Sepsis Prediction and Vital Signs Ranking in Intensive Care Unit Patients

Avijit Mitra[1], Khalid Ashraf[2]
[1]Semion, Mohakhali DOHS, Dhaka 1206, Bangladesh.
[2]Semion, Palo Alto, CA 94303, USA.
avipartho@gmail.com, khalid@semion.ai

Abstract

***Background**: Reliable prediction of sepsis in ICU remains a challenge. Machine learning(ML) based methods have performed better than rule-based models in recent studies. However, high accuracy, generalization and interpretability still remains a problem. Accurate and clinically explainable ML models will find application in hospital deployment and help reduce mortality in ICU patients.*

***Objectives**: The objective is to compare the performance of different single and ensemble ML model configurations with the rule-based scoring systems - quick sequential organ failure assessment (qSOFA), modified early warning score (MEWS), and sequential organ failure assessment (SOFA). We also rank features to find out which vital signs are most correlated with the onset of sepsis.*

***Materials and Methods**: In this study, we have compared several ML and rule-based models for detection and prediction of three categories of sepsis in ICU patients. We have designed an ensemble model using the single ML models and ranked the features for relative importance determination. We have used the retrospective Medical Information Mart for Intensive Care (MIMIC)-III dataset, restricted to intensive care unit (ICU) patients aged 18 years or more. Features for prediction were created from only six vital sign measurements and their changes over time.*

***Results**: We show significant improvement of area under the receiver operating characteristic curve (AUC) using neural network(NN)-based ensemble model compared to single ML and rule-based models. For detection of sepsis, severe sepsis and septic shock, our model achieves an AUC of 0.94, 0.91 and 0.89, respectively. Four hours before onset, it predicts the same three categories with an AUC of 0.80, 0.81 and and 0.84 respectively. Further, we find that using six vital signs consistently provides higher detection and prediction AUC for all the models tested.*

***Conclusions**: Our novel ensemble model outperforms existing severity scoring systems in detecting and predicting sepsis, severe sepsis, and septic shock, and is amenable to deployment in hospital settings.*

1. Introduction

Sepsis, a common systemic response to infection, is one of the leading causes of death in the United States [1]. Each year, approximately 750,000 hospitalized patients are diagnosed with severe sepsis in the United States alone and the mortality rate may go up to one-third of this population [2,3]. The treatment cost is very high, $20.3 billion dollars annually in USA hospitals [4]. Compared to any other condition, this is, on average, almost twice the cost [5]. Moreover, the occurence of severe sepsis is increasing by an approximate 13% per year [6]. With the progress of sepsis, organ failure and eventually death becomes evident. Previous studies have showed that through early diagnosis and treatment, it's possible to reduce mortality as well as the related medical expenditures [7–9].

According to the Sepsis-3 definition [10], sepsis is "life-threatening organ dysfunction caused by a dysregulated host response to infection ".Traditionally, a person is diagnosed with sepsis if he or she demonstrates two or more Systemic Inflammatory Response Syndrome (SIRS) [11] criteria with a suspected infection. This turns into severe sepsis in the event of organ dysfunction and finally, septic shock in case of refractory hypotension [11].

Rule-based disease severity scoring systems such as SOFA [12], qSOFA [10, 13], MEWS [14] etc. are often used in hospitals for quantitative definition of onset of sepsis but these scores lack credibility in sepsis diagnosis. With the increase in publicly available Electronic Health Records (EHR), it is possible to design an efficient and robust system to not only have an accurate detection but also a reliable prediction of sepsis.

In this paper, we study the performance of various ML models on the publicly available MIMIC-III dataset for detection and prediction of three categories of sepsis. We explore ML and DL architectures that perform well for sepsis detection and prediction and compare their performance against standard rule-based models. Finally, we rank fea-

tures to understand how the vital signs are correlated with onset of sepsis and what combination of vitals provides the best detection and prediction performance. Our contributions can be summarized as follows:

- We provide the first benchmark of three categories of sepsis detection and prediction on the MIMIC-III dataset with different rule-based and ML methods.
- We demonstrate a deep learning based ensemble method that achieves the highest AUC compared to the single models and other ensemble models.
- We show the highest improvement of AUC over the rule based methods in the literature.
- We provide a feature ranking for identifying which features are most important for sepsis prediction.
- We show that there is a universal improvement in AUC when six vital signs are used for both detection and prediction tasks.

The paper is organized as follows: in section 2, we review the relevant research in the literature, in section 3, we describe the details of the experiment from data preprocessing to network design. In section 4, we show the main results in the paper and in section 5, we discuss our results in comparison to other works done in the literature. Finally in sections 6 and 7 we discuss the limitations, future extensions and concluding remarks.

