Sepsis Prediction using Continuous and Categorical Features on Sporadic Data

Varsha Sharma, Chirayata Bhattacharyya, Tanuka Bhattacharjee,
Sundeep Khandelwal, Murali Poduval, Anirban Dutta Choudhury
TCS Research & Innovation, India
Email: (sharma.varsha1, chirayata.b, bhattacharjee.tanuka,
sundeep.khandelwal, murali.poduval, anirban.duttachoudhury)@tcs.com

***Abstract*—Sepsis is one of the most prevalent causes of mortality in Intensive Care Units (ICUs) and also one of the most expensive health-care problems. Delayed treatment is associated with increase in death and financial burden. This work proposes an early prediction of sepsis validated on Physionet Challenge 2019 dataset. The challenge is to extract continuous, categorical and domain-specific discriminating features from highly sporadic lab data and vital signals. We find that the imputation of extremely isolated data lower the prediction performance. In order to mitigate this, we use a sliding window on sporadic data to generate continuous features which capture the trend. We also devise a binning approach to generate categorical features from the aperiodic data in order to discriminate the deviation from normalcy. Lastly, we observe that a logical fusion of Random Forest and Logit Boost provides optimal performance. Normalized Utility Score (NUS) is used to benchmark the performance of the proposed baselines. Five-fold cross-validation of the best preforming pipeline across the data reveals high median NUS of 0.401.**

***Index Terms*—Sepsis, Predictions, Categorical, Sporadic**

I. Introduction

Sepsis is one of the leading life-threatening bacterial and viral infections that often turns fatal if not detected and treated in time. Sepsis refers to the syndrome wherein a previously known or unknown infection leads to immune system override and rapid progression to multi-organ failure. The Sepsis 3 working committee has emphasized on organ dysfunction as an essential part of the definition of sepsis [1]. The definition proposed therefore is *"life threatening infection caused by a dysregulated host response to infection"*. People can get affected by sepsis at any time, but those staying in Intensive Care Unit (ICU) are more susceptible to contract it. The surviving sepsis campaign quotes mortality rates from Europe and North America to be as high as 41% and 28.3% [2]. In U.S. hospitals, sepsis is the most expensive condition treated with an aggregate cost of USD $15.4 billion in 2009 [3], whereas non-specific diagnoses of sepsis account for another USD $23.7 billion each year [4]. The prevalence of sepsis is increasing, with a 17% increase in the number of documented cases between 2000 and 2010 [4], while sepsis-related deaths have surged to 31% between 1999 and 2014 [5]. Approximately 30,000 sepsis-related deaths occur annually in USA, with particularly high rates in critically ill patients admitted to ICUs [4].

Sepsis develops gradually and escalates to catastrophic multi-organ failure with a very high risk of mortality. However, no single laboratory test or clinical sign in itself can be considered diagnostic of sepsis. The diagnosis requires great clinical acumen and alertness, in combination with an astute analysis of laboratory results and physiological parameters like heart rate, mean arterial pressure and respiratory rate. A high index of suspicion coupled with known scores enables a clinician to institute treatment with antibiotics in time to save life. High risk populations especially include those with multiple co-morbidities like diabetes and heart disease, increasing age and intensive care admissions. Research reveals that mortality from severe sepsis and septic shock improves by 7.6% per hour with *early* and *appropriate* administration of antibiotics [6].

The use of standard culture techniques for the detection and isolation of pathogenic organisms from a sterile body fluid specimen is still considered the "gold standard" for diagnosis of infection and sepsis [7]. Routine blood cultures by this standard to detect sepsis can take 6 hours to 5 days to grow an organism to detectable levels, additional time is required to identify the organism and test for appropriate antibiotic susceptibility (24-48 hours). Various scoring systems like Sequential Organ Failure Assessment (SOFA) Score [8], Systematic Inflammatory Response Syndrome (SIRS) criteria [9] and Simplified Acute Physiology Score (SAPS II) [10]. These methods result in the well-structured tabulation of vital signs and lab data to generate indicative scores and risk assessments. However, they do not analyze trends in patient data or correlation between measurements.

A reliable means of annotated early prediction of sepsis using available lab data and vitals is an unsolved problem. The contributions in this paper are following:

1) Analyzing the effect of imputation and preprocessing on sporadic time series and categorical data
2) Devising novel domain specific features in consultation with medical experts.
3) Finally, we introduce an end-to-end pipeline to process a combination of vital signs and lab dataset to predict early diagnosis of sepsis.

