Comparative Study of different Machine Learning Classification models on Early Sepsis Prediction

Darshan Mehta
M. Tech, Data Analytics, Department of Computer Applications
National Institute of Technology
Tiruchirappalli, India
darshanmehta006@gmail.com

Dr. U. Srinivasulu Reddy
Assistant Professor, Department of Computer Applications
National Institute of Technology
Tiruchirappalli, India
usreddy@nitt.edu

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).

I. 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

II. RELATED WORKS

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 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. Six vital signs, including heart rate, diastolic blood pressure, respiratory rate, peripheral capillary oxygen saturation, systolic blood pressure, and body temperature, were used as inputs to the algorithm. Nemati et al. [8] presented the Artificial Intelligence Sepsis Expert (AISE) algorithm for early prediction of sepsis. AISE works by coupling low-resolution electronic health records, high-resolution blood pressures, and heart rates as its inputs. For the 4- to 12-h ahead prediction of sepsis, AISE achieved an accuracy and AUROC in the range of 63%–67% and 0.83–0.85, respectively. Scherpf et al. [9] predicted sepsis onset by applying Recurrent Neural Networks (RNN) to the gold standard definition proposed by Calvert et al. [10]. The proposed RNN consists of two 40-unit hidden Gated Recurrent Unit (GRU) layers. Despite achieving higher AUROC than InSight [6], this model may not be suitable for clinical implementation, mainly due to the low specificity (only 47%), generating a large proportion of false negatives. Recently, Lauritsen et al. [11] developed a deep network model for early detection of sepsis. They trained the model by a diverse multicenter data set and have compared results with a standard multilayer feedforward neural network in the form of a multilayer perceptron (MLP). They reported an AUROC of 0.84 and 0.79 for 3 and 10 h before sepsis onset. Collectively, there is an increasing demand for the development of autonomous systems that can aid-in identifying the early stages of diseases. Recently, Alireza Rafiei et al. [12] also, developed a Smart Sepsis Predictor (SSP) for sepsis prediction in patients admitted to ICUs using fully connected LSTM- CNN model. The SSP performance is much higher than existing methods, achieving an area under the receiver operating characteristic curve (AUROC) of 0.89 and 0.92, 0.88 and 0.87, and 0.86 and 0.84 for 4 h, 8 h, and 12 h before sepsis onset, respectively.

III. Data description and assumptions

A. Dataset

Data used in the research is sourced from ICU patients in two separate hospital systems and is obtained from Physionet.

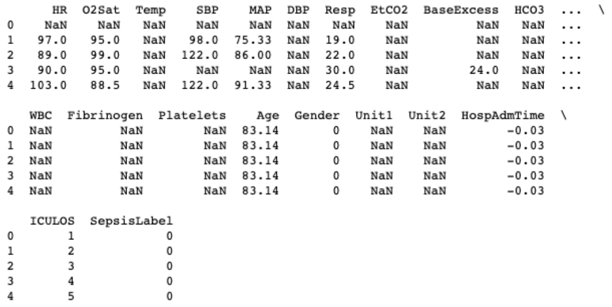
The data will be split into 75% Training and 25 % testing set.

The original data for each patient will be contained within a single pipe-delimited text file. Each file will have the same header and each row will represent a single hour's worth of data. Each hospital have 20,000 patients and hence 20,000 files.

Available patient co-variates 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)



```
      HR  O2Sat  Temp    SBP    MAP  DBP  Resp  EtCO2  BaseExcess  HCO3  ...  \
0    NaN    NaN   NaN    NaN    NaN  NaN   NaN    NaN         NaN   NaN  ...
1   97.0   95.0   NaN   98.0  75.33  NaN  19.0    NaN         NaN   NaN  ...
2   89.0   99.0   NaN  122.0  86.00  NaN  22.0    NaN         NaN   NaN  ...
3   90.0   95.0   NaN    NaN    NaN  NaN  30.0    NaN        24.0   NaN  ...
4  103.0   88.5   NaN  122.0  91.33  NaN  24.5    NaN         NaN   NaN  ...

   WBC  Fibrinogen  Platelets    Age  Gender  Unit1  Unit2  HospAdmTime  \
0  NaN         NaN        NaN  83.14       0    NaN    NaN        -0.03
1  NaN         NaN        NaN  83.14       0    NaN    NaN        -0.03
2  NaN         NaN        NaN  83.14       0    NaN    NaN        -0.03
3  NaN         NaN        NaN  83.14       0    NaN    NaN        -0.03
4  NaN         NaN        NaN  83.14       0    NaN    NaN        -0.03

   ICULOS  SepsisLabel
0       1            0
1       2            0
2       3            0
3       4            0
4       5            0
```

B. 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^{nd} approach.
Reason: Can predict sepsis without past patient data. More robust and need less resources.

