Utilizing Informative Missingness for Early Prediction of Sepsis

Janmajay Singh, Kentaro Oshiro, Raghava Krishnan, Masahiro Sato, Tomoko Ohkuma, Noriji Kato

Fuji Xerox Co., Ltd., Japan

Abstract

Aims: *Physicians have to routinely make crucial decisions about patients' health in the ICU. Sepsis affects about 35% of ICU patients, killing approximately 25% of the afflicted. In this paper, we aim to predict the occurrence of sepsis early by studying the missingness of physiological variables and using it with the overall trends in data.*
Methods: *We chose XGBoost as our base model and tried several variations by changing hyperparameters, window sizes and imputation methods. To further improve the model, we used masking vectors to represent the missingness of features in the dataset. Additional modifications include shifting the Sepsis Label to earlier time steps and tuning the classification probability threshold to further improve the model's performance.*
Results: *The XGBoost model with a sliding window of size 5, no imputation, utilizing informative missingness of all temporal variables and trained on labels shifted by 3 hours before $t_{optimal}$, achieved a Utility Score of 0.337 on the full test set. We identified as "CTL-Team" in the challenge and were officially ranked 5^{th} on the basis of this score.*

1. Introduction

Sepsis is a potentially life-threatening organ dysfunction caused by the body's extreme response to an infection. A recent study assessed the global incidence rate of sepsis at 31.5 million cases/year with 19.4 million cases of severe sepsis and potentially 5.3 million deaths annually [1].

In addition to a high occurrence rate, early detection and treatment of sepsis is essential for patient survival as each hour of delay leads to an average decrease in survival of 7.6% [2].

For early identification of highly susceptible patients, scoring metrics are commonly used which prove useful as track and trigger monitoring systems of patient health. For Sepsis monitoring, the SOFA (Sequential Organ Failure Assessment) score and qSOFA (quick-SOFA) are considered standards [3, 4].

More recently, the increased availability of healthcare data has opened new avenues to develop statistical models for data-driven representations of several illnesses.

The Physionet 2019 Challenge [5] was to encourage development of computational algorithms for early prediction of sepsis. The competition proposed a new metric, called the "Utility Score" for assessing time-wise prediction accuracy and favoured models which correctly predicted sepsis onset about 12 hours before a physician would diagnose the same.

For this purpose, we develop an XGBoost-based model for early prediction of Sepsis. A naive model does not prove suitable for the task because of certain characteristics of the dataset. Thus, we empirically prove the existence and importance of patterns in the missing data (called *informative missingness*) and leverage it to develop a more accurate model.

2. The Dataset

2.1. General Characteristics

The challenge used data comprised of three distinct hospitals in the United States. Data from corresponding hospitals were referred to as sets A,B and C respectively. The full training dataset provided by the organizers was a subset of data from sets A and B totalling ICU stay records of 40,336 patients. Data from set C remained hidden and was used to check generalization performance of models. Please see [5] for further details of the dataset.

In initial analysis, we noted the significant differences in ICU Length of Stay (ICULOS) of sepsis and non-sepsis patients. Almost all patients who were not diagnosed as having Sepsis, spent less than 60 hours in the ICU (mean stay of $\approx$ 37 hours with std of 15.8 hours). On the other hand, patients with Sepsis spent more time in the ICU (mean stay of $\approx$ 60 hours with std of 59.2 hours). This may be due to increased complications in medical care due to Sepsis.

Finally, there was also significant class imbalance. There were just 2932 (7.26%) Sepsis patients and only 27916 (1.8%) records labeled as data corresponding to Sepsis.

2.2. Informative Missingness

Several works from the Machine Learning in Healthcare domain utilize patterns in missing data to make more accurate predictions. At the same time, this serves to alleviate the sparsity of features in such datasets [6–8]. In [8], Lin et al. highlight the difference between Missing Completely at Random (MCAR) and Missing at Random (MAR) data and empirically prove that when data is not MCAR, including patterns of feature missingness (or *informative missingness*) is useful for machine learning models.

Inspired by the importance of informative missingness (**IM**) presented by the results of these earlier works, we analyzed our data in three stages to look for variables exhibiting IM.