2. Related Works

Use of EHR, laboratory results, and biomedical signals to track patients' sepsis progression from one stage to another to prevent fatal injury and death in the intensive care unit is a common approch [15–20]. Some studies focused on a viable way of calculating the mortality rate of sepsis patients [21–27] while Systemic Inflammatory Response Syndrome (SIRS) criteria [28–32] or high frequency heart rate variability [33, 34] was used by others to predict sepsis by analyzing before and after-onset symptoms.

In [28], blood analysis by RT-PCR expression and neural network analysis of related genes were performed to predict sepsis onset for 92 ICU patients. This study managed to predict 83.09% of cases 1 to 4 days prior to the clinical diagnosis with a sensitivity and specificity of 91.43% and 80.20% respectively. In another study, sepsis onset was predicted 2 to 3 days prior to the diagnosis through cell motion analysis using microfluidic devices [30]. [31] used lab tests, biomedical signals and SIRS scores to create a support vector machine (SVM) model that predicts the onset 0-24 h before diagnosis in 1,239 postoperative patients of which only 26 patients (2.1%) had sepsis, indicating a huge data imbalance that even 100 bootstraps could not address properly. The AUC ranged between 0.28 and 0.95 and the authors didn't report the accuracy of each group. These processes of sepsis detection i.e. using expensive medical equipment, acquiring daily blood sample, lab test results or performing gene analysis are not practical for regular usage.

[32] proposed a machine learning model with gradient tree boosting for 3 hours early prediction of sepsis, called InSight. This model takes nine items extracted from patient information, laboratory test results, and widely used vitals. They used 1,394 patients from a medical intensive care unit (MICU) of which 159 patients had sepsis. Their reported sensitivity, specificity, and AUC are 0.90, 0.81, and 0.83 respctively. [24] and [35] also used the same Insight model for severe sepsis detection and got an AUC of 0.89 at onset and 0.75, 4 hours prior to the onset. All these works either used MIMIC-III or MIMIC-II dataset. [36] validated the InSight primarily on a mixed-ward retrospective data set from the University of California, San Francisco (UCSF) Medical Center (San Francisco, CA) for detection and prediction of three sepsis related gold standards and got an AUC of .92 and .87 for detection of sepsis and severe sepsis respectively. This time InSight used six clinical items. [37] used deep learning models to make early sepsis prediction system and verification of its feature extraction capacity. The best result they got was an AUC of .929, using a variant of LSTM. They followed the feature extraction steps of [32].

3. Experiments

3.1. Dataset

This work uses the Medical Information Mart for Intensive Care (MIMIC)-III version 1.4 dataset [38], compiled from the patients admitted to the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA between 2001 and 2012. The MIMIC-III set includes comprehensive clinical data such as vital signs, medications, laboratory measurements, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data etc. from over 53,423 ICU stays for more than 40,000 patients. Two different critical care information systems CareVue (Philips) and Metavision (iMDSoft) were used for data entry, which handle and store some information differently. These systems were in place from 2001 to 2008 and 2008 to 2012, respectively. We used only the EHR-entered components of the MIMIC-III dataset, without any real-time waveform data or free text notes. Since the original MIMIC-III data collection did not impact clinical care and all data were deidentified in accordance with the Health Insurance Portability and Accountability Act (HIPAA) standards, the requirement for individual patient consent was waived by the Institutional Review Boards of BIDMC and the Massachusetts Institute of Technology.

3.2. Data Extraction and Imputation

The data were provided in the form of comma separated value (CSV) files and stored in a PostgreSQL [39] database. All the necessary CSV files were downloaded following the instructions mentioned in [40]. Several python [41] scripts were written to extract measurements and patient outcomes of interest cross-matching relevant CSV files. For each patient, all the measurements were binned by hour. For patient without any measurement in a given hour, the missing measurement was filled in using carry-forward imputation, that is applying the patient's last measured value to the following hour. When the patient didn't have any measurement prior to the missing hour, it was filled with the next available measurement. In the case of multiple measurements within an hour, the mean was used. After the data were processed, they were used to train, test and compare several machine learning classifiers to predict sepsis at onset and 4 hours prior to onset.

3.3. Gold Standards

For sepsis, severe sepsis and septic shock, we followed the gold standard and onset time definitions as mentioned in [36] to create our sepis, severe sepsis and septic shock data set for training and testing purposes. Some conditions such as organ dysfunction, chronic dialysis, pneumonia and kidney injury were ignored for severe sepsis to avoid complexity.

3.4. Comparators

We compared our best model's performance for each gold standard to three common severity scoring systems: SOFA, qSOFA and MEWS. To calculate the SOFA score, we took each patient's PaO2/FiO2, Glasgow Coma Score (GCS), mean arterial blood pressure, dobutamine, epinephrine, norepinephrine, dopamine dosage, bilirubin level, platelet counts, and creatinine level from five different CSV files. Then we categorized them in 6 categories and scored from 1 to 6 as described in [12]. The total sum of the category scores may go up to 24. For qSOFA, we used GCS, respiratory rate and systolic blood pressure as mentioned in [10]. The score ranges from 0 to 3. Finally, the MEWS score, which ranges from 0 (normal) to 14 (high risk), was determined from heart rate, systolic blood pressure, respiratory rate, temperature, and GCS. We used the scoring system presented in [42] to compute each patient's MEWS score.

