EasyChair Preprint

№ 7495

Prediction and Analysis of Sepsis by Using Machine Learning Algorithms (XG Boost & Light GBM)

Sivasankari Kannan, Priyadharshini Subramanian, Bharathi Arivalagan and Murugeshwari Adhiappan

February 23, 2022

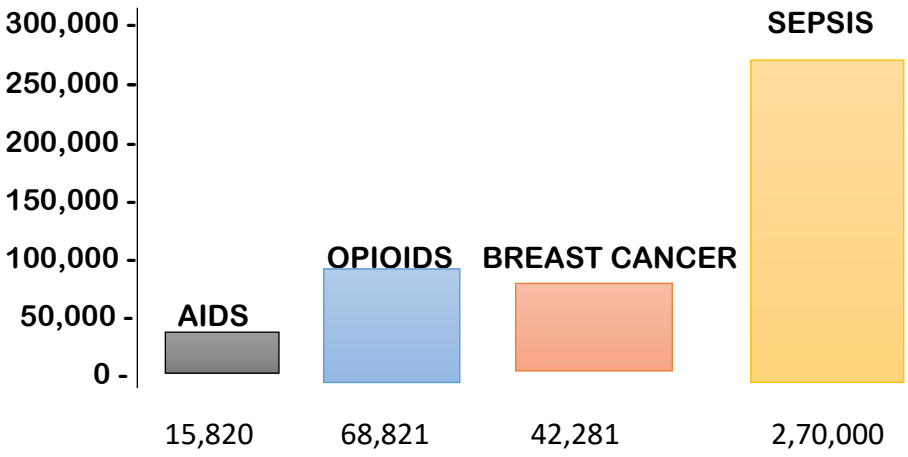
PREDICTION AND ANALYSIS OF SEPSIS BY USING MACHINE LEARNING ALGORITHMS (XG BOOST & LIGHT GBM)

ABSTRACT

Sepsis is a fatal condition that develops from blood poisoning. It occurs when the immune system attacks the body as it fights off infection. Sepsis is a medical emergency that should be treated soon as it develops. We like to frame two processing methods that are the mean processing method and the feature generation method by machine learning algorithms like XG Boost and Light GBM. These are designed to predict sepsis 6 hours in advance. XG Boost and Light GBM algorithm both play an admirable role in prediction performance (AUC:910~0.979), whereas Light GBM is the fastest acting in performance. It is powerful on multidimensional data. The key factor to predict early sepsis are WBC, platelets, and PTT.

INTRODUCTION

In 2016, Center for disease control and prevention (CDC) reported that sepsis takes the life of more Americans when compared to disease such as Aids, Breast Cancer and Opioid overdose.Sepsis kills 270,000 people in the United States for every year and 1.7 million people were get affected which is the most expensive influence of hospitalization in the United states



Sepsis is mainly caused by bacterial, virus, fungi infections. When an existing infection stimulates an excessive immune system response in our body develops sepsis. An infection has occurred, proteins and other chemicals are released into the bloodstream to fight against an infection trigger inflammatory response throughout the body can damage vital organ. Sepsis, severe sepsis, septic shock are the three stages of sepsis. Machine learning provides an efficient method in medical health care such as sepsis diagnosis earlier. Here the statistical strength feature, window feature, and medical feature are built by the feature generation method. Then to handle large missing data problems we are using the miceforest multiple interpolation methods.

LITERATURE SURVEY

By using linear regression, sepsis-related cases in the hospital were developed for all possible locations. The report of sepsis linked death as 11.0 million from that estimation of 48.9 million in the year of 2017 all over the world. Firmly, there was a standard for identifying the risk of sepsis on the patient by using Systematic Inflammatory Response Syndrome (SIRS). The ICU of critical care medicine and world federation of societies of intensive organized the audit of data throughout the world. Rationing a large core of database whereas to collect information finally the data is to identify the sepsis of third patients such as pathogen, an intercontinental difference of rate in assurance and outcomes.

