International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 12 | Dec 2022 www.irjet.net p-ISSN: 2395-0072

Sepsis Prediction Using Machine Learning

[1]Mohammad Ateeq, [2] Vineet K Joshi, [3]D Naga Praneeth, [4]Gugulothu Ravi

[4]Assistant Professor, Department of Computer Science and Engineering, SNIST, Hyderabad-501301, India
[1,2,3]B. Tech Scholars, Department of Computer Science and Engineering, SNIST, Hyderabad-501301, India

---

Abstract:- Sepsis is a blood poisoning condition that can increase the mortality risk in ICU patients when the body exhibits a dysregulated host response to an infection and results in organ failure or tissue damage. The expense of treating sepsis in hospitals is rising yearly as it develops into a serious health issue. Different techniques have been developed to monitor sepsis electronically, however in order to reduce the risk of death, it is crucial to forecast sepsis as soon as feasible before clinical reports or conventional techniques. The primary characteristics influencing the classifier's predictions have been outlined, making the model easier for medical professionals to understand. MLP Classifier has been used for the early diagnosis of sepsis, particularly in ICU patients been applied.This study demonstrates how machine learning algorithms, employing six vital signs taken from patient records over the age of 18, can reliably predict sepsis at the time of a patient's admittance into the intensive care unit. Sepsis may be predicted early, which can assist doctors administer supportive care and save mortality and medical costs. Unprecedented assessment measures have been obtained, and they can be very helpful in accurately and promptly predicting sepsis.

I. INTRODUCTION

Sepsis is a life-threatening condition that occurs when the body's response to an infection leads to inflammation and tissue damage throughout the body. It is a leading cause of death in hospitals and can progress rapidly if not properly diagnosed and treated. Early detection of sepsis is crucial for improving patient outcomes, and machine learning techniques have the potential to significantly improve the accuracy and speed of sepsis diagnosis. One approach to sepsis detection using machine learning is to use patient data, such as vital signs and laboratory results, to train a model to predict the likelihood of sepsis.

This data can be collected from electronic health records or other sources and may include demographic information, previous medical history, and current symptoms. The model can then be used to identify patients who are at high risk for sepsis, allowing healthcare providers to initiate early treatment and potentially prevent the progression of the condition.

Utilizing machine learning as a different strategy for analysis. patterns in patient data and identify early warning signs of sepsis. This can be done by analysing trends in vital signs over time or by looking for changes in biomarkers that are indicative of sepsis. By identifying these early warning signs, healthcare providers can intervene before the condition becomes severe and potentially save lives. Overall, the use of machine learning in sepsis detection has the potential to significantly improve patient outcomes by enabling early diagnosis and treatment. But it's crucial to pay close attention to the ethical implications of using machine learning in healthcare and ensure that the technology is used responsibly and in a way that benefits patients.

The development of artificial intelligence technologies has made it possible to diagnose sepsis early on. These techniques have been created to study and anticipate the health of the human body and acquire accurate prescription information to support doctors in making rapid and effective decisions. They integrate electronic medical records, medical imaging, pathophysiology, and other data.

II. \LITERATURE SURVEY

In several medical specialties, an AI-based diagnostic system has been proven to be efficient. In the domain of sepsis diagnosis, prognosis, and therapy machine learning algorithms used include supervised learning and reinforcement learning. For example, Beck et al. develop the C-Path (Computational Pathologist) system to automatically diagnose breast cancer and determine the likelihood that patients will survive by looking at breast tissue imaging.

The primary two difficulties in the current study involve the use of different physiological indicators and modelling efficient machine learning algorithms for the

diagnosis, prognosis, and treatment of sepsis. Similarly, Additionally important for predicting sepsis in advance is choose appropriate variables and design valuable algorithms in the clinical setting. The model's input variables are physiological indications, and the model's output parameter is the patient's condition would develop sepsis many hours later. In particular, input data include vital indicators such as heart rate, oxygen saturation, and body temperature; biomarkers such as procalcitonin and interleukin-6; laboratory values such as bicarbonate and creatinine; and demographic factors such as gender and age. In Most of the categories, a large number of missing values, such as those in MIMIC III (Intensive Care Medical Information Market Database), which has been utilised in several research. Most studies omit variable with a large number of missing values from predictors, resulting in the loss of useful information. To fill in missing information, some research utilize imputation and mean filling methods, although this might lead to selection bias or confounding factor mixes. The data preparation approach must be examined in light of the features of various data sets.

