IJCSEInternational Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-8, Issue-1, Jan 2020 E-ISSN: 2347-2693

Early Sepsis Prediction in Intensive Care Patients using Random Forest Classifier

Aparna Shenoy[1*], K.V. Viswanatha[2], Raju Ramakrishna Gondkar[3]

[1,2,3]CMR University, Bengaluru, India

[*]*Corresponding Author: aparnashenoy88@gmail.com, Tel.: +91-9741479025*

DOI: https://doi.org/10.26438/ijcse/v8i1.1722 | Available online at: www.ijcseonline.org

Accepted: 12/Jan/2020, Published: 31/Jan/2020

***Abstract*—** Sepsis is one of the most common causes of morbidity and mortality in the Intensive Care Unit (ICU) patients. The lack of sensitive and specific clinical and laboratory variables for early identification of sepsis in critically ill patients is the causative factor for needless and delayed or untimely interruption of a proper antibiotic therapy. The current work developed a machine learning-based early sepsis prediction model in intensive care patients with vital parameters which evaluated the goodness of model fit and its accuracy. The predictors were extracted from combinations of vital sign measure and their changes over time. The dataset consisted of 20,336 patients (medical and surgical) who were admitted in ICU. Random Forest Classifier was used as the Machine learning algorithm for developing a predictive model. For the early prediction of sepsis in ICU patients, the Random Forest Classifier achieved an AUROC curve of 0.58 for the data collected from the patients within 24 hours. Sepsis is being the common cause of admission in ICU worldwide, a machine learning technique adopting statistical methods to conclude relationships between patient features and outcomes in large data set was successfully applied to predict adverse events.

***Keywords*—**MachineLearning, EarlySepsisprediction, Random Forest Classifier, AUROC.

I. INTRODUCTION

Sepsis is a clinical condition defined by the presence of infection and systemic inflammatory response syndrome (SIRS). Sepsis is often a life threatening condition resulting in multi-organ failure and septic shock leading to a mortality rate of 25 to 40 percent. Sepsis is most frequently caused by systemic bacterial infection with resultant endotoxin release, but it can also be manifested by fungal and viral etiology [1]. In western countries, septic patients account for as much 25 percent of ICU bed utilization and 1-2 percent of overall hospital admissions. The mortality rates ranges up to 45.7 percent in patients with septic shock. According to Indian study, the sepsis rate is 6.2 percent of hospital admissions with an ICU mortality of 56 percent [2].

Determining mortality risk is vital for accepting critical decisions in intensive care units (ICU). Sepsis and multi-organ dysfunction syndrome (MODS), which is a common sequel of sepsis, use enormous ICU resources and are the leading causes of mortality. Since 1970, enormous advances have been made in our understanding of sepsis, and are translated into new approaches in its management. With progression to sepsis-associated organ failure, sever sepsis or septic shock has become frequently fatal and represents a significant health care burden with increased incidence of morbidity, mortality and cost.. Early detection and timely treatment of sepsis can reduce the mortality to a great extent in patients admitted with sepsis. The currently available methods such as early warning parameters like sequential organ failure assessment (SOFA) and supporting sepsis biomarkers are not enough to clearly identify sepsis patients and transfer their treatment to a higher level of care [3].

Machine learning technique is an emerging promising tool to clinicians avoiding the diagnostic uncertainty, identify sepsis patients in advance, select appropriate and timely antibiotics to prevent multi-organ damage and reduce mortality [4].

This current work was designed to adopt a machine learning algorithm to predict early sepsis based on clinical and administrative data generated through preliminary investigations carried out in patients admitted in ICU. The early detection of sepsis resulted in proper monitoring and management of the patient leading to significant reduction in mortality rate.

Rest of the paper was organized as follows, Section I included the introduction of early detection of sepsis in ICU patients using machine learning techniques, Section II consisted of the related work on sepsis detection using machine learning techniques, Section III depicted the

© 2020, IJCSE All Rights Reserved

methodology of developing the model, tools and algorithms used, Section IV described results and discussion, Section V concluded the research work with future directions.