Author, year (reference)	Countries	Design	Number of intensive care unit (ICU) admissions screened	Outcome	Relative frequency (%)	Mortality (%)
Alberti, 2002 (20)	Six European countries, Canada, and Israel	Prospective cohort study	14 364	Infectious episodes	21.1	22.1 vs. 43.6 (community- vs. hospital-acquired infection)
Padkin, 2003 (21)	England, Wales, and Northern Ireland	Administrative database	56 673	Severe sepsis	27.1	35 vs. 47 (ICU vs. hospital mortality)
Annane, 2003 (22)	France	Administrative database	100 554	Septic shock	8.2	60.1
EPISEPSIS, 2004 (23)	France	Prospective cohort study	3 738	Severe sepsis or septic shock	14.6	35 vs. 41.9 (30-day vs. 2-month mortality)
Finfer, 2004 (24)	Australia and New Zealand	Prospective cohort study	5 878	Severe sepsis	11.8	26.5 vs. 32.4 (ICU vs. 28-day mortality)

Ponce de León, 2000 (34)	Mexico	Cross-sectional	Admissions to 254 ICUs (n = 895)	1-day prevalence of infections	294/895 (ICU admissions)	33.6
Zapata, 2001 (35, 36)[d]	Colombia	Prospective cohort study	Patients with nontraumatic SIRS at 2 hospitals (n = 533)	Sepsis	Not reported	23.5
Sifuentes, 2001 (37)	Mexico	Cross-sectional	Patients with bacteremia (n = 600)	Characterization of bacteremic patients	3 428/19 530[g] (blood cultures)	28
Morales, 2001 (38)	Cuba	Cross-sectional	Hospitalized patients[h]	Nosocomial infections	4/100[c,h] (discharges)	Not reported
Bilevicius, 2001 (39)	Brazil	Retrospective case-series	Admissions to ICU (n = 249)	Sepsis	54/249 (ICU admissions)	56
Luján, 2002 (40)	Cuba	Surveillance	Patients with nosocomial infections at 3 hospitals[h]	Nosocomial infections	5.3/100[c,h] (discharges)	Not reported
Cordero, 2002 (41)	Cuba	Retrospective case-series	Patients with nosocomial infection (n = 1 241)	Nosocomial infections	219/1 241[c] (nosocomial infections)	Not reported
Notario, 2003 (42)	Argentina	Retrospective case-series	Patients with bacteremia (n = 596)	Characterization of bacteremic patients	596/6 605[g] (blood cultures)	Not reported
Jaimes, 2003 (5)	Colombia	Prospective cohort study	Patients admitted at two emergency rooms (n = 734)	Sepsis	657/734 (infection as cause for admission)	30.7
Jaimes, 2004 (43)	Colombia	Cross-sectional	Patients with request for blood cultures (n = 500)	Nosocomial bacteremia	89/500[g] (blood cultures)	22.6 vs. 36 (negative vs. positive blood cultures)

FRAMEWORK AND APPROACH

In physiological ICU database of three independent hospital systems has 1714 sepsis patients out of 22326 patients. Within 1 hour a data frequency shows 790,125 observations where it has 40 indicators such as 8 vital signs, 6 demographic indicators, and 26 laboratory values which appear below the table A. Per day, ninety percent of missing values are caused by measuring the laboratory values that table A contains vital sign, demographic, laboratory values and has a long gap of intervals and most of the values are missing. Suppose we ignore missing values, most of the information will be lost to predict sepsis. In this study feature generation and mean feature generation method solves the problems of missing values. The original data has 790215 observations there are 22336 patients among those 1714 patients are suffered from sepsis and so the observation is named sepsis label 0 and sepsis label 1. Therefore, the ratio of this category is 5:1

TABLE A:

Vital Signs	Unit	Missing Percentage
Heart Rate (HR)	Beats per minute	7.7%
Pulse Oximetry (O2Sat)	%	12.0%

