

Proximity of Cellular and Physiological Response Failures in Sepsis

Ali Jazayeri , Muge Capan , Julie Ivy, Ryan Arnold, and Christopher C. Yang

Abstract—Sepsis is a devastating multi-stage health condition with a high mortality rate. Its complexity, prevalence, and dependency of its outcomes on early detection have attracted substantial attention from data science and machine learning communities. Previous studies rely on individual cellular and physiological responses representing organ system failures to predict health outcomes or the onset of different sepsis stages. However, it is known that organ systems' failures and dynamics are not independent events. In this study, we identify the dependency patterns of significant proximate sepsis-related failures of cellular and physiological responses using data from 12,223 adult patients hospitalized between July 2013 and December 2015. The results show that proximate failures of cellular and physiological responses create better feature sets for outcome prediction than individual responses. Our findings reveal the few significant proximate failures that play the major roles in predicting patients' outcomes. This study's results can be simply translated into clinical practices and inform the prediction and improvement of patients' conditions and outcomes.

Index Terms—Sepsis prediction, proximate failures, sepsis feature selection.

I. INTRODUCTION

SEPSIS is defined as an organ dysfunction due to pathophysiological responses to infection [1], [2]. Previous studies show that not only is it widespread, but it has a relatively high mortality rate [3]. In the retrospective data of 2.9 million patients, 21% of adult patients have had clinical manifestations of sepsis, and 6% of them died in the hospital or were discharged to

hospice [4]. Despite its severity, studies have shown that early diagnosis and appropriate therapeutic management of patients with sepsis can significantly improve their outcome and reduce the rate of irreversible organ dysfunction [5], [6]. However, early detection of sepsis remains a complex problem due to the lack of a universally accepted definition of sepsis symptoms and misinterpreted clinical signs of sepsis due to shared features with other health conditions. Therefore, identifying reliable clinical indicators for early prediction of sepsis progression is an emerging focus discussed in several recent studies [7]–[10].

With regards to analyzing reliable data-driven clinical indicators, machine learning approaches leveraging big data have shown advantages, such as improved prognosis of complex medical conditions, increased diagnosis accuracy, and reduced costs over the continuum of health care. These advantages are attributed to the rapid emergence of electronic health records collection and storage systems [11], [12].

Sepsis has attracted attention by health care analytics due to its prevalence and complexity. For example, Gultepe *et al.* adopted a set of machine learning techniques such as naïve Bayes, support vector machines, and hidden Markov models to predict lactate levels and in-hospital mortality of sepsis patients. The features selected for the mortality prediction are lactate level, white blood cells (WBC) counts, respiratory rate, temperature, and mean arterial pressure (MAP) [13]. In another study, Tsoukalas *et al.* adopted an inferential statistics approach to predict mortality and length of stay (LOS) of sepsis patients. The features selected for the mortality prediction are temperature, respiratory rate, WBC, MAP, and lactate levels. For predicting the LOS, they also used positive blood culture data and the number of drug administrations during patients' hospitalization in addition to the five features used for mortality prediction [14]. Taylor *et al.* compared multiple approaches, including logistic regression, classification, and regression tree (CART) models and random forest model, to predict in-hospital mortality of patients with sepsis. The random forest model had a significantly higher area under the curve (AUC) in their study than other models. The most important features were identified as oxygen saturation, respiratory rate, blood pressure, Blood Urea Nitrogen (BUN), and Albumin [15].

Hornig *et al.* developed four models; vital model (vital signs and patient demographics), chief complaint model (vital signs, patient demographics, and chief complaint), bag of words, and topic model (both include vital signs, patient demographics, chief complaint, and nursing assessment) with two different text processing approaches for the chief complaint and nursing

Manuscript received November 6, 2020; revised March 9, 2021, April 23, 2021, and June 12, 2021; accepted July 12, 2021. Date of publication July 21, 2021; date of current version November 5, 2021. This work was supported in part by the National Science Foundation Smart and Connected Health (Award number: 1833538), and in part by the National Library of Medicine of the National Institutes of Health under Grant 1R01LM012300-01A1 and Award number: R01LM012300, in part by the National Science Foundation under the Grants NSF -1741306, IIS-1650531, and DIBBs-1443019. (Corresponding author: Ali Jazayeri.)

Ali Jazayeri and Christopher C. Yang are with the College of Computing and Informatics, Drexel University, Philadelphia, PA 19104 USA (e-mail: aj629@drexel.edu; ccy24@drexel.edu).

Muge Capan is with the Decision Sciences and MIS Department, LeBow College of Business, Drexel University, Philadelphia, PA 19104 USA (e-mail: muge.capan@drexel.edu).

Julie Ivy is with the Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695 USA (e-mail: jsivy@ncsu.edu).

Ryan Arnold is with the Santa Ynez Valley Cottage Hospital, Solvang, CA 93463 USA (e-mail: ryanarnold08@mac.com).

Digital Object Identifier 10.1109/JBHI.2021.3098428

assessment data analysis. The demographic data included age and gender, and vitals data include acuity, systolic blood pressure (SBP), diastolic blood pressure, heart rate, pain scale, respiratory rate, oxygen saturation, and temperature. The outcome of interest for prediction was infection and using different models, the discriminatory value of each set of data (four models) were quantified. They developed a support vector machine and compared its performance with logistic regression, naïve Bayes, and random forest for different models [16].

