Enhanced Early Sepsis Onset Prediction: A Multi-Layer Approach

Kevin Ewig

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2022

Reading Committee:

Dr. Juhua Hu

Dr. Ankur Teredesai

Katherine Stern, MD

Program Authorized to Offer Degree:
Computer Science and Systems (Data Science)

University of Washington

Abstract

Enhanced Early Sepsis Onset Prediction: A Multi-Layer Approach

Kevin Ewig

Chair of the Supervisory Committee:
Dr. Juhua Hu
School of Engineering and Technology

Sepsis is a life-threatening organ malfunction caused by the host's inability to fight infection. Without proper and immediate treatment, sepsis can lead to death. Early diagnosis and medical treatment of sepsis in critically ill populations at high risk for sepsis and sepsis-associated mortality are vital to providing the patient with rapid therapy. The mortality rate increases with each hour that antibiotic treatment is delayed. Studies show that advancing sepsis detection by 6 hours leads to earlier administration of antibiotics, which is associated with improved mortality. However, clinical scores like Sequential Organ Failure Assessment (SOFA) are not applicable for early sepsis onset prediction, while machine learning algorithms may be able to capture the progressing pattern for early prediction. Therefore, this thesis aims to develop a machine learning model that predicts sepsis onset 6 hours before it is suspected clinically. Although some machine learning algorithms have been applied to sepsis prediction, many of them did not consider the fact that six hours is not a small gap. To overcome this big gap challenge for early sepsis detection, this thesis explores a multi-layer approach in which the likelihood of sepsis occurring earlier than 6 hours is output from the 1st layer and fed to the 2nd layer as features to help predictions for the 6-hour horizon. Moreover, we use the hourly sampled data like vital signs in an observation window to derive a temporal change trend to further assist in sepsis prediction, which however is often ignored by previous traditional machine learning algorithms. Our empirical study shows that both the multi-layer approach to alleviating the 6-hour gap and the added features to capture the temporal trend can help improve the performance of early sepsis prediction.

ACKNOWLEDGMENTS

I would like to thank Professor JuHua Hu, my advisor, for her help and guidance with this thesis. Many thanks also to Dr. Katherine Stern for her input on the data and support with the clinical explanations. I would also like to thank Professor Ankur Teredesai for his overall guidance and Tucker Stewart for assistance with the sepsis data.

This thesis is partially supported by cloud computing credits awarded by UW eScience Institute, UW Research Computing, and Microsoft. It is also partially supported by NSF (IIS-2104270). All opinions, findings, conclusions and recommendations in this paper are those of the author and do not necessarily reflect the views of the funding agencies.

1 Introduction

Sepsis is a major public health concern. It is a life-threatening disease caused by a host's failed response to an infection [25]. The immune system of a sepsis patient becomes aggressive in its protection against infection in the body, which causes organ dysfunction and potential organ damage. Sepsis is still a common problem in modern medical settings, particularly Intensive Care Units (ICUs). According to a global survey conducted in 2018 [22], roughly 13.6% to 39.3% of patients admitted to ICU are impacted by sepsis. This share is 29.5% worldwide. Patients with sepsis also experience longer and more expensive hospital stays. For patients that survive, sepsis can cause an increased risk of permanent organ damage, and physical disability [12].

Early treatment before the formation of sepsis in patients has been shown to improve the chances of successfully treating and preventing the disease [14]. Early care can prevent 80% of sepsis-related deaths, and the chances of survival drop by 8% every hour if action is not taken [15]. In particular, studies have shown that treating sepsis 6 hours earlier before the onset significantly improves the patient's chance of recovering [9, 10]. Therefore, this thesis focuses on the problem of early sepsis prediction, that is, 6 hours in advance. Several clinical scores such as SOFA have been developed to indicate the onset of sepsis with clinical-based data [25, 3]. These measures are helpful when sepsis is onset already but have been limited for early prediction. For early sepsis onset prediction, predictions derived using supervised machine learning models such as Random Forest or Long Short-Term Memory (LSTM) models have vastly outperformed clinical scores [1].

Because of this, there has been ongoing research in developing a machine learning predictive model to detect the onset of sepsis early before it is suspected clinically. Some studies have used traditional machine learning models such as Logistic Regression and Random Forest [19, 18]. However, these studies did not capture the progressing temporal pattern that can be useful for early sepsis prediction. Recently, deep learning models have also been applied in sepsis prediction, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) [16]. Compared with traditional machine learning methods, deep learning often provides better prediction performance. However, training deep models can be tedious to obtain the optimal layers and parameters. In addition, none of them has considered the fact that 6-hour is not a small gap. Without the progressing information for the next 6 hours before onset, it is challenging to make an accurate prediction, which is explored in this thesis.

