Sepsis prediction using ensemble random forest

Conference Paper *in* AIP Conference Proceedings · April 2022

DOI: 10.1063/5.0072499

5 authors, including:

Kanaga Suba Raja Subramanian
SRM Institute of Science and Technology Tiruchirappalli
89 PUBLICATIONS **410** CITATIONS

Sepsis prediction using ensemble random forest

Cite as: AIP Conference Proceedings **2405**, 030027 (2022); https://doi.org/10.1063/5.0072499
Published Online: 05 April 2022

S. Kanaga Suba Raja, K. Valarmathi, S. Deepthi Sri, et al.

Sepsis Prediction Using Ensemble Random Forest

S.Kanaga Suba Raja [a)], K.Valarmathi, S. Deepthi Sri, S. Harishita, V. Keerthanna

Department of Information Technology, Easwari Engineering College Ramapuram, Chennai, India.

[a)]Corresponding author: skanagasubaraja@gmail.com

Abstract: Machine Learning is widely used technology in today's market, It plays an important role in healthcare industry. It is impossible to process thousands of medical reports manually but with machine learning algorithms millions of data can processed in no time. Using machine learning algorithms, it is possible to predict a disease within hours even before their onset, which is not possible in the real time. Sepsis is one among fatal disease brought by the body's reaction to an infection which in turn causes Systematic Inflammation Response Syndrome leading to a progressive organ dysfunction. With help of Ensemble approach, sepsis is predicted with improved accuracy (97.7%) using improved random forest. Our Proposed System is helpful for doctors to diagnose the sepsis very quickly compared to traditional way.

Keywords: - Sepsis, machine learning, Random Forest

INTRODUCTION

Sepsis occurs due to bacterial, viral orfungal infection within the blood. When a bacterial or viral infection occurs, an inflammation is provoked which leads to therelease of a chemical known as cytokines. These cytokines wherever released cause effects inside the tissues. The inflammatory effects are unpleasant but have the aim of recruiting immune cells to fight the infection. Certain cytokines affect the blood vessels and makes it leaky which causes neutrophils, a white blood cell to migrate from the blood vessels to the tissue. This is followed by the release of heat which is theresult of neutrophils being released into thetissues. It causes redness due to capillary dilation. The fluids that build up due to other cytokines and leaky blood vessels leads to tissue swelling. Then it undergoes a progression from severe sepsis to septic shock, which is a morbid condition.

As mentioned in[9] , Sepsis is causedby a poor host retaliation to infections. Sepsis still remains as major health condition, patients in ICU are likely to be affected, they are admitted due to several possibilities of infection which leads to a higher mortality rate. As patients in ICU aremore likely to be affected by sepsis, it is not easy to identify whether a patient is affectedor not. Overall cost to treat sepsis is rising 12-14 percent every two years. Even If the treatment is delayed by an hour the mortality rate increases by 4-8% [10,11]. Fora patient with a suspected infection, the qSOFA(Quick SOFA) score will be determined, where an increase in a point or 2indicates the severity of patients in ICU and they should be transferred to a higher level of care. There were several case studies which proposed that machine learning tools can be used to decrease the uncertainty in disease diagnosis and can easily identifypatients with sepsis by accessing the real time clinical data [12].

In India around 11.3m sepsis cases and around 2.9m sepsis-related deaths (2.9 million) were reported, as of January 2020.Though the number of deathcases are high, still it is under-reported. India has higher sepsis death rate it standssecond among south Asia. The evaluationof data depends only on the countries whose income is high, as no data can be acquired in the countries whose income isless than the average [13].

12th National Conference on Recent Advancements in Biomedical Engineering
AIP Conf. Proc. 2405, 030027-1–030027-7; https://doi.org/10.1063/5.0072499
Published by AIP Publishing. 978-0-7354-4178-1/$30.00

LITERATURE SURVEY

Several researches have been done to support the use of machine learning frameworks to detect sepsis and ElectronicHealth Records were used as the input to train the developed ML models.

[2] A machine learning framework was proposed and prediction was done with a bunch of deep learning algorithms. Thoughthe model was more specific it lacked sensitivity The accuracy of the system wasinferior.

[4] An analysis of positive LOS was done to determine the characteristics of the22 features that differentiate the control state and the sepsis state using Naïve Bayes algorithm. The positivity of LOS was determined from the blood culture. A change in HRV can possibly be a markerfor the onset of sepsis and a low specificitywas observed.

[3] A Decision Process was proposed where the features with missing values areanalyzed. The analysis was spectral based.The proposed method was found to improve the prediction outcome. Only 5000 patient records were considered.

[6] A machine learning framework was developed with a combination of sepsis positive and negative patient record. The feature selection was done by Random Forest to obtain high accuracy and to identify important biomarkers. The specificity is low.