II. RELATED WORK

Ghassemi et.al. [3] summarized the latest trends of machine learning in critical care with a large amount of effort invested in the processing and validation of data acquired in the intensive care unit. His work served as a benchmark on addressing the challenges and highlighted the opportunities for researchers of machine learning to collaborate with clinical staff and engage with clinical experts to identify and tackle important problems in healthcare.

Nemati et. al. [4] demonstrated that high performance models could be constructed to predict onset of sepsis by combining data available from the electronic health records (EHR) and high resolution time series dynamics of blood pressure (BP) and heart rate (HR). Patients who were incorrectly labeled as those who could develop sepsis conferred significant mortality, making this tool potentially useful in other clinical syndromes and disease processes unrelated to sepsis.

A.Vellido et al. [5] described the increasingly complex challenge posed by the current data availability surplus in medicine in general with critical care in particular. He exercised various machine learning techniques in extracting usable knowledge from the data.

Barajas et al. [6] proposed a framework exploiting the dynamic and heterogeneous information from the patients EHRs to predict subsystem failure. The framework considered into account the future uncertainty by training the model using the complete patient path from admission to discharge/death.

Thottakkara [7] compared the performance of risk prediction models for forecasting postoperative sepsis and acute kidney injury. The study assessed the impact of feature reduction techniques on predictive performance, Logistic regression, generalized additive model and support vector machines which performed better compared to Naïve Bayes model. Feature extraction using principal component analysis improved the predictive performance of all models.

Bhattacharya et al. [8] presented a binary classification algorithm designed to address the problem of imbalance that commonly appeared in clinical datasets. The following conclusions could be drawn from the above literatures.

a) Most of the existing algorithms were based on machine learning methods, indicating that it was effective tool for early sepsis prediction.

b) Missing value in medical data was a common observable fact, which became one of the main challenging factor affecting the classification result.

III. METHODOLOGY

A. MATERIALS

The database consisted of data collected on hourly basis from 20,336 patients who were admitted in ICU. Each column of the table provides a sequence of measurements over time (e.g. Heart rate over several hours). Non-Temporal approach was used in the present study.

B. TOOLS USED

Scikit-learn is one of the most popular open source machine learning libraries in Python which is built on other popular similar athenaeums. Numpy, matplotlib and. Scikit-learn provide machine learning algorithms which includes classification, regression, dimensionality reduction and clustering. Moreover, Scikit-learn also provides feature extraction and model evaluation.

Providing an equal sample of positive and negative instances to the classification algorithm will result in an optimal result. The dataset that is highly skewed toward one or more classes has proved to be a challenge. Imbalanced-learn in a python package offering a number of re-sampling techniques commonly used in datasets showed a strong relation between-class imbalance.

C. MEASUREMENTS AND PATIENT INCLUSION

The patient data were analyzed from each of the clinical vital parameters mentioned in Table 1. The vital signs information was frequently made available and routinely collected in the ICU. The collected data showed the unfavorable values for all the variables during the first 24 hours of evolution of sepsis.

Table 1. Sepsis Dataset Features

Sl. No.	Variable	Description
1	HR	Heart rate (beats per minute)
2	O2Sat	Pulse oximetry(%)
3	Temp	Temperature (Deg C)
4	SBP	Systolic blood pressure (mm Hg)
5	MAP	Mean arterial pressure (mm Hg)
6	DBP	Diastolic blood pressure (mm Hg)
7	Resp	Respiratory rate (breaths per minute)
8	HCO3	Bicarbonate (mmol/L)
9	FiO2	Fraction of inspired oxygen (%)
10	pH	Hydrogen ion concentration
11	PaCO2	Partial pressure of carbon dioxide from arterial blood (mm Hg)
12	SaO2	Oxygen saturation from arterial blood (%)
13	AST	Aspartate transaminase (IU/L)