First, it should be noted that training successful deep neural networks is expensive in terms of time, computational resources, and data. We aim to avoid using deep neural networks but still capture the progressing temporal change pattern for early sepsis prediction. We accomplish this by computing the hourly changes in feature values like heart rates within a given observation window. Doing this will allow us to consider the temporal changes in the feature even in a traditional machine learning model, where features are often considered independently. Second, we generate a temporal

change trend from hourly collected data such as vital signs in an observation window to aid in sepsis prediction. Earlier traditional machine learning methods generally overlook adding a temporal change to the model. Using a multi-layer technique to close the 6-hour gap and adding features to capture the temporal trend can enhance early sepsis prediction accuracy. The contributions of this thesis can be summarized as follows.

- We propose generating a series of delta values to capture the temporal change trend window for a set of features. The temporal change is incorporated into the feature set for training and prediction.

- We implement a multi-layer strategy in which the likelihood of sepsis arising earlier than 6 hours is generated from the first layer and provided to the second layer as an additional feature to aid 6-hour horizon predictions.

- Based on the proposed method, we apply an economical machine learning algorithm, that is XGBoost [5], to the trauma patients from year 2012 to 2015 at ICUs of UW Harborview Center, and obtain an AUROC of 98.2%, where the specificity for non-sepsis patients is 98.9% and the sensitivity for sepsis patients is 88.7%. More importantly, XGBoost is easy to interpret, which is essential for healthcare.

2 Related Work

Various machine learning models have been studied for sepsis prediction as follows.

2.1 Traditional Models

Some of the research done in sepsis prediction uses simple traditional machine learning models. For example, Zabihiet et al. [31] used a wrapper feature selection algorithm based on XGBoost to extract five different sets of features from clinical data to predict sepsis 6 hours before onset. Both valid and missing clinical data are used to derive the relevant attributes. Afterward, an ensemble model comprised of five XGBoost models is utilized to predict sepsis. The result was an average AUROC of 83.3%.

In addition, Firoozabadi et al. [6] created models to predict sepsis as part of the PhysioNet/CinC Challenge 2019. In this study, the authors took hourly samples of 40 features from each patient's data from three different ICUs. They processed the data to remove outliers, fill in missing data, and replace the remainder with the population average. They discovered similarly that an ensemble of bagged decision trees is effective for early sepsis prediction 6 hours before onset. A similar

observation is shown by Fu et al. [7]. The AUROCs obtained from these two studies were less than 80%.

However, these models did not consider that the temporal change trend can be helpful for early sepsis onset prediction and missing 6-hours of progressing information is challenging to make an accurate prediction. In this thesis, we will address both by adding temporal change trend features and proposing a multi-layer approach.

2.2 Deep Learning Models

Deep learning models have also been applied to address specific challenges of sepsis prediction. For example, Gilbertson et al. [11] used the Principal Component Analysis (PCA) to reduce the dimensionality of patient data before applying a simple RNN. The PCA method reduced the original 40 patient features to 10 main components. In addition to this, a fast Sequential Organ Failure Assessment (qSOFA) score generates one additional characteristic. These 11 characteristics are then loaded into a deep neural network (DNN) classifier. However, the characteristics connected with each hour are examined independently.

Later, Tsang et al. [26] applied LSTM which allows consideration of past time-step data embeddings within the prediction of the current time-step to capture the temporal pattern. Most recently, Shah et al. [24] proposed a DNN model to predict sepsis 6 hours in advance. The DNN model is used because of its feature learning capacity and its property of approximating functions. Better sepsis onset prediction performance were observed using these deep learning models with an AUROC from 88.8% to 89.3%. However, none of them has considered to address the 6-hour gap challenge, that is explored by this thesis. Moreover, training DNNs are often time, resource, and data expensive, which is avoided by this thesis.

3 Proposed Method

We assume that the Electronic Heath Records (EHRs) for a single patient can be represented as a $t \times f$ matrix X. There are t number of rows, where each row represents an hourly observation of a patient's features and where the t-th observation is the most recent. There is also f number of columns, where each column represents a physiological feature. Let $x_{t,f}$ denote the numerical entry in X for time t and feature f.

Our first challenge is reducing the 6-hour gap between the prediction and onset time. We propose a multi-layer approach. In the 1st layer, we use the observed data up to the prediction time to determine the probabilities that a patient will develop sepsis between the current time t and the sepsis onset time $t + 6$. For example, we can use the observed data to determine the probability of sepsis onset at $t + 3$. Then, this probability will be treated as an additional input

feature in the 2nd layer of our approach. After that, the big 6-hour gap can be shrunk using an in-between measure with a smaller gap of 3. We can further shrink the gap by providing sepsis onset probabilities of each hour in-between to feed more features in the 2nd layer. Finally, the 2nd layer output will provide the sepsis onset prediction in 6 hours.