In [17], a randomized controlled clinical trial was conducted to evaluate the average LOS and in-hospital mortality rate when machine learning-based approaches are used for severe sepsis prediction. The patients were randomly classified into two groups. In the first group, a traditional approach was used to monitor the conditions of patients. In the second group, a machine learning-based classifier algorithm was used to generate a score. In cases where the score exceeded a pre-defined threshold, the in-charge nurse was called. They showed that using the machine learning-based classifier significantly decreased LOS and in-hospital mortality. The features used in their study included systolic blood pressure, diastolic blood pressure, heart rate, temperature, respiratory rate, peripheral capillary oxygen saturation (SpO_2), and age. They also used glucose, BUN, pH, creatinine, lactate, fraction of inspired oxygen (FiO_2), WBC, International Normalized Ratio (INR), platelets, and bilirubin as optional features. In another study [18], an artificial intelligence-based expert system was developed to predict the onset of sepsis. They showed that their system can predict ICU patients' sepsis onset between 4-12 hours sooner than traditional methods. Their system uses 65 features, including six high-resolution dynamical features calculated using five hours sliding windows, with five hours overlap, such as standard deviation of respiratory rate intervals and MAP, ten clinical features (such as heart rate and SBP), 25 general laboratory features (such as WBC, creatinine, and bilirubin), five Arterial Blood Gas laboratory features (such as pH and partial pressure of carbon dioxide, pCO_2), 19 demographics/history/context features (such as surgery in the past 12 hours, number of antibiotics in the past 12, 24, and 48 hours, and age).

The typical approach in these studies is collecting data for a set of features, pre-processing the collected data to transform the features into appropriate input formats, and implementing the analytical methods to the different subsets of data to ensure that the derived insights are generalizable. However, these analytical approaches treat individual cellular and physiological responses as independent features, i.e., the failure of one response over a predetermined time window is assumed to have no effect on the failure of another response. This assumption is a significant shortcoming of previously developed models considering organ systems are not independent, and their failures contribute to changes in other organ systems' states. Furthermore, at the individual cellular and physiological responses level, the relative importance of responses for prediction of patient outcomes is not identical [15]. Addressing these shortcomings, this paper aims to verify whether there is a difference among proximate elevation or depression of cellular and physiological responses causing organ failures. If some of the proximities of cellular

and physiological responses occur more frequently in different subpopulations of sepsis, we can determine if they can be used more effectively than individual cellular and physiological responses to predict the patient's outcome. More specifically, the objectives of this paper are twofold: i) to explore the relationship between proximate cellular and physiological responses causing organ failure and in-hospital mortality of hospitalized sepsis patients, and ii) to analyze the frequency of observed physiological responses causing organ failure across different patient subpopulations.

This paper is organized as follows. In the next section, the study population is defined, sub-populations definition criteria are explained, and the modeling approach and techniques adopted to identify proximate cellular and physiological responses are discussed. Then, in the results section, the findings of the analysis are presented, including the significant proximate responses. Next, in the discussion section, the results' clinical relevance is discussed in more detail, and finally, the paper concludes with key insights and directions for future work.

II. MATERIAL AND METHODS

A. Study Population

This study includes 12,223 adult sepsis patients (corresponding to 16,327 hospital visits) admitted between July 2013 and December 2015. The data is sourced from two hospitals with 1100 total in-hospital beds of a single tertiary care health care system. The data set includes retrospectively collected Electronic Health Records (EHR) data with the following inclusion criteria: age ≥ 18 at arrival, arrival to the hospital between July 2013 and December 2015, and with visit types of inpatient, Emergency Department only (outpatient), or observational visits. The study was approved by the health system's Institutional Review Board.

B. Subpopulations Definition

The significance of proximate failures of cellular and physiological responses is studied in a subpopulation of infected patients who have experienced sepsis. Therefore, all the sepsis patients have met infection criteria, defined as being administered with anti-infective for at least four days or a positive viral PCR (polymerase chain reaction) test for influenza. The sepsis population is defined as infected patients who have experienced organ dysfunction in the period of 24 hours before the first anti-infective administration until the last administration. Different markers and physiological variables can represent organ systems and dysfunctions. For example, in [19], it is shown that some of these markers, such as systolic arterial pressure and its variability as a cardiovascular marker, are associated with mortality and can inform mortality risk stratification. The organ systems and their associated responses adopted in this study are shown in Table I. The criteria used to identify organ dysfunctions are provided in Appendix I-A or the case of vasopressor administration. These criteria were developed by synthesizing established Sepsis-3 guidelines [20] and subject matter experts'

TABLE I

ORGAN SYSTEMS AND THEIR ASSOCIATED CELLULAR AND PHYSIOLOGICAL RESPONSES AND INFLAMMATORY BIOMARKERS CONSIDERED IN THIS STUDY

Organ Dysfunction	Response	Abbreviation
Cardiovascular	Systolic blood pressure (SBP)	Sb
	SBP _{max} * - Systolic BP	Sd
	Mean arterial pressure (MAP)	Mp
Renal	Creatinine	Cr
	(Creatinine - C _{base} **)/(C _{base})	Cd
	Blood Urea Nitrogen (BUN)	Bu
Hematopoietic	WBC	Wb
	Platelet	Pl
Metabolic	Lactate	La
Gastrointestinal	Bilirubin	Bi
Respiratory	Fraction of inspired oxygen (FiO ₂)	Fi
	Pulse oximetry (SpO ₂)	Px
	SpO ₂ /FiO ₂	Or
	Oxygen (O ₂) Source	Os
Central Nervous	Glasgow Comma Score	Gc
	Glasgow Best Verbal Response	Gv
Biomarkers	Procalcitonin	pc
	C-Reactive Protein	cr
	Erythrocyte Sedimentation Rate	sr

*: Maximum SBP for each observation within 8-hour windows.

