Early Prediction of Sepsis Based on Patient Vital Signs And Laboratory Values Using XGboost

Anurag Dubey[1], Neha Khare[2]
[1, 2] Dept of CSE
[1, 2] TIET, Jabalpur, MP

Abstract- *Sepsis is a fatal disease with a high mortality rate, especially in intensive care patients. Early and accurate diagnosis of sepsis is important because delay in treatment increases mortality. Infectious Disease Prevention, Clinical Abnormalities and Early Warning Signs are the usual diagnostic criteria for diagnosing sepsis in practice. However, the score cannot provide an early prediction for sepsis, in which the mortality rate will decrease if patients receive emergency treatment. The person applying for isolation can predict the fact of sepsis 6 hours before the diagnosis of the disease. To achieve this, a patient's electronic medical record, demographic information, and vital signs are used. This study uses a data preprocessing strategy adapted to the dataset. This plan introduces a new outlier-based mean data evaluation method, increasing the value of existing data and thus improving the overall accuracy of the prediction. It is made easier for physicians to understand the model by providing an explanation of the main points that affect the distributor's estimate.*

Keywords- Sepsis detection, vital sign, Laboratory values, Machine Learning, Accuracy, XGBoost.

I. INTRODUCTION

Sepsis is a life-threatening medical emergency that can rapidly lead to tissue damage, organ failure, and death [1]. Sepsis is considered responsible for more than one-third of the hospital deaths in the United States and the increased incidence have been a growing concern [2]. It is one of the most expensive conditions to treat, representing 13% of the total U.S. healthcare cost. Additionally, statistics show that the average length of stay in hospitals for sepsis patients is nearly 75% longer than that of other medical conditions [3]. It has been reported that the early intervention and recognition of sepsis can significantly reduce the overall mortality and cost burden of sepsis. The importance of early prediction and treatment of sepsis is emphasized in the current clinical and observational studies that show a lower risk of mortality for sepsis patients who received antibiotics and intravenous fluids on time [4], [5]. In another study, it is reported that hourly delays in the initiation of antibiotic therapy can cause an average increase in the mortality rate by 7.6% [6]. In the context of sepsis diagnosis, the Systemic Inflammatory Response Syndrome (SIRS) criteria were considered to be central [7]. Recently, the third international consensus definition for sepsis and septic shock (Sepsis-3) was published. For diagnostic criterion, the Sequential Organ Failure Assessment (SOFA) scoring system was proposed. Furthermore, SIRS criteria have been criticized for inadequate specificity and sensitivity since SIRS may occur in several non-infectious scenarios [8]. The SOFA score is based on the degree of dysfunction of six organ systems, such as respiratory, coagulation, hepatic, cardiovascular, renal, and neurological systems [9]. According to the Sepsis-3 guidelines, patients with a SOFA score of 2 or more are associated with an organ failure consequent to the infection, meaning that a higher SOFA score indicates the increased mortality risk. The Modified Early Warning Score (MEWS) is another scoring system used for the determination or prediction of sepsis [10]. These updated definitions and gold standards have been adapted to facilitate the earlier identification and timely management of septic patients. However, sepsis is a dynamic condition and, hence, such criteria may not yield accurate outcomes. Consequently, early prediction of the onset of sepsis remains a challenging problem. There has been a significant surge in using deep learning as well as machine learning for solving multivariate, complex, and nonlinear problems. Training such models requires a significant volume of data. Meanwhile, the intensive care unit (ICU) patients are monitored

Early prediction of Sepsis based on patient Vital Signs and Laboratory Values using XGBoost consistently. This has generated an abundance of data, which allows for training DNNs for event prediction or decision support in critical care cases [11]. Recent studies have incorporated such DNN-based approaches using electronic health records (EHRs) for identifying the early stages of complex diseases [12], [13], [14].