The machine learning algorithms generally include support vector machines, gradient boosting trees, random forests, Logistic regression, and neural networks. Among them, MLP Classifier have shown good performance. The A model with improved prediction capabilities will be examined further rand improved results for clinical service. so, that Early sepsis choices can be better made by physician diagnosis.

The research have performed well in the area of sepsis prediction. The quantity of data utilised in these studies is, however, reduced because the majority of the missing values are handled by direct deletion or forward filling, and the model's explanatory power is therefore constrained. The following arguments in detail explain why it is difficult to implement these techniques in clinical settings. A comprehensive data set is lacking. Researchers make use of information from various patient groups, such as the MIMIC public database or other unbiased sources of hospital data. They choose different clinical factors to create their models, and the size of the data also varies considerably. Different clinical criteria for sepsis and assessment indicators are used as prediction settings' premise and indicators.

III. METHODOLOGY

1. DATA SET:-Dataset contains data of 36 thousand patients. Each patient is represented by 41 features.

	HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess	HCO3	...	WBC	Fibrinogen	Platelets	Age	Gender	Unit1	Unit2	HospAdmTime	ICULOS	SepsisLabel
0	0.0	0.0	0.00	0.0	0.00	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	83.14	0	0.0	0.0	-0.03	1	0
1	97.0	95.0	0.00	98.0	75.33	0.0	19.0	0.0	0.0	0.0	...	0.0	0.0	0.0	83.14	0	0.0	0.0	-0.03	2	0
2	88.0	99.0	0.00	122.0	86.00	0.0	22.0	0.0	0.0	0.0	...	0.0	0.0	0.0	83.14	0	0.0	0.0	-0.03	3	0
3	90.0	95.0	0.00	0.0	0.00	0.0	30.0	0.0	24.0	0.0	...	0.0	0.0	0.0	83.14	0	0.0	0.0	-0.03	4	0
4	103.0	88.5	0.00	122.0	91.33	0.0	24.5	0.0	0.0	0.0	...	0.0	0.0	0.0	83.14	0	0.0	0.0	-0.03	5	0
5	110.0	91.0	0.00	0.0	0.00	0.0	22.0	0.0	0.0	0.0	...	0.0	0.0	0.0	83.14	0	0.0	0.0	-0.03	6	0

Fig 1. Data Set

2. FEATURE SELECTION:- In the mean processing method, 41 variables were determined to participate in the training model, including (a) vital signs indicators (HR, O2Sat, Temp, SBP, MAP, DBP, Resp), (b) laboratory variables (HCO3, pH, PaCO2, AST, BUN, AlkalinePhos, Chloride, Creatinine, Lactate, Magnesium, Potassium,

Bilirubin_total, PTT, WBC, Fibrinogen, Platelets), and (c) demographic indicators (Age, Gender).

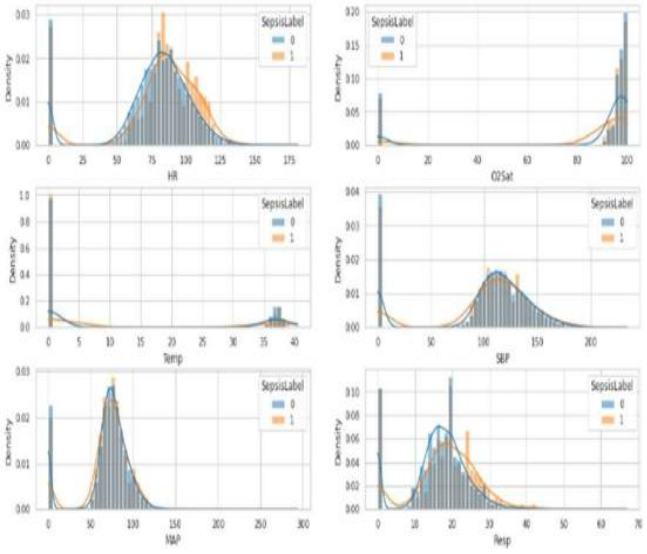

Fig 2. Vital Signs(a)

