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*— Disclaimer: This manuscript is a preliminary draft. Any comments or suggestions regarding potential modifications to our experimental approach are marked in red. Notations that need to be elaborated into full text are highlighted in green. Statements that have been highlighted in purple are those for which we are unsure of their scientific validity. Please note that the transitions between sections and paragraphs are still rough which makes the text seem disjointed. This will change until the final version. The rest of the abstract is filler text to test initial formatting. We appreciate your understanding and patience as we continue to refine this document.**

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum. [3]

***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 [10]. 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 [5, 9]. 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[10]. 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%. 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,

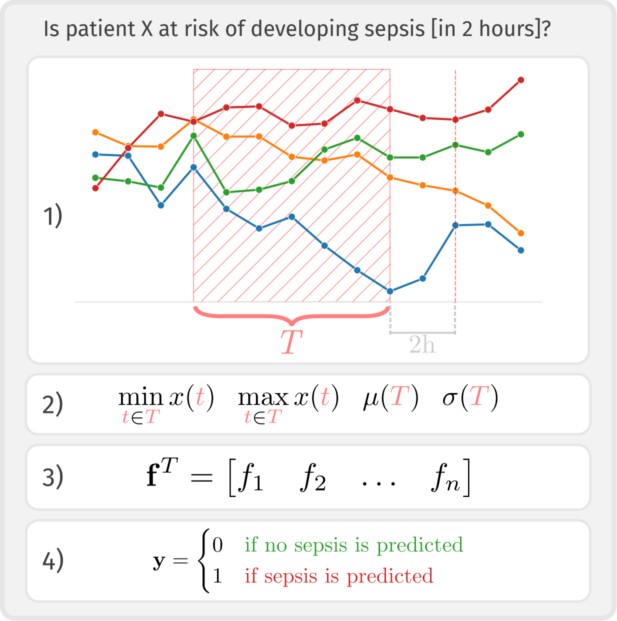
Is patient X at risk of developing sepsis [in 2 hours]?
1)
T
2h
2)
$\min_{t \in T} x(t)$
$\max_{t \in T} x(t)$
$\mu(T)$
$\sigma(T)$
3)
$\mathbf{f}^T = [f_1 \quad f_2 \quad \dots \quad f_n]$
4)
$\mathbf{y} = \begin{cases} 0 & \text{if no sepsis is predicted} \\ 1 & \text{if sepsis is predicted} \end{cases}$

Fig. 1. Workflow for predicting sepsis onset in patients, comprising 1) Feature Extraction from the MIMIC-IV database, 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.

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 [4]. 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.

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[6]. 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.[3]. 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.[8], 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

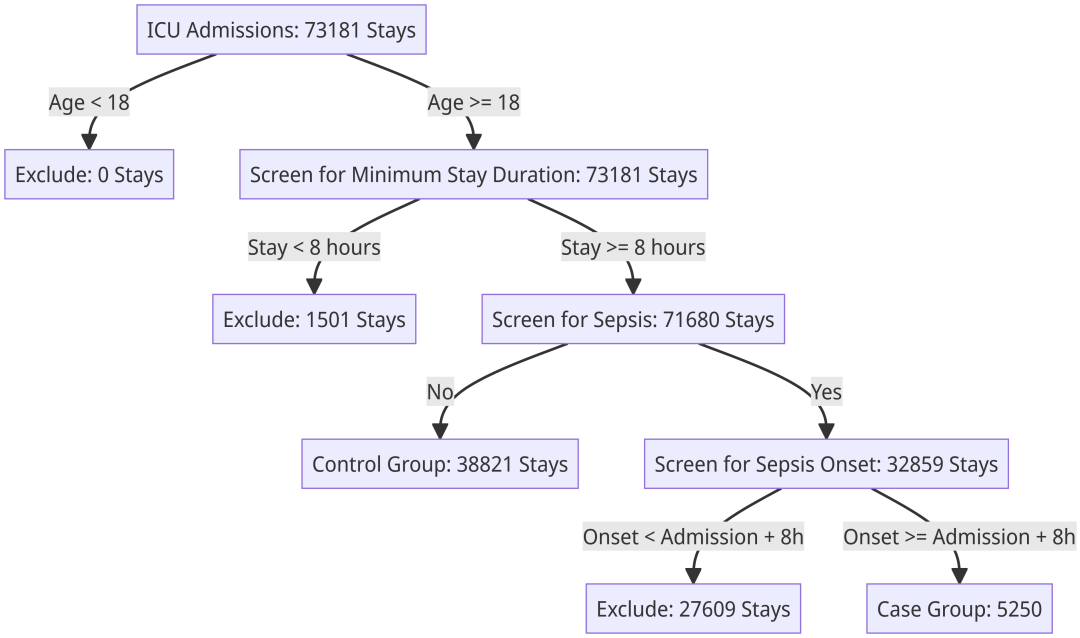

Fig. 2. Patient Inclusion Flow Diagram - The flow diagram represents the process of inclusion for the study, delineating the path of ICU stay patients. The case group is made up of 5250 ICU stays that last longer than eight hours and develop a sepsis no earlier than eight hours into the stay. The control 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%)

