Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 04, April 2021: 2310-2320

Research Article

Sepsis Prediction

Srijayanthi S[1], M Swetha[2], Sara S[3], Suchitra S[4], Sulaiha Shameena M[5], Swathi D[6]

[1]Assistant Professor, R.M.K Engineering College, R.S.M Nagar, Kavaraipettai
[2,3,4,5,6]Student, R.M.K Engineering College, R.S.M Nagar, Kavaraipettai.

Abstract

The intention of the proposed research is to create a machine learning model that predicts deadly disease at the earliest using the real time data and also to extend the research into a product-based approach. Generally, when an infection occurs in a body, when the immune system helps to overcome the infection by producing antibodies, but sometimes unusually the immune system overreacts to the infection. This will lead to a condition called Sepsis. India ranks in south asia in Sepsis death rate. The death rate of sepsis is 213 per lakh people in India. The major cause of death in the USA are Severe sepsis and septic shock. The early prediction of sepsis helps to quick recovery of the person. On the other hand, sepsis is very difficult to diagnose. The proposed solution contains the different methods of feature analysis and comparison between the classification model used. It can also reduce the death rate upto 13 per lakh people in India.

Keywords: *Machine learning, sepsis prediction, XGBoost, Hyperparameter tuning, Early Prediction*

Introduction

Sepsis is a deadly disease which occurs when a body's immune system immensely reacts to infections. It is the main pathway to death from infection. When an infection occurs in our body, our body's immune system will produce certain chemicals such as antibodies to fight against the infection. But sometimes the immune system will immensely react to infections, this causes Sepsis. This may also lead to death. Sepsis was previously classified as sepsis, severe sepsis and septic shock. Later, the sepsis was classified into sepsis and septic shock. The sepsis includes both sepsis and severe sepsis cases. The death rate due to sepsis is between 25% and 40%. In South Asia, India ranks second in Sepsis Death rate. It is estimated that approximately 213 per lakh people in India die every year due to Sepsis. There is no method to predict the sepsis at the earliest. The features such vital sign laboratory values can be used in machine learning models to predict the occurence of sepsis. In this paper, we

build a model using the XGBoost algorithm that predicts sepsis 6 hours before the clinical results.

Background (literature survey)

In 2013, the U.S spent $ 24 billion dollar [1]. In recent years, Medicare and Medicaid Services simulated an improvement in sepsis care and adoption of electronic health record (EHR) systems [2 3]. Machine learning based classification systems, plays an important role of exploration for sepsis research[4 5]. Using gradient boosted trees along with XGBoost classifier, a machine learning classifier was created using python[6]. After assessing the various scoring based prediction systems, SOFA and MEWS showed a little bias in identifying the disease[7]. Clinically, the efficiency of treatment can be improved by accurate identification of sepsis and predicting the early occurence of sepsis [8]. The possibility of delay in sepsis identification may be due to the poor sensitivity of the qSOFA [9].Inappropriate antibiotics use may be caused due to over diagnosis of sepsis [10]. The summary measure was performed using the AUROC, but it may fail due to imbalanced datasets[11]. The most recent consensus definition has to be used by all papers[12]. Although new advanced technology available for the treatment of sepsis like bundled early goal directed therapy (EGDT) [13], the dilemma regarding the onset of sepsis continues. The traditional clinical approaches were currently used. Omics techniques which screen disease-specific biomarkers in biological samples gives excellent results. Metabolomics is a notable approach out of the above mentioned methods. [17].

