

Sepsis Detection using Neural Networks

S. Babu

Faculty at the Department of CSE
V R Siddhartha Engineering College
Vijayawada, Andhra Pradesh
sbabu@vrsiddhartha.ac.in

Lalitha Anupama Annavarapu

Department of CSE
V R Siddhartha Engineering College
Vijayawada, Andhra Pradesh
anupama9618@gmail.com

Lakshmi Lahari Appala

Department of CSE
V R Siddhartha Engineering College
Vijayawada, Andhra Pradesh
lakshmi333lahari@gmail.com

Abstract—The immune system in our body, which is always working to stop the infection from entering, is what causes sepsis. The massive amount of synthetic compounds released into the blood at this stage produces widespread discomfort. An immune reaction to chemicals becomes out of balance which results in sepsis, bringing triggering reactions that can harm various organs. In the patient's case, an essential element is the feasibility of identifying sepsis disease development as the outcome. This work's main objective is to construct and train a Multi-Layer perceptron (MLP) model utilizing data that is currently accessible in the form of digital healthcare data and forecasts whether the result will be classified as a person having sepsis or not based on health records. Comparing the MLP's accuracy against that of many other models or classifier algorithms such as Gradient Boosting, Ada-Boost, and Linear Discriminant Analysis using log loss and metrics is the secondary objective. The proposed system comprises three stages. They are Pre-processing, Feature Importance, and Classification. The datasets used are PhysioNet Challenge data and the MIMIC-III dataset. In the pre-processing stage, the missing values are removed and a resampling technique is applied to the dataset. Feature importance is done with the help of the Xgboost framework to identify the important features. Then models are built using different classifiers to find out the best classifier with high accuracy and detect the sepsis disease in minimal time.

Keywords—Sepsis, Classifier algorithms, Multi-layer perceptron, Gradient Boosting, Linear Discriminant Analysis

I. INTRODUCTION

Sepsis is a serious disorder that can result in organ failure, tissue damage, or even death when the body reacts to toxins. The body often produces natural synthetics or chemicals in our circulatory system to lessen or balance the infection that is already present in our bodies [6]. Sepsis can result in organ failure, tissue damage, or even death. Sepsis is mostly caused by infection, which can affect anyone. Due to their weakened immune systems, infants, expectant mothers, older adults, and those with diabetes, lung infections, renal disease, or cancer are particularly affected. In terms of morbidity, costs associated with medical care, and mortality, this condition represents

a significant public health threat. Antibiotic treatment combined with early disease detection can improve the results. Even though several professional care associations have developed novel techniques for identifying sepsis, the crucial need for prompt recognition and treatment has gone unmet. If it is discovered in its early stages, it can be treated. Numerous studies have demonstrated that inadequate sepsis diagnosis and care can result in high fatality rates. As soon as a patient enters the emergency room or intensive care unit for treatment, our main goal is to diagnose sepsis. Even COVID-19 patients have more chances to get affected by sepsis disease.

Fever, defined as a temperature exceeding 38°C or even below the usual range of 36°C, is one of the signs of sepsis. Since the patients have low blood pressure, sepsis is treated with IV antibiotics to fight the infection. Controlling blood sugar levels involves injecting insulin into the blood. When blood-borne bacteria are the source of inflammation, corticosteroids are used to reduce it. Painkillers have also been used to ease discomfort during medical procedures. Sometimes, at extreme stages there are chances for damaged organs and tissues, in such cases, the Dialysis treatment is given for sustaining purposes.

Laboratory tests are used to detect sepsis in patients. Sepsis must be detected early to prevent mortality. Sepsis can also result from viral and bacterial infections. Urine testing, endotoxin tests, and blood cell count tests are some of the several blood test types.

Recurrent Neural Networks (RNN), Radial Basis Function Neural Networks, and Long Short-Term Memory (LSTM) is the current systems for sepsis diagnosis (RBFNN). An RNN variant that can recognize order dependence in sequence problems is the LSTM network. The disadvantage of LSTM is that training is a very challenging undertaking. Long sequences of processing are not possible [5]. The RNN model aims to identify the sequential properties of data and then use patterns to forecast future events. Due to its repetitive nature, the computation is inefficient and only 82% accurate [12]. Radial basis functions (RBFs) are used as activation functions in RBFNNs, which are Artificial Neural Networks (ANNs).

The disadvantage of RBFNN is that categorization takes longer.

