Limitations of Naïve Feature Aggregation in Predicting Sepsis Onset: A Study Using the MIMIC-IV Database

Claas de Boer
Institute for Medical Informatics and Biometry
Technische Universität Dresden
Dresden, Germany
claas.de_boer@mailbox.tu-dresden.de

Jonas Höpner
Institute for Medical Informatics and Biometry
Technische Universität Dresden
Dresden, Germany
jonas.hoepner@mailbox.tu-dresden.de

***Abstract*— In this study, we investigate a naive approach to feature engineering for the prediction of sepsis onset in Intensive Care Unit (ICU) patients using the Sepsis-3 criteria and the MIMIC-IV database. The naive approach involves the straightforward aggregation of features across a feature window into single scalar values, utilizing basic aggregation functions such as minimum, maximum, mean, and standard deviation. In addition, we explore the potential predictive power of the mere presence of diagnostic measurements, questioning whether the availability of specific measurements alone suffices to predict sepsis onset. Our findings reveal that the feature aggregation approach leads to only marginal improvements in prediction over random guessing with the simple classifiers deployed. Moreover, our analysis confirms that the presence of variables alone is not indicative of the onset of sepsis. The results cast doubts on the efficacy of such an elementary approach in a complex and critical task like sepsis prediction. The study serves as a reminder of the intricacies involved in handling time-series medical data and highlights the need for more sophisticated methods to harness the potential information embedded within.**

***Index Terms*— MIMIC-IV, Sepsis-3, Machine Learning, Precision Medicine**

I. Introduction

Sepsis is a life-threatening condition that arises when the body's response to infection causes injury to its tissues and organs [14]. It remains a leading cause of death in hospitals worldwide, contributing to nearly one-third of all hospital deaths. Despite the advances in healthcare technology and critical care medicine, sepsis incidence is still growing with a high mortality rate [7, 12]. Thus, the ability to predict its onset can significantly influence patient outcomes by enabling early and appropriate interventions.

The Sepsis-3 criteria, introduced in 2016 by the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3), represent a significant shift in how sepsis is defined and diagnosed[14]. The Sepsis-3 defines sepsis as a life-threatening organ dysfunction due to a dysregulated host response to infection. Organ dysfunction is identified as an increase in the Sequential (Sepsis-related) Organ Failure Assessment (SOFA) score by two points or more, which is associated with an in-hospital mortality greater than 10%.

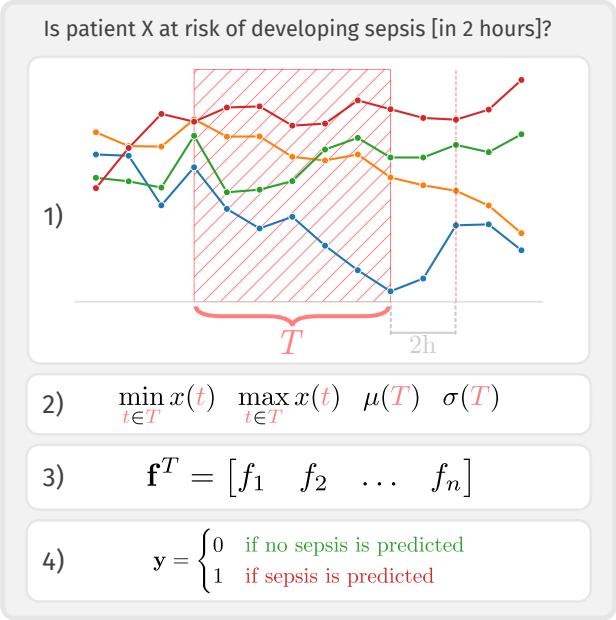

Fig. 1. Workflow for predicting sepsis onset in patients, comprising 1) Feature Extraction from the MIMIC-IV database, where line plots represent the temporal evolution of key features such as heart rate, blood pressure, and other vital signs, providing an overview of patient status during ICU stay, 2) Feature Engineering using the following aggregation functions over the feature window: minimum (min), maximum (max), mean (μ), and standard deviation (σ), 3) Compiling time series data into a single Feature Vector, and 4) Prediction of sepsis development within two hours (1) or a healthy outcome (0) through binary classification.

Septic shock, a subset of sepsis, is identified with a clinical construct of sepsis with persisting hypotension requiring vasopressors to maintain Mean Arterial Pressure (MAP) of 65mm Hg or higher, and having a serum lactate level greater than 2 mmol/L despite adequate volume resuscitation. The Sepsis-3 criteria underscore the importance of organ dysfunction in sepsis and move away from an infection-centered definition, thus providing a more precise method to identify and treat septic patients. The application of these criteria for early sepsis detection in intensive care units forms a part of our predictive

modeling approach discussed in this paper.

The rich and publicly available Medical Information Mart for Intensive Care (MIMIC-IV) database serves as the foundation of our investigation [6]. MIMIC-IV offers de-identified health data from 299,712 patients who were in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. This extensive dataset, encompassing demographic details, vital signs, laboratory results, medications, and more, is an invaluable resource for developing and validating predictive models.

For the purposes of this study and our predictive model, we do not differentiate between sepsis and septic shock. Although septic shock represents a more severe subset of sepsis according to the Sepsis-3 criteria, it is essential to understand that both conditions signify a dysregulated host response to infection leading to life-threatening organ dysfunction. Our primary objective is to predict the onset of this dysregulated response, irrespective of whether it progresses to the septic shock stage.

Recognizing the increasing complexity of machine learning models in sepsis onset prediction, our study explores a contrasting approach. We investigate the effectiveness of simpler classifiers using a straightforward, or naive, feature engineering approach. This method attempts to simplify complex time series data into single scalar values, using basic aggregation functions such as maximum, minimum, mean, and standard deviation - see Figure 1 for an outline of the predictive approach.

However, it's important to highlight that simplicity, while advantageous in terms of computational efficiency and interpretability, may have its limitations. In this paper, we critically evaluate the effectiveness of this naive feature engineering approach within the crucial two-hour window for sepsis onset prediction, employing the comprehensive MIMIC-IV database and the Sepsis-3 criteria. Our findings suggest that, at least in the specific setup of our study, this simplified approach does not appear to be well-suited to predict sepsis onset accurately. Through a detailed discussion of these results, we hope to shed light on the potential limitations and challenges of implementing such straightforward feature engineering approaches in the context of sepsis onset prediction.

