Non-Invasive Prediction Model to Detect Sepsis using Supervised Machine Learning Algorithms

Ruban S, Elreena Maria Pinto, Valerie Roselyn Cardozo, Kavya S.

Abstract— ***Sepsis is a life-threatening disease that causes tissue damage, organ failure and results in the death of millions of people. Sepsis is one of the highest risky diseases identified globally. A large proportion of these deaths occur in developing countries due to inaccessibility of hospitals or lack of resources. Blood samples are taken to confirm sepsis, but it requires the presence of laboratory and is time-consuming. The aim and objective of this study is to develop a practical, non-invasive sepsis prediction model that can be used to detect sepsis using supervised machine Learning algorithms. For this retrospective analysis, we used the data available from Physio-Net database.***

Keywords:Sepsis, Prediction model, Physio-Net dataset, Non-invasive.

I. INTRODUCTION

There are many diseases that are causing threat to human lives. One among them is Sepsis that causes tissue damage, organ failure and results in the death of millions of people annually worldwide. It not only affects the Adults but in neonates as well. Blood culture is a standard measure taken to diagnose sepsis, but it is time consuming. The untreated infection can lead to severe illness and gradually to death [1]. By using Machine learning we can help in providing better healthcare for sepsis patients. In this paper we have compared the performance of various machine learning algorithms in deciding whether the patient will have sepsis or not. The Sepsis data is obtained from PhysioNetdatabase[2] which is openly accessible This paper is organised as follows.

Section 2 highlights few works that are done in this area of sepsis using machine learning and section 3 describes the method we have used in comparing the performance of different classifiers for sepsis, and Section 4 provides the experimental results of our approach. Finally, Section 5 presents our scope and conclusion.

II. RELATED WORK

Though sepsis is considered as one of the deadliest diseases, there are not many aids to help detect sepsis in advance. To find whether a patient is suffering from sepsis, the physicians still rely on various clinical variables and physiological factors. Hence it is the need to develop certain tools that may help the doctor to diagnose sepsis in advance. After the advent of machine learning algorithms in recent past, few research studies have developed few scoring systems to detect sepsis in advance, before the clinical tests could conform. Henry et al [3] came out with a scoring system based on the data that was extracted from the widely used MIMIC-II [4] dataset. In another work that was put forward by Calvert et al[5] based on the same dataset, a unique method for early detection of sepsis was developed based on the Time series data. Based on the MIMIC-III dataset another method was proposed by Harutyunyan et al [6]. It is widely referred to as LSTM method and is based on Neural Network. Various other methods have also been proposed in recent times to detect sepsis in the earliest possible. In our work we have used the PhysioNet database and studied the impact of different classification algorithms to detect sepsis in advance.

III. MATERIALS AND METHODS

The Sepsis data is obtained from PhysioNetdatabase[2] which is openly accessible. Data is sourced from ICU patients in three separate hospital systems. Data from two hospital systems is publicly available and the complete training database consists of two parts: Training set A and Training set B. Training set A consists of 20,336 subjects and Training set B consists of 20,000 subjects. The data for each patient is contained within a single pipe-delimited text file which was converted to CSV (Comma-Separated Values) file using python code. Each file has the same header and each row represents a single hour's worth of data. Available patient covariates consist of Demographics, Vital Signs, and Laboratory values, which are defined in Table 1, Table 2 and Table 3.

Table- 1: Attribute Description with Vital Signs

·	Attribute	Values/Type
Vital signs	Heart Rate	Beats per minute
	Pulse oximetry	%
	Temperature	Deg C
	Systolic BP	mm Hg
	Mean arterial pressure	mm Hg
	Diastolic BP	mm Hg
	Respiration rate	Breaths per minute
	End tidal carbon dioxide	mm Hg

Revised Manuscript Received on January 30, 2020.

Ruban S*, Faculty Member, PG Department of IT, AIMIT, St Aloysius College, Mangaluru, India. Email: ruban@staloysius.ac.in.

Elreena Maria Pinto, Student, PG Department of IT, AIMIT, St Aloysius College, Mangaluru, India. Email: elreena_pinto@outlook.com.