al., we employ the method of gapped offline training [8]. In this method, the predictive models access data from the entire feature window for sepsis patients until a specified time frame 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 in 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 the 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?"

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

The primary task of feature selection and engineering revolves around the definition of the target variable, which is derived based on whether sepsis develops during an ICU stay. This dichotomous variable, 'sepsis', categorizes each ICU stay as a 'sepsis' or 'non-sepsis' case.

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 our methodology. 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 vital indicators in sepsis diagnosis and were therefore integrated into our analysis.

To condense the features selected for our predictive model within the defined feature window, we employ 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 data:

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

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

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

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

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.

In line with our aim of building a predictive model that captures the most influential factors by naïve aggregation, 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 this analysis, it seems to serve as a sensible cutoff point.

Finally, it's important to note that the Pearson correlation coefficients used for feature selection are calculated between these aggregate values and the target variable. This process ensures that the selected features have a demonstrable relationship with the outcome of interest, irrespective of the time at which individual measurements were taken.

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. Implementation

In the implementation of our analysis, we utilized the 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 [3]. The matched onset times and cohort selection tables were created by us, with the corresponding code available on GitHub. 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. Through the combination of these tools and methods, we were able to efficiently conduct our study and ensure the reproducibility of our work, thus laying the groundwork for future research.

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 characteristics, including demographical information, of both the sepsis case population and the control population are detailed in Table I. The distribution of sepsis onset times is illustrated in Figure 3.

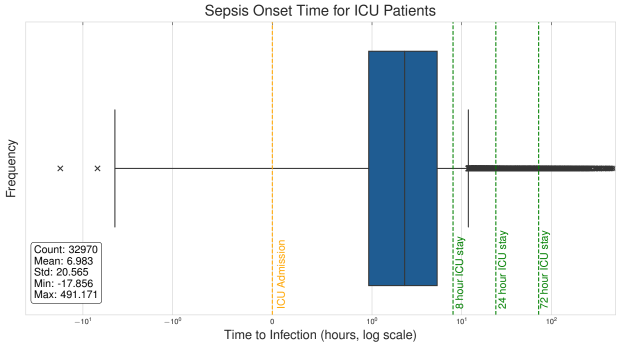

Fig. 3. Boxplot illustrating the distribution of sepsis onset times for ICU patients. The median time to sepsis onset, shown by the line within the box, represents the time point at which half of the patients developed sepsis. The box encompasses the interquartile range (IQR), representing the middle 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 0.1 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.

C. Explorative Data Analysis (EDA) Pittfalls

During our exploratory data analysis, we encountered several pitfalls that warrant consideration. Notably, when we included each ICU stay irrespective of its length, our data aggregation approach led to a distinctive null-value distribution between sepsis and non-sepsis cases. When employing presence encoding, where null features were denoted as 0 and actual values as 1, the XGBoost binary classifier achieved an MCC of 0.88. This outcome, however, was not satisfactory given that it was equivalent to the performance obtained with unencoded features. Such results allude to the potential issue of shortcut learning leveraged by the model, where it may be primarily learning from the null-value distribution instead of truly deciphering significant patterns within the feature space. As such, care must be taken to validate the source of model performance to ensure it is not artificially inflated by such artifacts in the data.

It is crucial to note that data collection frequency for ICU patients is significantly higher compared to non-ICU patients. By including all ICU stays, we inevitably encompass instances where sepsis developed immediately after admission. This introduces a complication as our time window for feature extraction ranges from eight hours before onset to two hours before onset, which in some cases precedes the patient's ICU admission. In such cases, particularly during the computation of standard deviations (where PostgreSQL's stddev, an alias for stddev_samp, requires at least two feature values), the available data might be insufficient. Furthermore, our method of randomly assigning sepsis onset from sepsis-positive cases to our control group can lead to control cases with sparse data, with up to 99% of rows potentially being null. This discrepancy in data distribution presents a considerable challenge for effective modeling and interpretation. Das wurde gefixt

D. Machine Learning

In our study, we utilized various machine learning (ML) models, including those available in the scikit-learn library such as RandomForestClassifier, ExtraTreesClassifier, DecisionTreeRegressor, GradientBoostingClassifier, LogisticRegression, HistGradientBoostingClassifier, GaussianNB, and KNeighborsClassifier, as well as the "binary:logistic" model from XGBoost. LogisticRegression served as our baseline model, acting as a reference for assessing the performance of the other classifiers.

