Representation Learning for Sepsis Prediction

Rohith Sathiamoorthy Pandian[1], Krish Rewanth Sevuga Perumal[1], Yotam Voskoboynik[2]
[1]Department of Computer Science, University of California San Diego, California;
[2]Bioinformatics and Systems Biology, University of California San Diego, California

Abstract
Sepsis is a life-threatening disease that leads to multi-organ failure and is caused as a result of a severe immune response to an infection. The world health organization made sepsis a global health priority and urged countries to improve methods of prediction, prevention and management of sepsis. In order to decrease the high cases of fatalities due to sepsis and morbidity for sepsis and septic shock, there is a need to increase the predictive rate of sepsis in an early stage. In this work we employ deep learning based neural-network models on both structured and unstructured data for sepsis prediction. We conduct experiments on the MIMIC-IV chest X-ray reports to evaluate the performance of the models using standard performance metrics such as positive predictive value, sensitivity and specificity. Our proposed representation learning technique exhibits a final precision of 80% in the predictive task, thus showing the effectiveness of the model.

Introduction
Sepsis is a life-threatening condition caused by an extreme immune response to an infection that leads to organ failure[1] and represents a substantial health burden, having accounted for $23.7 billion, or 6.2 percent of total hospitalizations costs in 2011 in the United States[2]. Tentative reports from high income countries have shown an annual estimate of 31.5 million sepsis cases, 19.4 million severe sepsis cases and 5.3 million deaths[3].

In 2017, the World Health Organization (WHO) made sepsis a global health priority and adopted a resolution in the seventieth World Health Assembly (WHA) to develop guidance and support countries on early diagnosis, prevention and management of sepsis[4]. As a result, strategies that incorporate early sepsis recognition, prevention and management of sepsis are being implemented[5,6,7].

The MIMIC-IV database[8] provides a rich list of clinical datasets that facilitates research and data driven solutions. In general, datasets consist of two types of data: structured and unstructured data. Structured data refers to data that has defined models and a fixed set of values (numerical or categorical). Some examples of structured data include vital features (Heart Rate, Systolic Blood Pressure, Diastolic Blood Pressure, etc), Diagnosis Features (Bicarbonate, Calcium, Glucose, etc) and comorbidities (Has Obesity, Has Hypertension, Has Diabetes, etc). Unstructured data, on the other hand, refers to data that does not have a predefined model or a fixed set of values such as physician reports, radiology and pathology reports and clinical progress reports. Structured data can directly be processed and used in models to infer information about some specific predictive tasks, but unstructured data has to be pre-processed, normalized and vectorized to a particular format to be used in the models. In this paper, we use the radiology reports from the MIMIC-IV chest X-ray database[9] to improve clinical sepsis predictions in conjunction with structured and unstructured data.

The goal of this project is to evaluate the performance of neural network-based deep learning models for predicting the presence or absence of sepsis using the structured data of patients who were admitted to the ICU along with their radiology reports.

Methods
The MIMIC-IV chest X-ray report gives a wide range of data about the patients admitted to ICU. We consider most of the structured information provided and test them on baseline logistic regression models. Later, we used deep learning models such as Conventional Neural Network (CNN) and Long Short-Term Memory (LSTM) to improve the model's performance. Since, the goal of the project was to be able to identify sepsis, we primarily focused on improving the Positive Predictive Value (PPV) of the model, since in such models it is okay to have a slightly higher false positive value than a false negative value.

Each patient's data consists of multiple readings. In the baseline method we try to predict the sepsis label by accumulating the readings of all patients in a single data frame. On the other hand, we convert these data into a time series input and predict the sepsis label on a patient level in the deep learning models rather than reading level.