II. LITERATURE SURVEY

Liu et al. [1] proposed a prediction model to find out sepsis disease built on the XGBoost framework. The developed model created several new features using the “time-dependent feature construction” method and “statistical count construction” method, which have better features to depict sepsis and considerably increase the proposed model's accuracy. Additionally, a suggested, “modified objective function” for training an XGBoost-based model was made.

Zhao et al. [2] proposed the algorithms used for machine learning to forecast early sepsis, six hours in advance, XGBoost processing method and LightGBM method are used to build two types of processing methods: the “mean processing approach” and the “feature generation” method. A variety of features, such as statistical strength, medical, and window features, are combined to create the feature generation methods. The “Miceforest Multiple Interpolation” method is used to solve situations with a lot of missing data. The feature-generating approach performs more than the mean processing method, according to the

results. Both the XGBoost algorithm and LightGBM algorithm have very good prediction performance.

Wang et al. [3] described that one of the important causes of morbidity and mortality in contemporary intensive care units is sepsis. Real-world medical data, however, frequently have a complex structure and a high level of uncertainty such as null values or missing values which are imbalanced data). To represent the intricately structured medical data for predicting reliable sepsis, a novel predictive framework using a "Multi-Branching Temporal Convolutional Network" (MB-TCN) is proposed in this research. According to experimental findings, MB-TCN performs better than approaches already in use in everyday practice.

Baldominos et al. [4] demonstrated that there are many different types of medical diseases known as infections that have a high risk of illness and even death. Similar to other diseases, having an early diagnosis can assist in giving patients the right care to lessen the harm the condition causes or to quarantine them to reduce the chance of spread. Computational intelligence can be helpful in this situation to foresee the possibility of infection and diseases in patients, setting off early alarms that can help medical teams react as quickly as feasible. The majority of works employ machine learning approaches, including random forest, logistic regression, SVM, and naïve Bayes the frequently compatible browser-based application that may be used without the most specialized hardware. Tensorflow in python is used to train all networks.

Al-Mualemi et al. [5] developed and put into use a machine learning (ML) type-based technique that can forecast cases of septic shock and severe sepsis and evaluate its impact on patient care and medical practice. By establishing the presence of toxins, the results of The laboratory tests act as the major early signal of septic shock. "Long short-term memory" (LSTM) CNN architecture is the primary deep learning network employed. Reducing the areas that need to be observed for clinicians' or doctors' decision-making.

Islam et al. [6] demonstrated that for identifying and forecasting sepsis patients, machine learning models for prediction outperformed the scoring systems for sepsis already in use, including SIRS, MEWS, SOFA, and qSOFA. Using machine learning algorithms to predict sepsis patients could help doctors monitor and take preventive measures to improve the status of patients. Additionally, it would help identify individuals who need assistance for medical the most, cut down on healthcare resource waste, and boost the necessary sensitivity or specificity, which would lower the number of false alarms.

Feng et al. [7] proposed that their study was to specifically look into the relationship between TTE and 28-day mortality in that demographic. The database MIMIC -III was used to distinguish between sepsis patients who had had TTE and those who had not. To confirm the reliability of findings, a variety of statistical techniques were used, including "multivariate regression" and the "inverse-probability-weighting" model.

Chen et al. [8] proposed an approach that has a recognition rate of 81.6%, a sensitivity of 89.57%, and a specificity of 65.77%. The "chaotic fruit fly optimization" was proposed to improve the measure of the kernel extreme learning machine. Before building the classification model,

they carried out the feature selection has done using the random forest to improve the diagnosis accuracy and determine the most significant biomarkers. Sepsis was successfully diagnosed using the final constructed model.

Kam et al. [9] aimed to construct early-stage sepsis detection models using deep learning algorithms and methodologies and to assess the viability and improvement of the novel deep learning approach against that of the regression method utilizing traditional "temporal feature extraction". They achieved enhanced performance with input to output layer that is fed forward in only direction neural networks utilizing extended short-term memory and enhanced performance with deep neural networks by comparing them to reference features.

Scherpf et al. [10] examined adult patients who were admitted to the critical care unit but did not meet the criteria for sepsis at the time of admission (from the MIMIC III database). To evaluate the performance of the prediction, look at the sequence length provided to the machine learning algorithms at various times before the beginning of sepsis. Additionally, the effect of the definition of sepsis onset is looked into. In comparison to similar efforts in the field, they evaluated the model using a rather large and thus more representative patient sample. The research proved that, when comparing prediction performance, a recurrent neural network outperforms insight. But further research is required for better results.