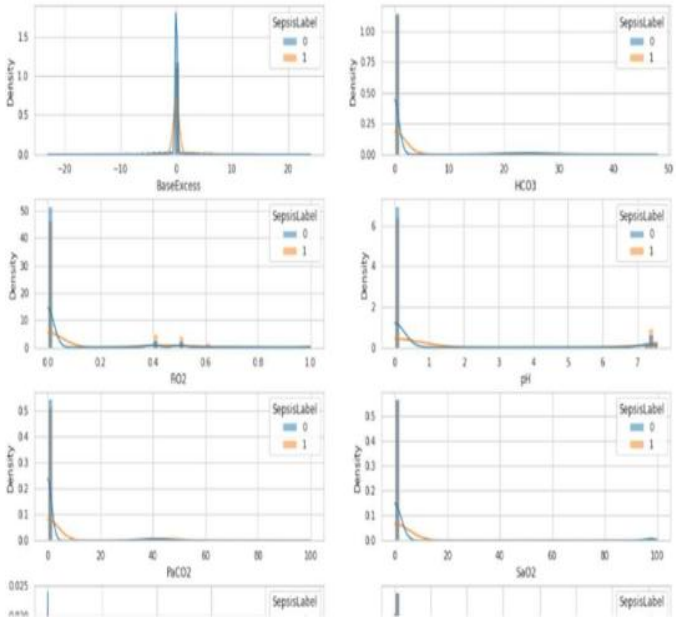

Fig 3. Laboratory Variables(b)

The variables that had missing proportions of greater than 98% were eliminated. The demographic metrics HospAdmTime (the interval between hospitalisation and ICU) and ICULOS (the interval between ICU hospitalizations) have been removed. HospAdmTime displays various numerical levels depending on the health of various patients, which may be connected to sepsis's extended incubation period. Since the primary goal of this study is to develop guidelines for predicting early sepsis from changes in certain physiological data, these variables are omitted. According to the statistics entered, patients with sepsis have a significant fatality rate. They frequently require prolonged ICU care, and the ICULOS value is typically excessively high. Contrarily, patients without sepsis typically get treatment in the ICU for a brief period of time before being discharged after their health has improved, resulting in a low ICULOS rating.

The variable ICULOS is eliminated because the variation in ICULOS value is caused by the different nature of the sickness situation, which is against the causal sequence of early sepsis anticipated from physiological data.

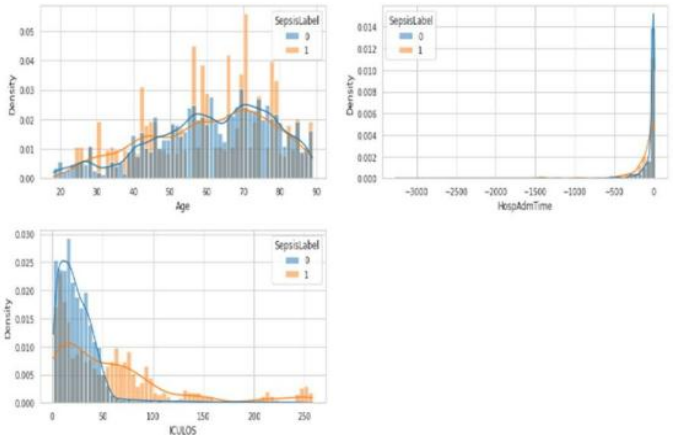

Fig 4: Demographics Indicator(c)

3. METHOD TO PREDICT SEPSIS:- MLPs are neural network models that work as universal approximators, i.e., they can approximate any continuous function, MLPs are composed of neurons called *perceptions*. a perceptron receives n features as input ($\mathbf{x} = x1, x2, \ldots, xn$), and each of these features is associated to a weight. Input features must be numeric. So, nonnumeric input features have to be converted to numeric ones in order to use a perceptron.

$$u(\mathbf{x}) = \sum_{i=1}^{n} w_i x_i.$$

The result of this computation is then passed onto an activation function f, which will produce the output of the perceptron. In the original perceptron, the activationfunction is a step function:

$$y = f(u(\mathbf{x})) = \begin{cases} 1, & \text{if } u(\mathbf{x}) > \theta \\ 0, & \text{otherwise,} \end{cases}$$

Thus, we can see that the perceptron determines whether $w1x1 + w2x2 + \cdots + wnxn - \theta > 0$ is true or false. The equation $w1x1 + w2x2 + \cdots + wnxn - \theta = 0$ is the equation of a hyperplane.

3.1 IMPROVEMENT OF PREPROCESSING

WARNING PERIODS:- The 6-hour warning period for each patient is immediately integrated to produce a single observation in the mean processing approach discussed above, however the model's performance in terms of prediction may not be sufficient. Further research is being done to determine whether or not higher performance results from segmentation time windows that are finer or denser. In order to calculate the mean vector, the 6-hour warning period is split into 2- or 3-hour time windows, and the mean processing procedure for the safe period and illness period is left alone. Figure 5displays further information. The improvement is compared with the original models of their generalisation capabilities based on MLP's Classifier and new datasets for training models are created in the same manner.

