Early Prediction of Sepsis Label using derived features and Bidirectional LSTM model

[1*]Rohan Shukla,[1*] Ekta Ghuse, [1*]Aniket Verma, [1*]Kalp Pawar,[1,2] Harsh Deshpande, [1*]Montu Saw

[1]Indian Institute of Information Technology, Nagpur
[*]Computer Science and Engineering Department
[2]Electronics and Communication Department

Abstract

Accurate data imputation increases training and prediction accuracy of the model when the clinical data used has 92% of missing values. The model is developed in order to predict the binary label at an optimal time of (t – 6) hours before determining it clinically at time t. Determining the onset before 6hours is chosen to be optimal than predicting it either too early which results in resource wastage or too late that increases the mortality rate. In this work, a new feature-based bidirectional recurrent neural network model i.e. Long-Short Term Memory (LSTM). The development is divided into sections- Imputation, processing, model-training performed on a time-series data set provided by SIH organizing team 2020.

1. Introduction

Sepsis is a life-threatening organ dysfunction syndrome caused by a dysregulated host response to infection. If not recognized early and managed promptly, it can lead to septic shock, multiple organ failure and death [1]. Any type of infectious pathogen can potentially cause sepsis. India has a meagre data on the medical condition; thus, it needs the attention of people who can comprehend the severity of the syndrome and make people aware about it.

The proposed model is designed to take into consideration all the clinical significance of the data and predict the accurate label. The methods used take the data and preprocesses it sequentially, followed by adding more features derived

2. Data

The data provided for SIH 2020 is obtained from physionet challenge 2019 and consists of patients' data from two hospitals. This data contains NaN values across 40 parameters for t hours of hospitalization.

3. Methods

a. Imputation

Multivariate imputation by chained equations (MICE), sometimes called "fully conditional specification" or "sequential regression multiple imputation" has emerged in the statistical literature as one principled method of addressing missing data. Creating multiple imputations, as opposed to single imputations, accounts for the statistical uncertainty in the imputations. In addition, the chained equations approach is very flexible and can handle variables of varying types (e.g. continuous or

binary) as well as complexities such as bounds or survey skip patterns [3]. In order to find maximum correlation between the available parameters, the independent variable chosen is the sepsis label that has to be predicted post model training. In case of testing, since the label is unavailable, we assume it to be 0 in the beginning for all the patients depending upon the number of hours they have been hospitalized. The patients who have stayed longer than a threshold value are assumed to have a sepsis label of 1. This column is later dropped before processing the data set further.

b. Features

The 40 features available through the data set are 40,336 patients retrieved from two hospital systems. The dataset had a separate file for each patient's data, and for each patient an hourly record of 40 variables was given. These variables were divided into 3 categories – Vital Signs, Demographics and laboratory values. The final column at each time t contains a Sepsis Label value that is either 0 or 1. The missing values corresponding to each column are filled using the imputation method (3.b).

In order to increase accuracy of the model, we have incorporated certain features derived from the provided clinical parameters. These include shock index, BUN/Creatinine ratio, PaO2/FiO2, Respiratory Quotient (RQ). Calculation of – qSOFA, SOFA (Sequential Organ Failure Assessment), SOFA deterioration, SIRS.

Since the dataset only had PaCO2 value available, we computed the value of PaO2 using (Normally, the value of RQ = 0.8).

$$PaO2 = 713 * FiO2 - \frac{PaCO2}{RQ}$$

While Imputation of the clinical parameters , we considered the Normal clinical range of the parameters which would ensure that imputation does not overshoot or undershoot the parameters, and hence enhance the accuracy.

Table :*Range Values used in our model*

Feature	**Min Val.**	**MaxVal.**
Temp	34.0 C	42.0 C
Base Excess	-3.0	3.0
HCO3	20mMol/lt	30mMol/lt
FiO2	0.0	1.0
pH	7.3	7.5
PaCO2	28mmHg	48mmHg
AST	0.0mU/ml	40mU/ml
BUN	5.0 mg/dL	25.0 mg/dL
Phos	10.0U/L	135.0U/L
Calcium	7.0 mg/dL	12.5 mg/dL
Chloride	95 mg/dL	110 mg/dL
Creatinine	0.4 mg/dL	1.6 mg/dL
Bilirubin	0.0 mg/dL	0.6 mg/dL
Glucose	50mg/dL	120mg/dL
Magnesium	1.2mg/dL	2.8mg/dL