II. Methods

A. Inclusion Criteria for ICU stays

This study primarily concentrates on singular ICU visits rather than individual patients. We encompass all ICU stays of patients aged eighteen years and older, as documented in the MIMIC-IV database. However, stays that have a duration less than eight hours, and those instances where sepsis develops within eight hours of the ICU stay, are excluded. The selection criterion adopted for this study is further detailed in figure 2. A comprehensive illustration of the configuration of our predictive approach's feature window can be found in section II-C.

B. Sepsis Onset

For the purpose of our study, we employ a distinct strategy for the time window and calculation of the Sequential Organ Failure Assessment (SOFA) score[8]. In brief, we use a rolling 24-hour window to calculate the SOFA score on an hourly basis. This method allows us to continually update the SOFA score based on the most recent 24 hours of patient data, providing an accurate reflection of a patient's dynamic physiological state. The query to provide the data for the rolling window was provided in *MIT-LCP/mimic-code* by Johnson et al.[4]. In the generated table was the average sofa score of the respectively last hour provided. The onset of sepsis in our study is defined as the first time the SOFA score increases by 2 points or more within this rolling 24-hour window.

For ICU stays that do not result in sepsis (forming our control group), it becomes crucial to designate a precise point in time for our prediction model. To accomplish this, we introduce a 'hypothetical' sepsis onset time. This hypothetical onset is determined through a random selection of an actual sepsis onset time from the pool of sepsis-positive ICU stays within our dataset. The sepsis onset time refers to the time at which the Sequential Organ Failure Assessment (SOFA) score first rises by two points.

To select this time, we compute the difference between the sepsis onset time and the time at which the patient is admitted to the ICU, referred to as the 'offset'. This offset is then added to the ICU admission time of a sepsis-negative patient to establish the hypothetical onset time. This can be illustrated by the following equation:

$$onset'_{control} = adm_{control} + (onset_{case} - adm_{case})$$

For a more detailed illustration of the process, please refer to Figure 4. In this figure, the sepsis onset time for a patient who tested positive for sepsis (referred to as 'Case') is represented by the solid red line. Conversely, the dashed green line indicates the randomly matched onset time for a sepsis-negative patient, or 'Control'.

This method ensures that we have a comparable time point for both sepsis and non-sepsis cases, thereby maintaining a uniform framework for analysis. Our approach closely mirrors the 'absolute onset matching' strategy described by Moor et al.[10], which refers to taking the absolute time since admission until sepsis onset for a case and assigning it as the matched onset time for a control. This consistent application of onset times across all cases ensures a consistent dataset concerning the used data and facilitates a more valid comparison of the predictive model's performance between sepsis and non-sepsis cases.

Still, it should be noted that the resulting onset time distribution varies between the case and control group as seen in Table I.

C. Time Window for Sepsis Prediction

For the prediction of sepsis and training our models, it is necessary to define a feature window from which we extract data to learn. Following the suggestion of Moore et al., we

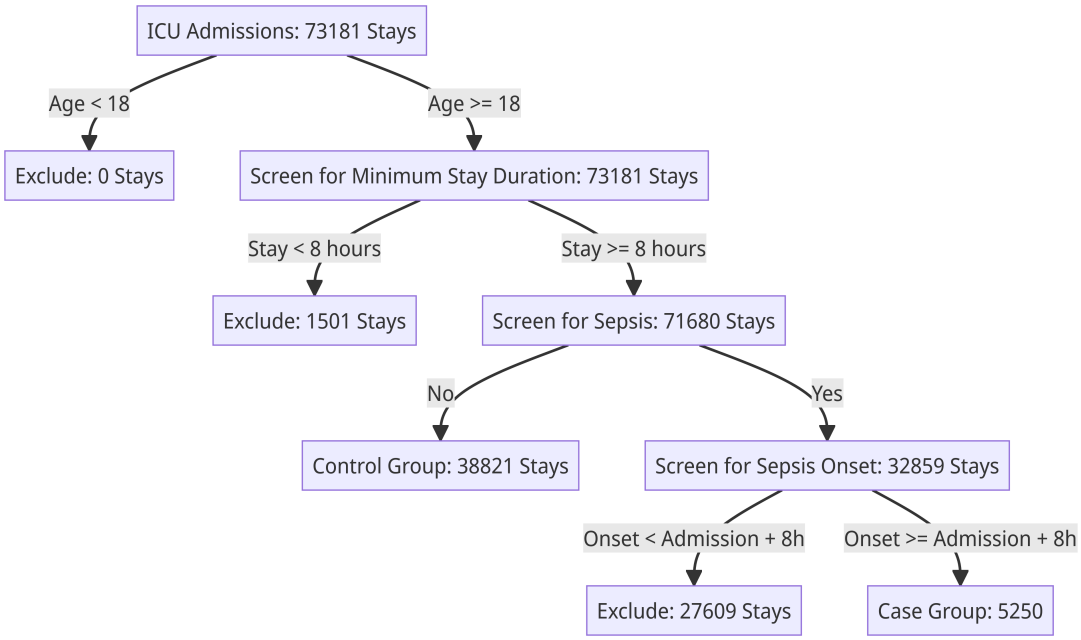

Fig. 2. Patient Inclusion Flow Diagram - The flow diagram represents the process of inclusion for the study, delineating the path of ICU stay patients. The case group is made up of 5250 ICU stays that last longer than eight hours and develop a sepsis no earlier than eight hours into the stay. The control group is made up of ICU stays that last at least eight hours and which do not develop a sepsis during their stay.

TABLE I
CHARACTERISTICS OF THE POPULATION INCLUDED IN THE DATASET. THE MEAN SEPSIS ONSET IS GIVEN IN HOURS SINCE ADMISSION TO THE ICU.

Variable	**Sepsis Cases**	**Controls**
n	5250	38821
Female	2350 (45.87%)	17806 (44.76%)
Male	2900 (55.24%)	21015 (54.13%)
Mean time to sepsis onset in ICU (median)	31.26h (15.35h)	15.90h (12.27h)
Age ($\mu \pm \sigma$)	64.8 ± 16.9	63.9 ± 17.4
Ethnicity		
White	3470 (66.8%)	26328 (68.3%)
Black or African-American	570 (10.9%)	4432 (11.5%)
Hispanic or Latino	189 (3.6%)	1504 (3.9%)
Asian	156 (3.0%)	1153 (2.9%)
Other	198 (3.8%)	1346 (3.5%)
Unknown	614 (11.8%)	3789 (9.8%)

employ the method of gapped offline training [10]. In this method, the predictive models access data from the entire feature window for sepsis patients until a specified point in time (prediction horizon) before the actual onset of sepsis occurs. This feature window's size can be further extended into the past as needed. Illustrations of this approach can be found in both Figure 1 (Box 1)) and Figure 4, which highlights the feature window of a specific ICU stay with a blue rectangle.