Data Processing

In this project, we have considered 141 different features including Vital Features (Heart Rate, Temperature, etc), Diagnosis Features (Calcium, Glucose, etc), MED Features (on_anesthesia, on_anticoagulants, etc), Comorbidities (has_pneumonia, has_shock, etc), sepsis labels and few other features (Age, ICULOS, etc). We had also parsed the anonymized patient ID from the data files and created a time column with entries 0, 1, …, etc for each reading for the patient to give it a time structured model which we use for the Deep Learning Models (CNN and LSTM). The report section was also parsed to create a 200 dimension word embedding representation and used to improve the sepsis prediction task. We have selected all the features, so that the model can make an informed decision about the relevant features for the sepsis prediction task rather than manually selecting the relevant features. The number of datapoints are shown in Table 1. While performing the exploratory data analysis of the features, it was seen that some features such as on_vent and FiO2 had more than 99% of the data as NaN i.e. missing from the report. As a result, we dropped 32 columns that had more than 20% of the data as NaN and were left with 109 features. There were still a significant number of missing values in other columns, so we had used mean imputation to fill in the values for the numeric columns. For the report text column, we used the forward fill method to replace the NaN values at a patient level and filled any remaining columns with an empty string. We then converted the integer columns to int values to ease computation. The structured data and the unstructured data (text report) were split for the analysis.

Table 1. Basic Statistics of the Data Set

Dataset	Total Number of Patients	Total Number of Records	Patients with sepsis to without sepsis ratio
Training Set	1,978	97,512	0.4684
Validation Set	1,462	110,182	0.5308

Baseline Model on Structured Data

The train and validation set were used to train and test a random forest logistic regression model on the records of all the patients concatenated into a single dataframe. Different metrics such as Accuracy, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Sensitivity and Specificity were used to evaluate the model's performance. We also selected the relevant features by using the feature importance metrics provided by the random forest classifier library and retrained the model with only the important features. The selected features and their importance value to the model are shown in Figure 1. Interestingly, the list of important features had the most relevant features used for sepsis prediction such as Body Temperature, Heart Rate, Respiratory Rate and Leukocyte Count[10].
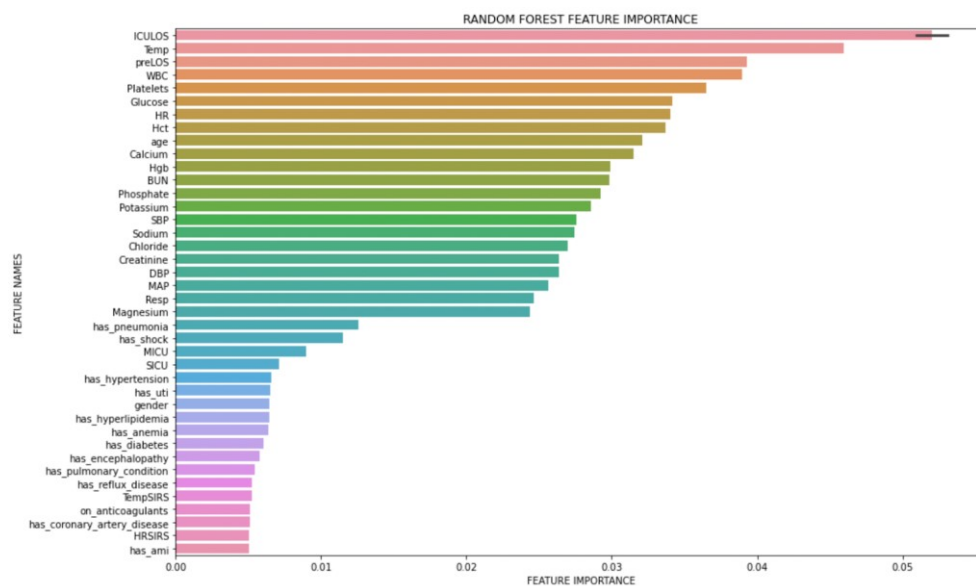

Figure 1. Random Forest Classifier Feature Importance

Though the baseline model would give a better PPV than the previous model, this will not be good enough for predicting sepsis labels, and hence we have to employ advanced neural networks based deep learning models namely a Sequential Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM) model to get a better precision value on the predictive task.