**: Initial creatinine value observed in each visit.

TABLE II

NUMBER OF VISITS AND PATIENTS IN EACH SUBPOPULATION

Population	Visits Count	Patients Count
Survivor	13,381	9,997
Non-survivor	2,946	2,946
Total:	16,327	12,223

TABLE III

A SUMMARY OF PATIENTS CHARACTERISTICS IN EACH SUBPOPULATION

	Sepsis Survivors	Sepsis Non-survivors
Gender		
Female	4,923	1,392
Male	4,497	1,411
Age in years (mean±SD)	64.3±17.1	73.5±14.4
Number of comorbidities (<i>n</i>) distribution		
<i>n</i> < 5	44%	29%
5 ≤ <i>n</i> < 10	43%	52%
<i>n</i> ≥ 10	11%	19%
unknown	2%	0%

input. Death was defined as any in-hospital death or discharge to hospice care. Based on these definitions, the sepsis patients were categorized into two subpopulations; sepsis survivors and sepsis non-survivors. If the patient is discharged to hospice or died while receiving an anti-infective, but before four days of anti-infective administration elapsed, the patient was also considered a sepsis non-survivor [4]. If a patient died during the study, just the last visit was considered in the non-survivor subpopulation for patients with multiple visits. Previous visits were categorized in the survivors subpopulation. The number of visits and patients included in each subpopulation are shown in Table II, and a summary of characteristics of the two subpopulations are provided in Table III.

C. Significant Proximate Failures

We consider two types of proximate cellular and physiological response failures, significant and non-significant. The non-significant failures are common in both survivors and non-survivors and cannot discriminate between the two subpopulations. We created two networks to identify the significant failures, one for each subpopulation. Each network is composed of cellular and physiological responses as nodes (column “Response” in Table I). Two nodes (c_i and c_j) are connected if their failures are proximate; i.e., their co-failures are observed in some pre-defined window in patients. To quantify the proximity of each pair of cellular and physiological responses, we use the following weight formulation:

$$wt_{ij} = \sqrt{P(c_i|c_j) \times P(c_j|c_i)} \quad (1)$$

$$= \sqrt{\frac{c_{ij}^{112}}{(c_{ij}^{11} + c_{ij}^{10}) \times (c_{ij}^{11} + c_{ij}^{01})}}$$

where, wt_{ij} represents the weight of the edge connecting nodes c_i and c_j , $P(c_i|c_j)$ is the probability of observing failure of response c_i given the failure of response c_j , and c_{ij}^{kl} (with $k, l \in \{0, 1\}$, 0: no failure 1: failure) shows the frequency of nodes c_i and c_j co-failure ($k = l = 1$) or failure of one without the other failing ($k \neq l$).

To quantify the proximity, we consider ten windows (aggregation times) from 30 to 300 minutes representing the predefined window where failures co-occur. For each aggregation time, α , we iterate over all the observations of each patient’s visit and consider α minutes before and after each observation. In other words, for each specific observation recorded at time O_t , all the failures recorded in the window of $[O_t - \alpha, O_t + \alpha]$ are assumed to be proximate. For each response pairs, this weight is separately computed in each subpopulation. For each response pair (represented by c_i and c_j in the network) that appeared simultaneously at least once in the patient’s data set, we add a weighted edge between nodes c_i and c_j with the corresponding weight of wt_{ij} . Based on the two subpopulations’ networks, we derive a third network, the *difference network*, composed of the same set of nodes and edges as the two previous networks. However, each edge’s weight in the difference network is computed as the difference between the same edge weights in the survivors and non-survivors networks. The edges in the difference network are used for predicting in-hospital mortality. We consider multiple thresholds to remove edges with weights less than the threshold to identify the significant proximate failures of responses. To accomplish that, we adopt a classification approach. In addition, because the limiting number of visits was from non-survivors, we randomly sample the same number of visits from the survivors subpopulation.

D. Model Development

To investigate the importance of proximate failures relative to other feature sets used commonly in the literature, we consider four sets of features; proximate failures of multiple cellular and physiological responses, individual failures of cellular and

Edges in difference network	0
Single cellular & physiological responses	1
Comorbidities	0
Age & Gender	1

→ 0101

Fig. 1. Model “0101” composed of features related to single failures of cellular and physiological responses and age and gender to predict patients’ outcome. The proximity of failures of cellular and physiological responses (first place holder) and comorbidities (third place holder) of patients are not included in this model and therefore are shown with a “0” in the model abbreviation.

physiological responses, age and gender, and comorbidities of patients. The comorbidities are defined as medical conditions recorded in the medical history of patients. The list of comorbidities considered in this study is provided in Appendix I-B.

To be able to compare the predictive power of these four feature sets in terms of patient in-hospital mortality, we developed 15 novel models, shown with a sequence of four binary digits. Each digit corresponds to absence (“0”) or presence (“1”) of each set of features. The individual failures of cellular and physiological responses feature set is composed of 19 variables (Table I). The maximum number of features represented by proximate failures is $\binom{19}{2} = 171$ (the maximum number of edges in the difference network). The age and gender are represented by two variables, and 30 comorbidities are considered in this study (Appendix I-B). Different combinations of these feature sets are used for predictive modeling. For example, Fig. 1 shows a model used to predict patients’ outcomes based on the single failures of cellular and physiological responses and the age and gender of patients.

E. Prediction of Patients’ Outcome

The outcome of interest in this study is the survival of patients. We applied multiple machine learning algorithms to compare the performance of models created based on different sets of features, including random forests, logistic regression, C5.0, and support vector machine (SVM). We used the caret implementation of these algorithms [21]. The random forests and SVM had slightly higher accuracy among these four algorithms, and we adopted the SVM for further analysis. For SVM, classifiers with polynomial kernel [22], [23] are used. The input data is composed of 70% of data for training purposes and a 30% untouched subset for testing purposes. In addition, we adopted 10-fold cross-validation for training and internal validation. The adopted algorithm was used to predict patients’ in-hospital mortality based on the available information.