The metabolomics which gives an observation based on the past medical experiments, is highlighted in recent studies [18-20]. A curated dataset for sepsis is developed by a study and descriptive and quantitative meta-analysis is performed on the dataset. The findings of meta analysis were also validated by prospective metabolomic cohort study [21]. When comparing to the system like the National Early Warning System (NEWS) which usually indicates the onset of sepsis using the score-based analysis, the ML- based prediction system is highly capable [22-24]. The benefits of using machine learning for the sepsis prediction is given by Shimabukuro et al., for which he developed a randomised clinical trial (RCT) and achieved 13.5% (p=0.018) in death rate and hospital stay was reduced from 13 days to 10 days [25]. On the other hand, the present studies model has its own drawbacks. The data required by the model mostly contains vital signs which should be collected in forehand in order to do the analysis. This case does not work for the real - time diagnosis [26]. A research conducted the research on both inside-ICU and outside-ICU scenarios. An accurate system

for classifying should be developed, any false alarm from a system may lead to incorrect diagnosis of the patients. [27].

A research tested the working of typical CNN classifiers and found that it produces excellent results in classifying the large medical datasets when compared to best fine- grained classifiers [28].

Materials and Methods

3.1 General Description:

Sepsis is when a body's immune system immensely reacts to infection. It's also called septicemia. There is no standard test method for sepsis. A delay in identification of sepsis leads to death of a person. We collected the data of both sepsis and non sepsis patients which basically contains Vital signs, Laboratory values, Demographics and Sepsis label. We analyzed the nature of data. We removed the row which contains less than 30

non-null values and remaining null values are removed by KNN imputer. The data balancing was done using an under-sampling technique called NearMiss algorithm. Then, we passed the data through classifiers namely DecisionTree Classifier, AdaBoost Classifier, Random Forest Classifier, Gradient Boosting Classifier, GaussianNB, Logistic Regression, XGB Classifier. After Hyper Parameter tuning using Randomized Search CV for XGboost Algorithm which gives 90% accuracy. The true negative rate is 0.06%. Later this model is pickled and used in web applications to provide API service to Labs using the flask framework. The hospital can use this model to predict real time data either as a single patient data or a bunch of patient data. This will be the first fastest method to identify the sepsis in a patient. The Sepsis diagnosis through Machine Learning model is completely new to India. Our product would create a new trend of successfully using machine learning in medical issues. Here, Innovation is the creation of a machine learning model which identifies the sepsis at the faster rate (at least 6 hours before the clinical test).

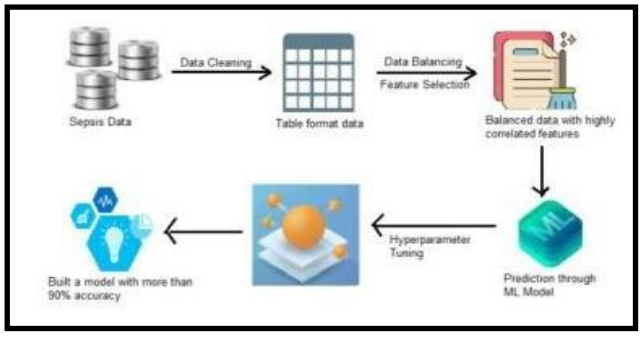

Figure 1 General workflow of our research

3.2 Importance Of Early Sepsis Detection:

Generally, Sepsis is identified by closely monitoring the patients EHR data and other related infections. But the challenge here is, the symptoms are not prominent at the earliest. A late prediction can lead to the death of a person. The Sepsis can kill the person as soon as 12 hours from the hours of development. The life time can also be extended from 30 days to 2 years.

The currently available clinical methods take days to predict the result, whereas each second count. Therefore, here is an approach to predict the sepsis at earliest provided the salient features are available. This approach also extends to prototype which will hopefully lead to a complete product which will eradicate Sepsis.

3.2 Data Pre-Processing:

The data contains a lot of null values and also the dataset is not balanced. Therefore, the data with more than 30 null values in a tuple were dropped. Sometimes, dropping the values may lead to data loss. However, imputing the missing values with less available values may lead to overfitting.

Therefore, the number of rows gets drastically reduced to 5975. Still the null values are not fully removed. It is not a good practice to use mean and median values to replace the null values in the clinical data. Hence, we used KNN imputer technique to replace the null values. The KNN imputer fills the missing values by calculating the nearest neighbour points.