We denote the ground truth[§] sepsis onset label as $\hat{y}$ that has an indication of onset hour t. More then often, the hourly physiological patient data contain missing values, which are imputed and filled in. In order to train the model to predict sepsis onset $h = 6$ hours into the future, for each patient, we first shift the values in $\hat{y}$ earlier by h hours and call this $\hat{y_h}$. This step creates a new target feature of sepsis that will occur in h hours. The leftover values on the table with no sepsis ground truth values are dropped from the data. This step is illustrated in Table 1 as an example. From this example, we can observe that if we want to predict sepsis onset in 3 hours, using observed patient data up to $t = 1$, sepsis onset prediction for $t = 4$ should be 0, and using data up to $t = 2$, sepsis onset prediction for $t = 5$ should be 1.

t	$\hat{y}$
1	0
2	0
3	0
4	0
5	1
6	1

$\rightarrow$

t	$\hat{y_3}$
1	0
2	1
3	1
4	-
5	-
6	-

$\rightarrow$

t	$\hat{y_3}$
1	0
2	1
3	1

Table 1: In this example, the table in the left contains the ground truth for sepsis at time t. In order to produce new ground truth values for sepsis at time $t + h$, the $\hat{y}$ values are shifted by $h = 3$ hours. The leftover rows are dropped.

After shifting the label for h hours to predict sepsis onset in h hours, for a certain hour t, we can use only the $x_{t,f}$ values across all of the features, and the new ground truth label $\hat{y_h}$ to train the sepsis onset classification model that will be able to make a prediction on new patients and provide

[§]The sepsis onset label has been given using the CDC's adult sepsis surveillance criteria with a priori modifications utilizing readily obtainable EMR data to improve specificity for the trauma population. It is required that all of the following be present: 1) an order for a new IV or qualifying oral antibiotic, not administered within the previous 48 hours and excluding antibiotics used for surgical prophylaxis, 2) a body tissue culture was ordered within 48 hours of antibiotic initiation, 3) a qualifying antibiotic was sustained for at least 4 consecutive days, or until death or discharge, and 4) a 2-point increase in the maximum daily sequential organ failure assessment (SOFA) score occurred within 3 days before and 3 days after the qualifying culture. The criteria is restricted to hospital-acquired infections, which is defined as cultures obtained on or after the third hospital day. Two subgroups were independently adjudicated before final sepsis assignments were made: culture-negative sepsis and patients meeting partial but not full criteria.

the probability of sepsis in h hours as y_h. This is illustrated as an example in Table 2. Specifically, given the hard label of sepsis onset or not, the trained model can provide the probability of sepsis onset of each time stamp, which contributes the 1st layer of our proposed method.

t	x_1	x_2	...	x_f
1	98	102	...	30
2	95	107	...	32
3	98	111	...	35
4	100	109	...	40
5	97	109	...	41

$+$

$\hat{y_3}$
0
0
0
0
1

$\rightarrow$

y_3
0.26
0.44
0.62
0.61
0.71

Table 2: 1st Layer: Probability of Sepsis in $h = 3$ hours. Features x_1, x_2, ..., x_f are used to train a model with the ground truth label $\hat{y_h}$, and the model is then used to predict the probability of sepsis onset in h hours as y_h.

After this, we include the probability of sepsis in h hours or y_h (i.e., $h = 3$), as a new feature of the dataset X to predict sepsis in 6 hours in the 2nd layer. This is shown in Table 3.

t	x_1	x_2	...	x_n	y_3
1	98	102	...	30	0.26
2	95	107	...	32	0.44

$+$

$\hat{y_6}$
0
1

$\rightarrow$

y_6
0.66
0.74

Table 3: 2nd Layer: A toy example of calculating the probability of sepsis onset in 6 hours with an additional feature of y_h, that is, the probability of sepsis onset at $h = 3$ output from the 1st layer.

Now, y_6 output from our multi-layer approach can be used to determine the presence or absence of sepsis in 6 hours. If y_6 shows that the probability of sepsis is more than 50%, we assume that the patient will develop sepsis in 6 hours. Otherwise, we determine that the patient will not develop sepsis in 6 hours. It is also important to note that multiple features y_h's output from the 1st layer's models can be added to enhance the prediction of y_6. For example, we could add the values of y_1, y_2, ..., y_5 to predict y_6.

3.1 Delta Values for Temporal Change Trend

In order to capture the temporal change trend that can be helpful for early sepsis onset prediction, we compute the delta values within the observation window, where the window size is denoted as w. To do this, we derive a new feature f' for each feature and set the value as the difference of

values for a feature between each nearby timestamp. This means that given feature f, we derive a new feature f' and then set the value as $x_{t,f'} = x_{t,f} - x_{t-1,f}$ within observation window w. The algorithm to derive this new feature can be found in Algorithm 1. The new features are added to help predict sepsis.