The four feature sets described in the previous section were used to create 15 models (Fig. 2). Considering that adopting different aggregation times and thresholds can impact the list of proximate failures derived from difference network, we implemented 407 different classifications. These classifications include eight models composed of the proximate failures with or without other feature sets; each is implemented for five support values and ten aggregation times (in total, 400 models). The other seven models do not depend on the aggregation times. These

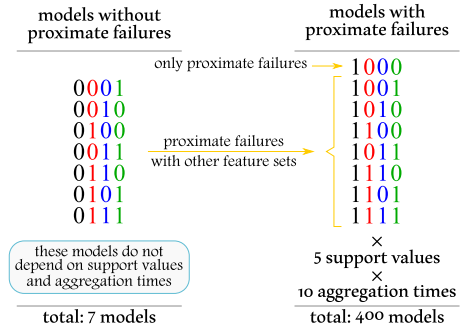


Fig. 2. The 15 models considered in this study composed of different combinations of feature sets. The seven models on the left do not depend on support values and aggregation times. Models including proximate failures of multiple cellular and physiological responses as one of their feature sets are shown on the right.

models are composed of different combinations of individual failures, age and gender, and patients’ comorbidities.

III. RESULTS

The results section provides information about the performance of models created for different thresholds at different aggregation times for varying levels of information (various combinations of feature sets). In the following, the results of the development, analysis, and classification models are described.

A. Network Development

To identify proximate failures, we created a weighted network for each subpopulation using the individual cellular and physiological responses as nodes and their co-occurrences as edges, as described in Section II.C. In these networks, the edges are weighted, representing the probability of observing proximate failures of physiological and cellular responses. Statistical analysis of weights of survivors and non-survivors networks found significant differences in the distributions of the weights of edges in the two networks (with $p - value < 0.001$). The statistical analysis included the Levene’s test for homogeneity of variances of weights, Shapiro-Wild test for normality test of weights distribution, and Kruskal-Wallis and Wilcoxon rank-sum tests for evaluating whether weights are from an identical distribution. From these two networks, a third weighted network was derived, in which the weights are the difference between the weights of the corresponding edges in the first two networks. We visualized the networks’ densities over the weight range, [0,1], to show the distribution of weights in these networks. In this context, the weight threshold served as a cutoff point such that in a given difference network, only the edges where the frequency of their corresponding responses simultaneously exceeded the selected threshold are recorded. The network’s density is computed as the ratio of the number of edges in the network divided by the total number of edges possible. Fig. 3 shows the network density and boxplot of the weights of the three networks. For any given weight threshold, the non-survivors’ network’s density is more than the density of the survivors’ network. It means that for any weight threshold, there are more edges in the non-survivors than survivors network. In other words, it is more probable to

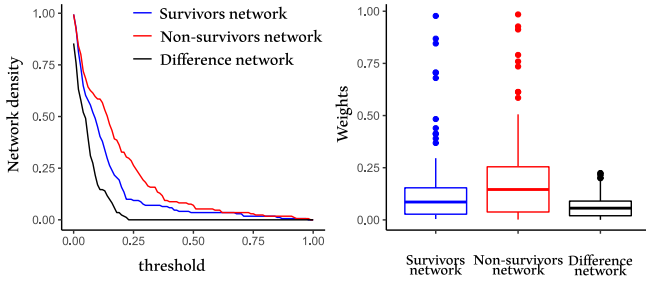


Fig. 3. Density of survivors, non-survivors, and difference networks at different weight thresholds.

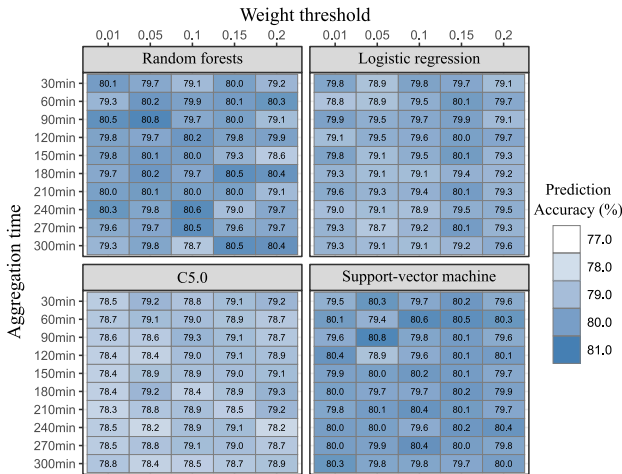


Fig. 4. The results of four different machine learning algorithms, random forests (RF), logistic regression, C5.0, and support vector machine (SVM), implemented on the model created from all the features (model “1111”) at different aggregation times. All the other aspects, such as train-test split and cross-validation, are identical among different implementations. The SVM and RF produced the highest, and C5.0 produced the lowest accuracy values.

observe proximate failures of responses in non-survivors than survivors. Considering that the two original networks’ weights are statistically different, we used the edges in the difference network for predictive modeling.

B. Predictive Modeling

As discussed in Section II.E, the SVM algorithm was adopted (Fig. 4) for comparing different sets of features represented by different models. The 15 models make it possible to compare the value of different patient information categories to predict their outcome. Among the 15 models, there are seven models that do not depend on the weight thresholds and aggregation times “0001,” “0010,” “0100,” “0011,” “0110,” “0101,” and “0111”. These models are composed of different combinations of individual failures of cellular and physiological responses, medical history of patients, and their age and gender. The results of the predictive modeling for these seven networks are shown in Table IV.