[5] A two-layer model was developed and the performance metrics were compared. Total of 586 patient records were taken for analysis and this approachhad low prediction rate in single layer.

[7] Model was developed with new approach which predicts sepsis with the physiological data obtained. It is inferred that this model handles missing values precisely. A validation method was developed where the performance was computed.

There are numerous literatures works available that put forth the advantage ofusing ML and Deep Learning Frameworks to detect fatal diseases. Without these advancements it would be a time- consuming task to detect these lethal conditions which would lead to a highermortality rate.

PROPOSED SYSTEM

The proposed system utilizes the components that are used to classify areillustrated in Table 1. Proposed system uses the data from Physionet 2019 challenge. It is obtained from ICU patients of two different hospital systems. The datasetconsists of 790216 rows and 42 features. The dataset feature comprises of Demographics, Vital Signs and Laboratory values. The Sepsis Label provides information of onset of sepsis, where 1 indicates presence of sepsis and 0 indicates absence of sepsis.

TABLE 1. Features in dataset

Vital Signs	Laboratory Values	Demographics	Outcome
HR	BASEEXCESS	AGE	SEPSIS LABEL
TEMP	HCO3	GENDER	
O2SAT	FiO2	UNIT 1	
SBP	PH	UNIT 2	
DBP	PaCO2	HOSPADMTIME	
MAP	SaO2	ICULOS	
RESP	AST		
EtCO2	BUN		
	ALKALINEPHOS		
	CALCIUM		
	CHLORIDE		
	CREATININE		
	BILIRUBIN DIRECT		
	GLUCOSE		
	LACTATE		

Vital Signs	Laboratory Values	Demographics	Outcome
	MAGNESIUM		
	PHOSPHATE		
	POTTASIUM		
	BILIRUBIN TOTAL		
	TROPONINI		
	Hct		
	Hgb		
	PTT		
	WBC		
	FIBRINOGEN		
	PLATELETS		

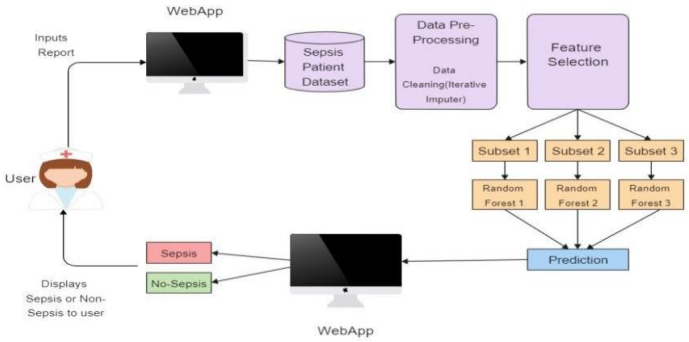

FIGURE 1. System Architecture

Data Preprocessing

The process of Data-Preprocessing provides a high quality of data for the proposed system to perform classification on the Sepsis label. It is most important to process the data before feeding the data intothe model. As the first step import necessarylibraries such as NumPy, pandas, matplotlib, and many more. While visualizing the dataset encountered missingvalues such as nan values in all the features figure 2. These missing values need to be handled for a getting a good performance from the model.

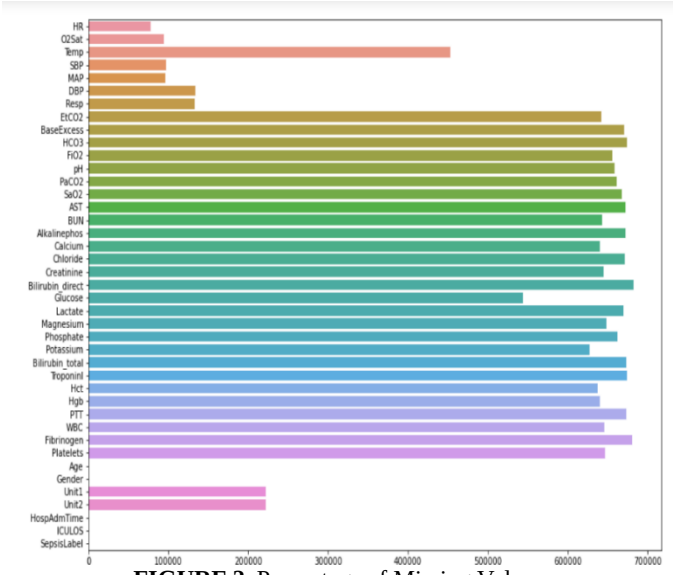

FIGURE 2. Percentage of Missing Values

After a brief analysis, it is found that imputation is the best strategy to handle themissing values. It is done by using a multivariate impute feature which is an Iterative Imputer. This imputer takes a feature with missing values as one variable and the remaining features as anothervariable which estimates and creates a model that uses a regressor to predict the missing values and it is done iteratively for each feature in a round-robin manner. The laboratory parameters with missing values are imputed, vital signs and demographic parameters are handled by replacing the missing values with the median value of thecolumn.