Algorithm 1: An algorithm to compute the delta values for f', new features added in X derived from a feature f.

Input: f: A feature in X used to compute the delta values.
Input: w: The observation window size.

```
// Initialize the new feature, f'
for r ← 1 to w do
    x_{r,f'} ← 0
end
// Populate the new feature f'
for r ← 1 to t do
    if r > w then
        for n ← 1 to w do
            x_{r,f'} ← x_{r-n,f} − x_{r-n-1,f}
        end
    else
end
```

4 Experiments

To evaluate the proposed method, we use trauma patients' data provided by UW Harborview Medical Center. It contains prospectively collected, de-identified data from injured adults ages 16 years and older who were admitted to Harborview Medical Center intensive care unit (ICU) between 2012-2015 and required at least three days of invasive mechanical ventilation. The dataset includes physiological data on 1,263 patients, 186 (14%) were identified to have sepsis during the first 14 days of admission. The patient data is sampled at different time intervals. Vital signs, for example, were sampled hourly, whereas laboratory tests were sampled daily or less frequently. As such, there are some features with many missing values. We use imputation strategies, which will be discussed later, to address this problem. The list of features used in this thesis is summarized in the Appendix.

4.1 Imbalanced-Class Problem in Data

One of the primary problems with the sepsis patient data we received is predicting an outcome that occurs in less than 50% of the study population using an hour-by-hour or day-to-day framework. Approximately 16.4% of the records in the dataset have a sepsis indicator, while the rest does not have it. Most of the patients admitted to the ICU never develop sepsis. Some patients develop sepsis within an hour or two, and some develop sepsis after a more extended period.

To account for the imbalanced nature of the data, we used the following methods. First, we only include hourly patient records from day two until day 14 of the patient's stay in the ICU. This is because the scope of this study is patients developing sepsis after being in the ICU. Limiting the data will reduce the number of hourly records with no sepsis that does not need to be included in the training model. Second, if a patient has an hourly record with sepsis, then we remove all the subsequent hourly records for that patient. For example, if we see a patient record with sepsis at 4:00, all following observations are removed. It is because we are only interested in the early sepsis onset prediction, that is, the first onset. We use 5-fold stratified cross-validation method for training. In addition, to further account for the imbalance of data, we use a random oversampling of 0.8. Finally, we use random under-sampling to make the number of minority class data the same as the number of majority class data in the sample. To illustrate this further, suppose there are 20 minority samples and 1000 majority samples. Random oversampling duplicates the 20 minority samples without replacing them until there are 800 samples. As a result, there are 800 minority samples and 1000 majority samples.

4.2 Missing Data Imputation

To handle the missing data problem, we apply and compare three state-of-the-art imputation strategies. The first is Carry Forward Imputation. This imputation implementation is the most straightforward approach where the previous data values are passed on to the next if the following values are missing. The second is Multiple Imputation by Chained Equations (commonly known as MICE) [28]. With the MICE approach, the missing values are filled iteratively using the mean and a linear regression model several times. The third is a machine learning model called Recurrent Imputation for Time Series (RITS) [4]. In this approach, a masking vector is constructed for the missing values. The time gaps between each record are also built. This is passed into a machine learning model for training. The resulting RITS model can then be used to impute the missing values.

4.3 Statistical Values

In our proposal, delta values are used to capture the temporal change trend within the observation window w. As a baseline, we can also add statistical information computed from a feature f within

the time window. Specifically, given a feature f within an observation window w, we can add six statistical features based on $S = [x_{t-w,f}, x_{t-w+1,f}, x_{t-w+2,f}, ..., x_{t,f}]$, that is $mean(S)$, $min(S)$, $max(S)$, the standard deviation of S, the skewness of S, and the kurtosis of S.

4.4 Evaluation Metrics

To evaluate the quality of this approach, we use the Area Under the Receiver Operator Curve, confusion matrix counts, Sensitivity and Specificity. The evaluation of the model for predicting the onset of sepsis in 6 hours will be calculated based on results from the confusion matrix shown.

	Sepsis Diagnosis		
	Absolutely Positive	Absolutely Negative	Total
Predicted Positive	TP	FP	$TP + FP$
Predicted Negative	FN	TN	$FN + TN$
Total	$TP + FN$	$FP + TN$	

Each entry in the table is as follows, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Note that the entries inside of a confusion matrix (TPs, FPs, FNs, TNs) are defined as follows:

- True Positives (TPs): the number of positive examples accurately classified by the model.
- True Negatives (TNs): the number of negative examples accurately recognized by the model.
- False Positives (FPs): the number of negative cases wrongly categorized as positive by the model (i.e., negative examples incorrectly classed as "positive").
- False Negatives (FNs): the number of positive cases wrongly categorized as negative by the model (i.e., positive examples incorrectly labeled as "negative").