Feature Selection is done using three techniques.

They are

3.3 Correlation Matrix

3.4 Select K best

3.5 Feature importance

1. Correlation Matrix:

On performing the feature analysis using correlation matrix, none of the features have correlation ratio greater than 0.5. Therefore, none of the features were selected from this method.

2. Select K best:

The Select K best algorithm is used along with the scoring function ‘chi2’. From the result obtained by this method, the features with scores greater than 50 were selected.

3. Feature importance:

The dataset was passed into a classification model called ExtraTreesClassifier which is an ensemble technique from sklearn. From the result obtained, the top 15 features with high

scores were selected. After trying out different feature analysis techniques, the following features were selected. 'O2Sat','Temp','SBP','EtCO2','HCO3','SaO2', 'BUN','PTT','Alkaline-phos', 'Calcium', 'Chloride', 'Creatinine', 'Glucose', 'Bilirubin_total','WBC','Fibrinogen', 'Platelets', 'SepsisLabel. The dataset is not balanced. The ratio of Sepsis patients record is very less that of the non-sepsis patients record. An under-sampling technique called NearMiss is used to balance the data. The number of sepsis patient records is 252 and that of non-sepsis patient records is 252.

Finally, a clean and balanced dataset is derived. The dataset contains 504 total records and 17 highly correlated features.

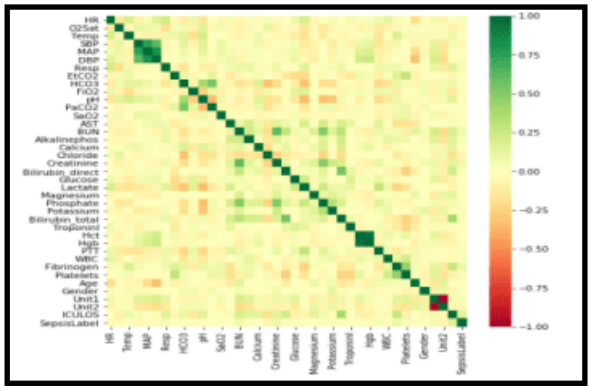

Figure 2 Correlation Matrix

	Title	Score
31	Fibrinogen	1968.677811
14	BUN	546.859690
32	Platelets	508.581713
15	Alkalinephos	456.873344
29	PTT	98.219382
3	SBP	52.011310
13	AST	49.127598
25	Bilirubin_total	42.597172
18	Creatinine	41.043881
6	Resp	30.468142
30	WBC	28.964834
20	Glucose	28.248489
0	HR	24.015329
4	MAP	19.996834
26	TroponinI	14.158456