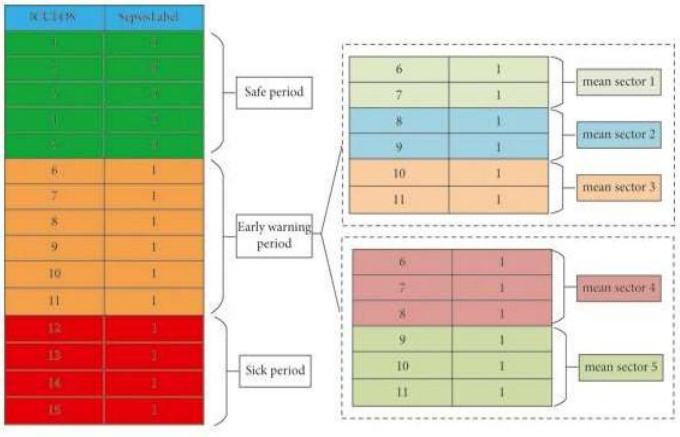

Fig 5. MLP Mean Calculator

4. FEATURE IMPORTANCE:- For the feature importance score, we take the MLP Classifier algorithm in the mean processing method as an example; the top 10 variables with feature importance scores are Temp, O2Sat, Resp, HR, Age, SBP, MAP, PTT, $PaCO_2$, and Potassium as shown in Figure 6. This means that these variables play an important role in predicting the risk of sepsis.

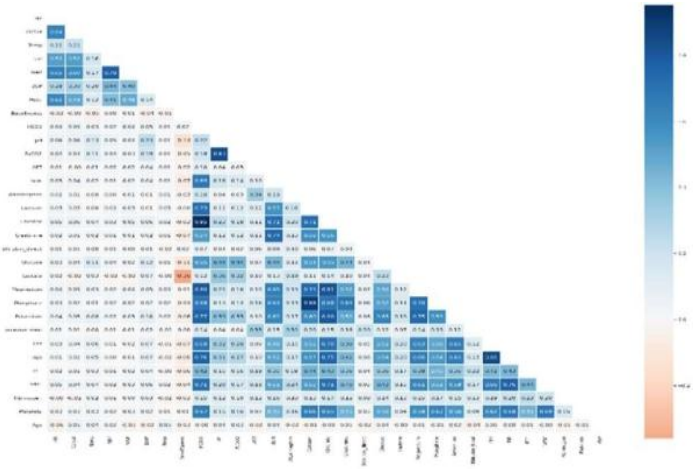

Fig 6: Feature Importance

IV. UML DIAGRAM

A sequence diagram (UML) is a visual representation of the flow of messages between objects during an interaction. A group of objects connected by lifelines and the messages they exchange throughout the course of an interaction make up a sequence diagram.

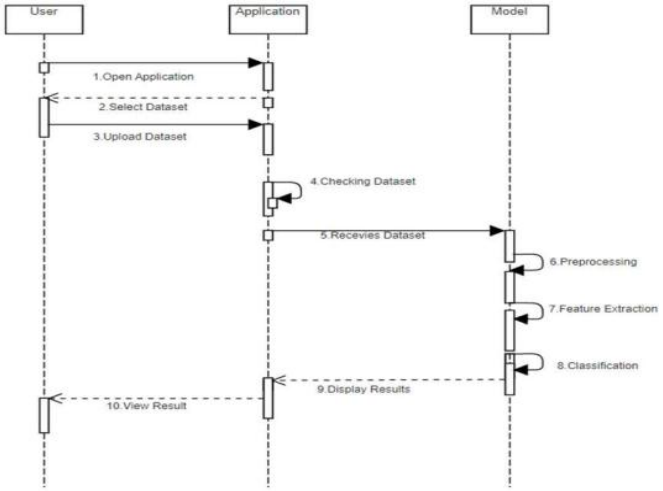

Fig 7. Sequential Diagram

V. RESULTS

1. MODEL PERFORMANCE:- In the mean processing method (method1), the MLP Classifier algorithms differ in performance. The MLP's algorithm has a Log loss rate of 0.15, with better distinction performance between 0-1 categories.