Deep Learning Model Architectures:
The sequential convolutional neural network model used had 7 layers as shown in Table 2. The first layer was a 1D convolutional layer with 16 filters, a kernel size of 3, stride length of 1 and a Rectified Linear Unit (ReLU) activation function. We then use a Global Max Pooling operation on the 1D temporal data from the conv1d layer, where the input representation is down sampled by considering the maximum value over the time dimension. Then a regular densely connected neural network layer with 280 hidden dimensions is used followed by another ReLU activation layer. Finally, another densely connected neural network with 1 dimension is used as the output layer and passed to a sigmoid activation layer to get the final prediction. The final model was compiled using a binary cross-entropy loss function and an adam optimizer[11].

The second model is also a sequential neural network model with 4 layers as shown in Table 2. The first layer is a Long Short-Term Memory (LSTM) with 50 neurons. This is then passed into a dropout layer with a frequency of 0.2 which is then passed to a flatten layer. The output is then finally passed into a densely connected neural network layer with 1 dimension and a sigmoid activation function. The resultant of the model is used as the final prediction value. The final model was compiled using a binary cross-entropy loss function and a rmsprop optimizer.

Table 2. Deep Learning Model Architecture

Layer	CNN			LSTM		
	Layer	**Output Shape**	**Number of Params**	**Layer**	**Output Shape**	**Number of Params**
First	Conv1D	(None, 400, 16)	5,152	LSTM	(None, 402, 50)	31,600
Second	GlobalMaxPooling1D	(None, 16)	0	Dropout	(None, 402, 50)	0
Third	Dense	(None, 280)	4,760	Flatten	(None, 20100)	0
Fourth	Dropout	(None, 280)	0	Dense	(None, 1)	20,101
Fifth	Activation	(None, 280)	0	-	-	-
Sixth	Dense	(None, 1)	281	-	-	-
Seventh	Activation	(None, 1)	0	-	-	-

Deep Learning Model on Structured Data
The patient data that had all their records as a time series data after cleaning the NaN values were used as the input dataset. The patient ID and the custom inserted time records were used to create a sequence of data. For each patient the values of the feature columns from the data frame were converted to a list resulting in a data shape of $1978 \times n \times 107$ for the train data frame and $1468 \times n \times 107$ for the validation data frame, where 1978 is the number of patients, n in the number of records for each patient which varies for each patient and 107 in the number of features. The output sepsis label was decided based on whether the patient had sepsis or not, it would be 1 if the patient had sepsis, 0 otherwise. To get the same dimension for all records, we truncate the second dimension of the train data with zeros to a maximum length of 402 (maximum number of records for a patient). We then reshape the model and the resultant shape is $1972 \times 402 \times 107$ for the train data frame and $1468 \times 402 \times 107$ for the validation data frame. In this project we use two models, namely a Sequential Convolutional Neural Network with 7 layers and a Long Short-Term Model with 4 layers. The total trainable parameter count for the structured train data is 10,193 for the convolutional neural network model and 51,701 for the Long Short-Term Memory model.

Deep Learning Model on Structured and Unstructured Data
The structured data used for the previous model is considered as it is and the additional doctor's report text is added as additional embeddings. The text is first processed to remove any numbers and punctuations and then vectorized

using the TreebankWordVectorizer the NLTK library. The tokens are then cleaned by removing any stop words. The cleaned words are then converted into a vector of 200 dimensions using the PubMed Word2Vec dataset. The vectorized tokens are then flattened out by taking the mean of each dimension of the total text report. This flattened vectorized data is then appended to the structured data. The new shape of the datasets after adding the 200 dimensioned text data is $1978 \times n \times 307$ for the train data frame and $1468 \times n \times 307$ for the validation data frame. The total trainable parameter count for the structured train data is 19,793 for the convolutional neural network model and 91,701 for the Long Short-Term Memory model.