For our study, we have set the prediction horizon to two hours before the detected onset of sepsis. Additionally, the feature window designated for prediction encompasses a six-hour period preceding this two-hour prediction horizon. This means that our model relies on data from the range of eight to two hours prior to the onset of sepsis.

The prediction horizon itself functions as the juncture at which predictions for individual cases are computed. At this stage, the model is tasked with answering the question: "Will patient X develop sepsis in the next two hours?". This task is formulated as a binary classification problem.

Importantly, we intentionally exclude data from the immediate two hours before sepsis onset (creating a gap of two hours). This deliberate exclusion aligns with real-world clinical scenarios where the early signs of sepsis might not be immediately identified or responded to. Thus, our approach ensures that the model reflects the actual conditions and constraints of clinical practice.

D. Feature Selection & Engineering

For feature selection, we commence by identifying possible predictor variables within the intensive care unit (ICU) data. These encompass vital signs, laboratory results, blood gas analysis, and composite scores such as the Glasgow Coma Scale. To evaluate the relevance of each potential feature, we compute the Pearson Correlation Coefficient with the

target variable, utilizing data from the feature window as defined in the previous subsection. This approach assists us in pinpointing key variables that may significantly influence the prediction of sepsis onset. Furthermore, we sought the expertise of a physician who recommended the inclusion of specific markers for sepsis. These markers, namely lactate levels, blood pH, blood CO2 concentration, and arterial-alveolar oxygen difference (AaDO2), have been identified as indicators in sepsis diagnosis and were therefore integrated into our analysis.

In line with our aim of building a predictive model that captures the most influential factors, we establish a correlation threshold and include only those features with an absolute correlation value greater than 0.10. This process ensures our model focuses on variables that have a comparatively reasonable correlation with sepsis onset compared to the other features. While a Pearson correlation coefficient of 0.10 is typically not viewed as representing a "strong" correlation, given the context of the other correlation values in our analysis, it seemed to serve as a sensible cutoff point. The correlations where computed between aggregated values and the sepsis target variable.

To condense the features selected for our predictive model within the defined feature window, we employ naïve aggregation methods. Specifically, for each feature within the window, we compute the minimum, maximum, mean, and standard deviation, thereby reducing time-series data into single-value representations. These aggregated values form the components of a feature vector, effectively capturing the range, central tendency, and variability of each feature's values over the specified window. By converting the time series data into scalar values, this method not only encapsulates key characteristics of the data but also reduces the dimensionality of our dataset. As illustrated in Figure 1, we applied the following functions and their corresponding PostgreSQL function equivalents to compute the maximum (Equation 1), minimum (Equation 2), mean (Equation 3), and standard deviation (Equation 4) of the time series $x(t)$ in the time interval T:

$$\max_{t \in T} x(t) \tag{1}$$

$$\min_{t \in T} x(t) \tag{2}$$

$$\mu(T) = \frac{1}{N} \sum_{t \in T} x(t) \tag{3}$$

$$\sigma(T) = \sqrt{\frac{1}{N} \sum_{t \in T} (x(t) - \mu(T))^2} \tag{4}$$

The rationale behind this aggregation strategy is twofold. Firstly, it significantly simplifies the data structure, reducing the high-dimensional time-series data to more manageable feature vectors. Secondly, it enables the use of simple, interpretable classifiers for our predictive model. Instead of resorting to more complex, computationally heavy time-series based machine learning approaches, we can apply well-established classification algorithms that are easier to interpret and validate.

E. Data Sparsity Analysis

The scarcity or absence of certain medical measurements can sometimes provide unexpected insights into patient conditions. In our study, we aimed to explore whether the sparsity of such data might carry hidden information pertinent to sepsis diagnosis.

We therefore implement an unconventional approach where we train the models on *presence encoded* data. This involved representing the actual values with a '1' and substituting any returned null values with a '0'. This encoding strategy enables us to create a binary framework representing only the presence or absence of specific measurements.

By employing this methodology, we sought to investigate whether the distribution of these presence markers was non-random and whether machine learning models might leverage this pattern to differentiate between sepsis and non-sepsis patients. Specifically, we examined if the scarcity or abundance of certain measurements, rather than their intrinsic values, could reveal distinguishing characteristics for sepsis onset prediction.

F. Machine Learning Setup

We employed a variety of machine learning (ML) models for the binary classification task, including well-recognized algorithms from the scikit-learn library. These encompassed the random forest classifier, extra trees classifier, decision tree classifier, gradient boosting classifier, logistic regression, histogram-based gradient boosting classifier (HistGradientBoostingClassifier), and the Gaussian Naive Bayes Classifier (GaussianNB) . In addition to these, we utilized the "binary:logistic" model from the XGBoost library. We selected Logistic Regression as our initial benchmark or baseline model to facilitate a comparison of the performance of the other models against a fundamental standard. This selection served as a benchmark to evaluate how effectively the other models performed relative to a basic criterion. All models were consistently trained using a fixed 80/20 training/test split, ensuring uniformity in assessment. We utilized grid search to identify the best hyperparameters for the XGBoost model, allocating 10% of the training set for validation purposes. However, we did not tune the hyperparameters of the classifiers that were made available through scikit-learn.

We also constructed a specialized Voting classifier that combined some of the models using a majority vote method. This method necessitated the selection of an odd number of models to avoid potential ties, and as such, we selected the top five models based on their Matthews Correlation Coefficient (MCC) scores[1, 9]. The decision to focus on these models was made to optimize the combined model's performance. The specific models that were chosen, based on their ranking in the top five according to the MCC score, will be detailed and reported in the results section for the given experiments.