In section II, we briefly describe the data-set and remark on some observations. In section III, we present our pipe-line with clear definition of multiple baselines. Section IV and Section

V convey the result and conclusion respectively.

II. Analysis of Sepsis Dataset

A. Dataset

The dataset used in this study are provided by Physionet 2019 challenge [11]. Organisers provided hourly data which consist of 8 vital signals, 26 laboratory parameters and 6 demographic details of 40,336 subjects from two ICU units - Medical-ICU (MICU) and Surgical-ICU (SICU). Fig. 1 visualizes the normalized raw data for an ICU patient transforming from non-sepsis to sepsis stage around 89^{th} hour of her MICU stay. Left vertical axis represents the vital signs and right vertical axis represents the laboratory investigations. Discontinuity in vitals and lab values indicate missing data. Respective variation of the lab values are represented by the bars where the shades demonstrate the value (the higher the darker). In order to preserve the practicality of the problem at hand, Physionet Challenge organisers imposes a *restriction on analysis of the future data*. For the rest of this paper, the same restriction i.e. "non-availability of future data at any given hour" is maintained. The organisers introduced Normalised Utility Score (NUS) [11], which is used as the performance criteria in this paper.

B. Correlation of Sepsis and SOFA

Sequential organ failure assessment (SOFA) is a clinical prediction score used to track patient's status during the stay in ICU. SOFA score calculation requires information about the Hepato-renal function and Coagulatory status. Recently, a simplified version of SOFA named as quick SOFA (qSOFA) [8] was introduced by the Sepsis-3 group [7]. qSOFA requires only three parameters, namely Respiratory Rate (RR), Systolic Blood Pressure (SBP) and Glasgow Coma Scale (GCS). t_{SOFA} is defined as a two point change in the patients SOFA score. This along with clinical suspicion of infection helps in identifying potential for end-organ damage. Our objective is to analyze the correlation of SOFA score with the sepsis labels. The challenge data provides all parameters except GCS and Dopamine required for Nervous system and Cardiovascular system information respectively. Hence, we approximate both SOFA and qSOFA score with the available parameters for all the patients at each hour and correlate it with the ground truth sepsis labels, As shown in Fig. 2, the correlation is extremely poor (peaked around zero). This may be attributed to the inconsistency in the reported data e.g. in the data shown in Fig. 1, Serum Bilirubin is an important indicator of hepatic function and is found missing in the dataset.

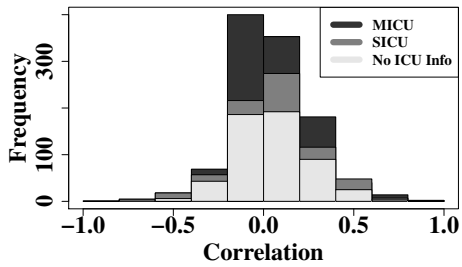


Fig. 2. Histogram of qSOFA and Sepsis correlation for different ICU units

C. Temperature and SBP at Normal HR

Abnormal HR is associated with sepsis [12]. But when HR is within normal range (60-100), we perceive some relation