Valerie Roselyn Cardozo, Student, PG Department of IT, AIMIT, St Aloysius College, Mangaluru, India. Email: cardozovalerie1997@gmail.com

Kavya S, Student, PG Department of IT, AIMIT, St Aloysius College, Mangaluru, India. Email:

Table- 2: Attribute Description with Lab Values

	Attribute	Values/Type
Laboratory Values	Excess Bicarbonate	mmol/L
	Bicarbonate	mmol/L
	Fraction of inspired oxygen	%
	Partial pressure of CO_2 from arterial blood	mm Hg
	Oxygen saturation from arterial blood	%
	Aspartate transaminase	IU/L
	Blood urea nitrogen	mg/dL
	Alkaline phosphatase	IU/L
	Calcium	mg/dL
	Chloride	mmol/L
	Creatinine	mg/dL
	Bilirubin direct	mg/dL
	Glucose	mg/dL
	Lactate	mg/dL
	Magnesium	mmol/dL
	Phosphate	mg/dL
	Potassium	mmol/L
	Total bilirubin	mg/dL
	Troponin I	ng/mL
	Hematocrit	%
	Hemoglobin	g/dL
	Partial thromboplastin time	Seconds
	Leukocyte count	count*10^3/µL
	Fibronogen	mg/dL
	Platelets	count*10^3/µL

Table- 3: Attribute Description with Demographics Values

	Attribute	Values/Type
Demographics	Age	Years
	Gender	Female(0) or Male(1)
	Hours between hospital admit and ICU admit	Hours
	ICU length-of-stay	Hours since ICU admit
Outcome	Sepsis Label	Yes(1) or No(0)

We have initially considered 1000 files of training set A for analysis and to train because of the data load on our system. In 1000 records, there were 20.6% of sepsis records and rest were non-sepsis. Since, there were many missing values in the parameters, we took the average mean of each columns of each patient record and applied feature selection using filter importance property to filter out the parameters that are not necessary. The following figure (Figure 1) gives the top 20 scores of each feature from the data. The higher the score, the more important or relevant the feature is to your output variable.

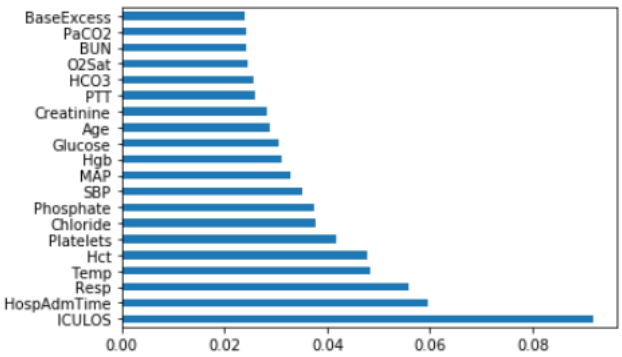

Fig 1: Feature Selection of the Attributes

Seven different classification algorithms were used to train models that could classify whether a patient is suffering with sepsis or not. The algorithms used for the classification problem are discussed below.

Logistic Regression:

It is one of the widely used classification algorithm since 20th century. Especially when the target variable is categorical, this method is used. In this study the variable that represents the class is a variable that is categorical dependent.

K-Neighbours Classifier:

This classifier used a very different way of classification. A method based on neighbours majority vote is being used. The neighbours are given weightage that are more than the others that are distant. The distance between vector and neighbour is assigned as d, and the weight that is assigned is represented as 1/d. It can also be considered as a technique that is used for learning based on instance. During the classification is when the estimation also occurs.

Decision Tree Classifier:

This classifier is also one among the classification techniques that used the Tree structure. As the name suggests a construction of a tree takes place for the target value that is represented using the leaves on the basis of the data's that is being observed on the branches. The features of the branches lead to the class labels whereas the class labels are represented as leaves.

Gaussian Naïve Bayes:

The method of classification based on probability leads to this classification method. This method is predominantly used when the features are strongly independent. According to the Gaussian distribution each class are assumed to hold values that are continuous. Initially the values are divided based on the class they fall. Later on every feature is extracted and then the average and the variance of them are calculated.

Support Vector Machines:

It is a classification method that is not based on probability but based on a hyperplane. To make it clear, if the training data that is labelled is given then the method gives an optimal value that will divide the data into parts. A line that categorizes the dataset into two categories in the 2-D spaces are generally referred to as Hyperplane.

Multi-level Perceptron Classifier:

This classifier is mostly used for modelling the relationship between the different inputs and the various outputs. It helps you in modelling the correlation and the data training allows you to differ various weights for the input pairs and biases to calculate errors that are later passed on using the back propagation method.

Random Forest Classifier:

This method is used to create decision trees based on the data that are selected from the training set. It is further used to summarize the data from the various other decision trees to find out the last class of the test set.

IV. RESULTS

In this area we examine the metric for evaluating different models:

Accuracy: Classification Accuracy is the ratio of the number of correct predictions to the total number of input samples.

$$\text{Accuracy}=\frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made}$$

Table II depicts the performance of the various classifiers dependent on classification accuracy of classification with cross validation.

