Early Prediction of Sepsis in EMR Records Using Traditional ML Techniques and Deep Learning LSTM Networks*

Mohammed Saqib, Ying Sha, May D. Wang

Abstract— **Sepsis is a life-threatening condition caused by infection and subsequent overreaction by the immune system. Physicians effectively treat sepsis with early administration of antibiotics. However, excessive use of antibiotics on false positive cases cultivates antibiotic resistant bacterial strains and can waste resources while false negative cases result in unacceptable mortality rates. Accurate early prediction ensures correct, early antibiotic treatment; unfortunately, prediction remains daunting due to error-ridden electronic medical records (EMRs) and the inherent complexity of sepsis. We aimed to predict sepsis using only the first 24 and 36 hours of lab results and vital signs for a patient. We used the Medical Information Mart for Intensive Care III (MIMIC3) dataset to test machine learning (ML) techniques including traditional methods (i.e. random forest (RF) and logistic regression (LR)) as well as deep learning techniques (i.e. long short-term memory (LSTM) neural networks). We successfully created a data pipeline to process and clean data, identified important predictive features using RF and LR techniques, and trained LSTM networks. We found that our best performing traditional classifier, RF, had an Area Under the Curve (AUC-ROC) score of 0.696, and our LSTM networks did not outperform RF.**

I. Introduction

Sepsis is a life-threatening disease with major consequences for both the nation and the world. Defined as organ dysfunction caused by an infection [1], sepsis is the leading cause of death in the ICU, affecting 18 million worldwide with mortality rates up to 30% [2]. Treatments accounted for $24 billion of cost in the US health system in 2013 [3]. Early detection remains a key to counter the high cost and mortality; latest standards urge administering antibiotics at most an hour after the development of the disease [4], as delaying antibiotic administration by an hour could increase mortality by 1.8% for severe cases [5]. Moreover, ruling out a diagnosis early helps prevent overuse of antibiotics [6]; as an example, only 71% of suspected cases of sepsis were bacterial for one hospital [7].

Sepsis is multifaceted, manifesting many preliminary signs, and causing symptoms of extreme fever, blood pressure drops, and heart rate spikes [6]. Doctors currently diagnose and predict by using patterns in patient vital signs based on these symptoms, as well as specialized biomarkers. However, the complexity of sepsis and massive organ dysfunction in different body systems could cause different molecular biomarkers for each case [8]. In other words, testing for a single biomarker is insufficient; only certain combinations provide confidence to predict sepsis. For example, acute phase proteins are associated with many types of inflammation, including sepsis, but are not accurate alone due to similar inflammatory complications [8]; only multiple biomarkers can ensure correct predictions.

Physiological features can also be integrated to predict sepsis. Large datasets have been published with "heterogenous data types", each with their own set of challenges [9]. Heart rate characteristics, in particular, are known to be correlated with sepsis; higher variation in heart rate is correlated with increased risk of sepsis [10]. Other signs like body temperature, blood pressure, and capillary refill time improve early prediction as well [11].

Models in literature have used the integration of different data types for prediction. In one study, researchers directly competed with existing clinical decision support systems (CDSS). 5,278 patient records, using ICD-9 codes to identify sepsis, were used to train RF, LR, and regression trees; the best model was RF, with Area Under the ROC Curve (AUC) of 0.860 [12]. Another study transformed high-resolution data into segmented averages. The researchers identified sepsis with blood lactate values and achieved an AUC of 0.882 with a Logistic Model Tree trained on lab and vital signs data [13]. In another study, LSTM models on MIMIC2 data using Severe Inflammatory Response Syndrome (SIRS) to identify sepsis achieved an AUC of 0.929 [14]. Different studies in literature used varying criteria to identify sepsis, with varying applicability to useful prediction.

Sepsis is not a disease with an actionable test, which has led to an explosion of scoring mechanisms, some of which are not truly appropriate for ML; we selected the Angus criteria, which is an ICD coding system, to identify sepsis for our dataset. Unlike other tests, the Angus criteria uses final ICD diagnoses of organ failure and infection, instead of values form the training data to prevent data leak issues. It does not use SIRS criteria, which was rejected for sepsis prediction by experts [1] and uses but does not depend solely on septicemia ICD codes, which have been accepted as too narrow for identifying all cases of sepsis. Retrospective analysis using a dataset with the gold standard of human physician labels has shown that the

* The work was supported by the U.S. Department of Health and Human Services Centers for Disease Control and Prevention under Award HHSD2002015F62550B, the NIH National Center for Advancing Translational Sciences under Award UL1TR000454, the National Science Foundation Award NSF1651360, and Children's Healthcare of Atlanta and Georgia Tech Partnership Grant.