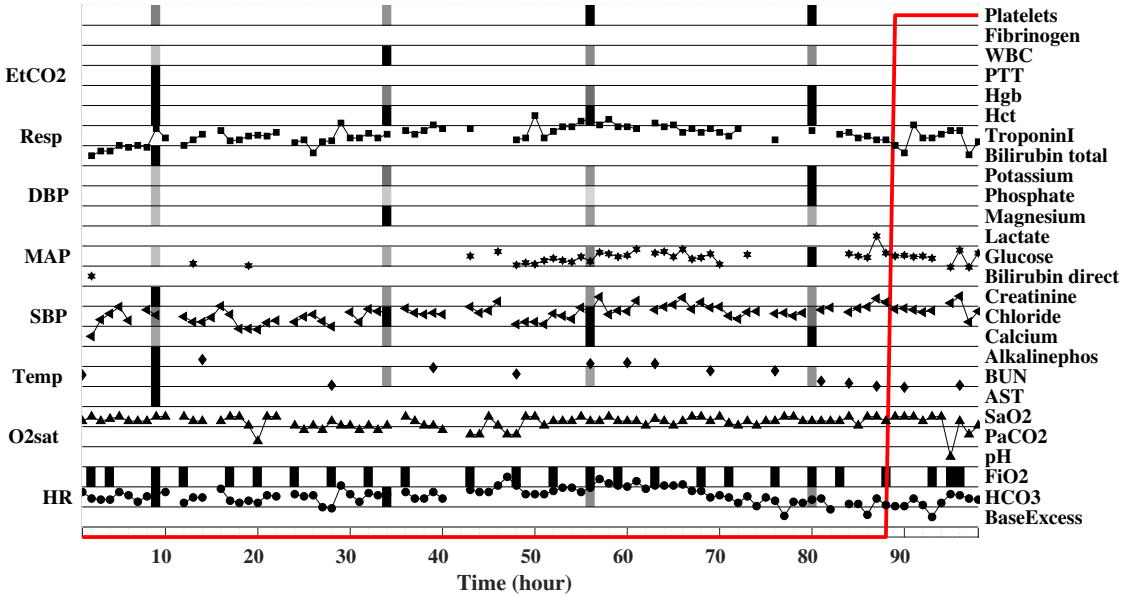
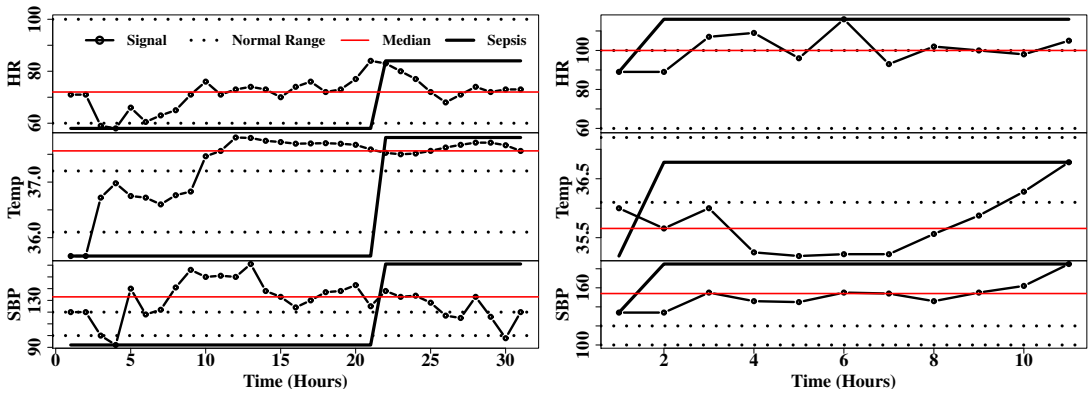


Fig. 1. Raw data plot of a 65 years old female subject staying in MICU.



(a) Direct relation (b) Inverse relation

Fig. 3. Relation between Temparature and SBP

between temperature and SBP. Both parameters are correlated bidirectionally (direct and inverse) with each other. When temperature is lower than its normal range, there exists an inverse correlation between temperature and SBP (Fig. 3(a)). When temperature is higher than its normal range, temperature and SBP exhibit a direct correlation (Fig. 3(b)). This trend is visible in 70% of the subjects.

D. Missing Data & Imbalance

The clinical dataset is *inconsistently inconsistent*. There is no specific trend in missing data across the ICU stay of a subject as well as across subjects. The laboratory tests are the most sporadic in nature. The vital signs, though measured more frequently, is far from a steady periodicity (Fig. 4).

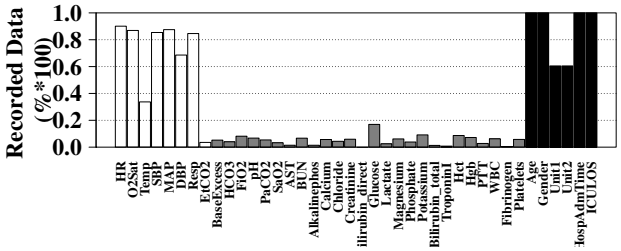


Fig. 4. Bar plot of recorded data percentage of vitals (white), lab data (grey) and demographic data (black).

Challenge dataset is highly imbalanced with 7.26% of the subjects heading to sepsis (2,932 out of 40,336). The sepsis occurrence in terms of hourly instances is much lower 1.8%.