IV. Procedure and Data Modelling

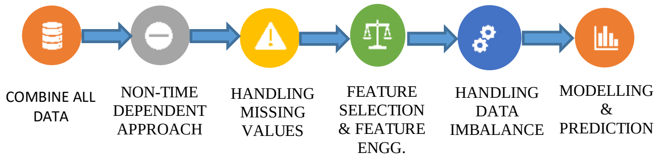


Fig.1

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.

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.

3. *EDA – Handling Missing Values*

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

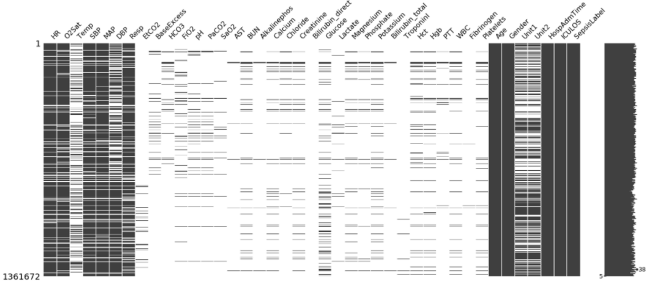


Fig.2

There are more than 90% of missingness in the dataset

To handle those missing values, we follow 2 approaches:

i. Remove features with missingness > 92%

ii. Categorically encode features to handle missingness.

4. *Feature Selection & Feature Engineering*

4.1 Feature Selection:

Two Approaches employed for Feature Selection:

i. Checked correlation of features contributing to the presence of Sepsis.

ii. Read health magazines and Research journals such as

- US National Library of Medicine [13]
- Centres for Disease Control and Prevention [14]
- Sepsis-The American Journal of Medicine [15]

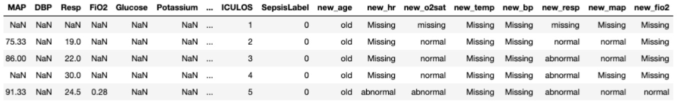
and filtered out the most named indicator of Sepsis Outcome: Heart rate, Pulse Oximetry, Body temperature, Blood Pressure (SBP, DBP), Mean Arterial Pressure, Respiration rate, Frac of inspired oxygen, Age, Gender, Hospital Admission Time and ICU length of stay.

4.2 Feature Engineering:

Developed 8 new features and are described:

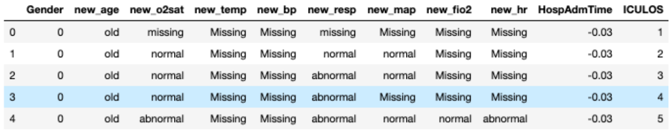
a) **new_age**: has 3 categorical values – old, young and adult

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



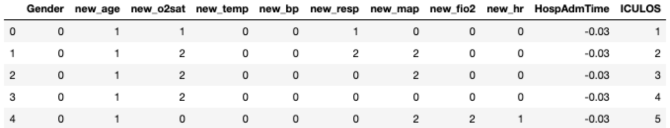
MAP	DBP	Resp	FiO2	Glucose	Potassium	...	ICULOS	SepsisLabel	new_age	new_hr	new_o2sat	new_temp	new_bp	new_resp	new_map	new_fio2
NaN	NaN	NaN	NaN	NaN	NaN	...	1	0	old	Missing	missing	Missing	Missing	missing	Missing	Missing
75.33	NaN	19.0	NaN	NaN	NaN	...	2	0	old	Missing	normal	Missing	Missing	normal	normal	Missing
86.00	NaN	22.0	NaN	NaN	NaN	...	3	0	old	Missing	normal	Missing	Missing	abnormal	normal	Missing
NaN	NaN	30.0	NaN	NaN	NaN	...	4	0	old	Missing	normal	Missing	Missing	abnormal	Missing	Missing
91.33	NaN	24.5	0.28	NaN	NaN	...	5	0	old	abnormal	abnormal	Missing	Missing	abnormal	normal	normal

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.



	Gender	new_age	new_o2sat	new_temp	new_bp	new_resp	new_map	new_fio2	new_hr	HospAdmTime	ICULOS
0	0	old	missing	Missing	Missing	missing	Missing	Missing	Missing	-0.03	1
1	0	old	normal	Missing	Missing	normal	normal	Missing	Missing	-0.03	2
2	0	old	normal	Missing	Missing	abnormal	normal	Missing	Missing	-0.03	3
3	0	old	normal	Missing	Missing	abnormal	Missing	Missing	Missing	-0.03	4
4	0	old	abnormal	Missing	Missing	abnormal	normal	normal	abnormal	-0.03	5

Since, all these are categorically values, we have **encoded** them using One-Hot encoding so that it is easier to run ML algorithms.