For the evaluation of these models, we selected the following metrics, chosen for their ability to offer a comprehensive insight into the classifiers' performance:

- MCC (Matthews Correlation Coefficient) provides a balanced measure of a binary classifier's quality, taking into account true and false positives and negatives, and ranges from -1 (complete disagreement between prediction and actual) to +1 (perfect prediction) [3, 2].
- F1 Score balances precision and recall, with a range of 0 to 1, where a higher value means better performance, particularly when dealing with imbalanced classes.
- ROC AUC (Receiver Operating Characteristic Area Under the Curve) is a summary measure that evaluates the model's ability to discriminate between positive and negative classes, ranging from 0.5 (no discrimination) to 1 (perfect discrimination).

G. Implementation

In the implementation of our analysis, we utilized MIMIC-IV v2.2, hosted on a PostgreSQL 15.2 database, with the concept queries provided by the MIMIC-Code repository for the rolling 24-hour SOFA window and sepsis3 diagnosis [4]. The matched onset times and cohort selection tables were created by us, with the corresponding code available on GitHub[1]. The psycopg2 2.9.6 package served as the interface between our Python code and the database, and for data preprocessing and manipulation, we used pandas 2.0.1 and numpy 1.23.2. The data visualization was performed using matplotlib 3.7.1 and seaborn 0.12.2, and the machine learning models were provided by scikit-learn 1.2.2 and xgboost 1.7.6.

III. Results

A. Inclusion Criteria

In the MIMIC-IV database, a total of 73,181 ICU stays are recorded, all of which involve adult patients aged 18 or above. Among these stays, 1,501 are excluded from the control group as they last shorter than the requisite eight hours. From the remaining 71,680 stays, 38,821 do not result in the development of sepsis, while 32,859 do. Notably, out of the 32,859 stays associated with sepsis, only 5,250 occur more than eight hours into the ICU stay. A comprehensive patient flow diagram detailing these statistics is provided in Figure 2, the distribution of sepsis onset times is illustrated in Figure 3. We can therefore report a sepsis prevalence of 11.91% in the study cohort as well as a prevalence of 44.90% in the MIMIC-IV database based on individual ICU stays.

The characteristics, including demographical information, of both the sepsis case population and the control population are detailed in Table I. The gender distribution is slightly skewed towards males, with 55.24% in sepsis cases and 54.13% in controls; females comprise 45.87% of the sepsis cases and 44.76% of the controls. The mean time to sepsis onset in the ICU for sepsis cases is 31.26 hours (with a median of 15.35 hours), while for controls it is 15.90 hours (with a median of 12.27 hours). The mean ages for the sepsis cases and controls are 64.8 ± 16.9 and 63.9 ± 17.4 years, respectively. In terms of ethnicity, the majority are White (66.8% in sepsis cases and 68.3% in controls), followed by Black or African-American, Hispanic or Latino, Asian, Other, and Unknown, with each group showing a slight variation between the sepsis cases and controls.

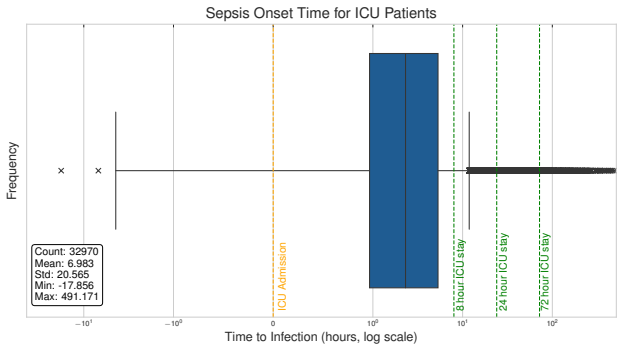

Fig. 3. Boxplot illustrating the distribution of sepsis onset times for ICU patients. The median time to sepsis onset, shown by the line within the box, represents the time point at which half of the patients developed sepsis. The box encompasses the interquartile range (IQR), representing the middle 50% of sepsis onset times. The whiskers extend to the most extreme data points not considered outliers, and any individual points outside this range represent potential outliers. Key time points such as 8 hours, 24 hours, and 72 hours are highlighted, indicating significant periods during the ICU stay.

B. Feature Selection & Engineering

The features that either achieve a Pearson Correlation Coefficient greater than 0.10 or were suggested as potential markers by medical personnel, and are therefore included in the feature set for our study, can be found in Table IV. In total, 73 features were selected, leading to a feature vector of 73 elements for each ICU stay. Among the selected features for the final predictive models are statistics of heart rate, systolic blood pressure (sbp), diastolic blood pressure (dbp), mean blood pressure (mbp), respiratory rate, body temperature, standard deviation of platelets count, and statistics of albumin level, among others.

C. Sepsis Onset Prediction

Upon training the classifiers with the selected features, the following observations were made. All classifiers fall short of exceeding an F1 score of 0.25 or an ROC AUC of 0.6 (see Table II) except for the XGBoost Classifier. Such values, specifically an F1 score under 0.25 and an ROC AUC less than 0.6, point to an underwhelming performance, signaling difficulties in maintaining both precision and recall, as well as a lack of efficacy in differentiating between the classes. Of the group, the XGBoost Classifier emerges as the standout, registering an F1 score of 0.34 and an ROC AUC score of 0.63. Yet, these figures further underline the broader trend of unsatisfactory performance in balancing precision with recall and in class separation. As per the MCC metric, the XGBoost Classifier again leads with a score of 0.26, which nonetheless

[1]https://github.com/cdboer/kp-medizinische-informatik-ss23

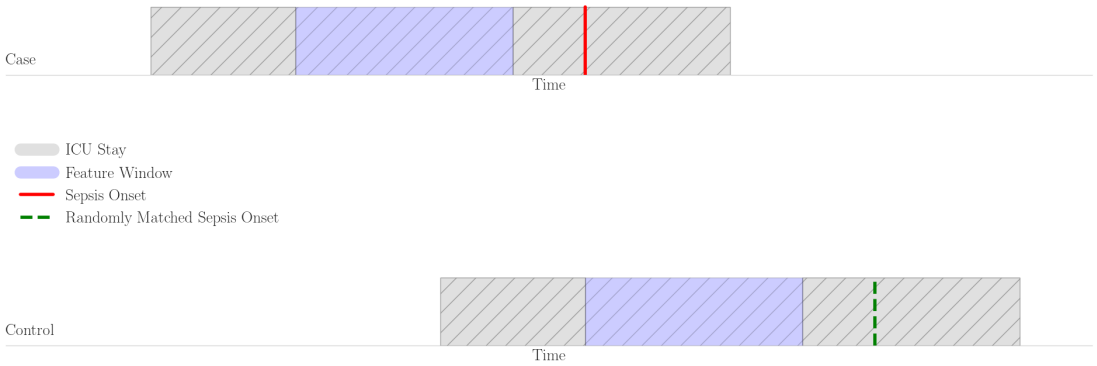

Fig. 4. Schematic representation of the feature window for positive and control cases in the study. Every case has a specific sepsis onset. For positive cases, the feature window is defined based on the time of sepsis onset. Given a random control, our method sets a random, fictitious sepsis onset time, replicating the timing structure of positive cases to ensure a fair comparison. Figure adapted from [10], see offline training scenario.

hints at a mere weak positive correlation between the predicted and actual labels. In a nutshell, the analysis reveals that the classifiers' performance is hardly distinguishable from random guessing, none of them were able to reliably predict the development of sepsis in ICU patients within the next two hours.