In the first stage of analysis, we separated the patients into Sepsis and non-Sepsis classes. Then we computed the overall observation rates for each variable (except those with 100% observation) and compared them for the 2 classes. This is illustrated in Fig.1.

In the second stage, we wanted to see the observation trends of variables identified as relevant (IM variables) to Sepsis patients. For this purpose we plotted the hourly probability of observation of such variables for both classes. This can be seen for two such variables (FiO2 and Lactate) in Fig.2 - Left. Both variables show persistent differences in observation probabilities between classes. Similar trends were also observed for other variables.

Finally, since our aim was to maximize the Utility score metric, we were most interested in a 6-hour time window around $t_{optimal}$ [5]. Variables which show clear trends in observation probabilities in this window should help a predictive model achieve a higher score. This can be seen on Figure 4 - Right. There is a marked increase in probability of observing both variables a little after $t_{optimal}$. Similar peaks for other IM variables were observed as well, but to varying extents.

The above analysis was done to assess whether including informative missingness would improve Utility score. Considering persistent differences in hourly observation probabilities and peaks before t_{sepsis}, we concluded that missingness patterns of IM variables would be strong early indicators of Sepsis onset. We will later include this in our model.

3. Methods

We chose our model to be based on the Extreme Gradient Boosting (XGBoost) algorithm. Several variants were trained and compared. For implementation, we chose the open source library and corresponding optimizations presented in [9].

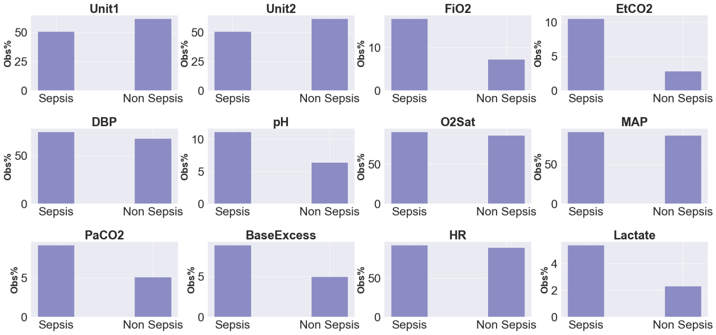

Figure 1: Top 12 features showing the largest difference in observation rates between Sepsis and non-Sepsis Patients.

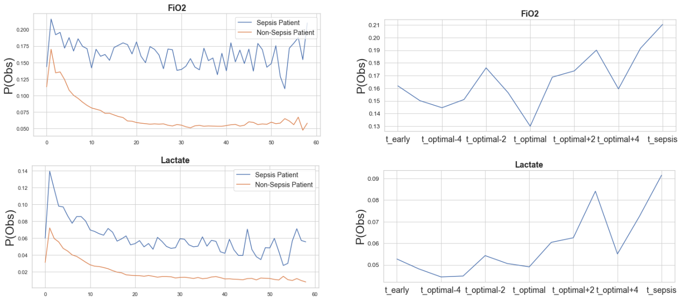

Figure 2: **Left:** Overall observation probability for patients with 20≤ICULOS≤70. **Right:** Observation Probabilities for all Sepsis patients around $t_{optimal}$.

3.1. Data Preprocessing

Formally, the dataset is a sequence of records where each observation is denoted by $x_t \in \mathbb{R}^d$ such that t is the time of observation and d is the number of features.

Windowing. Since XGBoost is not a sequence learning model, we made non-overlapping windows of w consecutive records and use as a single input. Thus, for an observation at t, we concatenate records from x_{t-w+1} to x_t ($(d * w)$ features) to predict the Sepsis Label at t. For each new record as input, the window moved by 1 to include new information (sliding window).

Informative Missingness. As mentioned in [5], there were 3 types of variables in the dataset. Two of them - Vitals and Laboratory variables were temporal in nature. Each observation x_t was used as input without any imputation, i.e. missing features were represented as NaNs. To represent missingness patterns, we used all temporal features of each x_t to create corresponding masking vectors m_t, similar to [6] and [7] such that; ${m_t}^d$ = 0 if ${x_t}^d$ is not observed and 1 otherwise.