III. Proposed Methodology

As mentioned in section II, the concerned dataset is unique in nature when compared to generic time series classification problems. In this section, we present a novel pipeline to handle this data. The flow chart of the proposed method is shown in Fig. 5 and the comprehensive details of various components are described in the following subsections.

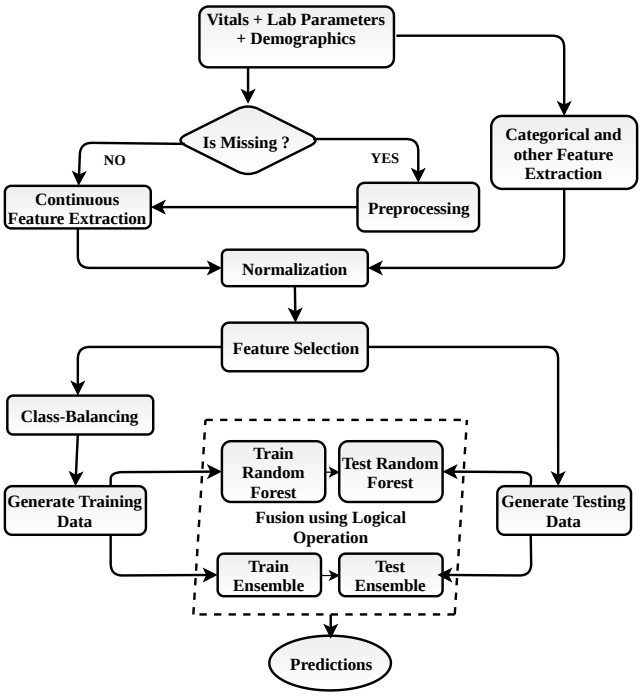


Fig. 5. Flow Chart

A. Preprocessing

We try to impute the missing data with the methods available in [13], like Last Observation Carried Forward (LOCF) and Next Observation Carried Backward (NOCB), mean imputation and interpolation. Fig. 6 shows the imputed data plots for a sample subject.

LOCF and NOCB are common statistical approaches to the analysis of longitudinal data where some follow-up observations may be missing. Longitudinal data tracks the same sample at different points in time. Computing the overall mean is an imputation method that takes no precedence of the time series characteristics or relationship between the variables. We perform the interpolation on all the 8 vital signs by fitting spline/linear if at least two values are present. In case there

exists only one value in the entire signal, we repeat the same value for the rest of the ICU stay and if there exists no value for the entire signal, we take random values within the normal range. Ranges for these random values are carefully selected so that variation between them is less. Imputation results are provided in section IV.

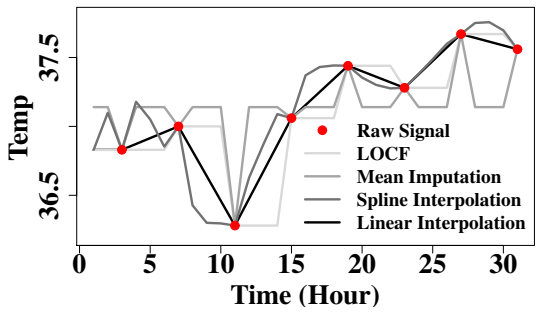


Fig. 6. Visual comparison of data imputation methods.

B. Feature Description

In the proposed technique, we extract 336 features from 34 signals (vitals and lab data) and 6 key demographics. Predominantly, variables or any observational data either represent measurements on some continuous scale, or they represent information about some categorical or discrete characteristics. We extract the information in both ways. The considered features can be categorized as follows:

1) Baseline 1 (Continuous Features): Sepsis progresses when the immune response to bacterial infection injures own tissues and organs. Given the fairly fast progression of this medical condition, it is important to observe recent patient history over a time window (say i hours). For first i hours, we consider the entire history available at that point of time i.e. the window size increases from 1^{st} hour to $(i-1)$ hours. From i^{th} hour, we take a sliding window of past $(i-1)$ hours and the i^{th} hour for feature calculation.

We observe that the occurrence and/or frequency of lab tests are related with the progress of sepsis. For example, in Fig. 1, during the initial hours, frequency of temperature measurement is low and after 56^{th} hour, it increases gradually till the person is labeled sepsis in the 89^{th} hour. Fig. 1 also reveals that Mean Atrial Pressure (MAP) becomes more frequent after 48^{th} hour. A possible reason can be that depending on the current progression of sepsis in a subject, the physicians may decide to carry out certain procedures and/or administration of medicine(s). To measure the effect of those, they need to frequently monitor certain lab data. Hence, frequencies of lab tests become indirectly connected to sepsis progression. Medical experts indicated some more statistical aggregators to capture the instability of 34 parameters (i.e. both lab data and vitals).