	Gender	new_age	new_o2sat	new_temp	new_bp	new_resp	new_map	new_fio2	new_hr	HospAdmTime	ICULOS
0	0	1	1	0	0	1	0	0	0	-0.03	1
1	0	1	2	0	0	2	2	0	0	-0.03	2
2	0	1	2	0	0	0	2	0	0	-0.03	3
3	0	1	2	0	0	0	0	0	0	-0.03	4
4	0	1	0	0	0	0	2	2	1	-0.03	5

5. *Handling Data Imbalance*

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

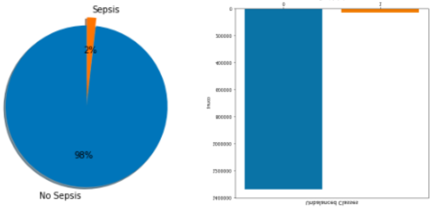


Fig. 3

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

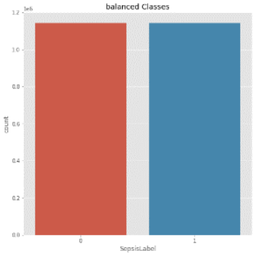


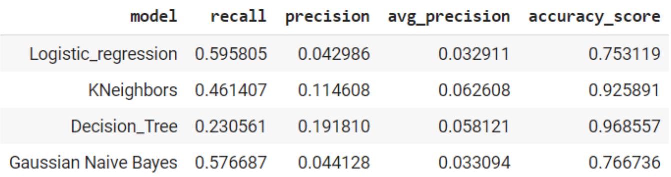
Fig. 4

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.



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

Table 1.

V. Comparison and Conclusion

A. *Comparison*

Fig. 5 shows the comparison of different models on the basis of accuracy on predicting sepsis.

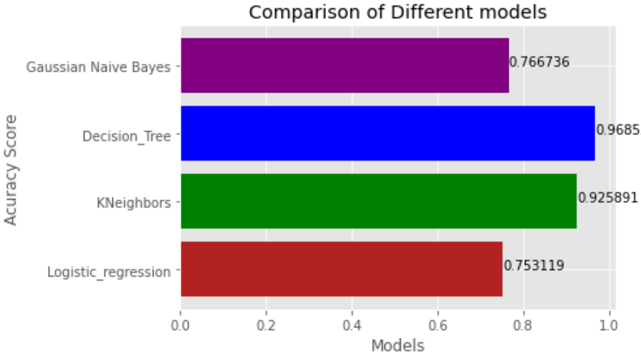


Fig. 5

B. *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] Epidemiology and Costs of Sepsis in the United States-Carly J Paoli, Mark A Reynolds, Meenal Sinha, Matthew Gitlin, Elliott Crouser

[2] Kumar *et al*., 2006; Seymour *et al*., 2017.

[3] Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)

[4] Association mapping for early sepsis prediction via statistically significant shapelet mining

[5] S. Horng, D.A. Sontag, Y. Halpern, Y. Jernite, N.I. Shapiro, L.A. Nathanson, Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning, PloS One 12 (2017).

[6] T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, D. Shimabukuro, U. Chettipally, M.D. Feldman, C. Barton, D.J. Wales, R. Das, Prediction of sepsis in the intensive care unit with minimal electronic health record aata: a machine learning approach, JMIR Med Inform 4 (2016) e28.

[7] Q. Mao, M. Jay, J.L. Hoffman, J. Calvert, C. Barton, D. Shimabukuro, L. Shieh, U. Chettipally, G. Fletcher, Y. Kerem, Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU, BMJ open 8 (2018), e017833.

[8] S. Nemati, A. Holder, F. Razmi, M.D. Stanley, G.D. Clifford, T.G. Buchman, An interpretable machine learning model for accurate prediction of sepsis in the ICU, Crit. Care Med. 46 (2018) 547–553.

[9] M. Scherpf, F. Gra¨ßer, H. Malberg, S. Zaunseder, predicting sepsis with a recurrent neural network using the mimic iii database, Comput. Biol. Med. 113 (2019) 103395.

[10] J.S. Calvert, D.A. Price, U.K. Chettipally, C.W. Barton, M.D. Feldman, J.L. Hoffman, M. Jay, R. Das, A computational approach to early sepsis detection, Comput. Biol. Med. 74 (2016) 69–73.

[11] S.M. Lauritsen, M.E. Kalør, E.L. Kongsgaard, K.M. Lauritsen, M.J. Jørgensen, J. Lange, B. Thiesson, Early detection of sepsis utilizing deep learning on electronic health record event sequences, Artif. Intell. Med. (2020) 101820.

[12]SSP: Early prediction of sepsis using Fully connected LSTM-CNN model by Alireza Rafiei, Alireza Rezaee, Farshid Hajati, Soheila Gheisari, Mojtaba Golzan

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

[14] Centers for Disease Control and Prevention

[15] Sepsis - The American Journal of Medicine