Early Sepsis Detection from ICU Clinical Data

Garvit Kochar (2022185)   Aditya Singh Yadav(2022039)

Abstract

Sepsis is a severe disease that arises when the body's response to an infection injures its own tissues and organs. It may lead to shock, multi-organ failure, and death-specially if not recognized early and treated promptly. One of the global health reports says that around 27-30 million people develop sepsis each year among which around 7-9 million people die, i.e., 1 death every 3.9 seconds approximately. Also, the survivors may face life-long consequences. In the U.S., sepsis costs more than any other health conditions i.e., around 25billion$/year and a majority of these costs are for sepsis patients that were not diagnosed at admission [1]. Globally, these expenses are even more. Altogether, it's a major public health issue responsible for significant mortality, morbidity, and healthcare expenses. Early detection and antibiotic treatment of sepsis are critical for improving sepsis outcomes, where each hour of delayed treatment has been associated with roughly an 4-8% increase in mortality [2]. To help overcome this problem, we are doing a comparative study of different ML classification techniques on sepsis prediction so that we can use our findings for developing a system to predict sepsis as early as possible.

Keywords–Sepsis, Mortality, Morbidity, ML (Machine Learning).

1 INTRODUCTION

Sepsis is defined as life-threatening organ dysfunction caused by a dysregulated host response to infection. For clinical operationalization, organ dysfunction can be represented by an increase in the Sequential [Sepsis-related] Organ Failure Assessment (SOFA) score of 2 points or more, which is associated with an in-hospital mortality greater than 10%. Septic shock should be defined as a subset of sepsis in which particularly profound circulatory, cellular, and metabolic abnormalities are associated with a greater risk of mortality than with sepsis alone. Patients with septic shock can be clinically identified by a vasopressor requirement to maintain a mean arterial pressure of 65 mm Hg or greater and serum lactate level greater than 2 mmol/L (>18 mg/dL) in the absence of hypovolemia. This combination is associated with hospital mortality rates greater than 40%. In out-of-hospital, emergency department, or general hospital ward settings, adult patients with suspected infection can be rapidly identified as being more likely to have poor outcomes typical of sepsis if they have at least 2 of the following clinical criteria that together constitute a new bedside clinical score termed quickSOFA (qSOFA): respiratory rate of 22/min or greater, altered mentation, or systolic blood pressure of 100 mm Hg or less [3].

There are currently two important approaches in early recognition of Sepsis through Machine Learning:

i) Association Mapping in Time Series

ii) Classification of Time Series

In our research, we have done a comparative study of different Machine Learning classification models for early prediction of sepsis according to their features and behaviors.

The outline of our article is as follow:

- Section II - The study of related works in the literature

- Section III - Data Description and Assumptions

- Section IV - Procedure and Data Modelling

- Section V - Comparison and Conclusion

2 RELATED WORKS

2.1 Use of Association Mapping For early Sepsis Prediction

Currently, most shapelet discovery methods do not rely on statistical tests to verify the significance of individual shapelets. Therefore, identifying associations between the shapelets of physiological biomarkers and patients that exhibit certain phenotypes of interest enables the discovery and subsequent ranking of physiological signatures that are interpretable, statistically validated and accurate predictors of clinical endpoints [4]. However, these techniques are out of our research.

2.2 Use of ML Techniques For early Sepsis Prediction

Recent studies have focused on utilizing machine learning techniques for early prediction of sepsis onset. Horng et al. [5] used Support Vector Machines (SVM) to create an automated trigger at emergency department triage for sepsis clinical decision support. Desautels et al. [6] developed a dedicated machine-learning algorithm called InSight, which uses vital signs, age, Glasgow Coma Score (GCS), and pulse oximetry as

input features. Their proposed method demonstrated accuracy and area under the receiver operating characteristics (AUROC) curve of 57% and 0.74 in predicting sepsis 4 h before onset. The InSight algorithm has shown better results than expert scoring in predicting sepsis at hospitals. In 2018, Mao et al. [7] proposed and validated the InSight algorithm for the detection and prediction of three sepsis-related gold standards (SIRS, SOFA, and Modified Early Warning Score (MEWS)) using four new institutions' datasets and transfer learning to evaluate generalizability.
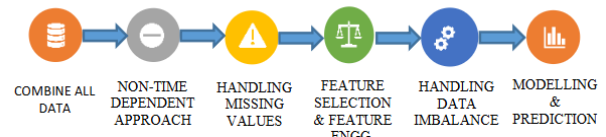
3 DATA DESCRIPTION AND ASSUMPTIONS