Figure 3 Feature Selection by Select K Best

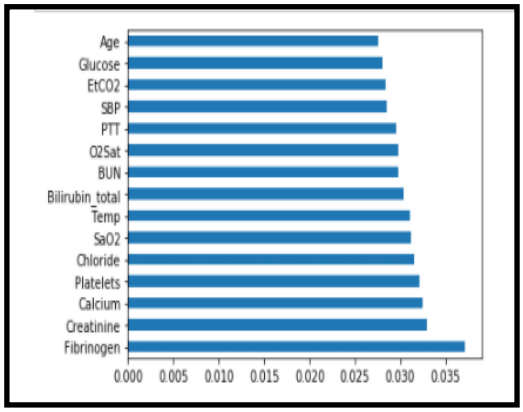

Figure 4 Feature Selection by Feature Importance

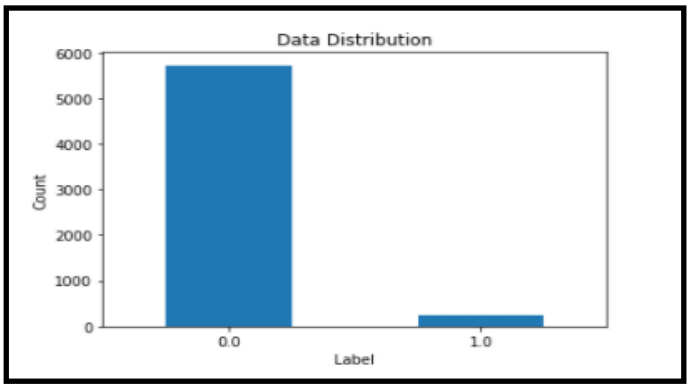

Figure 5 Data Imbalance

Modelling Of Sepsis Prediction

The dataset 80% and 20% which are train and test data respectively. These train data is passed through different classification models. The model used and accuracy obtained are as follow:

The accuracy of DecisionTreeClassifier is 0.81, AdaBoostClassifier is 0.85, RandomForestClassifier is 0.87, GradientBoostingClassifier is 0.89 , GaussianNB is 0.87, LogisticRegression is 0.78, XGBClassifier is 0.88. In order to increase the accuracy we used hyperparameter tuning using RandomizedSearchCV for XGBClassifier is 0.90.

Although the above-mentioned model gives high accuracy, further work was conducted to produce high accuracy. The hyperparameter tuning was done using GridSearchCV and RandomizedSearchCV. The hyperparameter tuning is generally done to identify the best parameter that gives high accuracy for a model.

Firstly, the GridSearchCV method is used for algorithm logistic regression with possible arameters [0.001,0.01,0.1,1,10] for the ‘C’ value. The results show that

best C = {'C': 1}

Highest accuracy obtained = 84%

Secondly, the GridSearchCV method is used for algorithm XGBoost with possible parameters [0.001,0.01,0.1,1,10] for the ‘C’ value. The results shows that

best C = {'C': 0.001}

Highest accuracy obtained = 87%

Lastly, the RandomizedSearchCV method is used for the algorithm XGBoost with possible parameters

"learning_rate" :[0.05, 0.10, 0.15, 0.20, 0.25,

0.30] ,

"max_depth" :[3, 4, 5, 6, 8, 10, 12, 15],

"min_child_weight" :[1, 3, 5, 7],

"gamma" :[0.0, 0.1, 0.2, 0.3, 0.4],

"colsample_bytree" :[0.3, 0.4, 0.5, 0.7]

The result show that the best parameters are 'min_child_weight': 3,

'max_depth': 12,

'learning_rate': 0.15,

'gamma': 0.3,

'colsample_bytree': 0.5

And accuracy obtained on train dataset is 93% and accuracy obtained on test dataset is 90%.

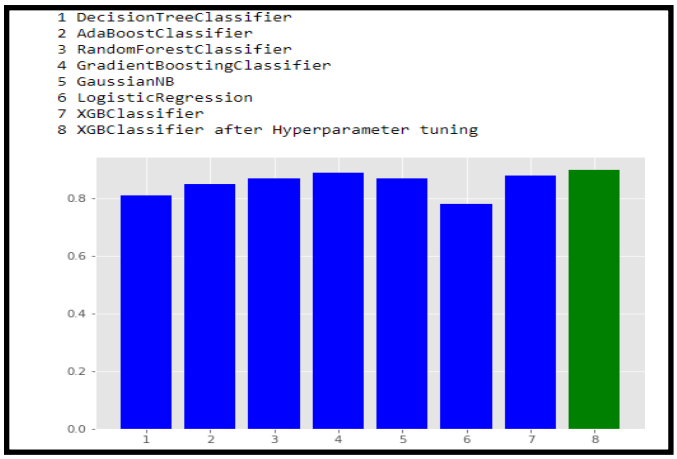

Figure 6 Comparison between accuracy obtained in different models.