M. Saqib is with the Department of Biomedical Engineering, Georgia Institute of Technology (email: saqibm128@gmail.com).

Y. Sha is with the School of Biology, Georgia Institute of Technology (e-mail: ysha8@gatech.edu).

M. D. Wang is with the Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA (corresponding author, phone: 404-385-2954; fax: 404-385-0383; e-mail: maywang@bme.gatech.edu).

978-1-5386-3646-6/18/$31.00 ©2018 IEEE

Angus criteria works well even against other similar ICD based criteria systems [15].

II. Dataset

The MIMIC3 Clinical Dataset records 61,532 ICU stays divided among 58,976 hospital admissions, themselves distributed among 46,520 subjects from Beth Israel Deaconess Medical Center, and maintained by MIT [16]. Records primarily consist of vital sign data, lab results, and the time of observation. This dataset, however, has multiple issues which need to be addressed. Events are irregularly sampled, include outlier data values, and can be entirely missing for some features and patients. Additionally, the same feature can be assigned multiple codes, which further complicates any processing.

III. Procedure

A. Data Preprocessing Pipeline

We began data preprocessing from MIMIC3 with data extraction. We selected and split data on each hospital admission, removing the admission entirely if the patient was younger than 18 years old. We then mapped medically coded events to a high-level feature and cleaned with custom functions for some features, including for outliers and physiologically impossible values [17]. The data was mapped to 47 features that we had outliers and values for, including blood features such as albumin and bicarbonate, as well as other vital signs, (i.e. diastolic and systolic blood pressure and heart rate).

Data was irregularly sampled and missing entirely for certain admissions, so we needed to resample events. We removed patients missing blood pressure and heart rate values entirely. We resampled events recorded at arbitrary times into 6-hour bins for each feature, measured from the first event for an admission. For each bin and feature, we averaged values and used the result as the representative data point for the bin. We filled in missing data for certain 6-hour bins for a feature through a combination of forward-filling (i.e. using averaged value of closest past bin in relation to the missing bin) then back-filling (i.e. using averaged value of closest future bin in relation to the missing bin) and with physiologically normal values for other entirely missing features; we assumed missing data was similar to the last known data point, and that completely missing values are likely due to a doctor believing the feature to be normal for a patient and useless to record.

Only 38,270 records remained after preprocessing. Among the 38,270 hospital admissions, the Angus criteria identified 10,071 positive for sepsis. On average, 2.30 bins out of 4 bins were missing and filled in for the 24-hour case for each variable, while 3.32 bins were missing out of 6 bins for the 36-hour case for each variable and each hospital admission.

B. Traditional Machine Learning Techniques

We trained LR and RF. We partitioned data into training, validation, and testing sets of 81%, 9%, and 10% of the data, respectively, stratified using the Angus criteria labeling. We ran a chi-squared test of independence between the features in the combined train and validation sets on the Angus labels for feature selection; the test set is excluded to prevent information leaking. We only included the top 34 features. Some of the features tested for a dependency with sepsis are shown in **Table 1**. After feature selection, we normalized the dataset based on mean and standard deviation for each feature, to provide a common baseline for analysis of weights of the logistic regression model. We chose hyperparameters based on a grid search and evaluated based on the best F1 score metric on the validation set. Traditional ML techniques cannot natively take in data for time events, so we used two different methods to input the data for n=24 hours and n=36 hours. First, we averaged the cleaned data in the first n hours of events before the resampling step in the data preprocessing pipeline is used. For the second approach, we used the data for each 6 hour bin in the n hours and averaged to create n/6 new features for each original high level feature and time bin.

Feature	Chi-Squared Value	P-Value
WHITE BLOOD CELL COUNT	1383.677	2.18E-304
HEART RATE	1149.133	2.04E-253
DIASTOLIC BLOOD PRESSURE	489.8446	4.53E-110
SYSTOLIC BLOOD PRESSURE	410.3756	8.94E-93
MEAN BLOOD PRESSURE	397.2578	6.41E-90
WEIGHT	282.3454	6.79E-65
ANION GAP	83.9072	1.53E-21
BICARBONATE	40.145	6.94E-12
OXYGEN SATURATION	7.076864	2.30E-04
HEIGHT	2.683175	2.98E-03
TEMPERATURE	0.446421	1.48E-02
PH	0.094232	2.23E-02

Table 1. Chi-Squared Values for Independence Testing of Selected Features Against Angus Criteria Labeling. While WBC count is the highest rated feature (most likely to have a relationship with sepsis), surprisingly, pH and Temperature rate are much lower.