Yuan et al. [11] proposed an ensemble model (lightgbm, xgboost, and random forest) that incorporated boosting and bagging tree models were created to predict sepsis. On a subset of the inner test data, they compared the outcomes of the ensemble model and calculated each model's evaluation metrics. AUC of 0.792 and ACC of 0.727 were the best performance numbers attained offline. Finally, the suggested model was assessed using all available test sets. The usefulness score for LightGBM's single model was merely - 0.036. In comparison to a single tree-based model, the ensemble model performed better since it made use of the pre-treatment data.

Ying et al. [12] intended to find, validate, and test potential machine-learning algorithms for the early prediction of sepsis. They performed on decision trees, random forests, ads, gradient-boosted trees, and multilayer perceptrons will be used to develop prediction models for the 15 hours before the onset of sepsis. But the models are less accurate.

III. PROPOSED SYSTEM DIAGRAM

Designing and creating a method to identify sepsis using Multi-Layer Perceptron is the main objective. In our suggested system diagram which is shown in Fig. 1 the three main processes for sepsis detection are pre-processing, feature importance, and classification. The data is pre-processed using a resampling technique at the pre-processing stage. The XGBoost Algorithm is used to complete the next

feature importance stage. The results of detection are provided by the suggested classifier, known as the Multi-Layer Perceptron. Python is used for the project's implementation. Accuracy and log loss are used to measure classifier performance. To determine which classifier is best at identifying the disease, the measured classifier is compared to the others.

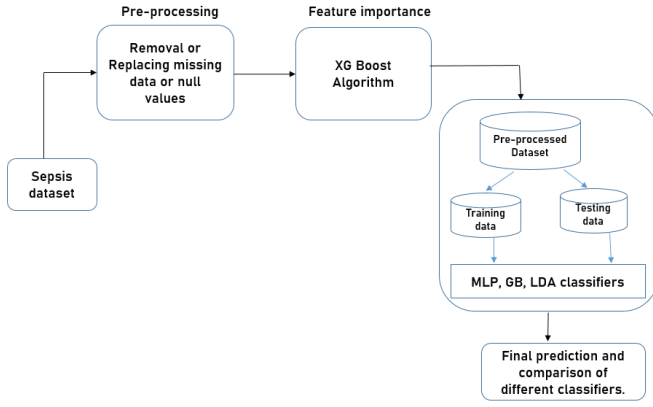


Fig. 1. Methodology of the proposed model

A. Dataset Preparation

Data is collected from the Physionet Challenge dataset where the patients in the ICUs of three different hospitals are the source of the dataset. Around 30,247 patients' clinical data from three separate hospitals have been gathered. There are 39 measurements of crucial signs, demographic information, and laboratory results in each person's clinical data. Crucial Signs include things like the heart rate, temperature, and mean arterial pressure. Platelet count, glucose, and calcium are a few of the laboratory values. Age, gender, and other variables are included. Sepsis is identified by the labels 0 (Non-sepsis) and 1 (Sepsis). Another dataset is MIMIC-III data which is also collected from the Physionet website. This database consists of information such as vital signs, laboratory test results medications, and demographics. It is notable for its diversity and encompasses a very high population of patients in the ICU.

B. Pre-processing

The dataset entries contain incredibly skewed data. The data is unbalanced. More than 97.8% of records had no sepsis labels, and only 2.2% have one as shown in Fig. 2. The dataset also has a high rate of missing data. And one way to deal with this is to ignore features that have more than 80% of their data missing. The data have been balanced using the resampling technique. Both the labels with 0 and 1 sepsis had the same number of records upon resampling as shown in fig. 3. And now, the suggested system diagram uses this resampled dataset in further steps.