The log-loss shows how accurate the forecast was the likelihood corresponds to the matching real or true value(0 or 1 in case of binary classification). The more the predicted the further the probability deviates from the actual value the log-loss value.

$$-\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1-y_i) \cdot log(1-p(y_i))$$

The Accuracy score is calculated by dividing the number of correct predictions by the total prediction number.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

```
Accuracy: 94.7885%
Log Loss: 0.15868214105819684
```

2. MLP ALGORITHM:-This model optimizes the log- loss function using LBFGS or stochastic gradient descent.

class sklearn.neural_network.**MLPClassifier**(hidden_layer_sizes=(100,), activation='relu', *, solver='adam', alpha=0.0001, batch_size='auto', learning_rate='constant', learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True, random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e- 08, n_iter_no_change=10, max_fun=15000)

Training of variables in algorithm for classification.

```
clf.fit(X_train, Y_train)
name = clf.__class__.__name__
```

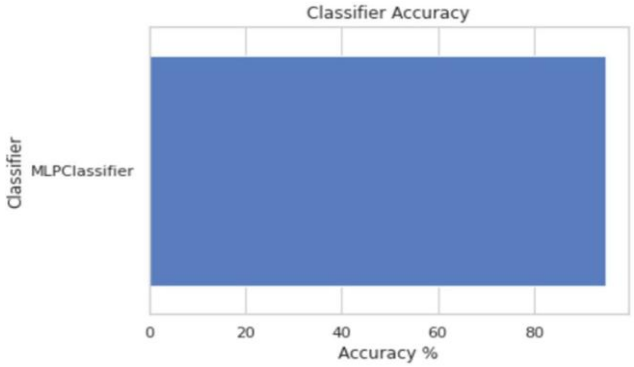

Fig 8. Accuracy

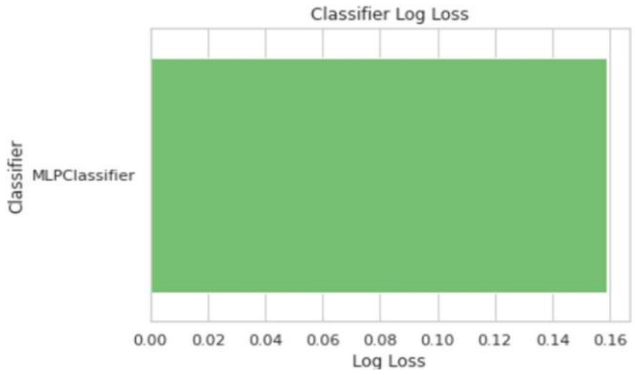

Fig 9. Log loss

VI.CONCLUSION

It is possible to use machine learning techniques for the detection of sepsis, a serious and potentially life threatening condition that can arise as a complication of infection. Sepsis is a complex and dynamic process thatcan be difficult to diagnose, and early identification and treatment are critical for improving patient outcomes. Several machine learning techniques exist that have been explored for the detection of sepsis, including supervised learning methods such as decision tree algorithms and support vector machines, as well as unsupervised learning methods such as clustering and anomaly detection.

One potential advantage of using machine learning for sepsis detection is the ability to analyse and interpret large amounts of patient data, including electronic health records, laboratory results, and vital signs, in order to identify patterns and correlations that may be indicativeof sepsis.

Overall, the use of machine learning for sepsis detection has the potential to improve the accuracy and timeliness of sepsis diagnosis, which can help to improve patient outcomes and reduce healthcare costs. However, more research is needed to fully understand the effectiveness and limitations of these approaches and to optimize their performance in real-world settings.

VII.REFERENCES

[1] K. E. Rudd, S. C. Johnson, K. M. Agesa et al., "Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study," *The Lancet*, vol. 395, no. 10219, pp. 200–211, 2020.

[2] L. Su, Z. Xu, F. Chang et al., "Early prediction of mortality, severity, and length of stay in the intensive care unit of sepsis patients based on sepsis 3.0 by machine learning models," *Frontiers in Medicine*, vol. 8, 883 pages, 2021.

[3] K. C. Yuan, L. W. Tsai, K. H. Lee et al., "The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit," *International Journal of Medical Informatics*, vol. 141, Article ID 104176, 2020.

[4] J. E. García-Gallo, N. J. Fonseca-Ruiz, L. A. Celi,and J. F. Duitama-Muñoz, "A machine learning-based model for 1 year mortality prediction in patients admitted to an intensive care unit with adiagnosis of sepsis," *Medicina Intensiva*, vol. 44, no. 3, pp. 160–170, 2020.

[5] J. Kim, H. Chang, D. Kim, D. H. Jang, I. Park, and K. Kim, "Machine learning for prediction of septic shock at initial triage in emergency department," *Journal of Critical Care*, vol. 55, pp. 163–170, 2020.

[6] A. H. Beck, A. R. Sangoi, S. Leung et al., "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," *Science Translational Medicine*,vol. 3, no. 108, 108ra113 pages, 2011.

[7] D. J. Stekhoven and P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[8] R Core Team, *R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing*, R Core Team, Vienna, Austria, 2014.

[9] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, vol. 27, no. 4, pp. 857–871, 1971.