LR and RF algorithms used the Python scikit implementation, and hyperparameters were chosen from the available options in the API. For LR, hyperparameters included ending tolerance, solver method, and regularization parameter. For RF, hyperparameters included split criteria, the number of trees, max number of features kept, max depth of each tree, and minimum fraction of the whole data for each leaf.

C. Deep Learning Techniques

We used the same pipeline for deep learning as our traditional ML techniques, with a training set, a validation set, and a testing set of sizes 81%, 9%, and 10%. Feature selection was done ahead of time using only the validation and training set. Due to the infeasibility of grid search techniques with deep learning, we used only a small subset of the combinations of hyperparameters on either an LSTM or Attentional LSTM model and evaluated models during the training phase with a log loss metric. For example, we tested 1, 2, and 3 layer networks and tested networks with 10, 50, 100, and 250 hidden nodes. We added a dropout layer with 20% probability of removing a signal above the LSTM or Attentional LSTM model to provide less dependency on each node during training, but removed the dropout layer for testing or validation. For each set of hyperparameters, we kept the learning rate at 0.5, decreasing by 50% every 20 epochs and momentum at 0.5

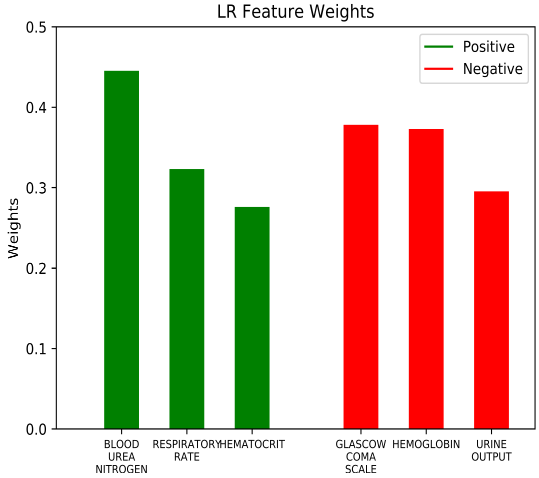

Figure 1. This figure represents the highest magnitude features of the 36-hour averaged LR model. LR weights cannot usually be interpreted for importance but because the data was normalized by mean and standard deviation, the weights of the features can provide a good way to compare different features' contributions towards a prediction.

	F1 Score	AUC-ROC	MCC	Prec.	Rec.
LR Avg. 24 Hrs.	0.4306	0.6278	0.3327	0.6508	0.3217
LR Avg. 36 Hrs.	0.4292	0.6269	0.3290	0.6446	0.3217
RF Avg. 24 Hrs.	0.5128	0.6684	0.3991	0.6682	0.4160
RF Avg. 36 Hrs.	0.5112	0.6672	0.3914	0.6535	0.4198
LR Binned 24 Hrs.	0.4862	0.6548	0.3805	0.6711	0.3811
LR Binned 36 Hrs.	0.4850	0.6541	0.3778	0.6667	0.3811
RF Binned 24 Hrs.	0.5559	0.6927	0.4410	0.6865	0.4670
RF Binned 36 Hrs.	**0.5669**	**0.6991**	**0.4564**	**0.7020**	**0.4755**

Table 2. Metrics for Traditional Machine Learning. MCC refers to Mathew's Correlation Coefficient while Prec. refers to precision and Rec. refers to recall. Input features were either averaged over the entire time span or were split into 6-hour bins and then averaged. The best model was RF with split data for the first 36 hours. Avg. refers to average, Hrs. refers to hours.