Results
Baseline Model
The performance of the baseline random forest logistic regressor is shown in Table 3. The baseline model has a low positive predictive value (PPV) of 0.76%. The redid model with the relevant features using the feature importance metrics from the random forest classifier provided a slightly improved performance. As seen from Table 3, the model gave a slightly better positive predictive value of 1.88%, yet it is not good enough for the predictive task in hand.

Table 3. Baseline Model Performance Metrics

Performance Metrics	Random Forest (With all features)	Random Forest (With only important features)
Confusion Matrix	$\begin{bmatrix} 26 & 3384 \\ 3944 & 102803 \end{bmatrix}$	$\begin{bmatrix} 64 & 3346 \\ 7903 & 98844 \end{bmatrix}$
Accuracy	0.9335	0.8979
Positive Predictive Value	0.0076	0.0188
Negative Predictive Value	0.9631	0.9260
Sensitivity	0.0065	0.0080
Specificity	0.9681	0.9672

Deep Learning Models on Structured Data
The deep learning models were first trained with 50 epochs to compare the model accuracy and loss for both the training and validation dataset. The plots are shown in Figure 2. The accuracy and the loss values for the model on both the training and validation set was analysed to choose the most optimal number of epochs, and the final convolutional neural network model was trained with 6 epochs and the long short-term memory model was trained with 4 epochs. The performance metrics of both the models are shown in Table 4. The CNN model resulted in a positive predictive value of 71.32% which is 69.44% better when compared to the baseline model. The long short-term memory model further improved the positive predictive value to 71.71% Thus as expected, the deep learning models seem to provide a better performance for the predictive task, which we further improve using the unstructured report data.

Table 4. Deep Learning Model Performance Metrics

Performance Metrics	Structured Data		Structured Data & Clinical Report	
	CNN	LSTM	CNN	LSTM
Confusion Matrix	$\begin{bmatrix} 363 & 146 \\ 95 & 864 \end{bmatrix}$	$\begin{bmatrix} 365 & 144 \\ 77 & 882 \end{bmatrix}$	$\begin{bmatrix} 373 & 136 \\ 70 & 889 \end{bmatrix}$	$\begin{bmatrix} 407 & 102 \\ 280 & 679 \end{bmatrix}$
Accuracy	0.8358	0.8494	0.8597	0.7398
Positive Predictive Value	0.7132	0.7171	0.7328	0.7996
Negative Predictive Value	0.9009	0.9197	0.9270	0.7080
Sensitivity	0.7926	0.8258	0.8420	0.5924
Specificity	0.8554	0.8596	0.8673	0.8694

Deep Learning Model on Structured and Unstructured Data

The deep learning models were also trained on the combination of the structured and unstructured text data with 50 epochs. The plots are shown in Figure 2. After analysing the accuracy and the loss values for the models, we chose 10 epochs for the convolutional neural network model and 3 epochs for the long short-term memory model. The performance metrics of both the models are shown in Table 4. The CNN model provided a positive predictive value of 73.28% This shows that the clinical report of the physician helps improve the performance of the deep learning models. The long-short term memory model on the structured and unstructured data resulted in a positive predictive value of 79.96%

CNN with Structured Data

LSTM with Structured Data

CNN with Structured Data & Clinical Report

LSTM with Structured Data & Clinical Report

Figure 2. Deep Learning Model's Accuracy and Loss Plot for the train and validation dataset

Discussion

Though the baseline random forest model seemed to give a very high accuracy of 93.34%, it had a very low positive predictive value (PPV) of 0.76% and thus this model will not be suitable for the task. Also, the training accuracy of the model was 100% which clearly showed that the model overfitted the data during training. Although the model with only the selected features using their importance to the model provided a slightly better performance with an accuracy of 89.79% and a positive predictive value of 1.88% which is 1.12% higher than the model with all the features used, it was not enough for the sepsis predictive task.

When training the model for 50 epochs, as expected, the training accuracy of both the CNN and LSTM model increases for the training set and decreases for the validation set as the number of epochs increases. Similarly, the loss decreases for the training set and increases for the validation set as the number of epochs increases. The convolutional neural network model resulted in an accuracy of 83.58% and a positive predictive value of 71.32% The long short-term memory model provided an accuracy of 84.94% and a positive predictive value of 71.71% (0.39% more than the convolutional neural network model).