3.5. Inclusion Criteria

For this study, we considered six clinical vital sign measurements: heart rate, peripheral capillary oxygen saturation (SpO2), respiratory rate, systolic blood pressure, diastolic blood pressure and temperature. We used only vital signs, which are frequently available and routinely taken in the ICU, ED, and floor units. Patient data were used from the course of a patient's hospital encounter. Patients in our final data sets were required to

- be adult (i.e. age $\geq$ 18).
- be admitted to medical Intensive Care Unit (ICU).
- have at least one measurement for each of the six vital signs.
- have at least 7 hours of data before onset.

Patients who didn't meet any of the above criterion, were excluded. Multiple hospital admissions (hadms) of a same patient were considered as separate entries.

After meeting all the above requirements and gold standards as mentioned in 3.3, the final data set contained 299 entries (out of 1240 hadms from 288 patients) for sepsis, 1046 entries (out of 3788 hadms from 1012 patients) for severe sepsis and 493 entries (out of 2520 hadms from 485 patients) for septic shock. We randomly chose 299, 1046 and 493 entries from non sepsis patient hadms to make a balanced data set for our classifiers.

3.6. Feature Selection

We took six raw vital sign data to generate our features. Following all the above-mentioned steps, we obtained three hourly values for each of the six vital sign measurements from the onset or target hour, the hour prior, and two hours prior. Then we took the two difference values for these three measurements. That made a total of five values from each vital sign resulting in a final feature vector of 30 elements for each entry of our final data set. This feature generation process was adopted from [36]. The difference values i.e. gradient information helped our classifiers to capture the temporal nature of the data.

3.7. Model Design

We performed two tasks - detection and prediction. Detection was done on the onset hour and prediction was done 4 hours prior to the onset. The onset criteria are defined in section 3.3. With our intuitive understanding of the data structures, we explored 4 classifiers for performance comparisons. They are Logistic Regression (LR), Random Forest (RF), XGBoost (XGB) and feedforward neural network or Multilayer Perceptron (MLP).

Logistic regression and XGBoost were implemented in python, using the scikit-learn [43] and XGBoost [44] python packages respectively. Random Forest was implemented in Weka version 3.8 [45]. We tuned the hyperparameters to obtain the most optimized models. We designed our neural networks in python. The chosen framework was tensorflow [46]. Keeping the sample size and

input feature dimension in mind, we designed a shallow 3 layer neural network. Due to the very small size of our data sets, more layers are redundant as that would eventually make the model overfit within very few epochs. The output value of our MLP denoted as o, can be expressed as,

$$o = f(W_{out}^T f(W_2^T f(W_1^T x + b_1) + b_2) + b_{out}) \quad (1)$$

where x is the input feature vector, W_1,W_2,W_out are the weight matrices and b_1,b_2,b_out are the bias terms of the 1st, 2nd and output layers respectively. The number of neurons on each layer, learning rate, optimizer, activation functions, regularization coefficient and dropout rate were tuned from multiple runs for each of the six cases (two tasks for each of the three categories). We found different sets of hyper-parameters working better for different cases. There was no universal set to work better on all cases.

Finally we explored ensemble of models using the best three models that is Random forest, XGBoost and MLP. There are a number of ways to perform ensemble on the trained models including linear averaging, bagging, boosting,stacked regression etc. In our previous work, we obtained the best results using simple linear averaging of the probabilities given by the individual models [47]. In this work we also explored a new architecture with a second 3 layer MLP that we trained on the 3 model's binary output probabilities. So this network takes a 6 dimensional probability vector as input rather then the 30 dimensional feature vector like any other single model. This model performed the best compared to all the other classifier models.

For all cases, we split our data sets in 70:10:20 ratio, that is we considered 70% of the data for training, 10% for validation to tune model hyper-parameters and rest 20% for testing.

4. Results

We will report our results in two steps. First, we will compare all the classifiers' performance and then we will compare the best model for each case with three standard scoring systems - SOFA, qSOFA and MEWS.

4.1. Evaluation Metrics

The performance of our classifier models was evaluated in terms of area under receiver operating characteristics curve (AUC). ROC curve is the graphical plot of true positive rate (TPR) vs false positive rate (FPR) of a binary classifier when classifier threshold is varied from 0 to 1. The number of positive instances that are correctly identified by the classifier is called true positive (TP) and the number of positive instances that are incorrectly classified by the classifier is called false negative (FN). The number of normal instances that are correctly classified as normal is called true negative (TN), and in a similar fashion, the number of normal instances that are incorrectly identified as positive instances is called false positive (FP). True positive rate (TPR) or sensitivity is the proportion of positive instances that are correctly identified as positive instances, while false positive rate (FPR) is proportion of normal instances that are incorrectly identified as positive instances.

