure of interest

☑ The authors declare that they have no known competing financial or personal relationships that could be viewed as influencing the work reported in this paper.

☐ The authors declare the following financial or personal relationships that could be viewed as influencing the work reported in this paper:

Funding

☑ This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

☐ This work has been supported by:

Journal Pre-proof

Author contributions

☑ All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

☐ All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship. Individual author contributions are as follows:

Journal Pre-proof

Journal Pre-proof

M. Chen : Data curation, Formal analysis, Writing- Original draft preparation, Software. **Alfredo Hernández**: Conceptualization, Methodology, Formal analysis, Resources, Writing- Reviewing and Editing.

Journal Pre-proof