Early Detection of Sepsis using Machine Learning

S.V. Evangelin Sonia, S. Sharanya, M. Sivaram, S. Vaishnavi, S. Sakthidevi

Abstract*: Sepsis is a global cause of the death due to infection and subsequent overreaction of the immune system. Mortality rates are highest in developed and developing countries in cases of septic shock. Sepsis is a clinical condition with an emergency referral that can be avoided by advance warning. SIRS is a normal immune response to any infection, and hospitals or antibiotics are not required by most people. Early sepsis prediction is possibly life-saving, and we are planning to predict sepsis 6 hours before clinical sepsis diagnosis. We used two types of Standard Scalar and Min Max Scalar pre-treatment methods to pre-treat the data. The Recursive Feature Elimination (RFE) method is used to pick the features most closely related to the predictive or efficiency factor. The ML algorithms used are the Logistic Regression algorithm and XGboost. Cross-validation 10 times with the train splitting method is used to validate attributes. we selected the best model from these two trained models.

Keywords: Sepsis, Machine Learning, Classifiers, Predictions.

I. INTRODUCTION

Sepsis is a leading reason for death, worldwide. The World Health Organization (WHO) reports that 30 million people develop sepsis and six million die per year from sepsis; an additional 4.2 million newborn babies including infants get infected. Passing on in septic shock is the strongest reason in developing and developed nations. Predicting presumptive sepsis is crucial to gaining the efficiency of the sepsis. Sepsis is a life-threatening condition that causes dysfunction of the organ due to dysregulated response of the host to infection. Sepsis is the main cause for in-hospital death rates and leads to an increased risk of physical organ failure, neurological damage and chronic illness in surviving patients. It has been shown that early and personalized care greatly boosts the sepsis outcomes.

Predictive models were employed to optimize treatment and be used in critical care settings, such as the Intensive Care Unit (ICU) potentially to identify septic patients early on [1].

Sepsis is a clinical condition with an emergency orientation which can be avoided by premiere warning. It's a SIRS, natural autoimmune condition to any infections and neither medication nor antibiotics are needed by most people.

Revised Manuscript Received on April 16, 2020.
Correspondence Author

S.V.Evangelin Sonia*, Assistant Professor, Computer science and engineering department, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India. Email: evangelinsonia.vs@gmail.com

S. Sharanya, Computer science and engineering department, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India

M. Sivaram, Computer science and engineering department, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India

S. Vaishnavi, Computer science and engineering department, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India

S. Sakthidevi, Computer science and engineering department, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India

The response to postoperative sepsis diagnosis inflammation is unclear as other infections without acute hypothermia, digestive, inflammatory and lung diseases and long-term autoimmune reactions (e.g. infections of the urinary tract) make it difficult to diagnose early. It is necessary to differentiate inflammation based on sepsis very early, and then to integrate successful treatment methods into the critical care plan [9].

Machine learning has emerged as a promising way to reduce diagnostic uncertainty, pick effective antibiotics and identify patients with severe sepsis [1]. Clinical variables in real time were used to develop an important predictive model that can accurately predict the beginning of sepsis in an intensive care unit (ICU) before clinical understanding. Electrocardiogram (ECG) is among the most specific signals to come from ICU patients. The signals could also be used to detect pulse rate changes. Heart rate variability (HRV) is a variable series of physiological differences in the time intervals of heart beat to beat [2].

II. OBJECTIVE

The aim of this work is to diagnose sepsis early on using clinical data. Early sepsis predictions help save lives, and we plan on predicting sepsis 6 hours prior hospital sepsis diagnosis. By contrast, late detection of sepsis is possibly life-threatening, and forecasting sepsis in patients with anti-sepsis (or detecting sepsis very early in sepsis patients) needs adequate medical support.