TPR or sensitivity shows the degree to which a model does not miss a positive instance. On the other hand, specificity indicates the degree to which a model correctly identifies normal instances as normal. The objective of a model is to attain high sensitivity as well as specificity so that it attains low diagnosis error.

4.2. Classifiers comparisons

The results have been reported in Table 1. As we can see, from the results, out of the 4 base classifiers, RF is the clear winner in most of the cases while LR did the worst. XGB and MLP lie in between. We believe it's lack of enough data that caused the MLP perform weakly compared to RF. But when we tested our ensemble model, it surpassed the best scores of the 4 classifiers. In detection task, it achieved an AUC of 0.94, 0.91 and 0.89 for sepsis, severe sepsis and septic shock respectively. Out of the three categories using our ensemble model, the best score obtained was for sepsis which is 0.94 . Here, we got a 2% increase than the best model (RF for AUC). In prediction task, our ensemble model got an AUC of 0.80, 0.81 and 0.84 for the three categories respectively which are again significantly better (4%, 1% and 3% increase compared to the best classifier scores).

Table 1. AUC using different classifiers for three sepsis gold standards. Here, Det = Detection, Pred = Prediction, S = Sepsis, SS = Severe Sepsis, SK = Septic Shock, MLP = Multilayer Perceptron, LR = Logistic Regression, XGB = XGBoost, RF = Random Forest.

Task	Category	MLP	LR	XGB	RF	Ensemble
Det	S	0.86	0.83	0.90	0.92	**0.94**
	SS	0.84	0.79	**0.91**	0.90	**0.91**
	SK	0.84	0.83	0.88	0.87	**0.89**
Pred	S	0.74	0.68	0.72	0.76	**0.80**
	SS	0.75	0.72	0.77	0.80	**0.81**
	SK	0.75	0.74	0.80	0.81	**0.84**

We can notice two interesting trends here. while for detection task, the best result was obtained for sepsis, septic shock got the best treatment for prediction task. Actually it's pretty obvious. Septic shock is the ultimate stage of septicemia. So it's no wonder that the vitals for septic hock are more informative for all the classifiers, even 4 hours prior to the onset, compared to the other two categories. For example a patient who has sepsis now (detection task), might

not develop any relevant symptom for sepsis 4 hours ago (prediction task). But that's not the case for septic shock, as a patient diagnosed with septic shock now has a high chance of developing sepsis or severe sepsis 4 hours ago. Another point worth mentioning is the very little improvement in severe sepsis. We believe this is due to not having the necessary organ failure measurements as inputs. More about this in section 5.

4.3. Comparison with Benchmarks

Next we compared our ensemble model with three standard severity scoring systems - SOFA, qSOFA and MEWS. We calculated AUC for all 3 systems following the procedures mentioned in section 3.4. The results have been presented in Table 2. Out of the three systems, MEWS did the best in detection task (0.70, 0.73 and 0.65 for sepsis, severe sepsis and septic shock respectively) and SOFA did the best in prediction task (0.52, 0.58 and 0.64 for sepsis, severe sepsis and septic shock respectively). It's clear that none of the scoring systems are reliable enough to detect sepsis, severe sepsis or septic shock let alone predict. In fact, even our ensemble model's prediction scores are significantly higher than the best detection scores of the three scoring systems and this holds true for all three cases.

Table 2. Comparison with rule-based scoring systems in term of AUC. Here, Det = Detection, Pred = Prediction, S = Sepsis, SS = Severe Sepsis, SK = Septic Shock.

Task	Category	SOFA	qSOFA	MEWS	Ensemble
Det	S	0.59	0.65	0.70	**0.94**
	SS	0.64	0.72	0.73	**0.91**
	SK	0.64	0.59	0.65	**0.89**
Pred	S	0.52	0.43	0.46	**0.80**
	SS	0.58	0.56	0.56	**0.81**
	SK	0.64	0.56	0.62	**0.84**

5. Discussions

There have been many works regarding early detection i.e. prediction of sepsis [15–20, 28, 30, 31] but they used either laboratory test results or expensive equipment for data analysis which significantly increase cost and delay the whole process of detection - making the systems practically unfeasible. There are also some works [21–27] where mortality rate was calculated for sepsis patients. Considering the choice of features or final objective, all these works are somewhat irrelevant to our study.

In [32], authors proposed an early sepsis detection system, called InSight and validated it on MIMIC-II data set. 3 hours prior to the onset, they achieved an average auc of 0.83 which is a bit higher than our AUC score for sepsis in prediction task. But we also need to account for the differences in our studies such as different data set (MIMIC-II vs MIMIC-III), different data preprocessing schemes and feature selection process and most importantly different hour look-ahead (3 vs 4). Motivated by [32], [37] experimented the affects of MLP and LSTM models on prediction tasks following the same data processing pipeline. They used three different feature vector sets of dimension 100, 109 and 209 which are a lot higher compared to our 30 dimensional feature vector for each patient. Both of these works reported results only for sepsis with a different gold standard definition, contrary to our results for all 3 categories.