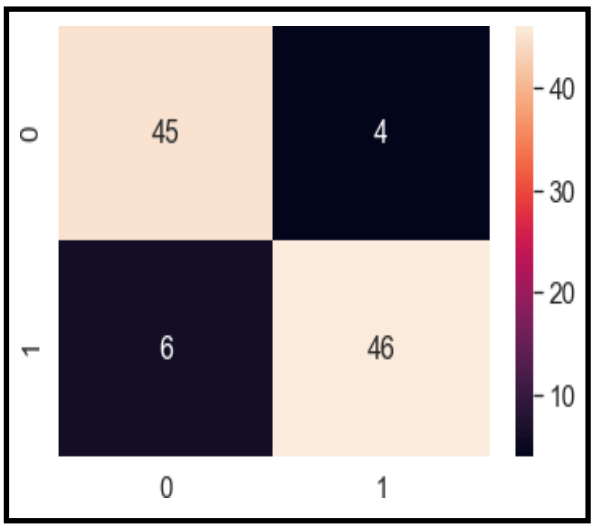

Figure 7 Confusion matrix

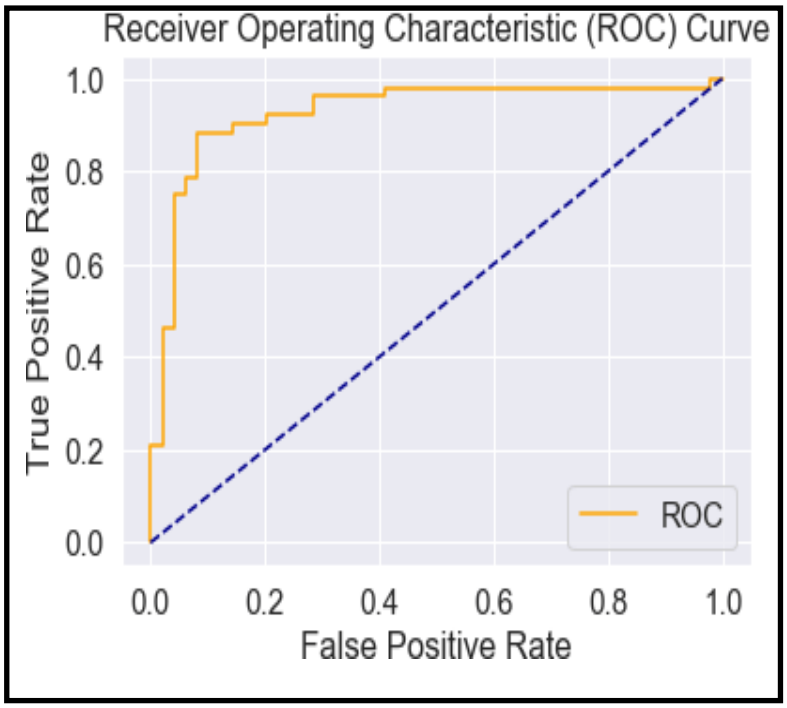

Figure 8 ROC Curve

Results and Discussion

The Data of sepsis and ono-sepsis are collected from ICU patients of three Hospitals, out of which data from only one hospital is used for this research work. The features in the dataset are generally Demographics, Vital Signs, and Laboratory values.

It contains 1172238 rows and 43 columns (features). The dataset also contains more null values. The features are given in the table below.

On the basis of research carried out, SelectKBest and Feature importance provide the highly correlated features. The XGBoost classifies more accurately than the other classifiers used with accuracy 90%.

If a sepsis patient is predicted as a non-sepsis patient by ML model, it will lead to a unexpectable results. This is called false- negative rate. The ratio of false negative to the total number of test dataset is 0.05 (relative very less value). When the sepsis patients data is correctly predicted, it is called True Positive rate.

The curve that is plotted against the false positive and true positive is known as ROC Curve (Receiver Operating Characteristic curve).