The logistic regression, which was utilized as our baseline model, yielded the lowest performance among all trained classifiers in terms of f1 and roc auc score. For the ensemble method using majority vote, the voting classifier incorporated the five best models according to MCC. These models included the XGBoost classifier, RandomForestClassifier, GradientBoostingClassifier, HistGradientBoostingClassifier, and ExtraTreesClassifier. The ensemble method performed marginally better than some of its individual constituent models, yet it was still surpassed in performance by the XGBoost classifier.

TABLE II
METRICS EVALUATING THE PERFORMANCE OF CLASSIFIERS TRAINED ON THE UNENCODED DATASET, INCLUDING MCC, F1 SCORE, AND ROC AUC SCORE. TOP VALUES ARE EMPHASIZED IN BOLD TEXT.

Classifier	MCC	F1	ROC AUC
RandomForestClassifier	0.23	0.16	0.54
DecisionTreeClassifier	0.13	0.24	0.57
LogisticRegression	0.12	0.07	0.52
GradientBoostingClassifier	0.22	0.15	0.54
ExtraTreesClassifier	0.19	0.16	0.54
HistGradientBoostingClassifier	0.22	0.17	0.55
GaussianNB	0.11	0.24	0.57
SupportVectorClassifier	0.18	0.09	0.52
XGBoost Classifier	**0.26**	**0.34**	**0.63**
VotingClassifier	0.23	0.17	0.55

D. Data Sparsity Analysis

In our examination of the distribution of retrieved measurements between the case and control cohorts, we identified several noteworthy patterns (refer to Figure 6).

Firstly, we observed that features computed using standard deviation often led to missing features. This phenomenon was predominantly due to the uneven distribution of standard deviation between the case and control groups. An underlying reason for this imbalance is found in the implementation detail of the standard deviation function in PostgreSQL. Specifically, the function requires at least two measurements to compute a value; otherwise, it returns NaN or Null, contributing to the observed discrepancies.

Secondly, when analyzing the null value percentages for various measurements, we found mixed patterns. The null values for vital sign measurements were roughly equally distributed between case and control groups, indicating a more balanced distribution. On the other hand, GCS scores displayed an initial uneven distribution, hinting at potential variances between the two cohorts.

Lastly, measurements taken from the blood gas or chemistry tables exhibited a starkly uneven distribution. Therefore, we would initially expect the presence encoded features to facilitate shortcut learning due to the uneven distribution of null values between the case and control groups.

Upon transforming the dataset using presence encoding, significant changes in the performance of various models were observed. In particular, the HistGradientBoosting, GradientBoostingClassifier and SupportVectorClassifier performed the best, attaining the highest MCC of 0.19 The HistGradientBoosting, RandomForestClassifier and XGBoost classifier lead the group in ROC AUC performance with a shared score of 0.53. For the F1 score, the GaussianNB classifier was the frontrunner, although it achieved a relatively modest score of 0.13. Contrary to our initial assumption, the use of presence encoding appears to have caused a decline in the overall performance across all classifiers. This suggests that the intrinsic values of the original dataset might have contained some minor but useful information for the classification task. While differences in null value distribution were discernible among selected features, this encoding strategy did not allow the classifiers to capitalize on any shortcuts. As a result, the models' ability to differentiate between classes only slightly surpassed that of pure chance. Similar to the results obtained with the unencoded data, the MCC, F1, and ROC AUC scores for each classifier are detailed (refer to Table III). The voting

ensemble for the voting classifier in this experiment consists of the HistGradientBoostingClassifier, the GradientBoostingClassifier, the SupportVectorClassifier, the ExtrasTreesClassifier and the LogisticRegression.

TABLE III
METRICS EVALUATING THE PERFORMANCE OF CLASSIFIERS TRAINED ON THE ENCODED DATASET (PRESENCE ENCODING), INCLUDING MCC, F1 SCORE, AND ROC AUC SCORE. TOP VALUES ARE EMPHASIZED IN BOLD TEXT.

Classifier	MCC	F1	ROC AUC
RandomForestClassifier	0.17	0.10	**0.53**
DecisionTreeClassifier	0.17	0.10	0.52
LogisticRegression	0.17	0.08	0.52
GradientBoostingClassifier	**0.19**	0.09	0.52
ExtraTreesClassifier	0.17	0.10	0.52
HistGradientBoostingClassifier	**0.19**	0.11	**0.53**
GaussianNB	0.08	**0.13**	0.52
SupportVectorClassifier	**0.19**	0.09	0.52
XGBoost Classifier	0.17	0.11	**0.53**
VotingClassifier	0.18	0.09	0.52

IV. DISCUSSION

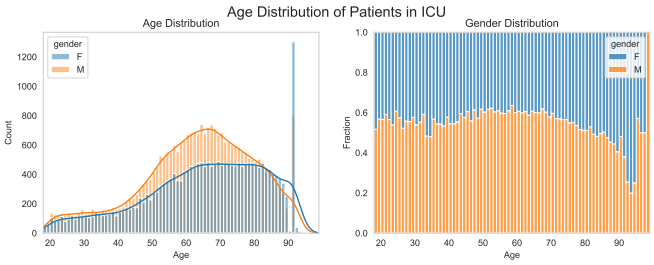

Fig. 5. Age Distribution and Gender Breakdown of ICU Patients: The left subplot displays the age distribution of ICU patients as a histogram, with bins representing each year of age. The right subplot illustrates the distribution of gender within the set of ICU patients, providing a percentage breakdown for each year of age to show how many patients are reported as female (F) and how many as male (M).