In addition to these confusion matrix counts, we will use the Sensitivity, Specificity and Area Under the Curve (AUC) to determine the performance of each model. $Sensitivity = \frac{TP}{TP+FN}$ is defined as a measure of how well the model can recognize positive examples. When evaluating model performance, sensitivity is frequently compared to specificity. $Specificity = \frac{TN}{FP+TN}$ is the fraction of real negatives properly detected by the model. Finally, AUC is calculated as the Area Under the $Sensitivity - (1 - Specificity)$ Curve.

4.5 Performance Comparison

To evaluate the proposed method, we first compare between with and without the 1st layer to shrink the gap. The one-layer setup did not include y_h values as features in the prediction process.

The two-layer setup did include the value of y_3 ($h = 3$) as a feature in the 2nd layer. We then used two different observation window sizes, $w = 6$ and $w = 12$. Finally, we compare different models: Logistic Regression (LR), Random Forest (RF), and XGBoost (XGB), to determine which one produces the best result. The complete result for all experiments is located in the Appendix.

Imputation Comparison

We first compare different imputation strategies for the missing data. We used one of the settings, a single layer with a 6-hour observation window running on XGBoost, to compare the outcomes of the three imputation algorithms. The results of the experiments are shown in Table 4.

Name	AUROC	Sensitivity	Specificity
Carry Forward	**0.958**	0.833	0.949
MICE	0.956	**0.845**	0.943
RITS	0.940	0.775	**0.973**

Table 4: Comparison of various XGBoost results

It can be observed that the results from Carry-Forward and MICE approaches were similar. Although the Carry Forward approach has the best AUROC, MICE and RITS had better results for Sensitivity and Specificity, respectively. Since the Carry Forward approach is easy to implement and there is not that big a difference between this and the other imputations, we used the Carry Forward approach in the following experiments.

Results and Discussions

Table 5 shows the top AUROC, Sensitivity, and Specificity results of each model. It can be observed that XGBoost performs better than other models in terms of AUROC and Sensitivity. Specifically, the best XGBoost model (2 Layers Size 6 XGB Delta) has obtained the best AUROC of 0.982. The Random Forest model provides a high Specificity value like 0.9999; however, it has very low Sensitivity, which indicates that the model is not effective in sepsis onset detection. Finally, LR has a balanced performance on Sensitivity and Specificity. However, they are both worse than XGBoost. Therefore, we will use only XGBoost for comparison in the following experiments.

It is also interesting to note that all of the highest values for AUROC, Sensitivity, and Specificity in XGBoost are in the models with 2 Layers and only use the Delta Values. Further examination of Table 6 shows that using the statistical values provides a higher AUROC for the one-layer setup. However, for the multi-layer design, using the Delta Values performs better than setups that use

Name	AUROC	Sensitivity	Specificity
1 Layer Size 6 LR Delta	0.828	0.736	0.789
1 Layer Size 6 LR Delta Stats	0.828	0.774	0.755
1 Layer Size 6 LR Stats	0.840	0.790	0.755
1 Layer Size 12 RF Delta	0.864	0.021	**0.9999**
1 Layer Size 12 RF Delta Stats	0.845	0.053	0.9998
2 Layers Size 6 RF Delta	0.891	0.016	0.9999
2 Layers Size 6 XGB Delta	**0.982**	0.887	0.989
2 Layers Size 12 XGB Delta	0.976	**0.892**	0.989
2 Layers Size 12 XGB Delta Stats	0.973	0.811	0.991

Table 5: Performance Comparison

Delta and Statistics or just Statistics. More importantly, 2-layer approaches provide significantly better performance compared to the 1-layer approaches, which demonstrates the effectiveness of our proposed method.

Name	AUC	Sensitivity	Specificity
1 Layer Size 6 XGB Delta	0.959	**0.838**	0.971
1 Layer Size 6 XGB Delta Stats	0.956	0.795	0.981
1 Layer Size 6 XGB Stats	**0.965**	0.784	**0.987**
1 Layer Size 12 XGB Delta	0.961	**0.844**	0.970
1 Layer Size 12 XGB Delta Stats	0.951	0.779	0.981
1 Layer Size 12 XGB Stats	**0.966**	0.801	**0.987**
2 Layers Size 6 XGB Delta	**0.982**	**0.887**	0.989
2 Layers Size 6 XGB Delta Stats	0.971	0.811	**0.990**
2 Layers Size 6 XGB Stats	0.974	0.822	0.986
2 Layers Size 12 XGB Delta	**0.976**	**0.892**	0.989
2 Layers Size 12 XGB Delta Stats	0.973	0.811	**0.991**
2 Layers Size 12 XGB Stats	0.973	0.811	0.991

Table 6: Comparison on XGBoost

Finally, we increase the number of features from the 1st layer added to the second layer. Specifically, we add the probabilities of sepsis onset in 1 hr, 2 hrs, to 5 hrs as additional features in the 2nd

layer. We then compare it with a two-layer setup with an observation window size of 12 hours using 3 hr probability only. The results are shown in Table 7. Based on this, we can observe that having multiple probabilities can further help shrink the gap and improve the prediction performance in terms of all metrics.