3.1 Dataset

Data used in the research is sourced from ICU patients in two separate hospital systems and is obtained from Physionet. The data will be split into 75% Training and 25 % testing set. Each hospital have 20,000 patients and hence 20,000 files. Available patient covariates consist of Demographics, Vital Signs, and Laboratory values.

Features:

- **8 Vital Signs:** Heart Rate, Temperature, Blood Pressure, Respiratory rate,

- **26 Laboratory Values:** Platelet Count, Glucose, Calcium etc

- **6 Demographics:** Age, Gender, Time in ICU, Hospital Admit time

- **1 Label:** 0 (Non-sepsis) and 1 (Sepsis)

3.2 Assumptions

- Combined dataset by appending all the patient files

- Total files: 43,765 .psv files

- Shape of original file: (1552287 * 41)

- The dataset is not time dependent.

2 approaches to solve it:

1. Add a time component and patient ID

2. Ignoring time component and consider each row independently

In our research we follow 2$^{\text{nd}}$ approach. **Reason:** Can predict sepsis without past patient data. More robust and need less resources.

4 PROCEDURE AND DATA MODELLING

Figure 1: Project Workflow

4.1 Combine all data

We have combined the dataset by appending all patient's .psv files and saved it as a .csv for later use.

4.2 Non-Time dependent approach

We have ignored the time component and consider each row independently. Using this approach, we can predict sepsis without past patient data. Also, it is more robust and need less resources.

4.3 EDA - Handling Missing Values

Most of Laboratory Data are having missing values (Fig 2).

Figure 2: Missing Data Bar Plot

To handle those missing values, we follow 2 approaches:

i. Remove features with missingness > 92%

ii. Categorically encode features to handle missingness.

4.4 Feature Selection & Feature Engineering

Developed 8 new features and are described:

a) **new_age:** has 3 categorical values – old, young and adult

b) **new_hr, new_temp, new_o2sat, new_bp, new_resp, new_map, new_fio2:** has 3 categorical values – normal, abnormal and missing

2

Next, we performed feature selection again on them and selected all above features, plus Gender, Hospital Admission Time and ICU length of Stay for further processing as a training set. Since, all these are categorically values, we have encoded them using One-Hot encoding so that it is easier to run ML algorithms.

4.5 Handling Data Imbalance

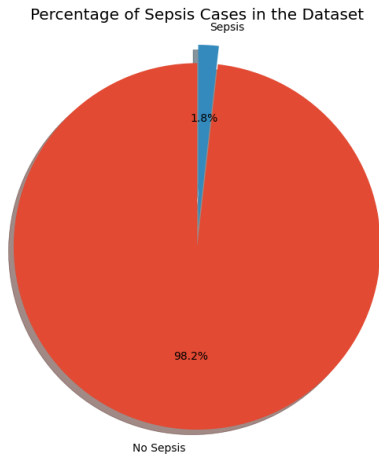Since, 98% of patients does not have sepsis and 2% have sepsis, it leads to a problem with Accuracy.

Figure 3: Data Imbalance

To deal with this data imbalance, we used oversampling technique (SMOTE). Fig. 4 shows the balance data.

Figure 4: Balanced Classes after SMOTE

4.6 Data Modelling and Predictions

Before modelling our data with different models, we split it into training (75% of data) and testing (25% of data) parts. For data modelling, we have used four different models, viz., Logistic Regression, KNN, Naïve Bayes, and Decision Trees. The prediction results of all these models are tabulated as below in Table 1.