A. Age Distribution of ICU Patients

Our cohort selection presented some unique characteristics. Since the MIMIC-IV ICU database only includes data for adult patients, the inclusion of underage patients was not an option. Additionally, we observed an unexplained spike in the number of patients aged 91 and 92. Interestingly, there were roughly twice as many patients in these age groups compared to those aged 90 or 93. This occurrence remained unexplained and stands as a peculiar feature of the MIMIC-IV dataset.

B. Imputation

In our analysis, we faced the challenge of handling missing data within the feature window. Rather than attempting to fill these gaps by extrapolating from existing data patterns outside of the feature window, we opted for a more straightforward approach. We chose to impute missing values by simply replacing them with ”-1.” This method allowed us to avoid making assumptions about the data outside the window of interest, keeping our calculations simple and focused on the data at hand. While this approach does have its benefits in terms of simplicity, it also means that we do not leverage as much data for training as we might have with a more complex imputation method.

C. Individual ICU Stays vs. Individual Patients

By considering each qualifying ICU stay as an independent instance, we recognize that a single patient may have multiple ICU stays, each representing a distinct episode of care with varying conditions, complications, and risk factors for sepsis. Other studies with the same premis While this approach enables the inclusion of all adult ICU stays meeting our criteria, it also introduces potential risks. Treating each stay as independent can overlook the cumulative effects of multiple stays on a patient’s overall health and may amplify the influence of patients with numerous stays in the dataset. These challenges necessitate careful consideration in the implementation of our stay inclusion criteria. Other studies consider individual patients and exclude subsequent ICU stays from their study cohort [11].

D. SOFA Score Computation

The computation of the hourly SOFA (Sequential Organ Failure Assessment) score for ICU patients requires careful evaluation. Typically, the SOFA score is calculated over a standard 24-hour period [14]. However, in this scenario, the data necessary for calculating the SOFA score only begins to accumulate after a patient’s admission to the ICU. Therefore, the SOFA score for the first 24 hours of a stay relies on data gathered from a timeframe less than the required 24-hour window. Additionally, the resolution of computing the SOFA score once per hour may be too low to accurately detect the actual onset point of sepsis, potentially affecting the precision of our analysis. Since the median time to sepsis onset for both the case and control groups occurs within this initial 24-hour period (see Table I), questions arise regarding the suitability of the SOFA score calculation as a reliable foundation for our study. Further investigation is needed to determine its applicability and relevance. For a discussion of this particular approach of SOFA score computation, see Johnson et. al [5].

E. Feature Engineering using Aggregation

We used aggregation processes to simplify the data by computing the minimum, maximum, mean, and standard deviation for each feature within the window. However, this aggregation process does not come without trade-offs. Specifically, this approach discards all time-specific data within the feature window, meaning that we lose any temporal patterns or sequences that might be present in the raw time-series data. For example, the maximum value of heart rate within the window could be due to a transient event, such as the patient exerting energy while changing position, which might not have significant predictive value for the onset of sepsis. Furthermore, the aggregation may obscure the relationship between fluctuations in the measurements and the onset of sepsis. Ultimately, under consideration of the low achieved scores, this method does not

prove as worthy compared to the savings of computational power, indicating that the benefits of computational efficiency do not compensate for the loss in potential predictive accuracy.

F. No External Validation

The approach chosen for this study is limited in its scope as it has only been evaluated using the MIMIC-IV dataset. This dataset, while comprehensive, may not fully represent the diverse range of patient profiles and conditions that could be encountered in other healthcare settings. The lack of external validation restricts the generalizability of our findings and raises questions about how the model might perform with different populations, in varying clinical environments, or under distinct protocols and procedures.

G. Circularity of Sepsis-3

Schamoni et al. [13] point out that using clinical measurements for predicting a sepsis label, which was itself derived from clinical measurements, could potentially be circular (a statistical term referring to the fact that one uses the same data for the selection of a model and its subsequent analysis). This runs the risk being unable to discover unknown aspects of the data, since classifiers may just confirm existing criteria instead of helping to generate new knowledge. In the worst case, a classifier would merely reiterate the guidelines used to define sepsis without being able to detect patterns that permit an earlier discovery. Sepsis-3 is a set of clinical criteria that outlines specific rules for measurements to identify a defined point in time as the onset of sepsis. As such, it could be considered a potentially circular measurement, since it uses certain parameters to define a condition that is then analyzed using those same parameters.

H. Skewed Data Distribution

The distribution of null values in our study demonstrates a pronounced skewness towards patients who are more severely ill (compare Figure 6), such as those who develop sepsis. This imbalance is not arbitrary but stems from the clinical practices often observed in Intensive Care Units (ICUs). Medical personnel tend to order more diagnostic tests for critically ill patients, naturally leading to a greater abundance of data for these individuals.

V. Future Work

Several potential directions for future research warrant exploration:

- Investigate alternative naive approaches, such as time series decomposition into trend, seasonality, and residual components. However, the effectiveness of such methods is questionable if we extrapolate from our own study and experience, suggesting that naive approaches may not be the most fruitful avenue to pursue.
- Explore the subdivision of the feature window into multiple smaller windows to reintroduce the time component into the machine learning approach.
- Consider the inclusion of time series data directly within the models, to possibly capture more nuanced patterns associated with sepsis development.
- Adjust the time window for prediction by conducting a horizon evaluation, such as examining 2, 4, 6, or 8-hour intervals. This approach, proposed by Moor et al.[10], could provide a more refined understanding of prediction dynamics over different time frames.

VI. Conclusions

The outcomes of the machine learning methodology have predominantly fallen short of expectations. None of the classifiers achieved evaluation scores that indicated a performance significantly superior to mere random speculation. As a result, the proposed technique of aggregating time series into scalar values using basic functions can largely be dismissed as unfit for the specialized task of predicting the onset of sepsis, at least within the confines of our research design. Furthermore, the distribution of clinical measurements between case and control groups failed to facilitate any discernible shortcut learning performance. However, we believe that our study lays down a straightforward and robust methodical framework for predicting sepsis using the sepsis-3 criterion. The preprocessing and feature extraction procedures established here may serve as a firm base for further inquiries into the development of sepsis in ICU patients.

Acknowledgments

We would like to thank Prof. Dr. med. Mario Menk of the University Hospital Carl Gustav Carus Dresden for his help and support in answering many of our domain specific questions. In addition, we would like to express our sincere gratitude and appreciation to our supervisors Dr. Markus Wolfien, Ian-Christopher Jung, Waldemar Hahn, and Katharina Schuler for their consistent support and guidance throughout the semester.