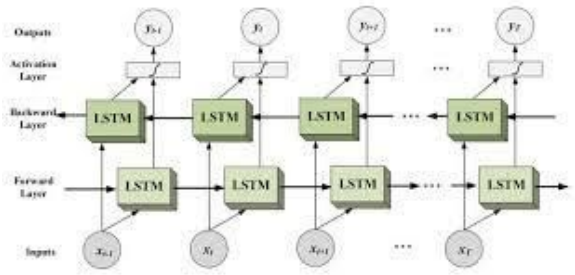
Potassium	3mMol/L	6mMol/L
Hct	31.0	56.0
HgB	6.5g/dL	19g/dL
WBC	2.0	15.0
Platelets	100/mu L	400/mu.L

c. Processing

In this section of method, the processing script takes the imputed data as input in the form of data frames and the derived features are added into each patient's data set sequentially. The output file then is of the size (no. of hours in the hospital, number of clinical and derived features). The processing of every patient helps in improving accuracy of the predicted labels because it is essential to understand that the onset of sepsis is dependent on various factors as defined in the SEPSIS-III definition. The previous definitions did not necessarily indicate the onset of sepsis. Since changes in white blood count, temperature and heart rate reflect inflammation, and are a result of some other infection than sepsis.

d. Model-Training

The model chosen is recurrent-neural network – Long Short-Term Memory (LSTM) Bi-directional model.



e. Prediction

The model produces Binary Classification of the Sepsis Output label , 0 indicating that the patient does not have Sepsis, and 1 indicating that the patient might be having Sepsis.

f. Regression

Linear Regression was used in order to map the predicted labels to Sepsis Score.

4. Early Detection of Sepsis

Sepsis is a systemic inflammatory state due to an infection, and is associated with very high mortality and morbidity. Early diagnosis and prompt antibiotic and supportive therapy is associated with improved outcomes. Our objective was to detect the presence of sepsis soon after the patient visits the ICU.

5. Results

As per the test set provided by GE Healthcare, we were able to generate the predicted scores and received the following scores as a result. The F-score and utility score were not up to the mark due to the skewness of data and the model will further be enhanced to take all these parameters into consideration.

	AUROC	AUPRC	Accuracy
0	0.5	0.054539	0.945461

6. Web App

The Web App is built using Streamlit, which is a python library. There are two Dashboards available for the Doctors to monitor all the patient records in a detailed manner, while the Patient Dashboard allows the patient to monitor

their personal records and get some insight on the various clinical parameters.

The doctor dashboard calculates the sepsis prediction when the patient data is loaded. IT also gives a graphical view of the vital parameters of a patient which helps the doctors and nurses get a quick insight.

The patient dashboard on the other hand is very minimalist and gives a brief clinical analysis on the vital parameters and tells the patient, if a visit for checkup is required.

7. Discussion

Bidirectional LSTMs are an extension of traditional LSTMs that can improve model performance on sequence classification problems. The idea of Bidirectional Recurrent Neural Networks (RNNs) involves duplicating the first recurrent layer in the network so that there are now two layers side-by-side, then providing the input sequence as-is as input to the first layer and providing a reversed copy of the input sequence to the second. In problems where all timesteps of the input sequence are available, Bidirectional LSTMs train two instead of one LSTMs on the input sequence. The first on the input sequence as-is and the second on a reversed copy of the input sequence. This can provide additional context to the network and result in faster and even fuller learning on the problem.

8. Future Scope

1. Integrating the model and the Webapp together, so that Real-time processing and visualization of data is possible at the same time.
2. Improvising the model further to make it possible to process even larger real world databases in a shorter amount of time

9. Acknowledgments

We would like to express our gratitude to our Mentors Dr. Anuradha Singh for helping us with the annotation of the Datasets and Dr. Tausif Diwan for guiding us in the process of model training. We would also like to Thank Evaluators and Mentors from GE Healthcare for their valuable feedback. Further, we would also like to thank all cl

10. References

1. *A computational Approach to early sepsis prediction, Jacob Cavert et al.*
2. *Survival in Septic Shock, Anand B*
3. *Targeted real-time early warning score for septic shock, Katherine E. Henry*
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540576
5. https://machinelearningmastery.com/develop-bidirectional-lstmence-classification-python-keras/
6. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4968574/
7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC30742