III. LITERATURE REVIEW

V. Prasad et.,al [13] notes that hemodynamic sepsis administration in the emergency service focuses primarily on endotracheal intubation and vasopressor care to uphold sufficient heart rate and overall organ blood flow. Although regular pastimes with excellent blood pressure levels (including 65 mmHg for venous blood pressure or 90 mmHg for systolic pressure [SBP]), little attention is paid to the conceptual nature of heart rate. In just the two hours following the onset of hypertensive moments (SBP: 90 mmHg) or instantly proceeding the onset of vasopressor treatment, he used unsupervised re-examination techniques. The findings showed some hypotensive patients who appeared to get quickly affected (within 40 minutes). Among them, patients who had hypotension in the preceding hour due to a major and severe reduction in daily SBP used to have a higher prevalence of successive vasopressor administration than with a higher gradual decrease in hypertension.

A. Davoudi et al [14] state that fast mobilization in the intensive care unit (ICU) of critically ill patients will avoid adverse consequences, such as delirium and physical disability after discharge. His research used granular actigraphy data to classify the behavior of patients with sepsis at ICU. The study characterized the operation of sepsis sufferers at ICU to assist with potential survivability measures. The actigraphy characteristics of 24 patients were compared in four groups: intensive care patients with chronic sepsis (CCI), patients with rapid recovery sepsis (RR) in intensive care unit, Patients with sepsis-free ICU (ICU-control) and stable patients. He used a total of 15 statistical and circadian rhythm characteristics derived from data obtained during a five-day period of patient actigraphy. The findings showed that the activity characteristics of the four groups were significantly different. Additionally, he found that patients with CCI and ICU regulation displayed less regularity in their circadian rhythm than patients with RR.

K. Gunnarsdottir et al[15] defines septicemia as a pervasive immune reaction to infection, a major public health problem affecting billions of sick people worldwide each year in intensive care units (ICUs). Given that patients with intensive care are extensively equipped with physical detectors, early intervention of sepsis continues to remain difficult, possibly whereas physicians diagnose septicemia in (i) individually utilizing variable results extracted from bedside readings and (ii) extracting these ratings at a much lower rate than even the rate with which patient information are gathered. In his study, he is constructing a generalized linear model (GLM) for the likelihood that an intensive care sufferer will have septicemia based on demographic and bedside measurements. More precisely, the models were trained in 29 patient records from the MIMIC II repository and tested on a separate set of tests, including 8 patient records. The accuracy of the classification of 62.5 per cent was achieved by demographic tests.

Roman Z. Wang et al[11], Septicemia is a systemic disordered reaction to infectious disease that causes inflammation in the organs. This is the major cause of hospital death, and it is accountable for America's largest and most costly hospitalizations. Prognostic sepsis modeling has shown the potential for optimizing medication and results for patients in intensive care. Though, older models depended on an outdated sepsis theory based on the Systemic Inflammatory Response Syndrome (SIHR) feature. Hence the above research sought to create predictive sepsis models using the new sepsis description, Sepsis-3. In total, three classification methods, which include logistic regression (LR), vector support (SVM) and logistic model tree (LMT), have been used to forecast the occurrence of septicemia in patients in the Intensive Adult Treatment Unit (CIS) using symptoms and blood culture tests.

IV. METHODOLOGY USED

The data sets are collected from PhysioNet Computing in Cardiology Challenge 2019. Preprocessing the data is an important phase in the process of data mining. We have used two types of pre-processing methods Standard Scalar and Min Max Scalar. The idea behind Standard Scalar is to convert the data such that its distribution has a median value of 0 and a variance of 1. The Min Max scalar[10] standardizes the functionality by scaling each function to a given range. This estimator scales and individually translates each feature, so that it is on the training set in the given range, i.e. between zero and one. Feature Selection is the mechanism where certain features are picked that most relate to the predictive variable or performance. The Recursive Feature Elimination (RFE) approach is used for selecting the features most relate to the predictive variable or performance. The selected attributes are: HR (Heart Rate), O2SAT (Pulse oximetry), TEMP (Celsius), SBP (Systolic BP), MAP (Mean Arterial Pressure), DBP (Diastolic BP). Machine learning algorithms create a mathematical model based on input data, known as "training data," such that forecasts or decisions can be made without complex task programming. The ML techniques used in this paper are the algorithms Logistic Regression[5] and XGboost[4]. The 10-fold cross validation with train split approach is used to verify the attrubutes. As training data, 80 percent data is used and 20 percent data are used as test data. The effectiveness of the models is assessed in a clinical environment where limitations on operational quantities, such as precision and reliability of the model, have to be met and sepsis detected correctly before impacting.