Furthermore, we implemented a Voting classifier that integrated several models, including RandomForestClassifier, GradientBoostingClassifier, ExtraTreesClassifier, and HistGradientBoostingClassifier. The selection of these specific models for the voting ensemble was informed by their individual Matthews Correlation Coefficient (MCC) scores[1, 7], with the top five models being taken into consideration for inclusion. Detailed MCC scores for each model can be accessed in Table **??**. TODO: update the shortcut values with the broken, but correct ones

TABLE II
PERFORMANCE METRICS FOR UNENCODED

Classifier	**MCC**	**F1**	**ROC AUC**
RandomForestClassifier	**0.225**	0.163	0.543
DecisionTreeClassifier	0.134	**0.243**	**0.570**
LogisticRegression	0.102	0.061	0.514
GradientBoostingClassifier	0.223	0.148	0.539
ExtraTreesClassifier	0.194	0.161	0.542
HistGradientBoostingClassifier	0.222	0.173	0.546
GaussianNB	0.106	0.237	0.567
KNeighborsClassifier	0.185	0.200	0.553
SupportVectorClassifier	0.180	0.087	0.522
VotingClassifier	0.233	0.158	0.542
XGBoost Classifier	0.223	0.209	0.557

E. Data Sparsity Analysis

IV. DISCUSSION

A. Inclusion Criteria

Our cohort selection entailed some peculiarities. The MIMIC-IV ICU database contains only data for adult patients, so no underage patients could be considered. On the other side, we found a spike in the number of people with ages 91 and 92. There were roughly twice as much patients than patients

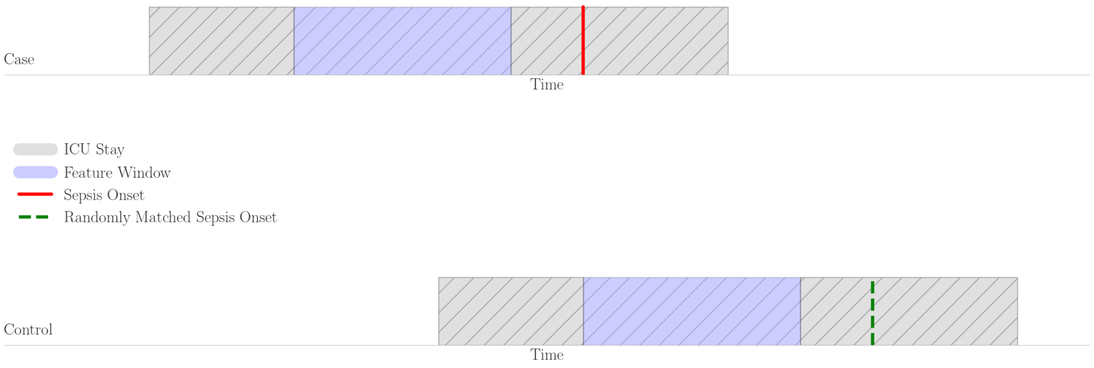

Fig. 4. Schematic representation of the feature window for positive and control cases in the study. Every case has a specific sepsis onset. For positive cases, the feature window is defined based on the time of sepsis onset. Given a random control, our method sets a random, fictitious sepsis onset time, replicating the timing structure of positive cases to ensure a fair comparison. Figure adapted from [8], see offline training scenario.

TABLE III
PERFORMANCE METRICS FOR PRESENCE ENCODING

Classifier	MCC	F1	ROC AUC
RandomForestClassifier	0.170	0.099	0.525
DecisionTreeClassifier	0.166	0.096	0.524
LogisticRegression	0.173	0.082	0.521
GradientBoostingClassifier	0.187	0.094	0.524
ExtraTreesClassifier	0.174	0.098	0.525
HistGradientBoostingClassifier	**0.194**	0.105	**0.527**
GaussianNB	0.078	**0.127**	0.525
KNeighborsClassifier	0.180	0.101	0.526
SupportVectorClassifier	0.187	0.094	0.524
VotingClassifier	—	—	—
XGBoost Classifier	—	—	—

with age 90 or 93. We could not explain this occurence, but this should not influence the outcome of this paper.

