IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

EARLY PREDICTION OF SEPSIS FROM CLINICAL DATA USING AI & ML

Anusha[#1], Nation Raj K Halli[#2], Ruksar M[#3], Yashas A S[#4], Shylaja B[#5]

1234 BE Students, 5Project Guide

Computer Science and Engineering Department, Dayananda Sagar academy of Technology and Management, Bangalore Karnataka, India

Abstract: Sepsis is a severe medical condition caused by body's extreme response to an infection leading to tissue damage, organ failure, and even death. The emergence of advanced technologies such as Artificial Intelligence and machine learning, allowed faster exploration of advanced way to recognize sepsis case. Sepsis occurs when chemicals released in the bloodstream to fight an infection trigger inflammation throughout the body. This paper provides a description about the methods which we are using to predict and diagnose the early detection of Sepsis. Here in the project we are mainly considering the data features such as vital signs, ECG signals, Pulse signals, Temperature and some other vital features.

Keywords: Sepsis, Artificial Intelligence & Machine learning, Systemic Inflammatory Response Syndrome (SIRS), multiple organ damage

I. INTRODUCTION

Sepsis is life-threatening organ dysfunction. As an irregular host response. Septic shock is a subset of sepsis. Circulatory, cellular and metabolic abnormalities are at greater scale, and therefore the risk of death is getting larger. According to World Health Organization (WHO) data, 27,000,000 cases of sepsis develop annually.

Sepsis is life-threatening organ dysfunction. As an irregular host response. Septic shock is a subset of sepsis. Circulatory, cellular and metabolic abnormalities are at greater scale, and therefore the risk of death is getting larger. According to World Health Organization (WHO) data, 27,000,000 cases of sepsis develop annually. The cost of treating sepsis is estimated to be $16.7 billion per year, making sepsis one of the most expensive conditions to diagnose and treat. Multiple studies have shown that accurate early diagnosis and treatment, including sepsis bundle compliance, can reduce the risk of adverse patient outcomes from severe sepsis and septic shock. Earlier detection and more accurate recognition of patients at high risk of developing severe sepsis or septic shock provide a valuable window for effective sepsis treatments. However, the heterogeneous nature of possible infectious insults and the diversity of host response often make sepsis difficult to recognize in a timely manner. Studies that have attempted to target the risk-factors associated with sepsis onset reveal that sepsis is not a uniform condition. For example, oncology patients are nearly ten times more likely to develop sepsis when compared to patients with no cancer history, and patients with sepsis that developed during hospitalisation experience a 23% higher mortality rate than patients with community-acquired sepsis.

New definitions intended to improve the clinical recognition of sepsis have been proposed because the previous use of screening based on Systemic Inflammatory Response Syndrome (SIRS) criteria was found to be nonspecific. However, SIRS-based sepsis screening is still used in many clinical settings. In addition to SIRS, other rule-based patient decompensation screening tools commonly used for the detection or prediction of sepsis in clinical practice include the Sequential (Sepsis Related) Organ Failure Assessment (SOFA) score and the Modified Early Warning Score (MEWS). These methods generate risk scores by manual tabulation of various patient vital signs and laboratory results and have been validated for severe sepsis detection in a variety of studies. Efficacy of these scores is limited in part because they do not leverage trends in patient data over time, or correlations between measurements. Some scoring systems, such as SOFA, are not widely applicable outside of the ICU and often require laboratory values that are not rapidly available.

While several major EHR systems now have automated sepsis surveillance tools available to their clients, these alert tools are rules-based and suffer from low specificity.

In response to the need for externally validated machine learning-based sepsis screening methods, this study evaluates the performance of our MLA which predicts and detects severe sepsis using data extracted from patient Electronic Health Records. It is important that sepsis prediction MLAs have generalizability to different clinical settings and are capable of high performance scores on a diverse dataset, without requiring extensive retraining.

The machine learning algorithm:

Here we will be going to construct our classifier using gradient boosted trees, implemented in Python with the XGBoost package.

Predictions were generated from patient age and the binned values for the vital signs of systolic blood pressure, diastolic blood pressure, heart rate, temperature, respiratory rate and SpO2 at prediction time. Where appropriate, we also concatenated the differences in measurement values between those time steps. In the data matrices, each clinical feature thus represented between 3 and 5 columns. Values were concatenated into a feature vector with fifteen elements. All data processing was performed using Python software. An ensemble of decision trees was constructed using the gradient boosted trees approach, after which the ensemble made a prediction based on an aggregate of these scores. In this way, at prediction time, the gradient boosted tree ensemble was able to access trend information and covariance structure with respect to the time window. This procedure of transforming time series problems into supervised learning problems has also been detailed in our previous work. XGBoost controlled for expected class imbalance in the data. Minority class scaling was employed within the algorithm, where instances of the minority class were given weight inversely-proportional to their prevalence, which effectively trained the models on approximately balanced data. Tree branching was determined by evaluating the impurity improvements gained from potential partitions, and patient risk scores were determined by their final categorization in each tree. We limited tree branching to six levels, included no more than 1000 trees in the final ensemble, and set the XGBoost learning rate to 0.1.These hyperparameters were chosen to align with previous work and justified in the context of the present data with a coarse grid search using training data.