[35] also used InSight to assess its performance for severe sepsis, on the population of MIMIC-III who were logged using MetaVision. Their reported AUCs are .89 in detection and .75 in prediction task which are lower compared to our findings (.91 in detection and .81 in prediction). However, their gold standard definition, patient inclusion procedures and feature vectors are significantly different than ours.

Most relevant to our work is the study presented in [36]. Our selection of features, and gold standard definitions were also inspired by this study. The authors reported sepsis detection scores for the three categories of sepsis and prediction score for severe sepsis only. However, their performance scores were reported on private dataset and hence, cannot be compared directly. Our work advances the state of the art in two ways compared to [36]. We provide comprehensive benchmark performance of various rule-based and ML models on the publicly available MIMIC-III dataset and demonstrate an ensemble model that performs better than any other single or ensemble model. Second, we do feature ranking to show that different vital sign signal is ranked higher for different single or ensemble models. However, there is a universal performance improvement for all single and ensemble models when the number of vital signs is increased from 5 to 6. We discuss these two observations in the two subsections below.

5.1. Benchmark for Three Categories of Sepsis

We report the first benchmark of three different ML methods i.e. LR, XGB and RF and three rule-based methods i.e. SOFA, qSOFA, MEWS on all three categories of sepsis for detection and prediction task. Availability of these benchmark numbers on a publicly available dataset enables future works to be compared against.

In addition, we applied neural network models for the first time on this task. We found other tasks i.e. mortality prediction or severity scoring from early admission data for which neural network almost always outperform logistic regression [48–50]. [26] and [27] worked extensively with deep neural networks on MIMIC-III for different prediction tasks such as in-hospital mortality, length of stay, ICD-9 code group etc. In particular, [26] used different sets of features and different types of algorithms including deep learn-

ing models to show the effectiveness of deep learning on such data sets. The authors showed that for multiple data modalities, specially when a large number of raw clinical time series data is used as input features, deep models learn better feature representations and this held true for all three tasks they performed. Both these works demonstrate the performance benefit of deep learning compared to other ML methods on this dataset. However, in our study, single neural network architectures performed poorly compared to RF. We varied the number of hidden units and layer numbers to make sure that the model has enough capacity to learn from the data features while avoiding over-fitting. However, for all the single NN architecture and sizes that we explored, we found RF to have higher AUC than NN in most of the cases.

In order to boost AUC, we explored standard ensemble techniques - averaging and weighting. None of them showed any significant improvement. So we designed a second 3 layer MLP that we trained on the 3 model's binary output probabilities. This ensemble model design performed better than any single model and provided higher AUC consistently for all the tasks studied in this work. For example averaging gave us an AUC of 0.91, 0.90, 0.87 for 3 categories of sepsis in detection task which is rather a bit degradation over the best model's result. Here our ensemble model achieved improvements of 3, 1 and 2 percentage points respectively compared to the averaging. Similarly, we got 2, 3 and 3 percentage improvements in prediction task compared to the averaging. Our ensemble model outperformed standard disease severity scores such as SOFA, qSOFA and MEWS for both the detection and prediction of sepsis, severe sepsis, and septic shock. This is one of the main contributions of this paper. To the author's knowledge, this is the first study that incorporates deep learning and ensemble design for the detection and prediction of all three sepsis categories, taking only six vital measurements as inputs.

A direct comparison of the performance of our ensemble model can't be performed with the results presented in [36] as those results are reported on private data collected from UCSF. One comment can be made about the overall improvement of AUC by using ML model compared to the rule-based model. Whereas [36] achieved AUC improvement of 5 to 11 percentage points compared to the rule-based models, our ensemble model achieves an improvement of 17 to 28 percentage points by using ensemble models compared to the rule-based models. This large margin of improvement is a strong motivator for using this type of ensemble models for sepsis detection and prediction in ICU patients.

5.2. Feature Ranking

We ranked the six vital sign input stream to gauge their individual effectiveness in sepsis detection and prediction. The results are shown in Table 3. We numbered the six vital measurements 1-6 in the following order - heart rate, spO2, respiratory rate, systolic blood pressure, diastolic blood pressure and temperature. Here we reported result for two cases- detection of sepsis and prediction of septic shock. We found feature 4 (systolic blood pressure) and feature 6 (temperature) as the most important vital signs for prediction and detection respectively. In [36], the authors also found systolic blood pressure as the most important vital sign on MIMIC-III for similar tasks. Since we performed both single model and ensemble model performance study, we found that even for single models, these two vital signs remain as the most important ones. For example, using XGBoost, in the sepsis detection task we got an AUC of 0.77, 0.68, 0.83, 0.64, 0.65 and 0.87 (highest AUC for temperature) respectively for the six vitals and in septic shock prediction task the AUCs were 0.70, 0.62, 0.73, 0.77 (highest AUC for systolic blood pressure), 0.63 and 0.68 respectively. These numbers also indicate systolic blood pressure and temperature as the most important vitals. Also a close observation reveals, feature 3 (respiratory rate) and feature 1 (heart rate) as the next most important features.