© 2020, IJCSE All Rights Reserved

14	BUN	Blood urea nitrogen (mg/dL)
15	Alkaline phospatase	Alkaline phosphatase (IU/L)
16	Calcium	(mg/dL)
17	Chloride	(mmol/L)
18	Creatinine	(mg/dL)
19	Bilirubin_direct	Bilirubin direct (mg/dL)
20	Glucose	Serum glucose (mg/dL)
21	Lactate	Lactic acid (mg/dL)
22	Magnesium	(mmol/dL)
23	Phosphate	(mg/dL)
24	Potassium	(mmol/L)
25	Bilirubin_total	Total bilirubin (mg/dL)
26	TroponinI	Troponin I
27	Hct	Hematocrit (%)
28	Hgb	Hemglobin (g/dL)
29	PTT	Partial thromboplastin time (seconds)
30	WBC	Leukocyte count (count*10^3/µL)
31	Fibrinogen	(mg/dL)
32	Platelets	(count*10^3/µL)
33	Age	Years (100 for patients 90 or above)
34	Gender	Female (0) or Male (1)
35	Unit1	Administrative identifier for ICU unit (medical ICU)
36	Unit2	Administrative identifier for ICU unit (surgical ICU)
37	HospAdmTime	Hours between hospital admit and ICU admit
38	ICU-LOS	ICU length – of- stay (hours since ICU admit)
39	SepsisLabel	For sepsis patients, SepsisLabel is 1 and for non sepsis patients SepsisLabel is 0

D. FLOW DIAGRAM

In this sub-section, a flow diagram was demonstrated which consisted of classification algorithm and model evaluation. In the initial steps after importing the data, the soiled or missing ones were replaced by a mean using imputation method. In this work 70 % of the entire dataset was retained for training purpose and the remaining 30% was utilized for testing the correctness of the classifier. The performance of the model was assessed on the basis of various performance criteria such as accuracy and recall.

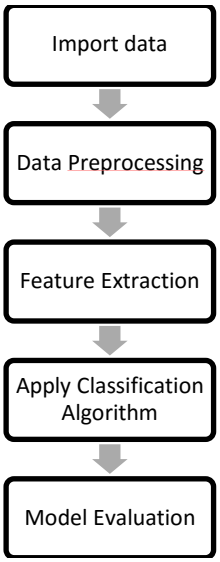

Figure 1. Flow diagram

E. CLASSIFICATION ALGORITHM USED

Random Forest

There are a number of classification algorithms which are used in developing model using large datasets [9]. Random forest is a supervised learning algorithm which is used for both regression and classification. Random forest classifier creates decision trees on data samples and provides the prediction from each of the sample and finally selects the solution by means of voting. Random forest is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result. Ensemble learning involves the combination of several models to solve a single prediction problem. Ensemble learning works by creating multiple models which learn to predict independently. Those predictions are then combined into a single prediction that is much better than a single classifier. Random forest is an ideal example for ensemble learning, as it relies on ensemble of several decision trees [10].

F. EVALUATION CRITERIA

The intent of this study was to predict the occurrence of sepsis in ICU patients at the earliest within 24 hours by applying machine learning techniques. The results of the developed model were analyzed based on certain evaluation criteria such as accuracy and recall as listed below.

Recall

It is the proportion of rightly categorized infected to the count of infected with the predicted Sepsis. It is calculated as given in equation 1.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP+FN})} \quad (1)$$

© 2020, IJCSE All Rights Reserved

Precision
It is described in terms of the proportion of accurately recognized sepsis infected to the count of people having sepsis. Precision is computed as in equation 2.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP+FP})} \quad (2)$$

Accuracy
It quantifies the precision of the learner. The accuracy is computed as in equation 3.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{Total data}} * 100 \quad (3)$$

F measure
It evaluates the test's accuracy in terms of both recall and precision. F measure is calculated as in equation 4

$$\text{F}_{\text{Score}} = \frac{2*(\text{Precision}*\text{Recall})}{(\text{Precision+Recall})} \quad (4)$$

TP signifies true positive: sepsis patients categorized as having sepsis.
TN signifies true negative: sepsis patients categorized as healthy.
FP signifies false positive: healthy patients predicted as sepsis infected.
FN signifies false negative: patients with sepsis wrongly classified as healthy..