On plotting the ROC curve, the area obtained under the ROC curve is very large, it indicates our model is predicting in a good way. From the ROC curve, the threshold value can be derived based on needablity of the project. The best threshold value for using XGBoost classification is 0.556292 with high accuracy 0.89.

prototype is also included, such that the real time working of the model can be estimated

Figure 9 Predicting for test case 11

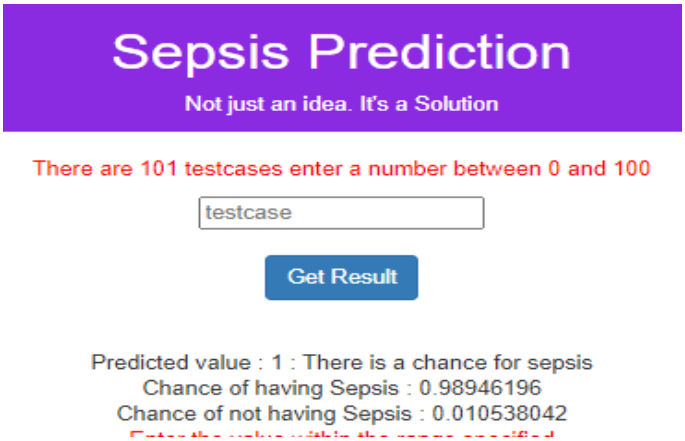

Figure 10 Result obtained for test case 11

Conclusion And Future Work

Our Machine learning model that was created is capable of prediction and estimating the probability of occurrence of sepsis. The model developed produces an accuracy of 90% and the threshold value 0.5 tends to produce highest accuracy. This will help to reduce the death rate from 213 per lakh to 13 per lakh in India. The future work should include analysis with different feature selection methods and classification methods. If some of the highly correlated features required for classification are available on time, the prediction can be delayed. Further development of the prototype is also included, such that the real time working of the model can be estimated.

References

[1] Torio C, Moore B. National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2013. HCUP Statistical Brief #204. Agency for Healthcare Research and Quality, Rockville, MD, 2016. Available: http://www.hcup-us.ahrq.gov/reports/statbriefs/sb204-MostExpensive-Hospital-Conditions.pdf

[2] Office of the National Coordinator for Health Information Technology. 'Office- based Physician Electronic Health Record Adoption,' Health IT Quick-Stat #50, 2016. Available: http://www.webcitation.or g/ 6rmdNMHPW

[3] HealthIT.gov. EMR Incentives & Certification, 2013. Available: https:// www.healthit.gov/providers-professionals/ ehr-incentive-programs

[4] Delahanty RJ, Alvarez J, Flynn LM, et al. Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. Ann Emerg Med 2019;73:334–44.

[5] Horng S, Sontag DA, Halpern Y, et al. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PLoS One 2017;12:e0174708.

[6] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Paper presented at the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[7] Innocenti F, Tozzi C, Donnini C, et al. SOFA score in septic patients: incremental prognostic value over age, comorbidities, and parameters of sepsis severity. Intern Emerg Med 2018;13:405–12.

[8] Møller MH, Alhazzani W, Shankar-Hari M (2019) Focus on sepsis. Intensive Care Med 45:1459–1461. https://doi.org/10.1007/ s00134-019-05680-4

[9] Serafm R, Gomes JA, Salluh J, Póvoa P (2018) A comparison of the quick-SOFA and systemic inflammatory response syndrome criteria for the diagnosis of sepsis and prediction of mortality: a systematic review and meta-analysis. Chest 153:646– 655. https://doi.org/10.1016/J.CHEST.2017.12.015

[10] Hiensch R, Poeran J, Saunders-Hao P et al (2017) Impact of an electronic sepsis initiative on antibiotic use and health care facility–onset clostridium difficile infection rates. Am J Infect Control 45:1091–1100. https://doi. org/10.1016/j.ajic.2017.04.005

[11] He Haibo, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21:1263–1284. https://doi.org/10.1109/TKDE.2008.239

[12] Singer M, Deutschman CS, Seymour CW et al (2016) The third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA 315:801. https://doi.org/10.1001/jama.2016.0287

[13] Rivers EP, Coba V, Visbal A, Whitmill M, Amponsah D. Management of sepsis: early resuscitation. Clin Chest Med. 2008;29(4):689–704 ix-x.