The list of features calculated from each those 34 parameters are (1) count of valid data records present, (2) out of range count, (3) difference between the value at i^{th} hour and $(i-1)^{th}$ hour, (4) difference between the i^{th} hour and the mean of the past $i-1$ hours and (5) difference between the final hour and the variance of the past $i-1$ hours; leading to a total of 170 features. We vary i from 6 to 96 in order to determine the sweet spot of the window size and as shown in Fig. 7, a sliding window length of 9 hours performs best.

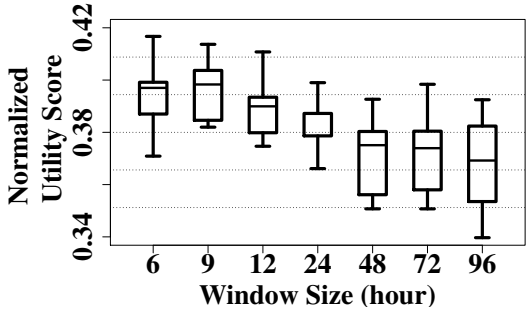


Fig. 7. The box-plot of 5-fold utility score with the variation of window size.

2) Baseline 2 (Categorical Features): We consider all 34 signals as qualitative data and bin each of them into 6 categories. Then we extract 4 features for each of the 34 signals, resulting into 136 categorical features. The features are (1) the category, (2) the upper limit (UL), (2) the lower limit (LL) and (4) one hot dummy encoding, elaborated in Algorithm 1.

Algorithm 1 Categorical Feature Extraction from Clinical Signals and following Limits: LL, UL, $LL2 = 0.8 \times LL$, $UL2 = 1.2 \times UL$

procedure CATEGORICAL($LL, UL, LL2, UL2$)
  switch S_i **do**
    case $(LL_i < S_i < UL_i)$
      return $(1, LL_i, UL_i, 1)$;
    case $(LL2_i < S_i < LL_i)$
      return $(2, LL2_i, LL_i, 1)$;
    case $(UL_i < S_i < UL2_i)$
      return $(3, UL_i, UL2_i, 1)$;
    case $(min(S_i) < S_i < LL2_i)$
      return $(4, min(S_i), LL2_i, 1)$;
    case $(UL2_i < S_i < max(S_i))$
      return $(5, UL2_i, max(S_i), 1)$;
    case $(S_i = NAN)$ **return** $(6, 0, 0, 0)$;
end procedure

Let us take a specific example. Table I reveals 6 categories and associate UL and LL for HR signal. In Table II, we randomly chose few cases from HR and the features calculated as per Algorithm 1. For example, in case 1, HR is 70 beats per minute (bpm). Therefore, the category for 70 is '1', UL is 100 and LL is 60.

3) Baseline 3 (Demographic Features): We include the following features for each subject at each hour - (1) current hospital admission duration, (2) ICU unit, (3) gender, (4) age and (5) current ICU stay duration, as compiled in [11].

TABLE I
CATEGORIES FOR HR SIGNAL

Input Range	Category	UL	LL
60-100	1	100	60
48-60	2	60	48
100-120	3	120	100
20-48	4	48	20
120-210	5	210	120
NAN	6	0	0

TABLE II
FEATURES EXTRACTED FOR SAMPLE HR CASES

Cases	HR	Extracted Features			
		Category	*UL*	*LL*	*One-Hot*
1	70	1	100	60	1
2	45	4	48	20	1
3	NAN	6	0	0	0
4	110	3	120	100	1

Moreover, ICU hours are divided into the non-overlapping bins of 10 hours along with their (6) LL and (7) UL. ICU stay duration is binned in steps of 10 hours e.g. < 10 hours, $10-20$ hours, $21-30$ hours etc. For example, if current ICU stay is 28, it is placed in the bin of $21-30$ hours with 20 as LL and 30 as UL. Similarly, age is divided into 4 categories, $0-25$, $25-50$, $50-75$ and $75-100$ years along with their (8) LL and (9) UL.