Temperature (Temp)	Deg C	66.2%
Systolic BP (SBP)	Mm. Hg	15.2%
…	…	…

Laboratory Variables	**Unit**	**Missing Percentage**
Base Excess (measure excess of HCO3)	Mmol/L	89.6%
Bicarbonate (HCO3)	Mmol/L	91.9%
Fraction of inspired oxygen (FiO2) pH	%	85.8%
Ph	/	88.5%
…	…	…

Demographics	**Unit**	**Missing Percentage**
Age	Years	0.0%
Gender	1 (Male) or 0 (Female)	0.0%
MICU (Unit1)	1 (No) or 0 (Yes)	48.9%
…	…	…

TABLE B

Basic Information	**Counts**	**Sum of Counts**	**Proportion (%)**	**Sum of proportion (%)**
Patients with Sepsis	20662	22336	92.33	100
Patients without Sepsis	1714		7.67	

0 (Sepsis Label)	773080	790215	97.83	100
1 (Sepsis Label)	17135		2.17	

MACHINE LEARNING ALGORITHM TO FORTELL EARLY SEPSIS

From the reference [10 - 14] the exact information of two structured tree algorithms which is XGBoost and Light GBM are estimated. Here we are explaining the model by engaging the SHAP value and feature importance score. For calculation of feature importance score.

$$\hat{I}^2_j = 1/M \sum_{n=0}^{N} \hat{I}^2_j(T_n)$$

N represents number of trees and T_n represent the nth tree, l-1 is the number of non-leaf nodes of the tree, f denotes the feature selected.

$$\hat{I}^2_j\ (T) = \sum_{s=1}^{l-1} \hat{I}^2_j I\ (v_{s=}j\ )$$

When the internal not S is split and After the split of internal nod S. X^2 is the reduction of square loss(MSE). If the larger value X^2, potential to lower the loss and more intense the ability to fit. There is a unrealializable to recognize the result of final production and the feature. So here the SHAP value can examine. Basically Shapley values is inspired to additive interpretation mode which is SHAP value. Let us assume that mean value of all the samples be Z base. Among which k th sample is h_k, the l th feature of the k th sample be h_k, l, the SHAP value of this feature be f(h_k, l), Therefore the predicted value is

$$Z_k = Z_{base} + f(h_k, 1) + f(h_k, 2) + \ldots + f(x_k, n)$$

If f(h_k, l) is greater than 0, then the feature prediction of target value is positive otherwise has negative effect. The SHAP value can be considered either dominant of prevalent characteristic of characters in every sample

MAKEOVERS OF WARNING PERIOD IN MEAN PROCESSING METHOD

To find out the specific observation, we directly integrated every patient for warning period of 6 hours in the previous mean processing method. Here the implementation of prediction model may not be acceptable. so we determine if the segmentation of window is better or firmer that can create best execution or not. Hence to evaluate the mean vector, we split up the warning period into two windows of 2 hours or 3 hours' time. The disease

period and the safe period in the mean processing method is remains unvaried.