The other eight models are composed of the proximate failures of cellular and physiological responses as one of their feature

TABLE IV

ACCURACY OF PATIENTS’ OUTCOME PREDICTION IN MODELS COMPOSED OF DIFFERENT COMBINATIONS OF INDIVIDUAL FAILURES OF CELLULAR AND PHYSIOLOGICAL RESPONSES, MEDICAL HISTORY OF PATIENTS, AND THEIR AGE AND GENDER

Model	Prediction accuracy (%)
0001	60.1
0010	67.0
0100	72.6
0011	67.4
0101	73.5
0110	75.3
0111	77.6

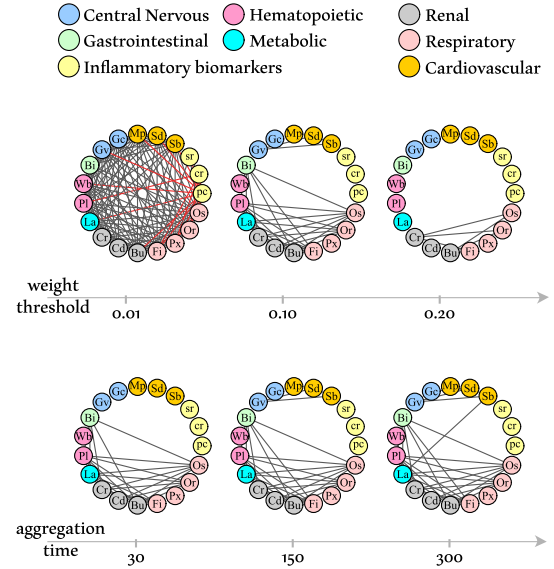


Fig. 5. Difference network at different weight thresholds (for aggregation time of 150 minutes) and different aggregation times (for weight threshold of 0.10). (Sb (systolic blood pressure); Sd (Sb (systolic blood pressure) - SBP); Mp (MAP); Cr (Creatinine); Cd ((Creatinine - C_base**)/(C_base)); Wb (WBC); La (Lactate); PI (Platelet); Bu (BUN); Bi (Bilirubin); Fi (FiO₂); Px (SpO₂); Or (SpO₂/FiO₂); Os (Oxygen Source); Gc (Glasgow Coma Score); Gv (Glasgow Best Verbal Response); pc (Procalcitonin); cr (CRP); sr (Sedimentation Rate). *: Maximum systolic blood pressure for each observation within 8-hour windows. **: Initial creatinine value observed in each visit).

sets. In general, the probability of having proximate failures of two responses increases with aggregation times. Two cellular and physiological responses failing at two different time points might be considered proximate if the aggregation window is wide enough.

The edge weights in the difference network show the difference in the magnitude of failure proximities in the survivors and non-survivors networks. These differences barely exceed 0.25 (Fig. 3). Although the maximum number of edges is 171, when we increase the threshold, the number of edges with weights more than the threshold decreases. The impacts of aggregation times and changes in thresholds are shown in Fig. 5. In this figure, the changes in the number of edges in the difference network are shown for three aggregation times and three threshold values. In this study, we performed predictive modeling for ten aggregation times and five threshold values. We considered eight different

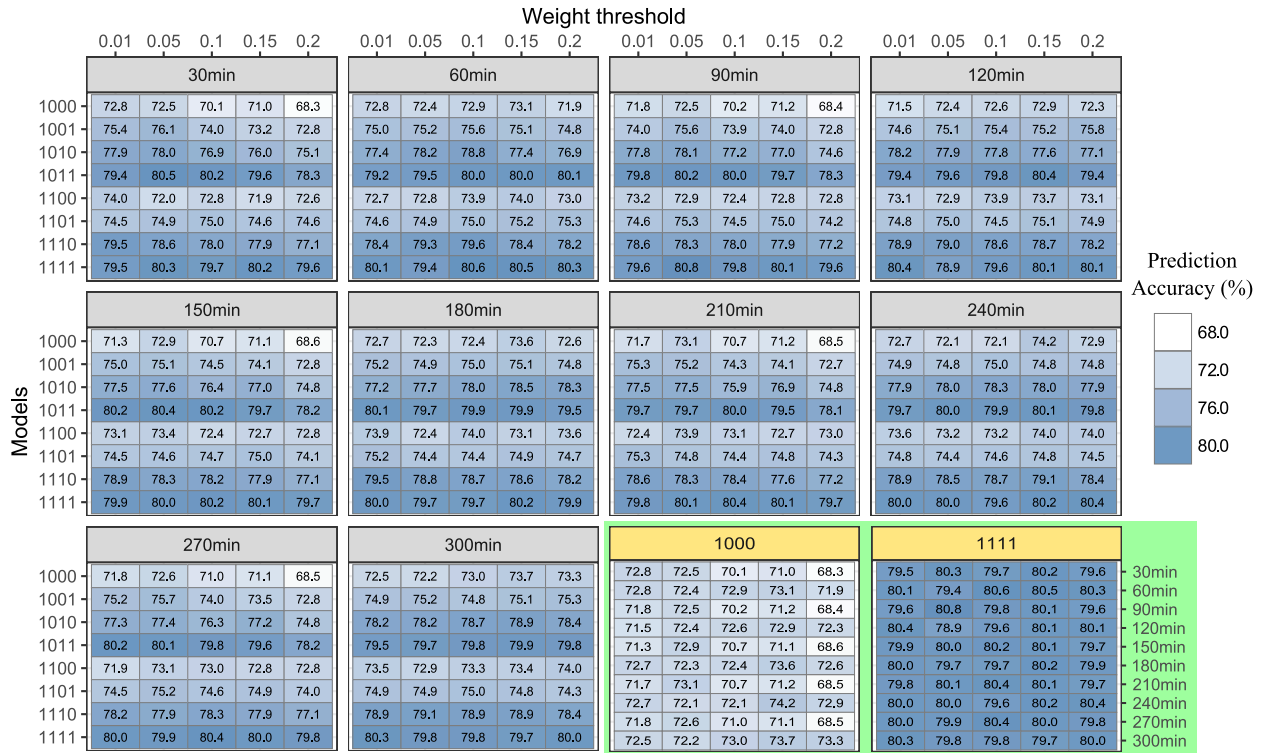


Fig. 6. Prediction accuracy of eight models composed of proximate failures of cellular and physiological responses as one of their feature sets at different aggregation times and weight thresholds. The heatmaps with grey header represent the prediction accuracies of eight models created with varying thresholds of weight. The two dimensions of heatmaps with yellow header are aggregation time and weight threshold shown for two “1000” and “1111” models.

combinations of feature sets by ten aggregation times by five threshold values creating 400 models. We visualized the results of prediction for these combinations of feature sets in Fig. 6.