5 COMPARISON AND CONCLUSION

5.1 Comparison

Fig. 5 shows the comparison of different models on the basis of accuracy on predicting sepsis.
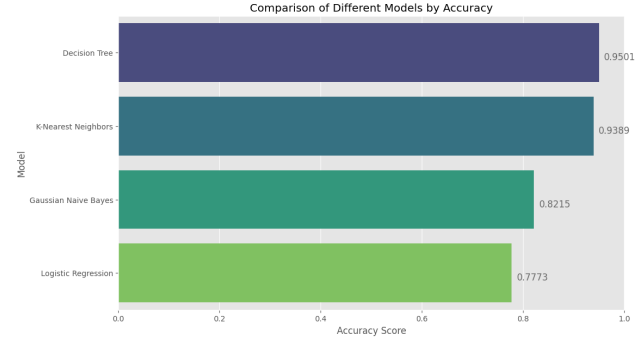
Figure 5: Comparison of Different models by Accuracy

5.2 Conclusion

Early prediction of sepsis is a challenging problem that, despite years of research and the introduction of several definitions, its onset is often indeterminate until later stages. In this study, we compare 4 different machine learning classification models viz., Logistic Regression, KNN, Naïve Bayes, Decision Tree.

Among the four models, Decision Tree comes out with the best results with accuracy of 96.8557 %. K-Nearest Neighbors Classifier also gives a good accuracy of 92.5891%. Our models were trained, tested, and evaluated using a dataset that was labeled based on the Sepsis-3 definition. Other Machine Learning algorithms like Support Vector Machine (SVM), XGBoost, and Deep Learning Techniques are left for the future scope of this research.

Sepsis has significant mortality among patients admitted to ICUs whose sepsis is not diagnosed or incorrectly marked. Therefore, using such tools and systems is potentially useful in clinical intervention, and it can significantly mitigate the economic burden, the ICU length of stay, morbidity, and mortality rate of sepsis. Our encouraging results indicate that Machine Learning and Deep Learning algorithms may aid physicians in predicting sepsis onset in ICUs after it has been trialed in a clinical setting.

References

[1] C. J. Paoli, M. A. Reynolds, M. Sinha, M. Gitlin, and E. Crouser, "Epidemiology and Costs of Sepsis in the United States," *Crit. Care Clin.*, vol. 34, no. 1, pp. 15-25, 2018.

[2] A. Kumar et al., "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic

Table 1: Model Performance Comparison

model	recall	precision	avg_precision	accuracy_score
Logistic_regression	0.595805	0.042986	0.032911	0.753119
KNeighbors	0.461407	0.114608	0.062608	0.925891
Decision_Tree	0.230561	0.191810	0.058121	0.968557
Gaussian Naive Bayes	0.576687	0.044128	0.033094	0.766736

shock," *Crit. Care Med.*, vol. 34, no. 6, pp. 1589-1596, 2006.

[3] 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.

[4] R. Kamaleswaran et al., "Association mapping for early sepsis prediction via statistically significant shapelet mining," *IEEE Int. Conf. Healthc. Inform.*, 2018.

[5] S. Horng et al., "Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning," *PloS One*, vol. 12, no. 4, p. e0174366, 2017.

[6] T. Desautels et al., "Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach," *JMIR Med. Inform.*, vol. 4, no. 3, p. e28, 2016.

[7] Q. Mao 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.

[8] S. Nemati et al., "An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU," *Crit. Care Med.*, vol. 46, no. 4, pp. 547-553, 2018.

[9] M. Scherpf, F. Gräßer, H. Malberg, and S. Zaunseder, "Predicting sepsis with a recurrent neural network using the MIMIC-III database," *Comput. Biol. Med.*, vol. 113, p. 103395, 2019.

[10] J. S. Calvert et al., "A computational approach to early sepsis detection," *Comput. Biol. Med.*, vol. 74, pp. 69-73, 2016.

[11] S. M. Lauritsen et al., "Early detection of sepsis utilizing deep learning on electronic health record event sequences," *Artif. Intell. Med.*, vol. 104, p. 101820, 2020.

[12] A. Rafiei et al., "SSP: Early prediction of sepsis using Fully connected LSTM-CNN model," *Int. J. Med. Inform.*, vol. 147, p. 104386, 2021.

[13] US National Library of Medicine, National Institutes of Health.

[14] Centers for Disease Control and Prevention.

[15] G. S. Martin, "Sepsis, severe sepsis and septic shock: changes in incidence, pathogens and outcomes," *Expert Rev. Anti Infect. Ther.*, vol. 10, no. 6, pp. 701-706, 2012.