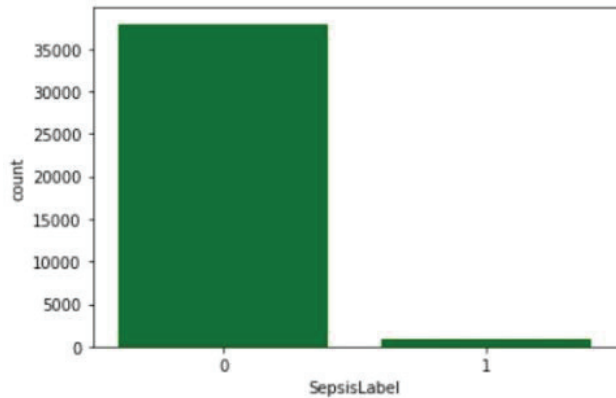


Fig. 2. Sepsis dataset before resampling

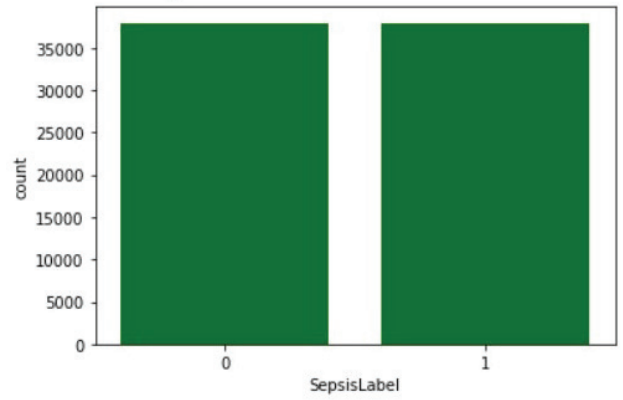


Fig. 3. Sepsis dataset after resampling

C. Feature Importance

Xgboost algorithm is used to choose the attributes that are more crucial for sepsis detection. The advantage of employing this is that the boosted trees are built, and it aids in obtaining significant scores for each of the dataset's features. It mostly offers a score to indicate how significant a feature is in the model. Based on an attributes comparison in the dataset, these are ranked.

D. Model Selection

For the best results, the model choice is a crucial element. The dataset contains 40 dependent variables and one independent variable, a performance measure that assesses whether or not a subject is sepsis- positive. MLP Classifier, Gradient Boosting Classifier, and Linear Discrimination Analysis are a few examples of classification algorithms.

Using a simple neural network, the MLP Classifier performs the classification task. It is ANN's property. MLPs are sometimes referred to as Vanilla networks, especially if they consist of a single hidden layer that avoids time-consuming lab tests and increases flexibility.

Gradient Boosting creates a prediction model from a group of weak prediction models. The model is created using several boosting techniques, which are then modified by awarding an improvement to any differentiable loss function.

Using linear discriminant analysis (LDA), one can reduce the number of dimensions. The dataset's overall number of dimensions is decreased.

E. Algorithm

Step 1: Take the sepsis dataset and perform re-sampling of the information by the re-sampling method to balance the dataset.

Step 2: Utilizing Xgboost Algorithm, get the component significance and take out the less significant highlights.

Step 2.1: Model the XGBClassifier() and reserve it in a variable.

Step 2.2: Fit the data using fit(X, Y) where input and output labels are X and Y respectively.

Step 3: Take this pre-processed dataset and isolate it into the Train and Test datasets as X-train, Y-train, and X-test, Y-test respectively.

Step 3.1: Train the MLPClassifier model by using the Train dataset(X)

Step 3.2: In the function fixing the following values for the parameters hidden layer sizes = (40, 10, 10, 10, 2), activation='tanh', solver='lbfgs', max iter=13000.

Step 3.3: Fit the information which was trained into MLPClassifier using fit(X, Y).

Step 3.4: Validate the prepared model with the Test dataset. **Step 3.5:** Print accuracy as result.

The above steps are also the same for different classifier algorithms but they have different function parameters.

IV. RESULTS

The Physionet challenge sepsis dataset is provided as input on which re-sampling is performed and feature importance is done to identify how important a feature is based on F-scores. Later trained the dataset on different classifiers and compared its accuracy and log loss. The results obtained are shown in figure 4.

```
=====
MLPClassifier
****Results****
Accuracy: 94.7885%
Log Loss: 0.15868214105819684
=====

GradientBoostingClassifier
****Results****
Accuracy: 91.3954%
Log Loss: 0.3111223602886631
=====

LinearDiscriminantAnalysis
****Results****
Accuracy: 72.4074%
Log Loss: 0.5490759157904654
=====
```

Fig. 4. Accuracy of different classifiers on the Physionet Challenge sepsis dataset

When MLP Classifier is used and got an accuracy 94.7% accuracy and log loss is 0.15. The gradient Boosting Classifier has given 91% accuracy and a log loss of around 0.3. The LDA has given 72% accuracy and the log loss is 0.54.

```
=====
MLPClassifier
****Results****
Accuracy: 82.6071%
Log Loss: 0.4102502176777522
=====

GradientBoostingClassifier
****Results****
Accuracy: 81.5357%
Log Loss: 0.40144150097604603
=====

LinearDiscriminantAnalysis
****Results****
Accuracy: 75.5000%
Log Loss: 0.496290799774722
=====
```

Fig. 5. Accuracy of different classifiers on the MIMIC-III dataset

The MIMIC-III dataset is provided as input on which re-sampling is performed and feature importance is done to identify how important a feature is based on F-scores. Later trained the dataset on different classifiers and compared its accuracy and log loss. The results obtained are shown in figure 5.