IV. DISCUSSION

Seven of the 15 models considered in this study were composed of different combinations of individual cellular and physiological responses, age and gender, and patients’ co-morbidities. These models did not consider the proximity of responses. Table IV shows the prediction accuracy for these seven models and the relative importance of each feature set’s information value. The model “0001” has the lowest accuracy. In this model, the age and gender of patients are used for predicting patient outcomes. This model’s accuracy (60.1%) shows that we can perform better than random guessing if the only information we have from sepsis patients is their age and gender. The model with only the history of comorbidities of patients (model “0010”) performs better (with accuracy 67.0%) than the model composed of age and gender. However, their combination does not change the accuracy compared with when we only know the patients’ comorbidities. Using only cellular and physiological information of patients, we could obtain higher prediction accuracy (72.6%) than any combination of age and gender and comorbidities. Complementing patients’ current status with their age, gender, and comorbidities (model “0111”) provides the highest accuracy (77.6%) among these seven models. Therefore, we

conclude that patients’ medical histories and clinical status are critical to achieving higher accuracy in predicting in-hospital mortality.

The other eight models are created based on the concept of the proximity of responses. We considered two response failures proximate when the failures co-occur in a set of various pre-defined time windows. The results of the implementation of these eight models (with varying aggregation times and weight thresholds) are shown in Fig. 6. In general, model “1000,” the first row of the ten heatmaps with grey headers in Fig. 6, has the lowest performance. We summarize this model’s accuracy at different aggregation times and weight thresholds in one heatmap (model “1000” with a yellow header in Fig. 6). The maximum accuracy that we could obtain using only the information provided by proximate failures of cellular and physiological responses was 73.7%. This accuracy is slightly higher than the accuracy we obtained using only individual cellular and physiological response failures. The general trend observed in these eight models shows that starting from a base model (“1000” or “1100”), the addition of age and gender (“1001” or “1101”) can improve the prediction accuracy slightly (by about 1-3%). However, the same base model complemented with patients’ comorbidities (in “1010” or “1110”) can improve the base model’s accuracy by about 6%. Simultaneous consideration of the age, gender and comorbidities feature sets results in the highest prediction accuracy in models “1011” and “1111”. These patterns are shown in Fig. 7 in which the mean and standard

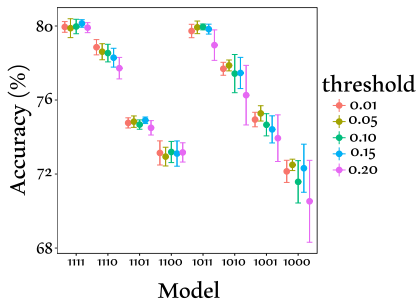


Fig. 7. Prediction accuracy of models composed of proximate failures of cellular and physiological responses as one of their feature sets. Each range includes mean $\pm$ standard deviation of prediction accuracies for the same threshold over different aggregation times.

deviation of accuracies are computed for different aggregation times of the same model and same weight threshold.

The impact of including proximate or individual failures in the prediction process can be further examined by comparing models “1011” and “0111”. The former has proximate failures without individual failures, and the latter includes individual failures without proximate failures. The accuracy obtained from “0111” is about 3% lower than the best performance obtained from “1011”.

Among all the models considered, the highest accuracy was observed in the model “1111” for an aggregation time of 90 minutes and a weight threshold of 0.05. In our previous work [24], we obtained 150 minutes as the best aggregation time. This difference in aggregation times can be attributed to the level of information available in each approach. In the current study, the model “1111” includes information regarding age, gender, comorbidities, and individual cellular and physiological responses, which was not considered in our previous work. The addition of this information reduces aggregation times. In other words, having more information about patients, such as their demographic or past medical histories, can reduce the period for which the patterns of cellular and physiological responses should be aggregated.

Using an appropriate machine learning technique that can provide a prioritized list of features based on their contribution to prediction, we could create a hierarchy of important failures and organ dysfunctions and their co-occurrences. Fig. 8 is derived from the first 20 most important features contributing to the prediction performance. Except for age and gender, the other 19 features are either individual or proximate failures of cellular and physiological responses. In Fig. 8, the important individual responses are shown with red borders. Based on this figure, the proximate failures of the responses associated with renal, respiratory, and metabolic organ systems in 90-minute periods are good predictors of patient outcomes. This representation of results, conducting the analysis at more granular levels, and reporting the derived insights at the aggregated levels, could enhance the interpretation of results and facilitate the translation of findings to the clinical practices.

In this paper, we used a threshold-based approach for selecting proximate failures from the difference network. Instead, we can adopt a weight-based approach where the edge weights of the

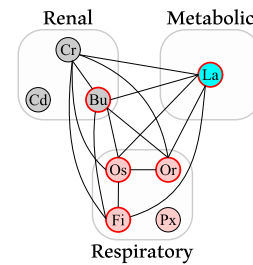


Fig. 8. Features with the highest relative importance in model “1111” at aggregation time 90 minutes and threshold 0.05. The cellular and physiological responses found to be individually important are shown with a red border. The important proximate failures are edges connecting the cellular and physiological response. (Cr (Creatinine - C_base)/(C_base); Cd ((Creatinine - C_base)/(C_base)); La (Lactate); Os (Oxygen Source); Or (SpO₂/FiO₂); Fi (FiO₂); Px (SpO₂)).