Table 3. Features ranking in term of AUC. Systolic blood pressure and temperature are the most important vital signs for sepsis prediction and detection respectively.

Features No.	Sepsis Detection	Septic Shock Prediction
1	0.86	0.70
2	0.81	0.67
3	0.87	0.71
4	0.81	**0.73**
5	0.80	0.66
6	**0.89**	0.67

Table 4. Features ablation in term of AUC. Sepsis is highly correlated with six vital signs. All models perform significantly better when six vital signs are used.

Features No.	Sepsis Detection	Septic Shock Prediction
1	0.89	0.73
2	0.88	0.75
3	0.90	0.75
4	0.89	0.77
5	0.90	0.77
6	**0.94**	**0.84**

We also increased the number of vital signs one by one to see if there is a specific trend in AUC improvement. In Table 4, the number of variables for set 2-6 have been chosen in a way so that the gradual change, upon addition of

new vital, becomes clear. We find that there is no particular trend up to five vital signs. We have performed different combinations of the vital signs to ensure that there is indeed no specific trend of AUC improvement when five vital signs are used as input. However, we find that there is a universal improvement in AUC when six vital signs are used for both detection and prediction tasks. It is not clear exactly why six vital signs provide a universal jump in AUC for all the models be it single or ensemble. This is a topic of future exploration.

6. Limitations and Future Scopes

In this section, we remark few shortcomings of the present study and possibility of future improvements. Most of the shortcoming stem from the collection and quality of ICU data.

- The MIMIC-III data set was derived from only one institution. So its not possible to claim universal adaptability of our models to other populations on the basis of this study alone. However, since neural networks learn from the data, similar architectures should perform well when trained on data from different demographics.
- Our gold standard references to determine sepsis, severe sepsis and septic shock rely on ICD-9 codes which might fail to capture all septic patients in the dataset if there were undetected sepsis patient. This is a limitation of the process itself, though ICD-9 codes have been used before for accuracy validation in the detection of severe sepsis [51].
- The sequence of laboratory tests mostly depends on physician suspicion. As a consequence, the gold standard of sepsis onset is highly subjective and dependent on individual physician. A consistent definition of sepsis onset time and proper gold standard generation is a task for future.
- Our imputation process and averaging of all measurements in an hour's interval may lead to the loss of some temporal information which might affect the performance of our models. For time-series data like this study, [25] proposed a better imputation mechanism that can capture missing information. Performance of our model on better imputed data stream is a topic of future exploration.
- Our trained models require at least three hours of data to predict or detect; thus eliminating the possibility of any first or second hour evaluation of any patient. We intend to work on these in our future work.

7. Conclusion

In this study, we have provided the first benchmark of various ML and rule-based models for three categories of sepsis detection and prediction task on the MIMIC-III data set ICU patient population. We then designed a particular ensemble model that outperformed all the single model results. Our ensemble model also showed a large margin of improvement over common rule-based sepsis detection methods. We rank the vital signs to find that six vital signs provide better performance than any other combination of vital signs for all the models. Since the model uses only six vital signs, we believe our model will be useful for application in real world hospital environment.

8. Acknowledgement

This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Thanks Leonid Oliker at NERSC for sharing his allocation on the OLCF Titan supercomputer with us on project CSC103. Thanks to Prabhat at NERSC for sharing his allocation on the NERSC computers. Thanks to Uli Chettipally, MD for some initial discussions regarding sepsis detection.

References

[1] S. L. Murphy, J. Xu, and K. D. Kochanek, "Deaths: final data for 2010," 2013.

[2] D. C. Angus, W. T. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, and M. R. Pinsky, "Epidemiology of severe sepsis in the united states: analysis of incidence, outcome, and associated costs of care," *Critical care medicine*, vol. 29, no. 7, pp. 1303–1310, 2001.

[3] E. K. Stevenson, A. R. Rubenstein, G. T. Radin, R. S. Wiener, and A. J. Walkey, "Two decades of mortality trends among patients with severe sepsis: a comparative meta-analysis," *Critical care medicine*, vol. 42, no. 3, p. 625, 2014.

[4] A. Pfuntner, L. M. Wier, and C. Steiner, "Costs for hospital stays in the united states, 2010: statistical brief# 146," 2006.

[5] "The cost of sepsis — — blogs — cdc," https://blogs.cdc.gov/safehealthcare/the-cost-of-sepsis/, (Accessed on 12/16/2018).

[6] D. F. Gaieski, J. M. Edwards, M. J. Kallan, and B. G. Carr, "Benchmarking the incidence and mortal-

ity of severe sepsis in the united states," *Critical care medicine*, vol. 41, no. 5, pp. 1167–1174, 2013.