Similar to the structured data model performance, as the number of epochs increased the training accuracy of both the CNN and LSTM model increased for the training set and decreased for the validation set. The loss decreased for the training set and increased for the validation set. The convolutional neural network model provided an accuracy of 85.97% and a positive predictive value of 73.28% (an increase of 1.96% from the CNN model using only the structured data). The long-short term memory model on the structured and unstructured data resulted in an accuracy of 73.98% (lowest of all the deep learning models) and a positive predictive value of 79.96% (an increase of 8.25% from the LSTM model using only the structured data and 6.68% better than the CNN model with both the structured and unstructured data). Though, the accuracy of the final model is lower than all the other models, the high precision value makes this model more suitable for the sepsis prediction task since, it is more important for such models to predict positives than reject negatives, i.e., a slightly higher value of true positive is better than a lower false negative value.

Conclusion

We performed a predictive task of sepsis prediction using structured and unstructured data of patients admitted in ICU from the MIMIC-IV database. We used different data processing and feature engineering techniques to select the most relevant features for the model. We evaluated the sepsis prediction task using standard metrics such as Precision or Positive Predictive Value, Negative Predictive Value, Sensitivity, Specificity and Accuracy. We initially used a baseline model on the structured data and then used deep learning based neural network models namely convolutional neural network and long short-term memory to improve the performance. We also used the same models to train the structured data along with the vectorized clinical report text and achieved a better positive predictive value. The performance of the predictive task can be further improved by better engineering the features, tuning the hyperparameters of the deep learning models and by incorporating better modeling strategies.

Future research directions include (1) Comparing the results with different neural network models and feature engineering techniques such as using the standard sepsis prediction chart to analyze the data and by manually selecting the standard features used for sepsis prediction. (2) Use different imputation strategies to fill NaN values for numeric features and the Report Text Column. (3) Use different techniques for flattening the vectorized features of the report text and combining them with the structured data.

References

1. Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):801-810.
2. Torio, Celeste M., and Brian J. Moore. National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2013. Agency for Healthcare Research, 2006.
3. Fleischmann C, Scherag A, Adhikari NK, et al. Assessment of Global Incidence and Mortality of Hospital-treated Sepsis. Current Estimates and Limitations. Am J Respir Crit Care Med. 2016;193(3):259-272.
4. WHA70. 7. Improving the prevention, diagnosis and clinical management of sepsis. In Seventieth World Health Assembly, Geneva, 29 May 2017. Geneva: World Health Organization; 2017.
5. Kim HI, Park S. Sepsis: Early Recognition and Optimized Treatment. Tuberc Respir Dis (Seoul). 2019;82(1):6-14.
6. Polat G, Ugan RA, Cadirci E, Halici Z. Sepsis and Septic Shock: Current Treatment Strategies and New Approaches. Eurasian J Med. 2017;49(1):53-58.

7. Pruinelli L, Westra BL, Yadav P, et al. Delay Within the 3-Hour Surviving Sepsis Campaign Guideline on Mortality for Patients With Severe Sepsis and Septic Shock. Crit Care Med. 2018;46(4):500-505.
8. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi L A, Mark R. MIMIC-IV (version 2.1). PhysioNet. 2022.
9. Johnson A, Pollard T, Mark R, Berkowitz S, Horng S. MIMIC-CXR Database (version 2.0.0). PhysioNet. 2019.
10. Oczkowski S, Al Mubarak Y, Alhazzani W, Rhodes A, Kübler A, Jankowski M. Sepsis and Septic Shock. McMaster Textbook of Internal Medicine. Kraków: Medycyna Praktyczna. https://empendium.com/mcmtextbook/chapter/B31.II.18.7. (Accessed 7 Dec, 2022).
11. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014 Dec 22.