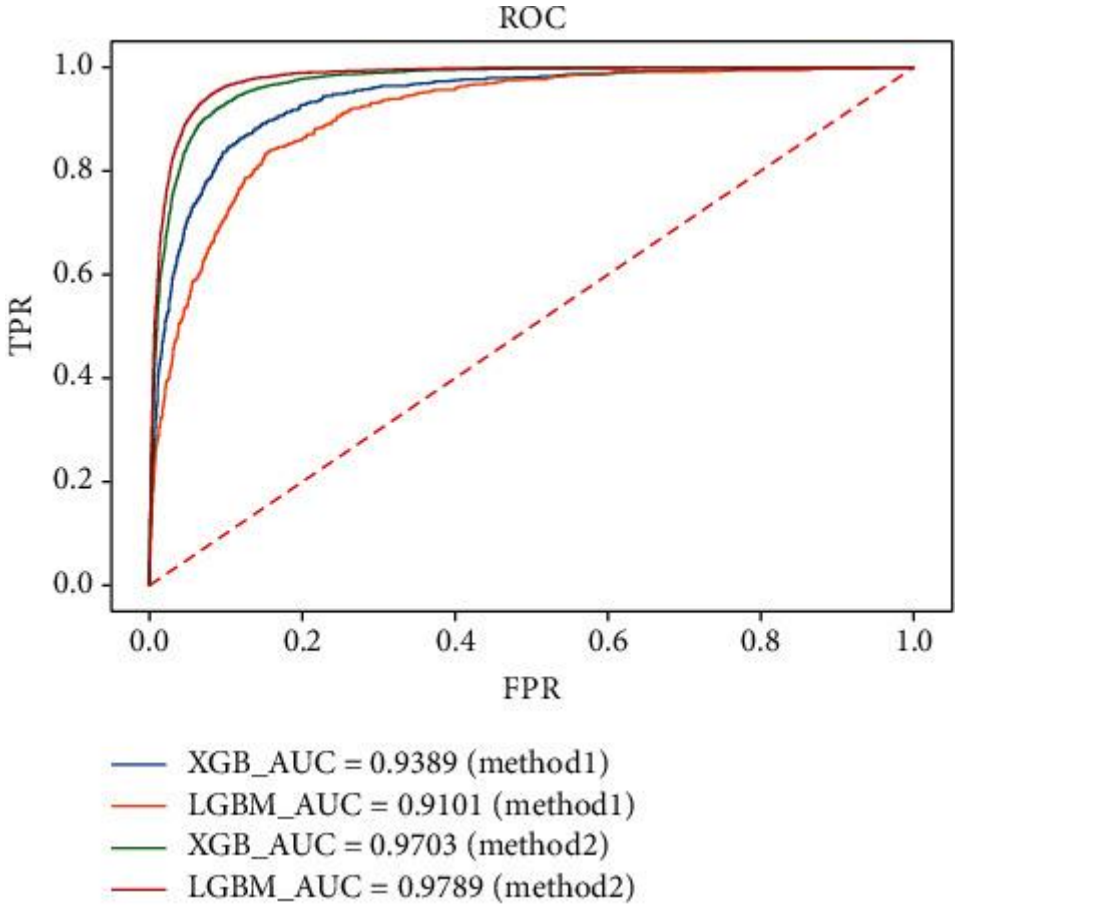
RESULT

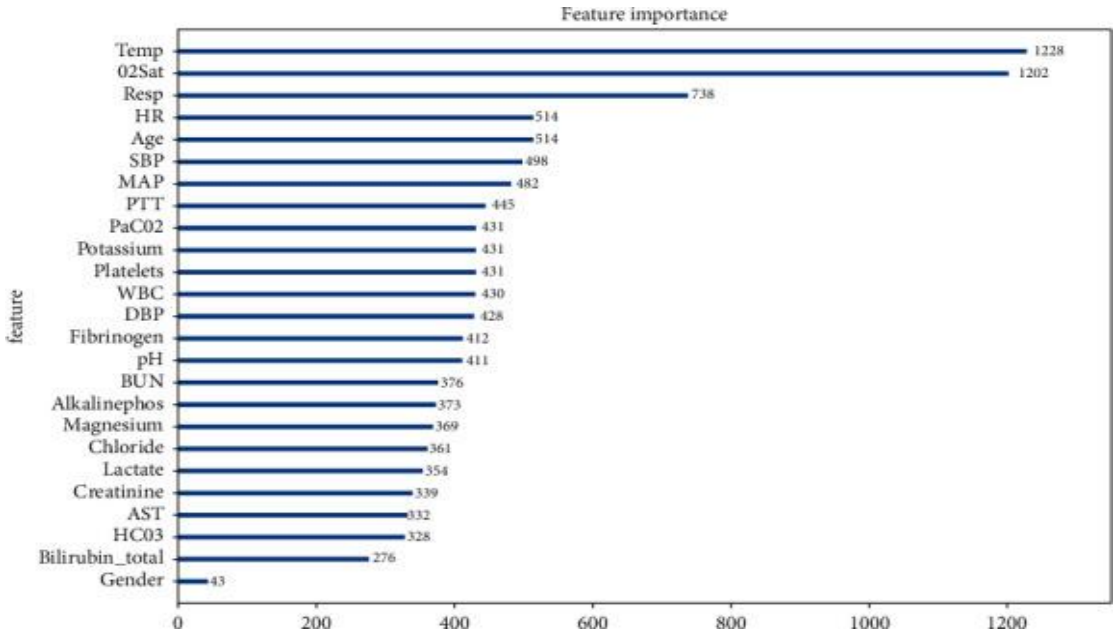
Here, We have chosen only 70% data for training, and a balance of 30% is for verification of test set and further evaluation in mean processing and feature generation method