[7] E. Rivers, B. Nguyen, S. Havstad, J. Ressler, A. Muzzin, B. Knoblich, E. Peterson, and M. Tomlanovich, "Early goal-directed therapy in the treatment of severe sepsis and septic shock," *New England Journal of Medicine*, vol. 345, no. 19, pp. 1368–1377, 2001.

[8] H. B. Nguyen, S. W. Corbett, R. Steele, J. Banta, R. T. Clark, S. R. Hayes, J. Edwards, T. W. Cho, and W. A. Wittlake, "Implementation of a bundle of quality indicators for the early management of severe sepsis and septic shock is associated with decreased mortality," *Critical care medicine*, vol. 35, no. 4, pp. 1105–1112, 2007.

[9] A. Kumar, D. Roberts, K. E. Wood, B. Light, J. E. Parrillo, S. Sharma, R. Suppes, D. Feinstein, S. Zanotti, L. 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*, vol. 34, no. 6, pp. 1589–1596, 2006.

[10] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 801–810, 2016.

[11] M. M. Levy, M. P. Fink, J. C. Marshall, E. Abraham, D. Angus, D. Cook, J. Cohen, S. M. Opal, J.-L. Vincent, G. Ramsay *et al.*, "2001 sccm/esicm/accp/ats/sis international sepsis definitions conference," *Intensive care medicine*, vol. 29, no. 4, pp. 530–538, 2003.

[12] J.-L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. Reinhart, P. Suter, and L. Thijs, "The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure," 1996.

[13] C. W. Seymour, V. X. Liu, T. J. Iwashyna, F. M. Brunkhorst, T. D. Rea, A. Scherag, G. Rubenfeld, J. M. Kahn, M. Shankar-Hari, M. Singer *et al.*, "Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 762–774, 2016.

[14] C. Subbe, M. Kruger, P. Rutherford, and L. Gemmel, "Validation of a modified early warning score in medical admissions," *Qjm*, vol. 94, no. 10, pp. 521–526, 2001.

[15] J. C. Ho, C. H. Lee, and J. Ghosh, "Imputation-enhanced prediction of septic shock in icu patients," in *Proceedings of the ACM SIGKDD workshop on health informatics (HI-KDD12)*, 2012, p. 18.

[16] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (trewscore) for septic shock," *Science translational medicine*, vol. 7, no. 299, pp. 299ra122–299ra122, 2015.

[17] S. W. Thiel, J. M. Rosini, W. Shannon, J. A. Doherty, S. T. Micek, and M. H. Kollef, "Early prediction of septic shock in hospitalized patients," *Journal of hospital medicine: an official publication of the Society of Hospital Medicine*, vol. 5, no. 1, pp. 19–25, 2010.

[18] K. Henry, C. Paxton, K. S. Kim, J. Pham, and S. Saria, "63: Rews real-time early warning score for septic shock," *Critical Care Medicine*, vol. 42, no. 12, p. A1384, 2014.

[19] D. Shavdia, "Septic shock: Providing early warnings through multivariate logistic regression models," Ph.D. dissertation, Massachusetts Institute of Technology, 2007.

[20] 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," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 315–325, 2013.

[21] P. Marty, A. Roquilly, F. Vallée, A. Luzi, F. Ferré, O. Fourcade, K. Asehnoune, and V. Minville, "Lactate clearance for death prediction in severe sepsis or septic shock patients during the first 24 hours in intensive care unit: an observational study," *Annals of intensive care*, vol. 3, no. 1, p. 3, 2013.

[22] P.-E. Charles and S. Gibot, "Predicting outcome in patients with sepsis: new biomarkers for old expectations," *Critical care*, vol. 18, no. 1, p. 108, 2014.

[23] D. W. Ford, A. J. Goodwin, A. N. Simpson, E. Johnson, N. Nadig, and K. N. Simpson, "A severe sepsis mortality prediction model and score for use with administrative data," *Critical care medicine*, vol. 44, no. 2, p. 319, 2016.

[24] J. Calvert, J. Hoffman, C. Barton, D. Shimabukuro, M. Ries, U. Chettipally, Y. Kerem, M. Jay, S. Mataraso, and R. Das, "Cost and mortality impact of an algorithm-driven sepsis prediction system," *Journal of medical economics*, vol. 20, no. 6, pp. 646–651, 2017.

[25] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, p. 6085, 2018.

[26] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmark of deep learning models on large healthcare mimic datasets," *arXiv preprint arXiv:1710.08531*, 2017.

[27] H. Harutyunyan, H. Khachatrian, D. C. Kale, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *arXiv preprint arXiv:1703.07771*, 2017.

[28] R. A. Lukaszewski, A. M. Yates, M. C. Jackson, K. Swingler, J. M. Scherer, A. Simpson, P. Sadler, P. McQuillan, R. W. Titball, T. J. Brooks *et al.*, "Presymptomatic prediction of sepsis in intensive care unit patients," *Clinical and Vaccine Immunology*, vol. 15, no. 7, pp. 1089–1094, 2008.