Feature Selection

Feature selection helps to select somespecific features from the set of featureswhich helps the model to makeclassification. The selected features are illustrated in figure 3.

TABLE 2. Feature selection

1	HR	Heart rate (beats per minute)
2	O_2Sat	Pulse oximetry (%)
3	Temp	Temperature (Deg C)
4	SBP	Systolic BP (mmHg)
5	MAP	Mean arterial pressure (mm Hg)
6	DBP	Diastolic BP (mmHg)
7	Resp	Respiration rate (breaths per minute)
8	Base Excess	Measure of excess bicarbonate (mmol/L)
9	BUN	Blood urea nitrogen (mg/dL)
10	Creatinine	(mg/dL)
11	WBC	Leukocyte count count *10^3/μL)
12	Platelets	(count * 10*3 / μL)
13	Age	Years (100 for patients 90 or above)
14	HospAdmTime	Hours between hospital admit and ICU admit
15	ICULOS	ICU length-of-stay (hours since ICU admit)

Predictive Modelling

There are many classificationalgorithms under supervised learning, one such algorithm is Random Forest, which is capable of performing both classification and regression. Instead of using single decision tree, it is preferable to use RandomForest, which is an ensemble approach to overcome over-fitting. Ensemble learning uses several different machine learning models together to produce optimized results.

One of the disadvantages of random forest is it increases the execution time if the number of estimators is more. To overcomethis disadvantage, three random forest weretrained parallelly with 1/3rd of the data to decrease the execution time. Training and Testing data were split from the dataset. Then the Training data is split into three andgiven to three Random Forest which aretrained parallelly. Ensemble approach is used for building the model.

PERFORMANCE ANALYSIS

The dataset which is used in the proposed system is fetched from Physio net2019 challenge. It comprises of 40000 records of patients who are in ICU. The performance of the predictive model is evaluated based on Precision, recall, accuracy, F1-Score, AUROC-Curve.

Precision

Precision is a vital performancemetric and is defined as,

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}$$

Recall

Recall being a metric to analyze theperformance of the model, it is defined as,

$$\text{Recall} = \frac{\text{TP}}{(\text{TP+FN})}$$

Accuracy

Accuracy is the ratio of true positive of the sepsis to the total number of records.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total data}} * 100$$

F1-Measure

F1-Measure is the average of recall orsensitivity and precision.

$$\text{F}_{\text{Score}} = \frac{2 * (\text{Precision} * \text{Recall})}{(\text{Precision+Recall})}$$

AUROC-Curve:

ROC curve plots the TP rate against the FP rate which is a probability curve andthe AUC is a curve which denotes the separability.

RESULT

By performing analysis on our data,it is found that the dataset had several missing values. The presence of missing values in a dataset will lower the accuracy of the prediction model. They also decrease the model's performance. Hence it is important to handle the missing values. It isdone by imputing the missing values present in laboratory parameters. The otherfeatures with missing values are handled byreplacing them with the median of the column. Then the data with the selected features is trained with the proposed model. In comparison with the existing system the proposed model has higher accuracy and a better performance. The proposed system differs from the existing system by using an ensemble approach to train the model. This approach decreases the execution time. The model is evaluated with the help of Recall, Precision, Accuracy, F1-Score and AUROC Curve which are given below.