Feature	RF Rank	RF Importance	LR Rank	LR Weight
BLOOD UREA NITROGEN	1	0.095779	1	0.456301
RESPIRATORY RATE	2	0.065888	4	0.325498
URINE OUTPUT	3	0.05061	7	-0.26937
SYSTOLIC BLOOD PRESSURE	4	0.048572	14	-0.07976
CREATININE	5	0.044741	20	-0.05493
HEART RATE	6	0.044682	9	0.197269
PLATELETS	7	0.043825	28	0.011166
WHITE BLOOD CELL COUNT	8	0.042957	15	0.078343
MEAN BLOOD PRESSURE	9	0.038063	29	-0.00645
GLUCOSE	10	0.038043	26	-0.01628

Table 3. Feature Ranking for the 10 Highest Magnitude Features for RF and Corresponding LR Ranking. These ranking were taken from models with average data input, for the first 36 hours. LR is ranked by absolute value of weight of feature in model. Features are arranged by highest RF relative importance.

throughout and ran the experiment for 1000 epochs, with a minibatch size of 1000 instances.

IV. Results

A. *Traditional Machine Learning Techniques*

Model results depended on length of window of data, as well as how the data was input into the model, as discussed in our methods. Models using 6-hour binned averages of the timeseries saw significant improvements over naïve average models over the entire period.

As illustrated in **Table 2** inputting binned averages of the first 36 hours for RF provided the best metrics. Most models using 36 hours had better results compared to using 24 hours of data. The only exception was LR models with binned data, which appeared marginally better with only 24 hours, though that appears likely to be due to chance.

Top features used by our models were the same as those known to be associated with sepsis, as shown in **Table 3.** RF models found that blood test features and blood pressure features were the most important towards the problem. LR weights were also taken as a proxy measurement of feature importance to prediction as shown in **Figure 1**. More interestingly, LR feature weights have positive and negative values, which provided analysis on direct or inverse relationships between features and sepsis. Heart rate and respiratory rate have a strong direct relationship with the occurrence of sepsis, as expected, while blood pressure values have a strong indirect relationship with the occurrence of sepsis.

B. *Deep Learning Techniques*

As shown in **Table 4**, the best LSTM models had varying success. Unfortunately, none of the LSTM models were as successful as the RF classifiers, except regarding the recall parameter, which is likely due to the use of log likelihood during training. LSTM models which generated subpar results were excluded from the results. LSTM models trained with attentional mechanisms only achieved

Attn.	h	Time (Hours)	Number Layers	F1 Score	AUC-ROC	MCC	Prec.	Rec
No	250	24	2	0.5112	0.6585	0.2988	0.462243	0.571698
No	250	24	3	0.5015	0.6583	0.3365	0.552154	0.459434
No	100	36	2	0.5017	0.6572	0.3249	0.528736	0.47735
Yes	250	24	1	0.5093	0.6602	0.3175	0.501832	0.516981
Yes	100	36	2	0.5043	0.6593	0.3322	0.539205	0.473585
Yes	250	36	2	0.5190	0.6664	0.3250	0.4991	0.5406

Table 4: Metrics for Best LSTM Models. h refers to dimension of hidden nodes in LSTM models while MCC refers to Mathew's Correlation Coefficient, Attn refers to whether attentional mechanisms were used, Prec. refers to precision, Rec. refers to recall.

a slight but noticeable improvement over LSTM models trained without attentional mechanisms. In general models performed better when supplied with data covering 36 hours than with data covering only 24 hours.

V. Discussion

Our LR and RF models after training provided insights into how each feature contributed to a prediction. While feature weights in the LR model showed both direct and inverse relationships between features and sepsis prediction, as shown in **Figure 1,** RF only showed relative importance, as shown in **Table 3**. Both LR and RF showed blood urea, which is indicative of renal failure, was more important than other features for prediction. Similarly, the Glascow Coma Scale (GCS) is the most negatively correlated with sepsis for LR, since a decrease represents neural dysfunction. Hemoglobin is almost equally negatively weighted; a decrease in hemoglobin would represent circulatory distress. Blood pressure values were also important, especially in RF, where 3 of the top 10 important features were blood pressure values, with total relative importance above 0.12. However, **Table 3,** shows that many of the features considered important by RF often had relatively lower absolute value weights in LR, usually due to RF relying on more complex relations than positive or negative correlation. Unlike LR or RF models, LSTM models could not provide similar insights.

Our results also suggest that an effective LSTM model required a threshold for both the number of layers as well as the hidden nodes (h). LSTM models with less than 2 layers and h < 100 had worse metrics than our other models, with AUC metrics below 0.65 and were excluded from our best results in **Table 4**. LSTM networks with Attention mechanisms improved slightly on metrics we tested for, compared to LSTM networks with none of the Attention mechanisms. Surprisingly, no LSTM models exceeded the metrics from our best RF model, except regarding the recall metric; the higher recall metric suggests that our LSTM networks could exclude possible false positive cases that our traditional ML models failed to deal with.