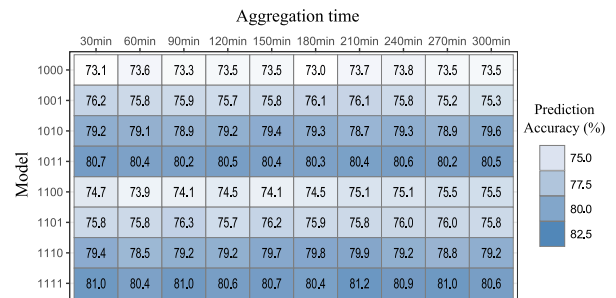


Fig. 9. Prediction accuracy of models comprises proximate failures of cellular and physiological responses as one of their feature sets at different aggregation times. In these models, instead of a subset of features obtained using different thresholds, all the edges with their associated weights are incorporated in the predictive modeling.

difference network are directly incorporated in the predictive modeling. The results for weight-based predictive modeling for different models and aggregation times are provided in Fig. 9. The weight-based approach results are slightly better than the threshold-based (about 1–2% for different models and about 1% for models with the highest accuracy). However, the same patterns among different models observed in Fig. 6 and among models with proximate failures of cellular and physiological responses and baseline models (Table IV) was observed in the weight-based approach.

V. LIMITATIONS AND FUTURE DIRECTIONS

One of the primary objectives of this study was to compare the relative information value provided by different sets of features: proximate and individual cellular and physiological responses, age and gender, and comorbidities of patients. We adopted the well-known SVM technique for predictive modeling to focus on this objective and prevent technical complexities associated with more advanced machine learning approaches. However, using other more advanced machine learning techniques could improve prediction accuracies. By adopting SVM, we were able to prioritize features based on their relative importance. Consequently, we could present the results at different granularity levels and increased the *interpretability* of the findings.

There are several promising future research directions that can address the limitations of this study. In this study, we used the data from a single healthcare system. However, to improve the generalizability of the findings, this approach should be applied to data from other health centers. Further, we used the sepsis definition provided by [1] along with the organ dysfunction thresholds provided by Sepsis-related Organ Failure Assessment (SOFA) [25], PIRO (predisposition, insult, response, organ dysfunction) [26] scores and clinical expertise. Comparing the model results using different sepsis definitions is another possible future direction. We used edges to represent proximate failures, and the edges in the difference network are composed of pairs of failures. In our future studies, we plan to adopt an association rule mining approach to mine frequent and significant failures of more than two responses. Another approach is to mine larger temporal and structural failures over time.

Furthermore, in this study, we focused on the simultaneous failures of cellular and physiological responses and biomarkers. The dynamics of these responses and biomarkers are affected by the clinical and pharmacological interventions. The significance of these types of interventions has been discussed in previous studies [19]. An avenue for future research would be consideration of these interventions and their stimulation roles in the dynamics of responses and biomarkers and study their impacts on the prediction performance.

APPENDIX I

A. Organ Systems' Components

See Table V

B. Co-Morbidities

TABLE VI
CO-MORBIDITIES CONSIDERED IN THE STUDY

Co-morbidity	Description: History of diagnosis with...
AIDS	AIDS
ALCOHOL	Alcohol Abuse
ANEMDEF	Anemic disorders
ARTH	Arrhythmias
BLDLOSS	Blood loss
CHF	Congestive heart failure (CHF)
CAD	Coronary artery disease (CAD)
CHRNUNG	Chronic pulmonary disease (COPD)
COAG	Coagulation disorders
DEPRESS	Depression
DM	Diabetes (without complications)
DMCX	Diabetes (with complications)
DRUG	Illegal drug use
HTN	Hypertension
HYPOTHY	Hypothyroidism
LIVER	Liver disease
LYMPH	Lymph disorders
LYTES	Electrolyte disorders
METS	Malignancy metastases
NEURO	Neurologic disorders
OBESE	Obesity prior to this visit.
PARA	Paralysis prior to this visit.
PERIVASC	Peripheral vascular disease
PSYCH	Psychological disorders
PULMCIRC	Pulmonary circulation disorders
RENLFAIL	Renal failure
TUMOR	Malignancy
ULCER	Ulcers prior to this visit.
VALVE	Valve disorders
WGHTLOSS	Significant weight-loss

TABLE V

CELLULAR AND PHYSIOLOGICAL RESPONSES AND BIOMARKERS RECORDED AS INDIVIDUAL FEATURES, AND THE CRITERIA RESULTING IN THE CORRESPONDING ORGAN DYSFUNCTION

Organ Dysfunction	Response	Abbreviation	Failure criteria
Cardiovascular	Systolic blood pressure (SBP)	Sb	<90 mmHg
	SBP _{max} * - Systolic BP	Sd	>40 mmHg within an 8-hour period
	Mean arterial pressure (MAP)	Mp	<65 mmHg
Renal	Creatinine	Cr	>1.2 mg/dL
	(Creatinine - C _{base} **)/(C _{base})	Cd	>50% from initial creatinine
	Blood Urea Nitrogen (BUN)	Bu	>20 mg/dL
Hematopoietic	WBC	Wb	<4,000 cells/mL
	Platelet	Pl	<100,000 cells/mL
Metabolic	Lactate	La	>2.0 mmol/L
Gastrointestinal	Bilirubin	Bi	>2 mg/dL
Respiratory	Fraction of inspired oxygen (FiO ₂)	Fi	>21%
	Pulse oximetry (SpO ₂)	Px	<90%
	SpO ₂ /FiO ₂	Or	<421
	Oxygen (O ₂) Source	Os	Mechanical ventilation required (bilevel positive airway pressure (BiPAP) or continuous positive airway pressure (CPAP) or ventilator)
Central Nervous	Glasgow Comma Score	Gc	<14
	Glasgow Best Verbal Response	Gv	<5
Biomarkers	Procalcitonin	pc	>0.15 ng/mL
	C-Reactive Protein	cr	>8 mg/L
	Erythrocyte Sedimentation Rate	sr	>20 mm/hr

*: Maximum systolic blood pressure for each observation within 8-hour windows.