Receiver Operating Characteristic (ROC)
The correctness of prediction models can also be anticipated by finding the total area under ROC curve (AUROC) that measures the quality of the classifier. The AUROC value ranges from 0 to 1. The model achieving more value of AUROC is observed as effective and also better performance.

IV. RESULTS AND DISCUSSION

Data analysis is a fundamental step to accomplish before applying machine learning algorithm in order to understand the data. Table 2 contains a statistical summary of the data features.

Table 2. Summary statistics of Independent variables

Variables	mean	std	min	25%	50%	75%	Max
HR	84.90	16.27	20	74	84	94	280
O2Sat	97.35	2.73	20	96	98	99	100
Temp	37.04	0.45	20.9	37.06	37.06	37.06	42.22
SBP	120.58	19.83	22	108	118.5	131	281
MAP	78.58	14.26	20	69	77	86	300
DBP	59.27	9.08	20	58	58.5	59	298
Resp	18.69	5.13	1	15	18	21	69
HCO3	24.00	1.24	0	24	24	24	55
FiO2	0.50	0.07	0	0.5	0.5	0.5	10
pH	7.38	0.02	6.62	7.39	7.39	7.39	7.93
PaCO2	40.10	2.68	10	40	40	40	100
SaO2	96.71	2.99	24	97	97	97	100
AST	61.47	130.5	3	57	57	57	9890
BUN	18.51	6.01	1	18	18	18	266
Alkalinephos	78.52	18.32	7	78	78	78	3833
Calcium	8.30	0.18	1.6	8.3	8.3	8.3	22
Chloride	105.98	1.71	26	106	106	106	145
Creatinine	0.93	0.41	0.1	0.9	0.9	0.9	46.6
Bilirubin_direct	1.40	0.19	0.1	1.4	1.4	1.4	37.5
Glucose	125.17	18.31	10	124	124	124	988
Lactate	1.82	0.44	0.2	1.8	1.8	1.8	31
Magnesium	2.00	0.10	0.2	2	2	2	9.7
Phosphate	3.40	0.32	0.2	3.4	3.4	3.4	18.8
Potassium	4.10	0.20	1	4.1	4.1	4.1	27.5
Bilirubin_total	0.92	0.61	0.1	0.9	0.9	0.9	46.6
TroponinI	4.30	0.43	0.3	4.3	4.3	4.3	49.3
Hct	30.25	1.67	5.5	30.2	30.2	30.2	71.7
Hgb	10.41	0.52	2.2	10.4	10.4	10.4	32
PTT	32.80	5.57	12.5	32.4	32.4	32.4	150
WBC	10.88	2.09	0.1	10.8	10.8	10.8	422.9
Fibrinogen	250.32	14.33	34	250	250	250	1760
Platelets	182.21	28.26	5	181	181	181	1783
Age	63.01	16.13	18.11	52.74	65.25	75.89	89

The figure 1 shows the distribution of patients based on the outcome variable 'SepsisLabel' which clearly shows class imbalance.

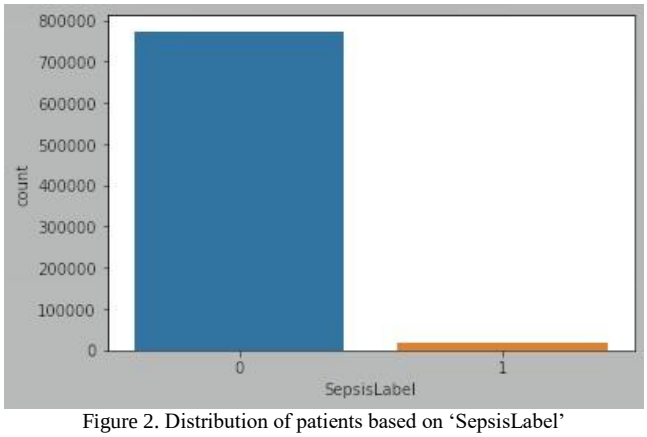

Figure 2. Distribution of patients based on 'SepsisLabel'

© 2020, IJCSE All Rights Reserved

The synthetic minority oversampling technique (SMOTE) preprocessing algorithm is used to treat the imbalanced data. The SMOTE algorithm carries out an oversampling approach to rebalance the original training dataset. The key idea of SMOTE is to introduce synthetic examples created by interpolation between several minority class instances that are within a defined neighborhood.