II. PROBLEM DEFINITION

With the existing classical tools of data analysis in the last decade, doctors are still facing a lot of limitations to detect sepsis early enough. Each time we detect sepsis patient one hour earlier, we get more 4-8% chance to save the patient's life. The proposed model will beat a lot of approaches. that has been tested in this challenge. The strength of the proposed model is the simplicity and speed of processing. It will be much more easier to implement it in the medical devices with lower capacity. So early detection of sepsis with increased accuracy will be the most vital factor that must be considered to implement the proposed model.

III.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. 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.

IV.DATA FEATURES

There are more than 40 health variables used to track the health situation of the patients in this challenge. The 40 columns are classified into three classes: vital signs, lab test, and static variables.

1)Vital signs:

Most of these features concern the main screening signal that would reflect continuous situation of the patient health. There are 8 vital signs provided in the data and we can cluster them into three main categories: ECG signals, Pulse signals, and Temperature.

• *ECG signals*: The main columns related are the heart rate, systolic blood pressure and diastolic blood pressure. The severity of sepsis is very correlated to the number of beats per minute. Many studies have shown that the more sepsis get worse, the more we observe ECG abnormalities. This can be explained by loss of excitability in cardiac tissue during the sepsis.

• *Pulse signals*: Respiration rate, O2Sat, and EtCO2 (End tidal carbon dioxide)

• *Temperature*: Symptoms of sepsis include: a fever above 101oF (38oC) or a temperature below 96.8oF (36oC) heart rate higher than 90 beats per minute. So, tracking the temperature is very important for the early detection task..

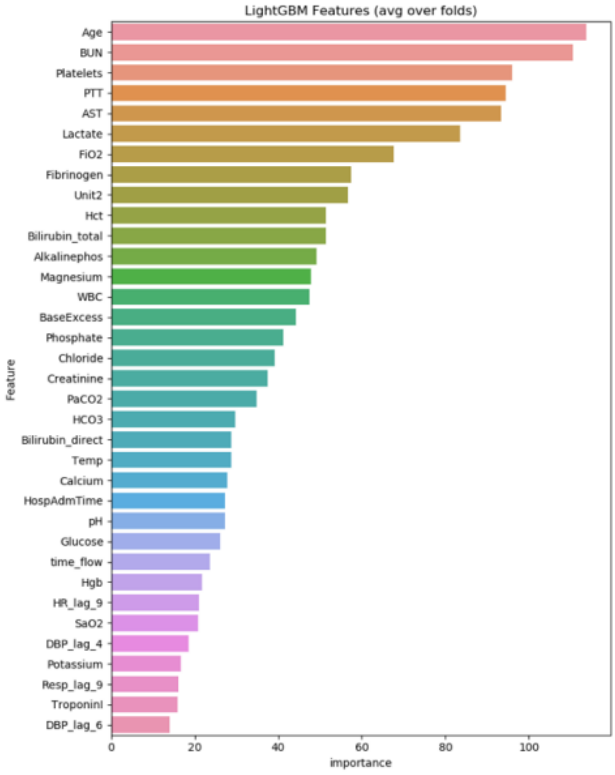
2)Lab test:

The role of laboratory tests is very significant for early detection of sepsis. Since the main definition of sepsis is the body's systemic inflammatory response to a bacterial infection. The spread of bacteria in the blood (bacteremia) make it a big indicator of sepsis infection. About 25 lab tests have been provided in the data for sepsis detection. These features include: White blood cells count, Blood urea nitrogen, Lactic acid, partial thromboplastin time, Leukocyte count, Platelets.

3)Demographic features:

There were some studies that have shown that sepsis mortality has a little thing to do with demographics. Especially age and length of stay. In our current model, we still find this is not the case. More results will be shown on the next paragraphs. The main demographic features in the data are: gender, age, and length of stay. Surprisingly, we found out in our model, that the age is a critical factor to predict sepsis risk.

Figure 1 presents the feature importance. From this figure, it can be seen that age is a crucial demographic factor to identify sepsis patients. Besides, lab tests come also on the top of the critical factors such as Blood urea nitrogen, platelets and partial thromboplastin time. These results are aligned very well with the biological description of sepsis.



V. PROCEDURE:

A. Data Pre-processing Pipeline:

Here we consider 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. 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 36-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.