Name	AUC	Sensitivity	Specificity
Multiple Probabilities XGB Delta	**0.988**	**0.897**	**0.991**
One Probability XGB Delta	0.976	0.892	0.989

Table 7: Comparison between multiple probabilities and one probability.

4.6 Model Explanation

To verify if these probabilities added from the 1st layer are helpful for making predictions, we apply LIME [21], a tool that attempts to explain the weights in a predictive model, and check the importance of features for sepsis onset prediction in the 2nd layer. Specifically, we took a random sample of the data and used LIME to get the importance of each feature for both models shown in Table 7. Top 5 important features for each model are listed in Table 8.

3 hour Probability Included

Feature	Value
sepsis-layer1-3hr	**0.026**
creatinine-6	−0.021
surgSum-8	−0.020
neutrophils-1	0.020
neutrophilis-9	0.019

1 hr to 5 hr Probabilities Included

Feature	Value
creatinine-3	−0.023
neutrophils-2	−0.021
neutrophils-5	0.019
sepsis-layer1-5hr	**-0.019**
creatinine-6	0.014

Table 8: Top values from LIME that influence the probability of sepsis in 6 hours

The left table shows the top values that affect the prediction for a two-layer model with only the 3 hr probability. The results show that the probability of sepsis occurring in 3 hours contributed significantly to the prediction of sepsis onset in 6 hours. The table on the right shows the top values that affect the prediction for a two-layer model with 1 hr to 5 hr probabilities. The results show that the likelihood of sepsis occurring within 5 hours contributed significantly to the prediction. This further demonstrates the importance of shrinking the gap as much as possible.

5 Conclusion

This thesis presents a new machine-learning-based prediction setup for early sepsis onset prediction for critically ill trauma patients. We constructed a series of delta values to record the temporal change trend within an observation window for a group of characteristics. We also used a multi-layer technique in which the first layer generates the risk of sepsis developing before 6 hours and provides it to the second layer as an additional feature to enhance the model's performance and close the 6-hour gap. Based on the experiments, adding the delta values within a specified observation window and adding more sepsis probability features improves sepsis prediction. In addition, in this study, we show that using multi-layers in conjunction with XGBoost produces the best results based on AUC and Sensitivity.

There are two future directions for this research. First, we would like to validate our developed models by using a new dataset to determine if this method of sepsis prediction will work as a whole. Finally, we would like to enhance the explainability of the model so that it can provide more valuable data to the clinical staff. This information, in turn, can identify discriminatory features and will be helpful to clinicians as this information can be used to monitor and screen for sepsis.

6 Appendix

6.1 Features Used in Generating Delta and Statistic Values

Vital Signs:
Heart Rate
Systolic Blood Pressure
Diastolic Blood Pressure
Mean Arterial Pressure (MAP)
Respiratory Rate
Temperature
Fraction of Inspired Oxygen (FiO2)
Cumulative exposures or Therapeutic Interventions:
bolusSum: IV fluid boluses (volume > 500mL in 1 hour)
Number of hours in Surgery
Number of days requiring invasive mechanical ventilation
Ventilated

Laboratory values and/or Other physiologic parameters:
Bicarbonate (bicarb)
StrongIon: Strong ion difference
Blood Urea Nitrogen (BUN)
Creatinine
Lymphocyte
Urine Output (UOP)
Acidosis, Hypoperfusion

6.2 Complete Results

The following tables (Tables 9, 10 and 11) are the complete results for each model (i.e., Logistic Regression, Random Forest, and XGBoost), with an observation window size of 6 and a size of 12.

Name	AUC	Sensitivity	Specificity
1 Layer Size 6 LR Delta	0.828	0.736	**0.789**
1 Layer Size 6 LR Delta Stats	0.828	0.774	0.755
1 Layer Size 6 LR Stats	**0.840**	**0.790**	0.755
1 Layer Size 12 LR Delta	0.827	0.763	0.783
1 Layer Size 12 LR Delta Stats	0.830	0.784	0.754
1 Layer Size 12 LR Stats	0.839	0.790	0.756
2 Layers Size 6 LR Delta	0.824	0.758	0.783
2 Layers Size 6 LR Delta Stats	0.824	0.763	0.756
2 Layers Size 6 LR Stats	0.838	0.784	0.756
2 Layers Size 12 LR Delta	0.825	0.747	0.781
2 Layers Size 12 LR Delta Stats	0.829	0.790	0.754
2 Layers Size 12 LR Stats	0.838	0.784	0.755