Additional Features. We found that binning the age demographic variable and encoding it as a one-hot vector led to slightly improved results. We also decided to use a similar severity scoring metric as described in Section 1. Due to the absence of the Glasgow Coma Scale variable in the dataset, SOFA score could not be used. Thus, we resorted to using the NEWS (National Early Warning Score) prescribed by the Royal College of Physicians (RCP) in the United Kingdom. [10] suggests that NEWS is more accurate for septic shock and sepsis related mortality prediction, even compared to SOFA and qSOFA scores. The NEWS was denoted by NaN whenever variables necessary for its computation were missing in the dataset. It was used as part of the masking vector in a similar way as the other variables.

Label Shifting. The Utility Score metric was defined such that a positive utility score could be achieved by making a correct sepsis prediction up to 6 hours before $t_{optimal}$ and about 12 hours before t_{sepsis}. Since we used a non-sequential model and our window size was relatively small, we decided to shift the Sepsis Labels further back to encourage even earlier predictions. That is for Sepsis patients, SepsisLabel = 1 if $t \geq t_{sepsis} - (6 + k)$ and SepsisLabel = 0 if $t < t_{sepsis} - (6 + k)$. We experimented with several values of k and found that $k = 3$ resulted in the maximum Utility Score on our local testing set.

Our preprocessing steps are illustrated in Figure 5.

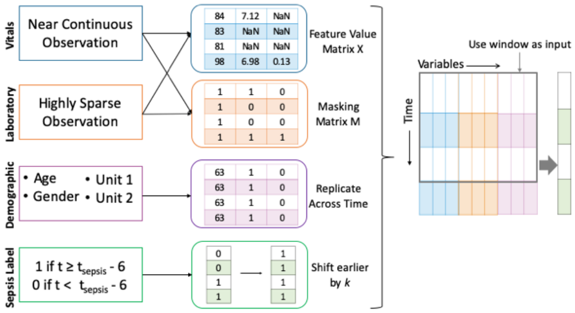

Figure 3: Data Preprocessing for the best model. The sliding window results in each record being used multiple times in the input.

3.2. Model Development

Several variants based on XGBoost were created and tested. Since any kind of imputation (mean, median, forward-filling) resulted in adverse results, no imputation was performed. No additional preprocessing was necessary since XGBoost automatically handles sparse inputs. We also augmented the input to our XGBoost model with a masking vector similar to [11]. The model was then trained on the shifted labels. The feature importances of the best model are summarized in Table 1.

Feature	Gain
Lactate indicator at t	4124.448
FiO2 indicator at $t-1$	3729.544
Lactate indicator at $t-1$	3352.092
FiO2 indicator at t	3307.474
FiO2 at $t-2$	2930.259

Table 1: 4 of the 5 most important features of the Best Model were indicator variables.

4. Results

Before model development, 20% of the training data was randomly selected (patient-wise) and kept aside for testing (8068 patients). Hyperparameters were tuned using k-fold stratified cross validation on the remaining training data. Performance of different XGBoost variants on the local testing set are summarized in Table 2. All models use a windowed input of size 5 as a default. The naive model simply used a windowed input and is called ”XGBoost-W”. All others are modifications of this model. As is evident from the results, mean imputation results in the worst performing model. This is probably due to important information being lost as a consequence of imputing missing values and bias introduced to data. A modification with no imputation and using masking vectors made from laboratory variables only was tried and it exceeded the baseline model (Sparse IM). Following this, a larger masking vector, using all temporal variables was created, which further improved the Utility score (All IM). Preprocessing steps described in Section 3.1 brought about additional improvements. This ”Best Model” achieved a mean AUROC score of 0.8406 and mean Utility of 0.404 on 5-fold stratified cross-validated data. The same were 0.8377 and 0.4402 on the local testing set. It also achieved the highest score on hidden set A with 0.401 Utility. For final rankings, the best model was run on other hidden sets; the results for which are summarized in Table 3.

Table 2: Comparison of several models.

Model	AUROC	Utility Score
XGBoost-W	0.8305	0.4074
+ Mean Imputation	0.8198	0.3969
+ No Imputation + Sparse IM	0.8328	0.4205
+ No Imputation + All IM	0.8318	0.4225
+ Best Model	**0.8377**	**0.4402**

5. Discussion

We performed some analysis to check the validity of our assumptions pertaining to the usefulness of IM in predic-

Table 3: CTL-Team's Best Model performance on different testing sets.