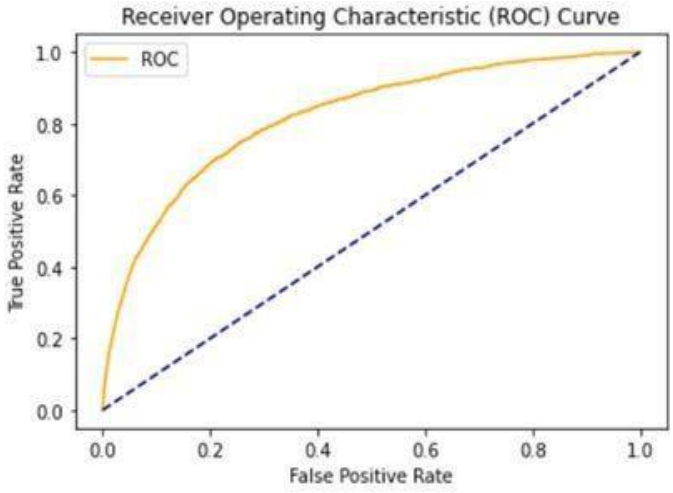

FIGURE 3. ROC Curve

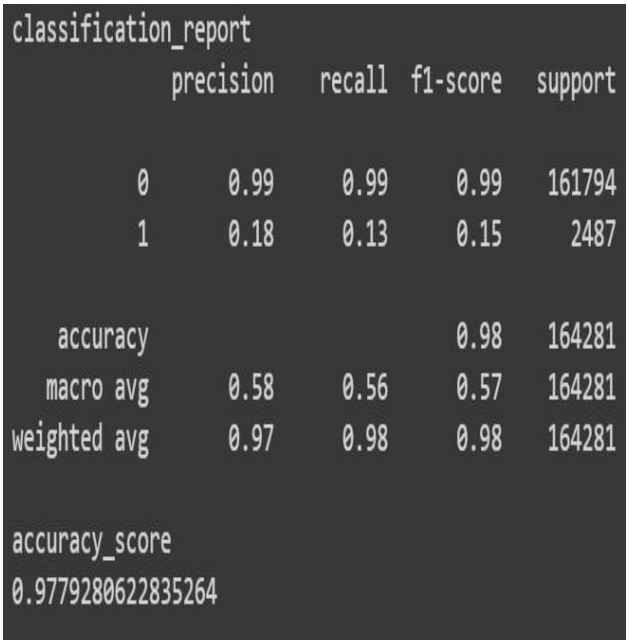

FIGURE 4. Classification report

TABLE 3. Comparison with existing system

	Accuracy	Sensitivity
Existing System	93.84%	93.22%
Proposed System	97.72%	98%

CONCLUSION

The primary intention of the developed system is to use a prediction model to classify the occurrence of sepsis from the ICU data. The proposed system provides a precision value of 0.97, a recall value of 0.98, a f1-measure of 0.98 and the accuracy value of 97%. This system also targets in reducing the execution time of the traditional random forest classification algorithm. The future work includes the development of a web-app to classify the disease and to provide an interface for the medics to predict sepsis onset.

REFERENCES

1. https://www.physionet.org/content/challenge- 2019/1.0.0/
2. Bilal Yaseen Al-Mualemi, Lu, A Deep Learning-Based Sepsis Estimation Scheme.IEEE Access(Volume: 9) ,10 December 2020.
3. R Murat Demirer, Oya Demirer Early Prediction of Sepsis from Clinical Data Using Artificial Intelligence.IEEE 2019 Scientific Meeting on Electrical- Electronics & Biomedical Engineering and Computer Science (EBBT) - Istanbul, Turkey (2019.4.24-2019.4.26).
4. Rohan Joshi; Deedee Kommers; Laurien Oosterwijk; Loe Feijs; Carola van Pul; Peter Andriessen Máxima Medical Centre, Veldhoven, The Netherlands, “Predicting Neonatal Sepsis Using Features of Heart Rate Variability, Respiratory Characteristics, and ECG-Derived Estimates of Infant Motion”, IEEE Journal of Biomedical and Health Informatics (Volume: 24, Issue: 3, March 2020).

5. Van Wyk, Franco; Khojandi, Anahita; Kamaleswaran, Rishikesan (2019). Improving Prediction Performance Using Hierarchical Analysis of Real-Time Data: A Sepsis Case Study. IEEE Journal of Biomedical and Health Informatics, doi:10.1109/JBHI.2019.2894570.

6. [6] Wang, Xianchuan; Wang, Zhiyi; Weng, Jie; Wen, Congcong; Chen, Huiling; Wang, Xianqin (2018). A new effective machine learning framework for sepsis diagnosis. IEEE Access. doi:10.1109/ACCESS.2018.2867728

7. Ioan Stanculescu; Christopher K. I. Williams; Yvonne Freer, "Autoregressive Hidden Markov Models for the Early Detection of Neonatal Sepsis", IEEE Journal of Biomedical and Health Informatics (Volume: 18, Issue: 5, Sept. 2014).

8. https://www.cdc.gov/

9. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3).JAMA 2016 Feb 23;315 (8):801- 810.[doi:10.1001/jama.2016.0287][Medline: 26903338]

10. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, Suppes R, Feinstein D, Zanotti S, Taiberg L, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical care medicine 2006; 34(6):1589–1596.

11. Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna TJ, Phillips GS, Lemeshow S, Osborn T, Terry KM, Levy MM. Time to treatment and mortality during mandated emergency care for sepsis. New England Journal of Medicine 2017;376(23):2235–2244.

12. Simon Lyra , Steffen Leonhardt , Christoph Hoog Antink.Early Prediction of Sepsis Using Random Forest Classification for Imbalanced Clinical Data. 2019 Computing in Cardiology Conference. December 2019.

13. Fahim Mahmud, Naqib Sad Pathan,Muhammad Quamruzzaman.Early detection of Sepsis in critical patients using Random Forest Classifier. 2020 IEEE Region 10 Symposium (TENSYMP) June 2020.