Table 9: Results from the Logistic Regression model

Name	AUC	Sensitivity	Specificity
1 Layer Size 6 RF Delta	0.869	0.021	0.9999
1 Layer Size 6 RF Delta Stats	0.864	0.053	0.9998
1 Layer Size 6 RF Stats	0.849	0.032	0.9998
1 Layer Size 12 RF Delta	0.864	0.021	**0.9999**
1 Layer Size 12 RF Delta Stats	0.845	**0.053**	0.9998
1 Layer Size 12 RF Stats	0.855	0.043	0.9998
2 Layers Size 6 RF Delta	**0.891**	0.016	0.9999
2 Layers Size 6 RF Delta Stats	0.878	0.048	0.9998
2 Layers Size 6 RF Stats	0.860	0.021	0.9998
2 Layers Size 12 RF Delta	0.889	0.021	0.9999
2 Layers Size 12 RF Delta Stats	0.859	0.053	0.9998
2 Layers Size 12 RF Stats	0.863	0.037	0.9998

Table 10: Results from the Random Forest model

Name	AUC	Sensitivity	Specificity
1 Layer Size 6 XGB Delta	0.959	0.838	0.971
1 Layer Size 6 XGB Delta Stats	0.956	0.795	0.981
1 Layer Size 6 XGB Stats	0.965	0.784	0.987
1 Layer Size 12 XGB Delta	0.961	0.844	0.970
1 Layer Size 12 XGB Delta Stats	0.951	0.779	0.981
1 Layer Size 12 XGB Stats	0.966	0.801	0.987
2 Layers Size 6 XGB Delta	**0.982**	0.887	0.989
2 Layers Size 6 XGB Delta Stats	0.971	0.811	0.990
2 Layers Size 6 XGB Stats	0.974	0.822	0.986
2 Layers Size 12 XGB Delta	0.976	**0.892**	0.989
2 Layers Size 12 XGB Delta Stats	0.973	0.811	**0.991**
2 Layers Size 12 XGB Stats	0.973	0.844	0.987

Table 11: Results from the XGBoost model

BIBLIOGRAPHY

[1] Time-specific metalearners for the early prediction of sepsis. December 2019.

[2] David Andaluz and Ricard Ferrer. SIRS, qSOFA, and organ failure for assessing sepsis at the emergency department. *J Thorac Dis*, 9(6):1459–1462, June 2017.

[3] Robert A Balk. Systemic inflammatory response syndrome (SIRS): where did it come from and is it still relevant today? *Virulence*, 5(1):20–26, November 2013.

[4] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018.

[5] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[6] Reza Firoozabadi and Saeed Babaeizadeh. An ensemble of bagged decision trees for early prediction of sepsis. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, 2019.

[7] Mengsha Fu, Jiabin Yuan, and Chen Bei. Early sepsis prediction in icu trauma patients with using an improved cascade deep forest model. In *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, pages 634–637, 2019.

[8] Satoshi Gando, Atsushi Shiraishi, Toshikazu Abe, Shigeki Kushimoto, Toshihiko Mayumi, Seitaro Fujishima, Akiyoshi Hagiwara, Yasukazu Shiino, Shin-Ichiro Shiraishi, Toru Hifumi, Yasuhiro Otomo, Kohji Okamoto, Junichi Sasaki, Kiyotsugu Takuma, Kazuma Yamakawa, Atsumi Hoshino, Atsushi Shiraishi, Toshiaki Abe, Manabu Sugita, Yoshihiro Hanaki, Akiyoshi Hagiwara, Shin-Ichiro Shiraishi, Yasukazu Shiino, Masahiro Harada, Hideaki Yoshihara, Kiyotsugu Takuma, Yasuhiro Otomo, Kazuma Morino, Yoshihiro Shimizu, Hiroyasu Ishikura, Toru Hifumi, Yoshizumi Deguchi, Sho Nachi, Satoshi Gando, Kohji Okamoto, Masato Kawakami, Seitaro Fujishima, Junichi Sasaki, Junichi Maehara, Kunihiko Okada, Kazuma Yamakawa, Kazuya Kiyota, Yasuo Miki, Kaoru Koike, Takashi Muroya, Hisashi Yamashita, Toshihiko Mayumi, Hideaki Anan, Tadashi Kaneko, Hirotada Kittaka, Hiroyuki Yamaguchi, and The Japanese Association for Acute Medicine (JAAM) Sepsis Prognostication in Intensive Care Unit and Emergency Room (SPICE) (JAAM SPICE) Study Group. The SIRS criteria have better performance for predicting infection than qSOFA scores in the emergency department. *Scientific Reports*, 10(1):8095, May 2020.

[9] Robert L Gauer. Early recognition and management of sepsis in adults: the first six hours. *Am Fam Physician*, 88(1):44–53, July 2013.