MODELS/ALGORITHM	ADVANTAGES	DISADVANTAGES	ACCURACY
RNN (Recurrent Neural Network)	Faster training speed and higher efficiency. Lower memory usage.	The computation is slow. Data leak.	Got 0.05 as normal utility score with ranking of 69.
Random Forest and Logic Regression model.	Versatility- it can be used for regression as well as classification. Reliable – method to find out the variables that has impact on the topic of Interest.	Data must be sensitive to outliers .	0.74 - HR to systolic ratio occurred for 69% of overall ability i.e predictive ability.
Logistic regression (LR), support vector machines (SVM) with radial kernel, and logistic model trees (LMT)	SOFA scores are used for accessing the ICU patients QSOFA : for accessing Both the ICU and Non ICU patients.	SOFA score used as a mortality prediction model underperformed compared APACHE-IV and SAPS-II	The sensitive values for LR , SVM and LMT are 0.752,0.56 0.671 respectively.
Partially observed Markov decision process.	SERA algorithm, which uses both.	AI is capable of learning over time with pre-fed data and past experiences but cannot be creative in its way of approach.	**63%** effective at detecting sepsis early compared to the developer prediction of 77-83%.

Table1: Comparison of various model

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-hor case for each variable and each hospital admission.

B. Data processing and features:

Before the learning step, the clinical dataset we have for the challenge contain a lot of inconsistencies. For our model we have performed many preprocessing techniques to ensure the consistency of the data and create new features. The processing pipeline can be described as follow:

(1) Handling missing values: Several lab tests features have up to 90% of missing values. The data resolution is hourly based, and it is difficult to perform lab experiments every hour for every patient. Which explain the high ratio of missing values in lab test columns. However, the values in these columns are very important and removing them would not be the best idea.

In order to handle missing values:
1. We have performed interpolation which fill a missing value with the mean of the two consecutive non missing values.
2. The remaining missing values are filled using backward and then forward values

(2) Lag features:
The purpose of this features is to capture the long- and short-time dependencies. We performed different sliding windows with mean of the last 1,2,...,9 hours.

(3) Data binning:
In order to reduce some variance in the signal columns. We create new features that aggregate the signal value into ranges and intervals. We have based the data binning on the max and min of each signal. There is an option to define range limits using Random Forest. We expect to explore it in future work. (4) Count Encoding of Categorical features: applied mainly on the demographics features such as: age and gender

Table 1: shows the various model that are used for early prediction of sepsis along with their advantages and disadvantages with their accuracy

VI. CONCLUSIONS

This paper provides an overview of how different models are used to predict and diagnose the early detection of sepsis. In this analysis, we have taken four distinct methods to consideration that is RNN model, POMDP model, Random forest and Logistic Regression model, support vector machines (SVM) with radial kernel, and logistic model trees (LMT).
Traditional techniques are sometimes adequate , but not always. Hence we are considering the above mentioned models which will be helpful for handling large volume of data. More efforts can be made in developing a simple and efficient method to predict sepsis at early stage.

VII.ACKNOWLEDGEMENT

We would like to thank Dr Kavitha and Ms. Shylaja B for their guidance and support . We thank the MIMIC3 Benchmark Repo group for part of our data preprocessing code .

VIII. REFERENCES

[1] Mohammed Saqib, Ying Sha, May D. Wang
" Early Prediction of Sepsis in EMR Records Using Traditional ML Techniques and Deep Learning LSTM Networks"

[2] Roman Z. Wanga , Catherine H. Sunb , Philip H. Schroederc , Mawulolo K. Amekod, Christopher C. Mooree , Laura E. Barnesd "Predictive Models of Sepsis in Adult ICU Patients" Department of Computer Sciencea , University of Virginia, Charlottesville, VA rw5dc, chs8wr, phs5eg, mka9db, ccm5u, lbarnes@virginia.edu

[3] R Murat Demirer " Early Prediction of Sepsis from Clinical Data Using Artificial Intelligence "Uskudar University Industrial Engineering Department İstanbul, Turkey murat.demirer@uskudar.edu.tr Oya Demirer Arel University Electrical-Electronic Engineering Department Istanbul, Turkey oyademirer@arel.edu.tr

[4] Christopher W. Seymour et.al , "Time to Treatment and Mortality during Mandated Emergency Care for Sepsis", N Engl J Med. 2017 June 08; 376(23): 2235–2244

[5] Automated Mortality Prediction in Critically-ill Patients with Thrombosis using Machine Learning V. Danilatou, D. Antonakaki, C. Tzagkarakis, A. Kanterakis, V. Katos, T. Kostoulas Bournemouth University, Faculty of Science and Technology, Bournemouth.

[6] Andrew Critch, "Toward negotiable reinforcement learning: shifting priorities in Pareto optimal sequential decision-making", arXiv:1701.01302

[7] Early Prediction of Sepsis from Clinical Data: the PhysioNet/Computing in Cardiology Challenge 2019

[8] R. Vio and P. Andreani, "Spectral analysis of unevenly sampled signals: an effective alternative to the Lomb-Scargle periodogram", Astronomy &Astrophysics,2018

[9] J. L. Gall, S. Lemeshow, and F. Saulnier, "A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study," JAMA, vol. 270, no. 24, pp. 2957–2963, 1993.

[10] P. A. Fuchs, I. J. Czech, and Ł. J. Krzych, "The pros and cons of the prediction game: the never-ending debate of mortality in the intensive care unit," Int. J. Environ. Res. Public Health, vol. 16, no. 18, 2019.