4) Baseline 4 (Domain Features): Some commonly used scoring systems for sepsis are added as features - National Early Warning Score (NEWS) (6 features), Modified Early Warning Score (MEWS) (5 features) and Acute Physiology and Chronic Health Evaluation II (APACHE II) (10 features) [14]. Major limitations of applying these scoring systems in our dataset are missing data such as level of consciousness and emergency oxygen therapy information - awareness, verbal & painful response etc. We can only select the signals which are available in the dataset and made separate categories, which yields 21 features.

IV. RESULT

Data imputation is performed subject-wise in training phase and window-wise in testing phase. However, in Table III, we observe that all imputation methods introduce bias in analysis and performs poorly. Hence we drop imputation from the pipeline shown in Fig. 5.

TABLE III
(NORMALIZED UTILITY SCORES) $\times 10^{-3}$ AFTER PREPROCESSING WITH MEDIAN VALUE IN BOLD

Imputation Method	Five-Folds					Variance
	1	*2*	*3*	*4*	*5*	
LOCF & NOCB	169	181	**175**	187	144	276
Mean	142	184	**176**	193	110	1185
Spline	186	**211**	213	208	251	550
Linear	192	**209**	220	220	163	578

One possible reason could be the unknown activities in ICUs e.g. a lot of medicines and external fluid get administered, other body conditions (fasting for some tests etc.) are imposed. Without those details, in a sporadic dataset, imputation is incomplete, resolution into lowering Signal to Noise Ratio (SNR) in the imputed signal.

Getting rid of the feeble and redundant features is an essential step to reduce variance in testing performance. Prior to feature selection, train data is normalized with the respective standard deviation and mean values and test features are normalized with the stored values of standard deviation and mean from training set. We rank all the extracted features using the Minimum Redundancy Maximum Relevance (MRMR) technique [15], and visualize the cross validation performance with increasing number of ranked features in Fig. 8. We choose the high bias, low variance combination i.e. the box having least stretch, high minimum and high median value. Fig. 8 shows the top 320 features give us the optimal performance. The most prominent features as shown in Table IV, got picked from different categories mentioned throughout this section, thereby validating the differentiating nature of the novel features.

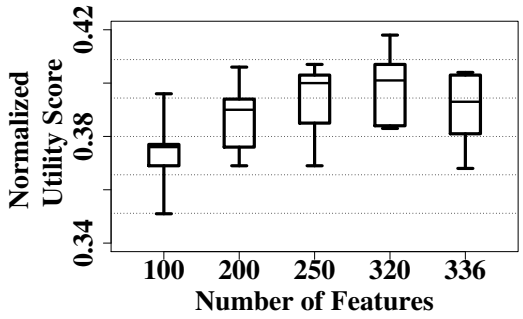


Fig. 8. The box plot of 5-fold utility score with number of selected features.

TABLE IV
LIST OF 10 PROMINENT FEATURES SELECTED FROM MRMR

Features (1-5)	Features (6-10)
ICU stay	$(i^{th}$-$(i-1)^{th})$ hr of O2Sat
Non-NAN count of FiO2	Out of Range for WBC
NEWS cutoff for HR	Out of Range for PTT
$(i^{th}$-$(i-1)^{th})$ hr of SaO2	Hosp. Adm. Time
Category of Resp	Category of Glucose

To mitigate class imbalance, we randomly under-sample the hourly instances of the majority class. Then We train two models; one is the Random Forest (RF) and other is Adaptive Logistic Regression of Ensemble learning algorithm (LB). Hyper parameters of both the models are tuned by Bayesian optimization [16] considering the error of median of NUS as the minimizing optimization function. Predictions from both the models are fused together by *conjunction operation* to obtain the final predictions.

To demonstrate the potential efficacy of the different baseline algorithms detailed in section III, 5-fold cross-validation is performed on all the windows (multiple windows are extracted from each subject as detailed in subsection III-B. In order to avoid over-fitting, the partitions for cross-validation are done in such a manner that windows from the same subject never

appear in both train and test data. The function for calculating NUS generously rewards the classifier for initial predictions of sepsis and penalizes it for late or missed predictions. According to the evaluation function, classifier is rewarded when sepsis forecast lies between 12 hours before and 3 hours after t_{sepsis}. Sepsis labels are already shifted ahead by 6 hours in the provided data. We shift them ahead again by 6 hours (only in training data) for optimal prediction. The performance evaluations in various baselines are detailed in Table V with highlighted median values.