Our research shows that ignoring the time dependent nature of sepsis will lead to significantly lower prediction metrics. Naively integrating higher frequency data from EMR records alongside single lab tests with an average over time did not capture minute details; sepsis can quickly progress from a simple infection, but an average of a timespan would fail to provide details of the progression. In fact, when we averaged all data over time, our LR and RF models would have lower metrics. Our implementation of a pipeline, where we instead segmented data from both EMR records and lab results into 6-hour bins, allowed for more minute analysis and provided more sophisticated analysis of time dependencies by our models. Accordingly, we found that models which could handle the time dependencies well, like RF and LSTM networks, generally had higher scores. Integration of lab results with vital signs data from a time series proved to be important for better prediction.

VI. Conclusion

Our experiments successfully used longitudinal data in a variety of models to predict sepsis. We showed that both LR and RF models could learn key features, both biomarkers and vital signs, already known to be correlated with sepsis, and that recurrent neural networks such as variants of LSTM models could achieve some success, though not as much as RF. Future improvements will likely come from integrating additional data from the MIMIC3 dataset such as waveforms and demographic data, testing additional models, and more thoroughly exploring the LSTM hyperparameter space.

Acknowledgment

We thank Johnny Chen and Hang Wu of MIBLAB for help editing our paper. We thank the MIMIC3 Benchmark Repo group for part of our data preprocessing code [17].

References

[1] M. Singer *et al.*, "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA,* vol. 315, no. 8, pp. 801-810, 2016.

[2] E. Slade, P. S. Tamber, and J.-L. Vincent, "The Surviving Sepsis Campaign: raising awareness to reduce mortality," *Critical Care,* journal article vol. 7, no. 1, p. 1, Jan. 8 2003.

[3] Department of Health and Human Services. (2016). *204, National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2013*. Available: Department of Health and Human Services

[4] D. De Backer and T. Dorman, "Surviving sepsis guidelines: A continuous move toward better care of patients with sepsis," *JAMA,* vol. 317, no. 8, pp. 807-808, 2017.

[5] V. X. Liu *et al.*, "The Timing of Early Antibiotics and Hospital Mortality in Sepsis," (in English), *American Journal of Respiratory and Critical Care Medicine,* vol. 196, no. 7, pp. 856-863, Oct. 1 2017.

[6] F. Lupu, ""Crossroads in Sepsis Research" Review Series Overview of the pathophysiology of sepsis," *Journal Of Cellular And Molecular Medicine,* vol. 12, no. 4, pp. 1072-1073, May 9 2008.

[7] T. C. Minderhoud, C. Spruyt, S. Huisman, E. Oskam, S. C. E. Schuit, and M. D. Levin, "Microbiological outcomes and antibiotic overuse in Emergency Department patients with suspected sepsis," (in eng), *Neth J Med,* vol. 75, no. 5, pp. 196-203, June 28 2017.

[8] C. Pierrakos and J.-L. Vincent, "Sepsis biomarkers: a review," *Critical Care,* vol. 14, no. 1, p. R15, Feb. 09 2010.

[9] P. Y. Wu, C. W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang, "Omic and Electronic Health Record Big Data Analytics for Precision Medicine," *IEEE Transactions on Biomedical Engineering,* vol. 64, no. 2, pp. 263-273, 2017.

[10] K. D. Fairchild and T. M. O'Shea, "Heart rate characteristics: physiomarkers for detection of late-onset neonatal sepsis," *Clin Perinatol,* vol. 37, no. 3, pp. 581-98, Sept. 2010.

[11] M. F. Alqahtani, L. E. Marsillio, and R. A. Rozenfeld, "A Review of Biomarkers and Physiomarkers in Pediatric Sepsis," *Clinical Pediatric Emergency Medicine,* vol. 15, no. 2, pp. 177-184, June 1 2014.

[12] R. A. Taylor *et al.*, "Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach," *Acad Emerg Med,* vol. 23, no. 3, pp. 269-78, March 2016.

[13] J. Guillén *et al.*, "Predictive models for severe sepsis in adult ICU patients," in *2015 Systems and Information Engineering Design Symposium*, 2015, pp. 182-187.

[14] 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, Oct. 1 2017.

[15] T. J. Iwashyna *et al.*, "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, no. 6, pp. e39-e43, 2014.

[16] A. E. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci Data,* vol. 3, p. 160035, May 24 2016.

[17] 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.