Dataset	AUROC	Utility Score
Test Set A	0.806	0.401
Test Set B	0.846	0.407
Test Set C	0.805	-0.094
Full Test Set	N/A	0.337

tive models. Since adding IM features led to higher AUROC and Utility scores, we assumed these models would have higher probabilities of predicting True Positives in the t_{early} to t_{sepsis} range with possibly earlier peaks. Trends in Figure 4 showed that this assumption was wrong. The reason these models performed better was a much lower number of False Positives in their predictions. Figure 5 shows prediction results on an overall basis (predicting at least one Sepsis record for a Sepsis patient). A similar trend of decreasing False Positives with IM addition was seen for record-wise predictions as well.

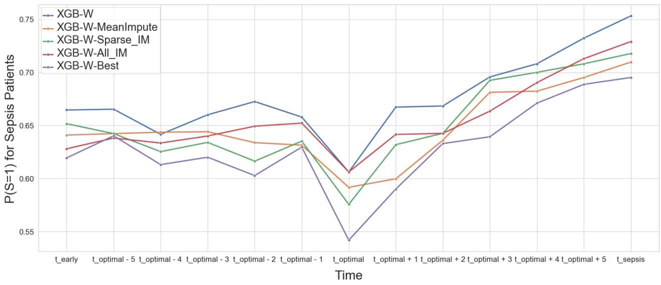

Figure 4: Sepsis prediction probabilities for various models around $t_{optimal}$.

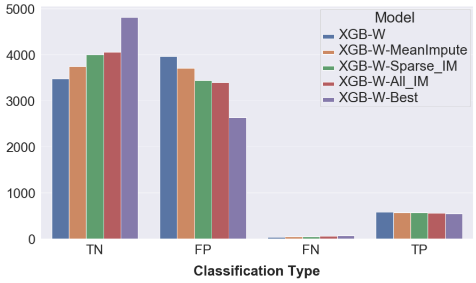

Figure 5: Patient-wise prediction comparison for various models.

References

[1] Fleischmann C, Scherag A, Adhikari NK, Hartog CS, Tsaganos T, Schlattmann P, Angus DC, Reinhart K. Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limitations. American journal of respiratory and critical care medicine 2016;193(3):259–272.

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

[3] Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart C, Suter P, Thijs L. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. Intensive care medicine 1996; 22(7):707–710.

[4] McLymont N, Glover GW. Scoring systems for the characterization of sepsis and associated outcomes. Annals of translational medicine 2016;4(24).

[5] Reyna M, Josef C, Jeter R, Shashikumar S, Westover MB, Nemati S, Clifford G, Sharma A. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. Critical Care Medicine (In Press);.

[6] Lipton ZC, Kale DC, Wetzel R. Modeling missing data in clinical time series with rnns. arXiv preprint arXiv160604130 2016;.

[7] Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Scientific reports 2018;8(1):6085.

[8] Lin JH, Haug PJ. Exploiting missing clinical data in bayesian network modeling for predicting medical problems. Journal of biomedical informatics 2008;41(1):1–14.

[9] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016; 785–794.

[10] Usman OA, Usman AA, Ward MA. Comparison of sirs, qsofa, and news for the early identification of sepsis in the emergency department. The American journal of emergency medicine 2019;37(8):1490–1497.

[11] Chen R, Jankovic F, Marinsek N, Foschini L, Kourtis L, Signorini A, Pugh M, Shen J, Yaari R, Maljkovic V, Sunga M, Song HH, Jung HJ, Tseng B, Trister A. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19. New York, NY, USA: ACM. ISBN 978-1-4503-6201-6, 2019; 2145–2155. URL http://doi.acm.org/10.1145/3292500.3330690.

Address for correspondence:

Janmajay Singh
Research & Technology Group,
Fuji Xerox Co., Ltd.,
6-1 Minatomirai, Nishi-ku, Yokohama-shi,
Kanagawa-ken, Japan. 220-8668
janmajay.singh@fujixerox.co.jp