**: Initial creatinine value observed in each visit.

ACKNOWLEDGMENT

The author would like to acknowledge that this work has been performed on behalf of the Sepsis Early Prediction Support Implementation System Collaborative. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

- [1] A. Rhodes *et al.*, "Surviving sepsis campaign: International guidelines for management of sepsis and septic shock: 2016," *Intensive Care Med.*, vol. 43, no. 3, pp. 304–377, 2017.
- [2] M. Cecconi, L. Evans, M. Levy, and A. Rhodes, "Sepsis and septic shock," *Lancet*, vol. 392, no. 10141, pp. 75–87, 2018.
- [3] C. Fleischmann *et al.*, "Assessment of global incidence and mortality of hospital-treated sepsis current estimates and limitations," *Amer. J. Respir. Crit. Care Med.*, vol. 193, no. 3, pp. 259–272, 2016.
- [4] C. Rhee *et al.*, "Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009–2014," *JAMA*, vol. 318, no. 13, pp. 1241–1249, Oct. 2017.
- [5] A. Kumar *et al.*, "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," *Crit. Care Med.*, vol. 34, no. 6, pp. 1589–1596, Jun. 2006.
- [6] E. Rivers *et al.*, "Early goal-directed therapy in the treatment of severe sepsis and septic shock," *New England J. Med.*, vol. 345, no. 19, pp. 1368–1377, 2001.
- [7] A. Javed *et al.*, "Clinical predictors of early death from sepsis," *J. Crit. Care*, vol. 42, pp. 30–34, 2017.
- [8] J. S. Calvert *et al.*, "A computational approach to early sepsis detection," *Comput. Biol. Med.*, vol. 74, pp. 69–73, 2016.
- [9] F. F. Larsen and J. A. Petersen, "Novel biomarkers for sepsis: A narrative review," *Eur. J. Intern. Med.*, vol. 45, pp. 46–50, 2017.
- [10] C. Henriquez-Camacho and J. Losa, "Biomarkers for sepsis," *BioMed Res. Int.*, vol. 2014, 2014.
- [11] Z. Obermeyer and E. J. Emanuel, "Predicting the future big data, machine learning, and clinical medicine," *New England J. Med.*, vol. 375, no. 13, pp. 1216–1219, 2016.
- [12] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: Using analytics to identify and manage high-risk and high-cost patients," *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014.
- [13] E. Gultepe, J. P. Green, H. Nguyen, J. Adams, T. Albertson, and I. Tagkopoulos, "From vital signs to clinical outcomes for patients with sepsis: A machine learning basis for a clinical decision support system," *J. Amer. Med. Informat. Assoc.*, vol. 21, no. 2, pp. 315–325, 2014.
- [14] A. Tsoukalas, T. Albertson, and I. Tagkopoulos, "From data to optimal decision making: A data-driven, probabilistic machine learning approach to decision support for patients with sepsis," *JMIR Med. Informat.*, vol. 3, no. 1, p. e11, Feb 2015.
- [15] R. A. Taylor *et al.*, "Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data driven, machine learning approach," *Academic Emerg. Med.*, vol. 23, no. 3, pp. 269–278, 2016.
- [16] S. Horng, D. A. Sontag, Y. Halpern, Y. Jernite, N. I. Shapiro, and L. A. Nathanson, "Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning," *PLoS One*, vol. 12, no. 4, pp. 1–16, Apr. 2017.
- [17] D. W. Shimabukuro, C. W. Barton, M. D. Feldman, S. J. Mataraso, and R. Das, "Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: A randomised clinical trial," *BMJ Open Respir. Res.*, vol. 4, no. 1, 2017.
- [18] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman, "An interpretable machine learning model for accurate prediction of sepsis in the ICU," *Crit. Care Med.*, vol. 46, no. 4, p. 547, 2018.
- [19] A. Porta *et al.*, "Association between autonomic control indexes and mortality in subjects admitted to intensive care unit," *Sci. Rep.*, vol. 8, no. 1, pp. 1–13, 2018.
- [20] M. Singer *et al.*, "The third international consensus definitions sepsis septic shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, Feb. 2016.
- [21] M. Kuhn, "Building predictive models in R using the caret package," *J. Stat. Software, Articles*, vol. 28, no. 5, pp. 1–26, 2008.
- [22] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, 1999.
- [23] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab-an S4 package for Kernel methods in R," *J. Stat. Softw.*, vol. 11, no. 9, pp. 1–20, 2004.
- [24] A. Jazayeri, M. Capan, C. Yang, F. Khoshnevisan, M. Chi, and R. Arnold, "Network-based modeling of sepsis: Quantification and evaluation of simultaneity of organ dysfunctions," in *Proc. 10th ACM Int. Conf. Bioinformatics, Comput. Biol. Health Informat., Ser. BCB '19*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 87–96.
- [25] J.-L. Vincent *et al.*, "The SOFA (Sepsis-Related Organ Failure Assessment) score to describe organ dysfunction/failure," pp. 707–710, 1996.
- [26] F. Rubulotta, J. C. Marshall, G. Ramsay, D. Nelson, M. Levy, and M. Williams, "Predisposition, insult/infection, response, and organ dysfunction: A new model for staging severe sepsis," *Crit. Care Med.*, vol. 37, no. 4, pp. 1329–1335, 2009.