METHODS		PRECISION	RECALL	F1-SCORE	KAPPA COEFFICIENT	MATTHEWS COEFFICIENT
MEAN PROCESSING METHOD	XGBoost	0.78*	0.55*	0.65*	0.60*	0.67*
	Light GBM	0.70	0.42	0.53	0.46	0.48
FEATURE GENERATION METHOD	XGBoost	0.89	0.61	0.72	0.67	0.69
	Light GBM	0.91*	0.65*	0.76*	0.72*	0.73*

PERFORMANCE OF MODEL

The performance of the model will differ by using lightGBM and XG boost algorithm in the mean processing method. The XGboost algorithm has a rate of o.55 so it will have better performance around 0 to 1 category. A comparison between the result of Mattews coefficient and kappa coefficient was made, therefore it seemed to be that the XGBoost algorithm has the most balanced on the result of the test Conversely, LightGBM has a better performance in the model between 0 to 1 categories. This shows that LightGBM is more excellent and overall the performance of LightGBM in the feature generation method is the best for this model to predict sepsis more easily.


ROC
TPR
FPR
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
XGB_AUC = 0.9389 (method1)
LGBM_AUC = 0.9101 (method1)
XGB_AUC = 0.9703 (method2)
LGBM_AUC = 0.9789 (method2)


Feature importance
feature
Temp 1228
02Sat 1202
Resp 738
HR 514
Age 514
SBP 498
MAP 482
PTT 445
PaC02 431
Potassium 431
Platelets 431
WBC 430
DBP 428
Fibrinogen 412
pH 411
BUN 376
Alkalinephos 373
Magnesium 369
Chloride 361
Lactate 354
Creatinine 339
AST 332
HC03 328
Bilirubin_total 276
Gender 43
0
200
400
600
800
1000
1200

DISCUSSION AND CONCLUSSION

The best way to train the feature generation method which is one of the data processing methods we use the LightGBM algorithm for it. Further SHAP value clearly defines the prediction result.

Mainly we are using the main processing method to remakes the complex data and ignore the occurrence of missing data one problem in this. we can only retrieve the information from different states of data. Sometimes the important information will be lost which deficient the capability to predict the model. Model's AUC peaks to 0.97 and determine the mean vector of the divided warming period data from new per 2 hours or 3 hours by using the improved mean processing method.

From the feature generation, method 1,00,000 observations are underscored, which came by a large amount of data. Once we have filled up the missing values by mice forest it will automatically arrive AUC of original data to 0.971 therefore we have seen the greatest improvement that AUC reaches 0.979.

LightGBM has the power to predict the model in a better way and also train the model effectively fast when compared to XGBoost. The reason behind this LightGBM takes up a low amount of memory and implements a strategy of leaf-wise based growth.

WBC, PTT, and platelets provide a helpful way to predict sepsis accurately, to analyze the organ function the PTT and Platelets are coagulating indicators. When the count of White Blood Cell(WBC) are amplified it causes changes in long-lasting bacterial infections.