A. Model Diagram

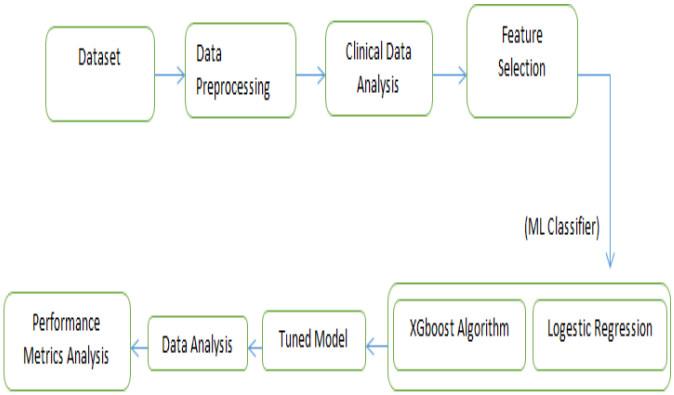

Fig 1: Model Diagram

B. Data Set

Data used in the study is from PhysioNet Computing in Cardiology Challenge 2019[16].The description of the dataset attributes are given below:

Vital signs

HR	Heart rate (beats p/m)
O2Sat	Pulse oximetry (%)
Temp	Temperature (^{0} C)
SBP	Systolic BP (mm Hg)
MAP	Mean arterial pressure (mm Hg)
DBP	Diastolic BP (mm Hg)
Resp	Rapid respiration (breaths p / m)
EtCO2	End of marsh Co2 (mm Hg)

Laboratory values

Base Excess	Measure of excess Bicarbonate (mmol/L)
HCO3	Bicarbonate (mmol/L)
FiO2	Fraction of inspired O_2 (%)
pH	N/A

PaCO2	Partial pressure of CO_2 from arterial blood (mm Hg)
SaO2	O_2 saturation from arterial blood (%)
AST	Aspartate transaminase (IU/L)
BUN	Blood urea nitrogen (mg/dL)
Alkalinephos	Alkaline phosphatase (IU/L)
Calcium	(mg/dL)
Chloride	(mmol/L)
Creatinine	(mg/dL)
Bilirubin_direct	Bilirubin direct (mg/dL)
Glucose	Serum glucose (mg/dL)
Lactate	Lactic fluid (mg/dL)
Magnesium	(mmol/dL)
Phosphate	(mg/dL)
Potassium	(mmol/L)
Bilirubin_total	Total bilirubin (mg/dL)
TroponinI	Troponin I (ng/mL)
Hct	Hematocrit (%)
Hgb	Hemoglobin (g/dL)
PTT	partial thromboplastin time (seconds)
WBC	White Blood Cell count (count*10^3/µL)
Fibrinogen	(mg/dL)
Platelets	(count*10^3/µL)
Demographics	(columns 35-40)
Age	age in years
Gender	Female (0) or Male (1)
Unit1	Administrative ICU unit code (MICU)
Unit2	Administrative ICU unit code (SICU)
HospAdmTime	Time between admission to hospital and admission to ICU
ICULOS	ICU length-of-stay
Outcome	
SepsisLabel	SepsisLabel is 1.For non-sepsis patients, SepsisLabel is 0.

C. Data Preprocessing

Data processing is a key step in making it suitable for ML. A large quantity of data is usually needed for the most common types of ML. In order to obtain better outcomes from the model implemented in Machine Learning projects, the data format must be right[3].