References

[1] Pierre Baldi et al. “Assessing the accuracy of prediction algorithms for classification: an overview”. In: *Bioinformatics* 16.5 (May 2000). _eprint: https://academic.oup.com/bioinformatics/article-pdf/16/5/412/48836094/bioinformatics_16_5_412.pdf, pp. 412–424. ISSN: 1367-4803. DOI: 10 . 1093 / bioinformatics/16.5.412. URL: https://doi.org/10.1093/bioinformatics/16.5.412.

[2] Davide Chicco and Giuseppe Jurman. “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation”. In: *BMC Genomics* 21.1 (Dec. 2020). Number: 1 Publisher: BioMed Central, pp. 1–13. ISSN: 1471-2164. DOI: 10.1186/s12864-019-6413-7. URL: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6413-7 (visited on 07/12/2023).

[3] Davide Chicco and Giuseppe Jurman. "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification". en. In: *BioData Mining* 16.1 (Feb. 2023), p. 4. ISSN: 1756-0381. DOI: 10.1186/s13040-023-00322-4. URL: https://biodatamining.biomedcentral.com/articles/10.1186/s13040-023-00322-4 (visited on 08/06/2023).

[4] Alistair Johnson et al. *MIT-LCP/mimic-code: MIMIC Code v2.2.1*. Version v2.2.1. July 11, 2022. DOI: 10.5281/ZENODO.6818823. URL: https://zenodo.org/record/6818823 (visited on 05/02/2023).

[5] Alistair E. W. Johnson et al. "A Comparative Analysis of Sepsis Identification Methods in an Electronic Database*:" in: *Critical Care Medicine* 46.4 (Apr. 2018), pp. 494–499. ISSN: 0090-3493. DOI: 10.1097/CCM.0000000000002965. URL: http://journals.lww.com/00003246-201804000-00002 (visited on 05/02/2023).

[6] Alistair E. W. Johnson et al. "MIMIC-IV, a freely accessible electronic health record dataset". In: *Scientific Data* 10.1 (Jan. 3, 2023), p. 1. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01899-x. URL: https://www.nature.com/articles/s41597-022-01899-x (visited on 05/02/2023).

[7] Kirsi-Maija Kaukonen et al. "Mortality Related to Severe Sepsis and Septic Shock Among Critically Ill Patients in Australia and New Zealand, 2000-2012". In: *JAMA* 311.13 (Apr. 2, 2014), p. 1308. ISSN: 0098-7484. DOI: 10.1001/jama.2014.2637. URL: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2014.2637 (visited on 07/21/2023).

[8] Simon Lambden et al. "The SOFA score—development, utility and challenges of accurate assessment in clinical trials". In: *Critical Care* 23.1 (Nov. 27, 2019), p. 374. ISSN: 1364-8535. DOI: 10.1186/s13054-019-2663-7. URL: https://doi.org/10.1186/s13054-019-2663-7.

[9] B. W. Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405.2 (1975), pp. 442–451. ISSN: 0005-2795. DOI: https://doi.org/10.1016/0005-2795(75)90109-9. URL: https://www.sciencedirect.com/science/article/pii/0005279575901099.

[10] Michael Moor et al. "Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review". In: *Frontiers in Medicine* 8 (2021). ISSN: 2296-858X. URL: https://www.frontiersin.org/articles/10.3389/fmed.2021.607952 (visited on 05/02/2023).

[11] Shamim Nemati et al. "An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU:" in: *Critical Care Medicine* 46.4 (Apr. 2018), pp. 547–553. ISSN: 0090-3493. DOI: 10.1097/CCM.0000000000002936. URL: http://journals.lww.com/00003246-201804000-00010 (visited on 05/02/2023).

[12] Kristina E Rudd et al. "Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study". In: *The Lancet* 395.10219 (Jan. 2020), pp. 200–211. ISSN: 01406736. DOI: 10.1016/S0140-6736(19)32989-7. URL: https://linkinghub.elsevier.com/retrieve/pii/S0140673619329897 (visited on 07/21/2023).

[13] Shigehiko Schamoni et al. "Leveraging implicit expert knowledge for non-circular machine learning in sepsis prediction". en. In: *Artificial Intelligence in Medicine* 100 (Sept. 2019), p. 101725. ISSN: 09333657. DOI: 10.1016/j.artmed.2019.101725. URL: https://linkinghub.elsevier.com/retrieve/pii/S0933365718305700 (visited on 08/06/2023).

[14] Mervyn Singer et al. "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)". In: *JAMA* 315.8 (Feb. 23, 2016), pp. 801–810. ISSN: 0098-7484. DOI: 10.1001/jama.2016.0287. URL: https://doi.org/10.1001/jama.2016.0287 (visited on 05/04/2023).

APPENDIX A
TABLES & FIGURES

TABLE IV

OUTCOME OF FEATURE SELECTION PROCESS: THIS TABLE PRESENTS THE 73 UNIQUE FEATURES CHOSEN FOR THE FINAL PREDICTIVE MODELS. THE SOURCE MIMIC-IV TABLE FOR EACH FEATURE IS INDICATED, ALONG WITH THE AGGREGATION METHODS I.E., MEAN (μ) OR STANDARD DEVIATION (σ), UTILIZED TO GENERATE THE FEATURES. A CONCISE EXPLANATION FOR EACH FEATURE IS PROVIDED IN THE DESCRIPTION COLUMN. PLEASE NOTE THAT A LINE (—) SIGNIFIES A CONTINUATION FROM THE LAST EXPLICITLY STATED VALUE IN THE ROWS ABOVE. WE HAVE ADOPTED THE POSTGRESQL SCHEMA NOMENCLATURE RECOMMENDED BY ALISTAIR JOHNSON IN THE MIMIC-CODE REPOSITORY FOR THEIR CONCEPT QUERIES [4]. HENCE, EACH TABLE UNDER THE MIMICIV_DERIVED SCHEMA HAS BEEN CONSTRUCTED BASED ON A CONCEPT QUERY FROM THE MIMIC-CODE REPOSITORY.