REFERENCE

1. Stekhoven D. J. Buhlmann P. Miss Forest—non-parametric missing value imputation for mixed-type data. *Bioinformatics.* 2012;**28**(1):112–118. DOI: 10.1093/bioinformatics/btr597

2. Lin C., Zhang Y., Ivy J., et al. Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-LSTM. Proceedings of the 2018 IEEE International Conference on Healthcare Informatics (ICHI); June 2018; New York, NY, USA. IEEE; pp. 219–228.

3. Reyna M., Josef C., Jeter R., et al. Early prediction of sepsis from clinical data-the Physio Net computing in cardiology challenge 2019 (version 1.0.0) *Physio Net.* 2019 DOI: 10.13026/v64v-d857.

4. Stekhoven D. J., Buhlmann P. Miss Forest—non-parametric missing value imputation for mixed-type data. *Bioinformatics.* 2012;**28**(1):112–118. DOI: 10.1214/07-sts242.

5. Li X., Xu X., Xie F., et al. A time-phased machine learning model for real-time prediction of sepsis in critical care. *Critical Care Medicine.* 2020;**48**(10):e884–e888. DOI: 10.1097/ccm.0000000000004494.

6. Rudd K. E., Johnson S. C., Agesa K. M., et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet.* 2020;**395**(10219):200–211. DOI: 10.1016/s0140-6736(19)32989-7.

7. Taneja I., Reddy B., Dam horst G., Zhao S. D. Combining biomarkers with EMR data to identify patients in different phases of sepsis. *Scientific Reports.* 2017;**7**(1):1–12. DOI: 10.1038/s41598-017-09766-1.

8. Raaijmakers Q. A. W. Effectiveness of different missing data treatments in surveys with liker-type data: introducing the relative mean substitution approach. *Educational and Psychological Measurement.* 1999;**59**(5):725–748. DOI: 10.1177/00131649921970116.

9. Garcia-Gallo J. E., Fonseca-Ruiz N. J., Celli L. A., Dui Tama-Muñoz J. F. A machine learning-based model for 1year mortality prediction in patients admitted to an intensive care unit with a diagnosis of sepsis. *Medicine Intensive.* 2020;**44**(3):160–170. DOI: 10.1016/j.medin.2018.07.016.

10. Chen T., He T., Benesty M., Khotilovich V., Tang Y., Cho H. Boost: an extreme gradient boosting. *R Package Version 0.4-2.* 2015;**1**(4):1–4

11. Kim J., Chang H., Kim D., Jang D. H., Park I., Kim K. Machine learning for prediction of septic shock at initial triage in the emergency department.

Journal of Critical Care. 2020; **55:163**–170. DOI: 10.1016/j.jcrc.2019.09.024.

12. Buhlmann P., Hot horn T. Boosting algorithms: regularization, prediction and model fitting. *Statistical Science.* 2007;**22**(4):477–505. DOI: 10.1214/07-sts242.

13. Hari Haran S. *Real-Time Sepsis Prediction Using an End-to-End Multi-Task Gaussian Process RNN Classifier.* Durham, NC, USA: Duke University; 2017.

14. Kumar, A. et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit. Care Med.* **34**, 1589–1596 (2006).

15. Ohno-Machado, L. Realizing the full potential of electronic health records: the role of natural language processing. *J. Am. Med. Inform. Assoc.* **18**, 539–539 (2011).

16. Sweeney, T. E. et al. A community approach to mortality prediction in sepsis via gene expression analysis. *Nat. Common.* **9**, 694 (2018).

17. Rhodes, A. et al. Surviving sepsis campaign: international guidelines for the management of sepsis and septic shock: 2016. *Intensive Care Medication.* **43**, 304–377 (2017).

18. Shashi Kumar, S. P. et al. Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. *J. Electrocardiol.* **50**, 739–743 (2017).

19. Batista, G. E., Prati, R. C. & Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explore. News.* **6**, 20–29 (2004).

20. Jaime's F. A literature review of the epidemiology of sepsis in Latin America. Rev Panam Salud Public a. 2005;18(3):163–71.