Following steps were carried out in the experiment:

(1) Conduct the classification using Random Forest classifier with the oversampled data developed using SMOTE algorithm.
(2) Evaluate the results against a set of indicators such as accuracy, precision and recall.

The evaluation indicators such as accuracy, precision, recall were defined according the confusion matrix shown in Figure 3 and Table 3 respectively.

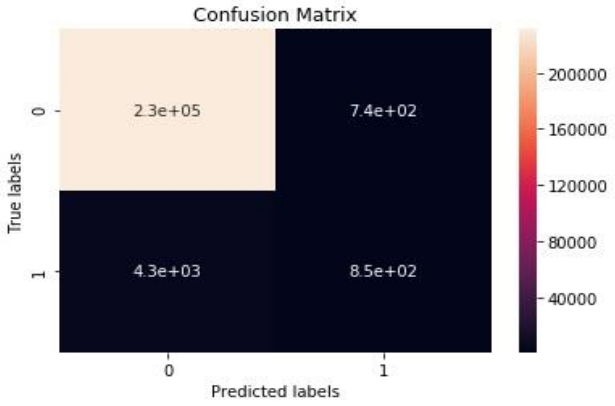

Figure 3. Confusion matrix

According to the confusion matrix in Figure 3, there are 2,31,985 true predictions and 5,080 false predictions.

Table 3. Model Evaluation

	Precision	Recall	F score	Accuracy
0	0.98	1.00	0.99	0.98
1	0.54	0.16	0.25	

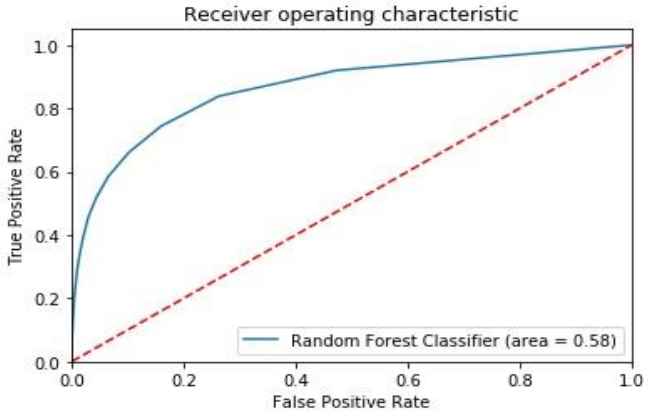

Figure 4. ROC curve for Random Forest Classifier

The Receiving Operating Characteristic in Figure 4 is a measure of classifier performance. Using the proportion of positive data points that are correctly considered as positive and the negative data points that are erroneously considered as a positive; we generated a graph that exhibited the rate of incorrectly predicting outcome. The metric range from 0.5 to 1 indicated that the model does a good job in differentiating between the two categories which comprise our outcome variable. ROC curves can examine both outcome – independent variable pairings and outcome-independent-model performances. The AUROC for this model was 0.58 which indicated a moderate fit. The study observed an overall accuracy of 0.98 and identified more non – infected patients compared to infected ones. By applying this model developed in the present study, one can predict occurrence or non- occurrence of sepsis using vital organ parameters.

V. CONCLUSION AND FUTURE SCOPE

The main aim of this study was to provide an overview of the Random Forest Classifier which is popular in the field of data driven prediction. This study can help the medical fraternity to predict early occurrence of sepsis in ICU patients. The random Forest Classifier provided an accuracy of 0.98 and AUROC of 0.58. The recall is 0.16 which means the model can identify 16 percent of all sepsis patients. The model has a precision of 0.54. In other words, when it predicts a patient is infected with sepsis, it is correct 54 percent of the time. This study also targeted the challenging characteristics of missing values and class imbalance seen in the sepsis dataset. There are also future plans to extend and improve the algorithm and also apply other classification algorithms to predict and diagnose sepsis.