Feature	Aggregation	Description	MIMIC-IV Table	MIMIC-IV Table Description
heart_rate	min, max, μ, σ	Statistics of heart rate	mimiciv_derived.vitalsign	Vital Sign Measurements
sbp	min, max, μ, σ	Statistics of systolic blood pressure	—	—
dbp	min, max, μ, σ	Statistics of diastolic blood pressure	—	—
mbp	min, max, μ, σ	Statistics of mean blood pressure	—	—
resp_rate	min, max, μ, σ	Statistics of respiratory rate	—	—
temperature	min, max, μ, σ	Statistics of body temperature	—	—
platelets	σ	Standard deviation of platelets count	mimiciv_derived.complete_blood_count	Results of Complete Blood Count Tests
albumin	min, max, μ	Statistics of albumin level	mimiciv_derived.chemistry	Biochemistry Panel (also known as Chemistry Labs)
globulin	min, max, μ	Statistics of globulin level	—	Biochemistry Panel (also known as Chemistry Labs)
abs_neutrophils	σ	Standard deviation of absolute neutrophils count	mimiciv_derived.blood_differential	Results of Differential Blood Tests
atyps	σ	Standard deviation of atypical neutrophils count	—	—
imm_granulocytes	min, max, μ, σ	Statistics of immature granulocytes count	—	—
abs_eosinophils	σ	Standard deviation of absolute eosinophils count	—	—
abs_monocytes	σ	Standard deviation of absolute monocytes count	—	—
bands	σ	Standard deviation of band cells count	—	—
metas	σ	Standard deviation of metamyelocytes count	—	—
gcs_motor	min, max, μ, σ	Statistics of Glasgow Coma Scale motor response score	mimiciv_derived.gcs	Results of Glasgow Coma Score Assesments
gcs_verbal	min, max, μ, σ	Statistics of Glasgow Coma Scale verbal response score	—	—
gcs_eyes	min, max, μ, σ	Statistics of Glasgow Coma Scale eye opening score	—	—
gcs_total	min, max, μ, σ	Statistics of Glasgow Coma Scale total score	—	—
aado2_calc	min, max, μ, σ	Statistics of calculated A-a gradient	mimiciv_derived.bg	Results of Blood Gas (BG) Analysis
lactate	min, max, μ, σ	Statistics of lactate level	—	—
totalco2	min, max, μ, σ	Statistics of total CO2 level	—	—
ph	min, max, μ, σ	Statistics of blood pH value	—	—

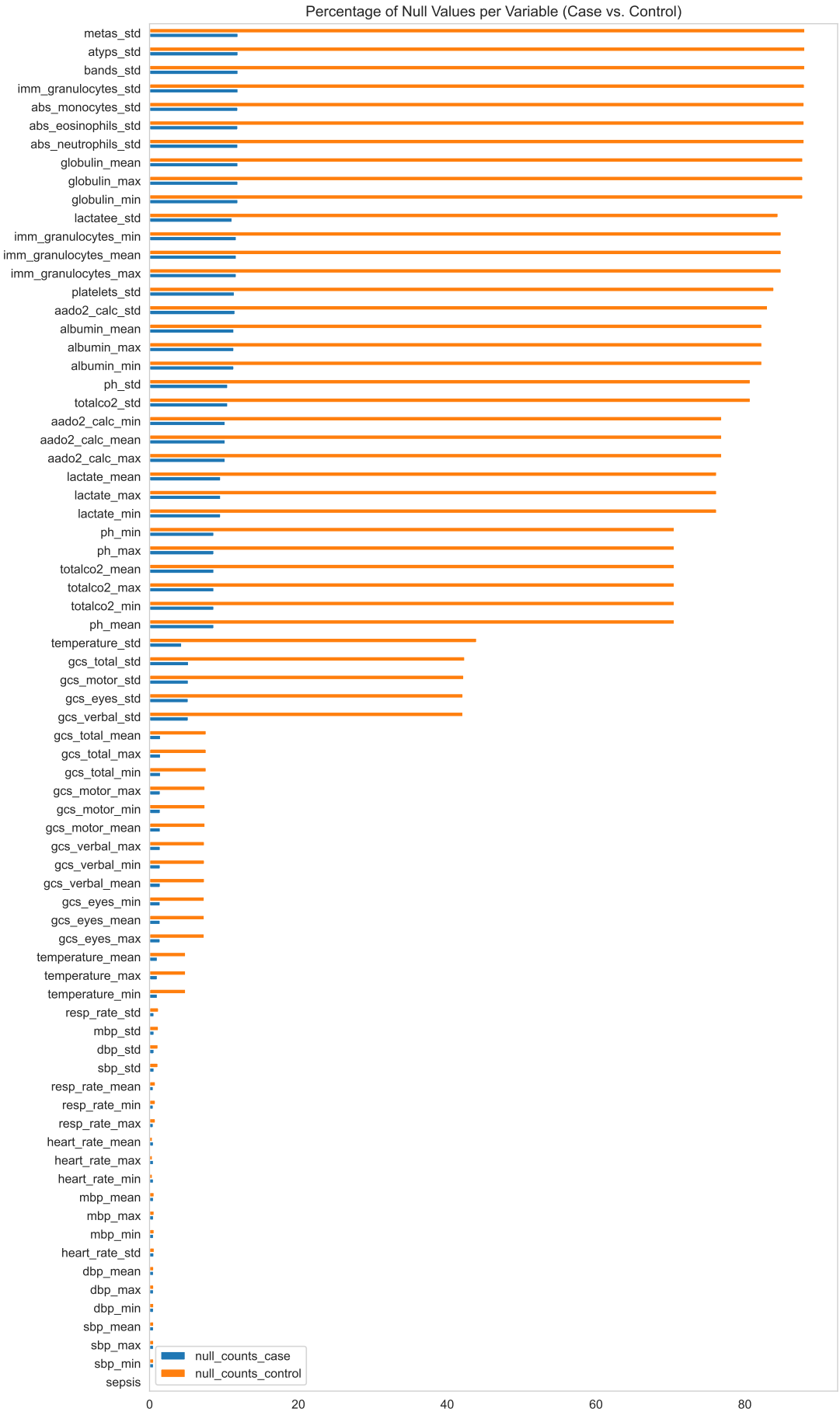

Fig. 6. Bar Chart: The chart illustrates the distribution of retrieved null values for each selected feature, with the values being shown as percentages. These are sorted by the total percentage counts across both cases and controls. A distinct shift in the null value distribution is observable specifically for controls. For instance, the uppermost part of the selected features unveils up to 90% missing data for this group.