Machine Learning largely depends on the test results. Preprocessing the data is an important phase in the process of data mining. Particularly applicable to data mining and machine learning ventures is the expression "garbage in, garbage out". Data collection techniques are often poorly regulated, resulting with the out of-range results (e.g. revenue: −100), unusual data variations (e.g. sex: male, pregnant: yes), missed values, etc. Review of data that was not carried out carefully analyzed for these problems may yield misleading results. Therefore, the interpretation and accuracy of the data is mainly before an experiment is carried out. Preprocessing data is always the most crucial step of a deep learning task, especially in the field of computational modeling. If there's a number unnecessary and redundant information, or noisy and inaccurate data, then it is more difficult to discover knowledge during the training process. The preparation and retrieval of data steps will take considerable processing time[8]. The final training set is the result of preprocessing the data. Pre-processing of data includes cleaning, selection of cases, normalization, transformation, extraction and selection of functions, etc. The preprocessing methods used here are Standard Scalar and Min Max Scalar. The concept behind Standard Scalar is to transform the data so its distribution has a mean value of 0 and a SD of 1. The Min Max scalar standardizes the functionality by scaling each function to a given range. This estimator scales and individually translates each feature, so that it is on the training set in the given range, i.e. between zero and one.

D. Feature Selection

Feature Selection is the mechanism where certain features are picked that most relate to the predictive variable or performance, automatically or manually. With irrelevant features in the data, the model's accuracy can be decreased and the model learned based on irrelevant features. The Recursive Elimination Function (RFE) algorithm is used for selecting features in the implementation. The RFE approach is a selection approach to features. Works by recursively eliminating attributes, and creating a model on the remaining attributes. The model accuracy is used to classify which attributes (and combinations of attributes) most contribute to the prediction of the target attribute [6, 7].

E. Machine Learning Algorithms

Machine learning (ML) is the theoretical analysis of algorithms and mathematical systems used by computing devices without the description of specific instructions to do a particular function. This is considered to be a branch of AI. Machine learning algorithms create a mathematical model based on sample data, called "training data," so predictions or decisions can be made without complex task programming. The ML algorithms used in this paper are logistic regression, the algorithm XGboost.

E.1 Logistic regression (LR)

Logistic regression is named for the function employed, the logistic equation, at the heart of the system. Mathematicians created the logistic equation, also known as the sigmoid equation, to explain the features of ecological population development, gradually extending and maximizing the ecosystem's capacity. This is an S-shaped curve that can accept any true-evaluated number and assign it to a value between 0 and 1, but never precisely at those limits [5].

1/(1 + e^-value) Where e is the basis for the natural logarithm (Euler's integer or EXP) and where e is the actual values you want to convert.

E.2 XGBoost Algorithm

XGBoost is a Machine Learning algorithm based on an ensemble of decision-trees using a gradient boosting method. When problem solving artificial neural networks involving unstructured data (images, text, etc.) consistently outperform all the other models or systems.

Nonetheless, Decision tree-based algorithms are considered beneficial-in-class when it comes to structured / table small and medium data. A wide range of applications: could be used to solve regression, classification, ranking and user-defined prediction problems. It can work properly on Windows, Linux and OS X, taking portability into account [4].

XGBoost and Gradient Boosting Machines (GBMs) are both set tree methods that use the gradient descent design concept to increase poor learners (CARTs in general). Whereas XGBoost builds on GBM's base architecture by optimizing architectures and improving algorithms.