[14] Innocenti F, Tozzi C, Donnini C, De Villa E, Conti A, Zanobetti M, Pini R. SOFA score in septic patients: incremental prognostic value over age, comorbidities, and parameters of sepsis severity. Intern Emerg Med. 2018; 13(3):405–12.

[15]. Ho KM, Dobb GJ, Knuiman M, Finn J, Lee KY, Webb SA. A comparison of admission and worst 24-hour Acute Physiology and Chronic Health Evaluation II scores in predicting hospital mortality: a retrospective cohort study. Crit Care. 2006;10(1):R4.

[16] Moreno RP, Metnitz PG, Almeida E, Jordan B, Bauer P, Campos RA, Iapichino G, Edbrooke D, Capuzzo M, Le Gall JR, et al. SAPS 3--from evaluation of the patient to evaluation of the intensive care unit. Part 2: development of a prognostic model for hospital mortality at ICU admission. Intensive Care Med. 2005;31(10):1345–55.

[17] Liu X, Ren H, Peng D. Sepsis biomarkers: an omics perspective. Front Med. 2014;8(1):58–67.
[18] Kiehntopf M, Nin N, Bauer M. Metabolism, metabolome, and metabolomics in intensive care: is it time to move beyond monitoring of glucose and lactate? Am J Respir Crit Care Med. 2013;187(9):906–7.
[19] Zurfluh S, Baumgartner T, Meier MA, Ottiger M, Voegeli A, Bernasconi L, Neyer P, Mueller B, Schuetz P. The role of metabolomic markers for patients with infectious diseases: implications for risk stratification and therapeutic modulation. Expert Rev Anti-Infect Ther. 2018;16(2):133–42.
[20] Dos Santos CC. Shedding metabo 'light' on the search for sepsis biomarkers. Crit Care. 2015;19:277.
[21] Wang, Jing, et al. "Prediction of sepsis mortality using metabolite biomarkers in the blood: a meta-analysis of death-related pathways and prospective validation." BMC medicine 18 (2020): 1-15.
[22] R.P.Dellinger,''Incidence Surviving Sepsis Campaign:International Guidelines for management of severe sepsis and septic shock, 2012,'' Intensive Care Med., vol. 39, no. 2, pp. 165–228, 2013.
[23] H. M. Giannini, J. C. Ginestra, C. Chivers, M. Draugelis, A. Hanish, W. D. Schweickert, B. D. Fuchs, L. Meadows, M. Lynch, P. J. Donnelly, K.Pavan,N.O.Fishman,C.W.Hanson,andC.A
[24] .Umscheid,''A machine learning algorithm to predict severe sepsis and septic shock:Development, implementation, and impact on clinical practice,'' Crit.CareMed.,vol.47, no. 11, pp. 1485–1492, Nov. 2019.
[25] O. A. Usman, A. A. Usman, and M. A. Ward, ''Comparison of SIRS, SOFA, and NEWS for the early identification of sepsis in the emergency department,'' Amer. J. Emergency Med., vol. 37, no. 8, pp. 1490– 1497, Aug. 2019.
[26] B.S.dosSantos,M.T.A.Steiner, A.T.Fenerich, and R.H.P.Lima, ''Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018,'' Comput. Ind. Eng., vol. 138, Dec. 2019, Art. no. 106120.
[27] Halligan,D.G.Altman,andS.Mallett,''Dis advantages Of Using The Area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach,'' Eur. Radiol., vol. 25, no. 4, pp. 932–939, Apr.2015.
[28] Al-Mualemi, Bilal Yaseen, and Lu Lu. "A Deep Learning-Based Sepsis Estimation Scheme." IEEE Access (2020).
[29] T.Sainath, A.-R.Mohamed, B.B. Kingsbury and K.Ramabhadran,''Deep convolutional neural networks for LVCSR,'' in Proc. Acoust., Speech Signal Process., May 2013, pp. 8614–8618