REFERENCES

[1] V. Ribas, A.Vellido, J. Rodriguez, J. Rello, *"Severe sepsis mortality prediction with logistic regression over latent factors"*, Expert Systems with Applications, pp. **1937-1943, 2012.**

[2] S. Chatterjee, M. Bhattacharya, S. Todi, "Epidemiology of Adult population Sepsis in India: A single centre 5 year experience", *Indian Journal of Critical CareMedicine*, **pp. 35-39, 2017.**

[3] A. Johnso, M. Ghassemi, S. Nemati, K. Niehaus, D. Clifton, G. Clifford, "Machine learning and Decision Support in Critical Care", In the Proceedings of 2016 IEEE . Institute of Electrical and Electronics Engineers, no. 2 (vol. 104), **pp. 444-466.**

[4] S. Nemati, A. Holder, F. Razmi, M. Stanley, G. Clifford and T. Buchman, "An In- terpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU", Critical Care Medicine, vol. 46, no. 4, **pp. 547–553 ,2018**.

[5] A.Vellido, F. Schleif and M. Biehl, "Advances in machine learning and computational intelligence", Neurocomputing, vol. 72, no. 7-9, **pp. 1377-1378, 2009**.

[6] K. Barajas, R. Akella, "Prediction of Pysiological Subsystem Failure and its Impact in the prediction of Patient Mortality", In the Proceedings of IEEE, Internaltional Conference on Big Data, **2015**.

[7] P. Thottakkara, T. Ozrazgat-Baslanti, B. Hupf, P. Rashidi, P. Pardalos, P. Momcilovic, A. Bihorac, *"Application of Machine*

© 2020, IJCSE All Rights Reserved

learning Techniques to High Dimensional Clinical to Forecast Postoperative Complications", PLOS ONE, vol. 11, no. 5, **2016**.

[8] S. Bhattacharya, V. Rajan, H. Shrivastava, "ICU Mortality Prediction: Classification Algorithm for Imbalanced Datasets",In the Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-17), **pp. 1228-1294, 2017**.

[9] Soni, HK., "Machine Learning – A New Paradigm of AI", *International Journal of Scientific Rsearch in Network Security and Communication,* **pp. 31-32,2019**

[10] Ghosh, S. & Waheed, Sajjad, "Strategic Analysis of classification algorithms for liver disease diagnosis", *Journal of Science Technology and Environment Informatics*,**pp. 361-370**, **2017**.

Authors Profile

Aparna Shenoy obtained her B.E. in Electronics and Communication from Canara Engineering College and M.Tech. in Digital Electronics and Advanced Communication from Manipal Institute of Technology Currently she is pursuing a Ph.D. in Computer Science in CMR University.Her research interests include various ClassificationAlgorithm,Disease Prediction, Machine Learning and Data Mining. The main idea behind her research is to provide ease to the doctors, various classification techniques can be utilized for early detection of Sepsis in ICU patients so that immediate intervention and essential precautions could be taken to prevent adverse events.

Vishwanatha K.V. has completed B.E. in Electrical Communication Engineering and M.E. degree in Electronics from Indian Institute of Science, Bangalore. He also possesses a Ph.D. in Electronics from IISc. Bangalore. He has more than two decades of teaching experience and has guided several students. He has over thirty publications to his credit. His research inerests are Alagorithms, Pogramming Languages, Network Security, Computer Graphics and Operating System.

Raju R Gondkar has 22 years of teaching and training experience with active research in the area of E -Learning, Image processing, Cloud Computing and Technology Management. He played the main role in drafting the Project report for establishing an Incubation Center to nurture entrepreneurship in IT sector with the support of State Government in 2004, which became the base for Incubation Centers sanctioned by Government of Karnataka. He was instrumental in establishing an IT incubation center with IBM and AIT Bangalore. He is currently working as professor at CMR University.

© 2020, IJCSE All Rights Reserved