V. EXPERIMENTAL RESULTS

The dataset is pre-processed using standard scalar and min max scalar. The missing values obtained are 8% of the total data and hence the null va lue attributes are ignored .The RFE feature selection algorithm selects the important features and those are passed to the ML classifiers to analyse the training data. The selected attributes are HR (Heart Rate), O2SAT (Pulse oximetry), TEMP (temperature), SBP (systolic BP), MAP (Mean Arterial Pressure), DBP (diastolic BP). The 10-fold cross validation with train split approach is used to verify the attrubutes. As training data, 80 percent data is used and 20 percent data are used as test data. With the selected validation set, the threshold was optimized for increasing cross-validation, where the optimum thresholds were defined as maximizing the F2 value. Using precision, sensitivity, specificity, and Precision, Negative Predictive Value (NPV), and Receiver Operating Curve (AUC) region, model performance was evaluated. Table 1 represents the performance metrics analysis for sepsis data. Fig 2 represents the ROC curve for XGboost algorithm. The LR gives an accuracy of 92% where xgboost gives an accuracy of 97%.The XGboost model is best when compared with LR model when it is analyzed with the performance.

Table 1: Performance metrics analyses

Metrics	LR	XGboost
Accuracy	92%	97%
Sensitivity	0.22	0.32
Specificity	0.98	0.96
Precision	0.5	0.6
NPV	0.98	0.97
AUC	0.77	0.86
Threshold	0.88	0.92

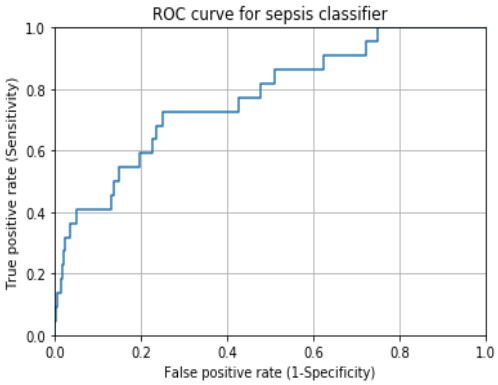

Fig 2: ROC curve for XGboost Classifier

VI. CONCLUSION

The Sepsis-3 description is used in this study to create two predictive models of sepsis in ICU patients. Essentially, the effectiveness of the models is measured by their efficacy in a clinical setting where constraints on operational quantities, such as model accuracy and efficiency, have to be met and sepsis correctly detected before affecting. Compared with the LR, the XGboost algorithm provided superior performance in classification. For this study, the time-window from which predictive values were chosen randomly for each patient was substantially wider than in previous work. Consequently, in a clinical environment, the models achieve greater usefulness. Then we will decide the onset time of sepsis and relative goals of joint influence of two convictions. Sepsis and non-sepsis credence's increase the degree to which the actions of the system influence the utilities. Future studies should investigate different methods of selecting features and collecting and processing data to enhance the model's clinical utility.

REFERENCES

1. R. M. Demirer and O. Demirer, "Early Prediction of Sepsis from Clinical Data Using Artificial Intelligence," 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), Istanbul, Turkey, 2019, pp. 1-4.
2. R. Z. Wang, C. H. Sun, P. H. Schroeder, M. K. Ameko, C. C. Moore and L. E. Barnes, "Predictive Models of Sepsis in Adult ICU Patients," 2018 IEEE International Conference on Healthcare Informatics (ICHI), New York, NY, 2018, pp. 390-391.
3. J. Thakur, S. K. Pahuja and R. Pahuja, "Neonatal Sepsis Prediction Model for Resource-Poor Developing Countries," 2018 2nd International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), Kolkata, 2018, pp. 1-5.
4. C. Yu, G. Ren and J. Liu, "Deep Inverse Reinforcement Learning for Sepsis Treatment," 2019 IEEE International Conference on Healthcare Informatics (ICHI), Xi'an, China, 2019, pp. 1-3.
5. Y. Li, W. Zhou, J. Zhang and C. Yan, "Clinical analyses of neonatal sepsis caused by Listeria monocytogene," Proceedings 2011 International Conference on Human Health and Biomedical Engineering, Jilin, 2011, pp. 388-390.
6. P. Ghasemi and M. R. Raoufy, "Prediction of mortality in patients with sepsis using detrended fluctuation analysis of Heart Rate Variability," 2016 23rd Iranian Conference on Biomedical Engineering and 2016 1st International Iranian Conference on Biomedical Engineering (ICBME), Tehran, 2016, pp. 150-154.
7. J. R. Moorman, D. E. Lake and M. P. Griffin, "Heart rate characteristics monitoring for neonatal sepsis," in IEEE Transactions on Biomedical Engineering, vol. 53, no. 1, pp. 126-132, Jan. 2006.
8. R. Gómez, N. García, G. Collantes, F. Ponce and P. Redon, "Development of a Non-Invasive Procedure to Early Detect Neonatal Sepsis using HRV Monitoring and Machine Learning Algorithms," 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), Cordoba, Spain, 2019, pp. 132-137.
9. K. Gunnarsdottir, V. Sadashivaiah, M. Kerr, S. Santaniello and S. V. Sarma, "Using demographic and time series physiological features to classify sepsis in the intensive care unit," 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, 2016, pp. 778-782.
10. G. Kukreja and S. Batra, "Analogize process mining techniques in healthcare: Sepsis case study," 2017 4th International Conference on Signal Processing, Computing and Control (ISPCC), Solan, 2017, pp. 482-487.
11. R. Z. Wang, C. H. Sun, P. H. Schroeder, M. K. Ameko, C. C. Moore and L. E. Barnes, "Predictive Models of Sepsis in Adult ICU Patients," 2018 IEEE International Conference on Healthcare Informatics (ICHI), New York, NY, 2018, pp. 390-391.