[29] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. M. Schein, and W. J. Sibbald, "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis," *Chest*, vol. 101, no. 6, pp. 1644–1655, 1992.

[30] C. N. Jones, M. Moore, L. Dimisko, A. Alexander, A. Ibrahim, B. A. Hassell, H. S. Warren, R. G. Tompkins, S. P. Fagan, and D. Irimia, "Spontaneous neutrophil migration patterns during sepsis after major burns," *PloS one*, vol. 9, no. 12, p. e114509, 2014.

[31] J. Kim, J. Blum, and C. Scott, "Temporal features and kernel methods for predicting sepsis in postoperative patients," Citeseer, Tech. Rep., 2010.

[32] J. S. Calvert, D. A. Price, U. K. Chettipally, C. W. Barton, M. D. Feldman, J. L. Hoffman, M. Jay, and R. Das, "A computational approach to early sepsis detection," *Computers in biology and medicine*, vol. 74, pp. 69–73, 2016.

[33] K. D. Fairchild, "Predictive monitoring for early detection of sepsis in neonatal icu patients," *Current opinion in pediatrics*, vol. 25, no. 2, pp. 172–179, 2013.

[34] A. Bravi, G. Green, A. Longtin, and A. J. Seely, "Monitoring and identification of sepsis development through a composite measure of heart rate variability," *PLoS One*, vol. 7, no. 9, p. e45666, 2012.

[35] T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, D. Shimabukuro, U. Chettipally, M. D. Feldman, C. Barton *et al.*, "Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach," *JMIR medical informatics*, vol. 4, no. 3, 2016.

[36] Q. Mao, M. Jay, J. L. Hoffman, J. Calvert, C. Barton, D. Shimabukuro, L. Shieh, U. Chettipally, G. Fletcher, Y. Kerem *et al.*, "Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and icu," *BMJ open*, vol. 8, no. 1, p. e017833, 2018.

[37] H. J. Kam and H. Y. Kim, "Learning representations for the early detection of sepsis with deep neural networks," *Computers in biology and medicine*, vol. 89, pp. 248–255, 2017.

[38] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[39] "Postgresql: The world's most advanced open source database," https://www.postgresql.org/, (Accessed on 11/28/2018).

[40] "Mimic-iii clinical database," https://physionet.org/works/MIMICIIIClinicalDatabase/files/, (Accessed on 11/28/2018).

[41] G. Van Rossum and F. L. Drake, *Python language reference manual*. Network Theory United Kingdom, 2003.

[42] T. Lam, P. Mak, W. Siu, M. Lam, T. Cheung, and T. Rainer, "Validation of a modified early warning score (mews) in emergency department observation ward patients," *Hong Kong Journal of Emergency Medicine*, vol. 13, no. 1, pp. 24–30, 2006.

[43] "scikit-learn: machine learning in python — scikit-learn 0.20.1 documentation," https://scikit-learn.org/stable/, (Accessed on 12/03/2018).

[44] "Python package introduction — xgboost 0.81 documentation," https://xgboost.readthedocs.io/en/latest/python/python_intro.html, (Accessed on 12/03/2018).

[45] "Weka 3 - data mining with open source machine learning software in java," https://www.cs.waikato.ac.nz/ml/weka/, (Accessed on 12/03/2018).

[46] "Tensorflow," https://www.tensorflow.org/, (Accessed on 12/03/2018).

[47] M. T. Islam, M. A. Aowal, A. T. Minhaz, and K. Ashraf, "Abnormality detection and localization in chest x-rays using deep convolutional neural networks," *arXiv preprint arXiv:1705.09850*, 2017.

[48] G. Clermont, D. C. Angus, S. M. DiRusso, M. Griffin, and W. T. Linde-Zwirble, "Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models," *Critical care medicine*, vol. 29, no. 2, pp. 291–296, 2001.

[49] L. A. Celi, S. Galvin, G. Davidzon, J. Lee, D. Scott, and R. Mark, "A database-driven decision support system: Customized mortality prediction," *Journal of Personalized Medicine*, vol. 2, no. 4, pp. 138–148, 2012. [Online]. Available: http://www.mdpi.com/2075-4426/2/4/138

[50] R. Caruana, S. Baluja, and T. Mitchell, "Using the future to "sort out" the present: Rankprop and multitask learning for medical risk evaluation," in *Advances in Neural Information Processing Systems 8*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds. MIT Press, 1996, pp. 959–965.

[51] T. J. Iwashyna, A. J. Odden, J. T. Rohde, C. A. Bonham, L. B. Kuhn, P. N. Malani, L. M. Chen, and S. A. Flanders, "Identifying patients with severe sepsis using administrative claims: patient-level validation of the angus implementation of the international consensus conference definition of severe sepsis." *Medical care*, vol. 52 6, pp. e39–43, 2014.