Due to keeping data and calculations simple, we decided to not use imputations on missing data but instead replace them with ”-1” in the data. Therefore we have the advantage of using real data everywhere with the downside of not having as much data to train on. For more information about this missing data, see IV-C.

By considering each qualifying ICU stay as an independent instance, we account for the fact that a single patient can have multiple ICU stays, with each stay representing a unique episode of care that could present different conditions, complications, and risk factors for sepsis. Consequently, all adult ICU stays meeting these criteria are included, regardless of the number of ICU stays a single patient may have had. The implementation of our stay inclusion criteria is necessary for the efficiency of our predictive model; this is primarily due to the feature window used in our machine learning approach, which necessitates a minimum time frame of 8 hours prior to the onset of sepsis for reliable predictive outcomes.

B. Feature Selection & Engineering

In this paper, 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 tradeoffs. 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. That said, these limitations are balanced against the increased simplicity and computational efficiency of using aggregate features in a conventional classification model. Ultimately, under consideration of the low achieved scores this does not prove as worthy compared to the savings of computational power and so on.

As we did not use imputation to fill in missing data, we had multiple features, which did not contain any data for some patients. To keep the ML model from learning from too much missing data, we had to filter features, which contained more than waren es 50%? null values over all patients. This removed mainly features from the blood gas analysis and laboratory data from our features, as those are not regularly taken, as for example vital signs are measured. Additionally, if a patient develops a sepsis, his condition can be described as generally worse, which implicates more necessary care and also more measured data for him. We could see that sepsis negative patients had way less data measured as sepsis positive patients.

C. Null value analysis

D. Limitations

- Only hourly onset calculation (resolution may not be precise enough, best we could get)
- Circular dependency of sepsis-3 definition on actions of medical personal
- Generality questionable since we only perform the analysis on data from a single hospital

- Time information is lost when using aggregates, changes in vitalsigns mostly captured compositional dimensions, not directly measured
- Limitierung: wir haben für die analysen kein Testdatenset abgespalten! Damit keine perfekte unabhängigkeit in den korrelationsmatrizen zu unseren (zukünftigen) Testdaten Stimmt das? Wir haben doch jetzt nen Testdatensatz mit immer gleicher random seed
- no time data in query, only max/min/avg
- Wo nichts gemessen, können wir nichts entscheiden

Additionally to the low scores, there are other inherent limitations in our methodology that require consideration. Firstly, the use of Sepsis-3 as a definition for sepsis onset establishes a circular dependency that may obscure the true onset of the condition. Doctors have to measure score relevant for sofa, depending on the time at which they choose to do so, the actual sofa score may have increased before the one on paper/in the database does. There was a paper discussion issues with sepsis-3, if only i could find it again... This reliance on medical personnel's swift actions for detection means that any delay in measuring scores essential for the SOFA computation could subsequently postpone the Sepsis-3 defined sepsis onset. Secondly, the SOFA score's hourly update frequency may not be sufficiently granular to effectively trace the real-time progression of the infection or illness. A higher resolution could reveal features more closely related to the evolving condition rather than rigidly adhering to predefined criteria. Additionally, our machine learning approach has only been validated on the MIMIC-IV dataset, thus it is tailored to a single hospital's data. The models we have developed may not be applicable or accurate elsewhere due to variations in data collection practices across different institutions. Furthermore, our feature aggregation approach, which compiles features across a given timeframe into a single feature vector, loses crucial time-specific data. Changes in vital signs or lab results over time may be essential in predicting sepsis, but our current methodology only accounts for the standard deviation or mean of a feature across an entire timeframe. Incorporating time-series information could lead to improved and more generalizable outcomes.

E. 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., could provide a more refined understanding of prediction dynamics over different time frames.
- Utilize our preliminary work (queries, code, exploratory data analysis) as a foundation for further inquiries into sepsis-related questions.

V. Conclusions

TODO: comparision with related work.

Acknowledgment

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 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] 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).

[4] 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).

[5] 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).

[6] 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.

[7] 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.

[8] 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).

[9] 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).

[10] 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

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 [3]. 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	—	—

A. Abbreviations

GCS	Glasgow Coma Scale
ICU	Intensive Care Unit
EDA	Explorative Data Analysis
ROC	Receiver Operating Characteristic Score
MCC	Matthews Correlation Coefficient [2]
ML	Machine Learning
MIMIC	Medical Information Mart for Intensive Care