12. M. Saqib, Y. Sha and M. D. Wang, "Early Prediction of Sepsis in EMR Records Using Traditional ML Techniques and Deep Learning LSTM Networks," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, 2018, pp. 4038-4041.
13. V. Prasad, J. C. Lynch, M. R. Filbin, A. T. Reisner and T. Heldt, "Clustering Blood Pressure Trajectories in Septic Shock in the Emergency Department," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 494-497.
14. A. Davoudi et al., "Activity and circadian rhythm of sepsis patients in the Intensive Care Unit," 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), Las Vegas, NV, 2018, pp. 17-20.
15. K. Gunnarsdottir, V. Sadashivaiah, M. Kerr, S. Santaniello and S. V. Sarma, "Using demographic and time series physiological features to classify sepsis in the intensive care unit," 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, 2016, pp. 778-782.
16. Matthew A. Reyna , Christopher S. Josef , Russell Jeter , Supreeth P , Shashikumar , M. Brandon Westover ,Early Prediction of Sepsis from Clinical Data: the PhysioNet/Computing in Cardiology Challenge 2019, 14 October 2019

AUTHORS PROFILE

S.V.Evangelin Sonia was born in Nagercoil, India, in 1987.She received the B.E degree in Computer Science and Engineering from the Anna University, India, in 2009, and the M.E. degree in Computer Science and Engineering from the Anna University, India, in 2011 and pursuing Ph.D degree in Computer Science and Engineering from the Anna University, India.In 2011,She joined as an Assistant Professor in Vins Christian College of Engineering ,Nagercoil,India. Since June 2016, she has been with the Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology,Coimbatore, where she is an Assistant Professor. Her current research interests include Data Mining,Data Analytics and Machine Learning.She is a Life Member of Institution of Engineers (India) [IEI]

S.Sharanya was born in Tirupur, India, in 1999.She is currently pursuing her final year B.E degree in Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology,India. She is expertise in C,Python.She is doing her internship at Skava.Her current research interest is in Machine Learning.

S.Vaishnavi was born in Coimbatore,India, in 1999.She is currently pursuing her final year B.E degree in Computer Science and Engineeringin Sri Shakthi Institute of Engineering And Technology,India.She is expertise in C,Python.She is doing her Internship at Amazon.Her current research interest is in Machine Learning.

S.SakthiDevi was born in Erode, India, in 1999.She is currently pursuing her final year B.E degree in Computer Science and Engineering in Sri Shakthi Institute of Engineering and Technology,India.She is expertise in C,Python.She is doing her internship at SoftCrylic.Her current research interest is in Data Analytics..