Table- II: Performance Evaluation of different Classifiers

Type of Classifiers	Accuracy	
LogisticRegression	Standard Deviation	0.977000
	Mean	0.011874
K Neighbors	Standard Deviation	0.975000
	Mean	0.016279
Decision Tree Classifier	Standard Deviation	0.974000
	Mean	0.011136
Support Vector Machine	Standard Deviation	0.974000
	Mean	0.014967
Gaussian NB	Standard Deviation	0.844000
	Mean	0.036111
MLP Classifier	Standard Deviation	0.975000
	Mean	0.016279
Random Forest	Standard Deviation	0.979000
	Mean	0.015133

Figure 2 and Figure 3 shows the accuracy comparisons of prediction score in machine Learning Algorithm.

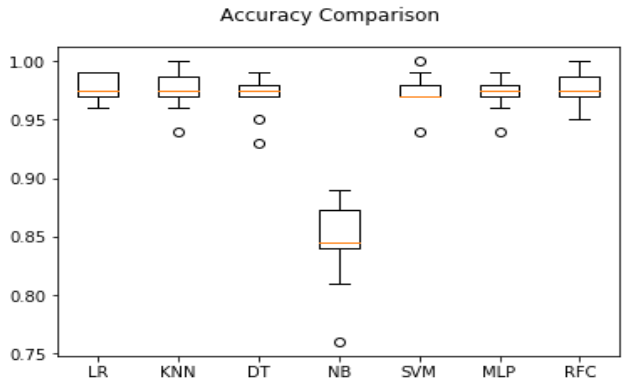

Fig. 2: Box Plot Showing the Accuracy Comparisons of Ml Algorithms

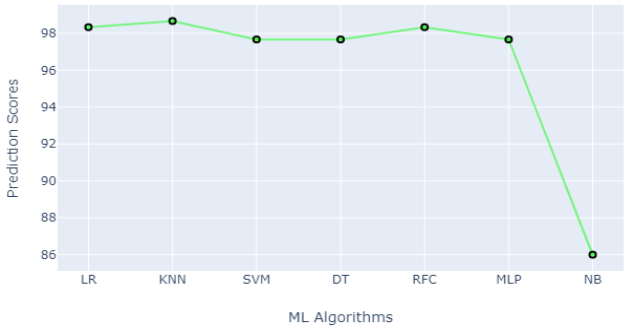

Fig. 3: Scatter Plot Showing the Accuracy comparisons of Ml Algorithms of Prediction Score

The above results show that Gaussian Naïve Bayes gives the least accuracy result (84.4%) where as Random Forest Classifier with 97.9% accuracy gives more accurate results.

V. CONCLUSIONS

Our study shows that supervised learning algorithms can play a valuable role in healthcare. With reliable training and testing datasets, real-time prediction and classification can be a living reality in future. This implementation can further be improved by checking the severity of sepsis and also early detection of sepsis which would help to take measures in advance. Our studies show that Logistic Regression, Random Forest and Support Vector Machines show promise in developing Artificial Intelligence Based Healthcare. In this paper we compared and contrasted the performance of various Supervised Learning Algorithms in the classification of Sepsis Patients.

REFERENCES

1. Martin, G. S. Sepsis, severe sepsis and septic shock: changes in incidence, pathogens and outcomes. Expert review of anti-infective therapy 10, 701–706 (2012).
2. Johnson, T. Pollard, L. Shen, L. Lehman, and M. Feng, "The MIMIC III Clinical Database",01-Oct2015.[Online]. vailable:https://physionet.org
3. Henry, K. E., Hager, D. N., Pronovost, P. J. & Saria, S. A targeted real-time early warning score (trewscore) for septic shock. Science translational medicine 7, 299ra122–299ra122 (2015).
4. Saeed, M. et al. Multiparameter intelligent monitoring in intensive care ii (mimic-a public-access intensive care unit database. Critical care medicine 39, 952 (2011).
5. Calvert, J. S. et al. A computational approach to early sepsis detection. Computers in biology and medicine 74, 69–73 (2016).

AUTHORS PROFILE

Ruban S PhD, is a Faculty in the Department of IT, AIMIT, St Aloysius College, Mangalore, India. He earned his Ph.D in the area of Information Retreival. His research interest includes Big Data Analytics,Health Informatics, Machine learning and Deep learning.

Elreena Maria Pinto, presently is a student of MCA, St Aloysius Institute of Management and Information Technology. Her research interests include Machine Learning, Data science and Health Informatics.

Valerie Roselyn Cardozo, presently is a student of MCA, St Aloysius Institute of Management and Information Technology. Her research interests include Machine Learning, Data science and Health Informatics.

Kavya, presently is a student of MCA, St Aloysius Institute of Management and Information Technology. Her research interests include Machine Learning, Data science and Health Informatics.