When MLP Classifier is used and got an accuracy 82.6% accuracy and log loss is 0.41. The gradient Boosting Classifier has given 81% accuracy and a log loss of around 0.40. The LDA has given 75% accuracy and the log loss is 0.49. Table 1 and Table 2 show the comparison of different classifiers on the Physionet challenge and MIMIC-III dataset.

TABLE I. COMPARISON OF CLASSIFIERS FOR THE PHYSIONET CHALLENGE DATASET

Physionet Challenge dataset	MLP	Gradient Boosting	Linear Discriminant Analysis
Accuracy	94.8%	91.4%	72.4%
Log Loss	0.15	0.31	0.54

TABLE II. COMPARISON OF CLASSIFIERS FOR THE MIMIC-III DATASET

MIMIC-III dataset	MLP	Gradient Boosting	Linear Discriminant Analysis
Accuracy	82.6%	81.5%	75.5%
Log Loss	0.41	0.40	0.49

V. CONCLUSION AND FUTURE WORK

In this work, different classifiers were used to compare the accuracy of sepsis detection. Among those classifiers, the MLP Classifier has given the highest accuracy i.e., 94.7% for the Physionet challenge sepsis dataset and 82.6% for the MIMIC-III dataset. This model helps to detect sepsis disease at earlier stages with high accuracy. Thus the real-world applicability of detecting sepsis has been proposed.

Our future work is to design an application that is user-specific and to deploy our model on the hospital website so that the doctors can identify the disease in minimal time with lesser false results.

REFERENCES

- [1] Liu, Shuhui, et al. "Dynamic sepsis prediction for intensive care unit patients using XGBoost-based model with novel time-dependent features." *IEEE Journal of Biomedical and Health Informatics* (2022).
- [2] Zhao, Xin, Wenqian Shen, and Guanjun Wang. "Early prediction of sepsis based on machine learning algorithm." *Computational Intelligence and Neuroscience* 2021 (2021).
- [3] Wang, Zekai, and Bing Yao. "Multi-branching temporal convolutional network for sepsis prediction." *IEEE Journal of Biomedical and Health Informatics* 26.2 (2021): 876-887.
- [4] Baldominos, Alejandro, et al. "Predicting infections using computational intelligence—a systematic review." *IEEE Access* 8 (2020): 31083-31102.
- [5] Al-Mualemi, Bilal Yaseen, and Lu Lu. "A deep learning-based sepsis estimation scheme." *Ieee Access* 9 (2020): 5442-5452.
- [6] Islam, Md Mohaimenul, et al. "Prediction of sepsis patients using machine learning approach: a meta-analysis." *Computer methods and programs in biomedicine* 170 (2019): 1-9.
- [7] Feng, Mengling, et al. "Transthoracic echocardiography and mortality in sepsis: analysis of the MIMIC-III database." *Intensive care medicine* 44.6 (2018): 884-892.
- [8] Wang, Xianchuan, et al. "A new effective machine learning framework for sepsis diagnosis." *IEEE Access* 6 (2018): 48300-48310.

- [9] Kam, Hye Jin, and Ha Young Kim. "Learning representations for the early detection of sepsis with deep neural networks." *Computers in biology and medicine* 89 (2017): 248-255.
- [10] Scherpf, Matthieu, et al. "Predicting sepsis with a recurrent neural network using the MIMIC III database." *Computers in biology and medicine* 113 (2019): 103395.
- [11] Fu, Mengsha, et al. "An ensemble machine learning model for the early detection of sepsis from clinical data." *2019 Computing in Cardiology (CinC)*. IEEE, 2019.
- [12] Ying, Teh Xuan, and Asma Abu-Samah. "Early Prediction of Sepsis for ICU Patients using Gradient Boosted Tree." *2022 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*. IEEE, 2022.