[10] Robert L Gauer. Early recognition and management of sepsis in adults: the first six hours. *Am Fam Physician*, 88(1):44–53, July 2013.

[11] Erik H Gilbertson, Khristian M Jones, Abigail M Stroh, and Bradley M Whitaker. Early detection of sepsis using feature selection, feature extraction, and neural network classification. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, 2019.

[12] C Jones and RD Griffiths. Mental and physical disability after sepsis. *Minerva anestesiologica*, 79(11):1306—1312, November 2013.

[13] Anahita Khojandi, Varisara Tansakul, Xueping Li, Rebecca S Koszalinski, and William Paiva. Prediction of sepsis and In-Hospital mortality using electronic health records. *Methods Inf Med*, 57(4):185–193, September 2018.

[14] Hwan Il Kim and Sunghoon Park. Sepsis: Early recognition and optimized treatment. *Tuberc Respir Dis (Seoul)*, 82(1):6–14, September 2018.

[15] Anand Kumar, Daniel Roberts, Kenneth E Wood, Bruce Light, Joseph E Parrillo, Satendra Sharma, Robert Suppes, Daniel Feinstein, Sergio Zanotti, Leo 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):1589–1596, 2006.

[16] Xin Li, G André Ng, and Fernando S Schlindwein. Convolutional and recurrent neural networks for early detection of sepsis using hourly physiological data from patients in intensive care unit. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, 2019.

[17] Karthigha M and V.S. Akshaya. A xgboost based algorithm for early prediction of human sepsis. In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 1643–1647, 2022.

[18] Fahim Mahmud, Naqib Sad Pathan, and Muhammad Quamruzzaman. Early detection of sepsis in icu patients using logistic regression. In *2019 3rd International Conference on Electrical, Computer Telecommunication Engineering (ICECTE)*, pages 173–176, 2019.

[19] Fahim Mahmud, Naqib Sad Pathan, and Muhammad Quamruzzaman. Early detection of sepsis in critical patients using random forest classifier. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 130–133, 2020.

[20] Chanu Rhee, Raymund Dantes, Lauren Epstein, David J Murphy, Christopher W Seymour, Theodore J Iwashyna, Sameer S Kadri, Derek C Angus, Robert L Danner, Anthony E Fiore, John A Jernigan, Greg S Martin, Edward Septimus, David K Warren, Anita Karcz, Christina Chan, John T Menchaca, Rui Wang, Susan Gruber, Michael Klompas, and CDC Prevention Epicenter Program. Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014. *JAMA*, 318(13):1241–1249, October 2017.

[21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[22] Yasser Sakr, Ulrich Jaschinski, Xavier Wittebole, Tamas Szakmany, Jeffrey Lipman, Silvio A Ñamendys-Silva, Ignacio Martin-Loeches, Marc Leone, Mary-Nicoleta Lupu, Jean-Louis Vincent, et al. Sepsis in intensive care unit patients: worldwide data from the intensive care over nations audit. In *Open forum infectious diseases*, volume 5, page ofy313. Oxford University Press US, 2018.

[23] Soodabeh Sarafrazi, Rohini Choudhari, Himanshi Mehta, Chiral Mehta, Omid Japalaghi, Kinjal Mehta, Hyunyoung Han, and Patricia Francis-Lyon. Cracking the "sepsis" code: Assessing time series nature of ehr data, and using deep learning for early sepsis prediction. 12 2019.

[24] Nemil Shah, Jay Bhatia, Nimit Vasavat, Rishi Desai, and Pankaj Sonawane. Early sepsis detection using machine learning and neural networks. In *2021 2nd Global Conference for Advancement in Technology (GCAT)*, pages 1–6. IEEE, 2021.

[25] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, Richard S Hotchkiss, Mitchell M Levy, John C Marshall, Greg S Martin, Steven M Opal, Gordon D Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C Angus. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, February 2016.

[26] Gavin Tsang and Xianghua Xie. Deep learning based sepsis intervention: The modelling and prediction of severe sepsis onset. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8671–8678, 2021.

[27] My Chau Tu, Dongil Shin, and Dongkyoo Shin. A comparative study of medical data classification methods based on decision tree and bagging algorithms. In *2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*, pages 183–187, 2009.

[28] Stef Van Buuren and Catharina GM Oudshoorn. Multivariate imputation by chained equations, 2000.

[29] Franco van Wyk, Anahita Khojandi, and Rishikesan Kamaleswaran. Improving prediction performance using hierarchical analysis of Real-Time data: A sepsis case study. *IEEE J Biomed Health Inform*, 23(3):978–986, January 2019.

[30] Andrew Wong, Erkin Otles, John P Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med*, 181(8):1065–1070, August 2021.

[31] Morteza Zabihi, Serkan Kiranyaz, and Moncef Gabbouj. Sepsis prediction in intensive care unit using ensemble of xgboost models. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, 2019.