TABLE V
PERFORMANCE COMPARISON FOR ALL BASELINES IN SUBSECTION III-B

Baselines	**NUS ($\times 10^{-3}$) for 5 Folds**					**Variance**
	1	***2***	***3***	***4***	***5***	
1 (RF)	225	250	**242**	248	230	122
2 (RF)	230	286	**272**	281	241	626
3 (RF)	356	384	**363**	371	331	390
4 (RF)	359	387	**370**	378	343	292
4 (LB)	374	397	395	**387**	371	141
4 (RF + LB)	383	418	**401**	407	384	227

As evident, a gradual improvement in median can be observed as we continue to add features while moving to higher baselines. However, the same is not true for the variance parameter. It is rather high in baseline 2 and 3. As revealed by the performance Baseline 4 variation in Table V, LB provides two-fold improvement - reducing variation as well as improving bias over RF. Reason could be high dimension of features. The fused approach of baseline 4 provides further improvement in bias in each fold. But the percentage of improvement is not uniform throughout all the folds, resulting into increase in variance.

V. CONCLUSIONS

In this paper, we devise an algorithm for early detection of sepsis and present a pipeline to analyze the lab and vital data in ICU. The algorithm is a fusion of two classifiers trained on diverse domain-specific novel features extracted from the clinical data. The method yields a 0.401 median NUS on random 5 fold cross-validation of the training dataset. A fusion of RF and LB classifiers yield high bias and low variance. Imputation performs poorly on the clinical data.

REFERENCES

[1] S. Riedel *et al.*, "Early identification and treatment of pathogens in sepsis: molecular diagnostics and antibiotic choice," *Clinics in chest medicine*, vol. 37, no. 2, pp. 191–207, 2016.

[2] F. Gül, M. K. Arslantaş, İ. Cinel, and A. Kumar, "Changing definitions of sepsis," *Turkish journal of anaesthesiology and reanimation*, vol. 45, no. 3, p. 129, 2017.

[3] A. Elixhauser *et al.*, "Septicemia in us hospitals, 2009: statistical brief# 122," 2006.

[4] C. Torio *et al.*, "National inpatient hospital costs: the most expensive conditions by payer, 2011: statistical brief# 160," 2006.

[5] L. Epstein, "Varying estimates of sepsis mortality using death certificates and administrative codes—united states, 1999–2014," *MMWR. Morbidity and mortality weekly report*, vol. 65, 2016.

[6] R. M. Klevens *et al.*, "The impact of antimicrobial-resistant, health care–associated infections on mortality in the united states," *Clinical infectious diseases*, vol. 47, no. 7, pp. 927–930, 2008.

[7] M. Singer *et al.*, "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)Consensus Definitions for Sepsis and Septic ShockConsensus Definitions for Sepsis and Septic Shock," *JAMA*, vol. 315, pp. 801–810, 02 2016.

[8] J.-L. Vincent *et al.*, "The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure," *Intensive care medicine*, vol. 22, no. 7, pp. 707–710, 1996.

[9] C. W. Seymour, , *et al.*, "Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)Assessment of Clinical Criteria for SepsisAssessment of Clinical Criteria for Sepsis," *JAMA*, vol. 315, pp. 762–774, 02 2016.

[10] L. Gall *et al.*, "A new simplified acute physiology score (saps ii) based on a european/north american multicenter study," *Jama*, vol. 270, no. 24, pp. 2957–2963, 1993.

[11] M. Reyna, Josef, *et al.*, "Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019," *Critical Care Medicine*, 2019.

[12] M. P. Griffin *et al.*, "Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness," *Pediatric research*, vol. 53, no. 6, p. 920, 2003.

[13] H. Kang, "The prevention and handling of the missing data," *Korean journal of anesthesiology*, vol. 64, no. 5, p. 402, 2013.

[14] W. C. Yuan *et al.*, "The significance of national early warning score for predicting prognosis and evaluating conditions of patients in resuscitation room," *Hong Kong Journal of Emergency Medicine*, vol. 25, no. 6, pp. 324–330, 2018.

[15] H. Peng *et al.*, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 8, pp. 1226–1238, 2005.

[16] Snoek *et al.*, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, pp. 2951–2959, 2012.