II. LITERATURE REVIEW

Sepsis is a clinical syndrome of physiologic, pathologic, and biochemical abnormalities induced by infection leading to life-threatening acute organ dysfunction

[15]. Incidence of sepsis in US hospitals account for 5.9% (1.7 million) of the hospitalizations and 15.6% (over a quarter of a million) of the in-hospital deaths in 2014 [16]. Furthermore, infectious etiology including sepsis is found to be the most common cause for 30-day hospital readmission. Sepsis readmission costs about $3.5 billion annually within the United States [17]. Thus, sepsis is very common, often fatal and requires rapid and timely interventions to improve overall clinical outcomes and more importantly enhance patient survival [18]. According to the Third International Consensus on sepsis-3 criteria, sepsis is now defined as 'life threatening organ dysfunction caused by deregulated host response to infection', and its clinical diagnosis is based on acute changes in Sequential Organ Failure Assessment (SOFA) score of ≥ 2. The quick SOFA (qSOFA) score was developed as a bedside tool to rapidly screen the patients at risk of sepsis outside critical care [16]. Although sepsis-3 criteria is increasingly common in intensive/critical care unit clinical trials for retrospective analysis, there are concerns regarding the complexity of the SOFA score, the lack of clinical evidence for the validity of sepsis-3 criteria, its applicability for widespread clinical practice, and the gap in recommendations that prompt necessary measurements and laboratory tests [19]. The prognostic accuracy of systemic inflammatory response syndrome (SIRS), SOFA, qSOFA for sepsis prediction vary widely among various retrospective clinical trials [19], [20] and are limited for early detection of sepsis as compared to the early warning score methods [20]. The timing of sepsis diagnosis is important, as it profoundly affects the clinical outcomes of patients as well as healthcare utilization and costs [21]. Therefore, accurate and reliable early diagnosis of sepsis is critical to lower the sepsis related mortality and healthcare costs. The PhysioNet/Computing in Cardiology Challenge 2019 focused on the development of automated machine learning (ML) algorithms for early sepsis detection from clinical and physiological data sourced from ICU patients. The challenge given to the research community was to predict sepsis 6 hours before the clinical prediction of sepsis [22]. Previous studies explored the potential for ML-based approaches to enhance ICU patient outcomes with sepsis [23]; however, the effectiveness and utility of these algorithms on clinical practice and patient outcomes, particularly outside of ICU settings, are yet to be established [24]. It is important to consider how sepsis prediction changes with the data available in these different settings, how patients need to be monitored across settings, as well as the multitude of factors that can influence the performance of sepsis prediction in diverse populations. This thesis delineates the influence of (i) feature selection among objective vital measures and laboratory biometrics and expert inputs and (ii) the choice of more homogenous/heterogeneous patient populations in the performance of ML algorithms for early detection of sepsis. Even though there have been many attempts to use machine learning (ML) to find sepsis early, it remains a significant concern for healthcare stakeholders worldwide. ML accelerates data processing and analysis, which can greatly aid in early prediction. With minor changes in deployment, predictive analytic approaches that use machine learning could be trained on increasingly larger data sets and provide deeper analysis on a variety of aspects. Early prediction aims at identifying the onset of sepsis well before a physician can do it. An accurate early predictive model would help the physicians to have a closer look at the patients well before the onset of sepsis and could reduce the morbidity rate as well as the financial cost of treating the patient. The four major categories of ML algorithms are supervised, unsupervised, semi-supervised, and reinforcement learning

- Supervised algorithms use labeled data sets to train the predictive or classification model and then the trained model is tested on unlabeled data sets to measure the effectiveness of the model. Supervised learning is used for sepsis prediction
- An unsupervised algorithm uses unlabeled data set and is especially used for clustering applications.
- Semi-supervised learning is the hybrid of supervised and unsupervised learning algorithms that use both labeled and unlabeled data sets. Using additional information from unlabeled data will surely improve the prediction outcome and is attempted by a few authors in early sepsis prediction [4].
- Reinforcement Learning (RL) does not need any data set and the learning agent continuously interacts with the environment to derive an optimum strategy for sequential-decision problems. RL algorithms are used by researchers for finding patient specific sepsis treatment strategies [9].

III. PROPOSED WORK

The problem remains important since the high rates of missing data can result in a potential bias leading to an inaccurate diagnosis and treatment as well as poor modeling and statistical analyses. Common approaches such as omitting the missing values along consecutive terms may result in information loss. It can also dramatically shorten sample size which is not feasible in order to produce reasonable results for DNNs. For instance, mean substitution is a common and simple practice of replacing missing values but it disturbs the variance of completed data and the correlation of other variables. There are two main drawbacks to the current sepsis prediction models; 1) Inadequate performance for longer prediction time and 2) Limited usage of data sets. researcher

feels confident about their work and takes a jump to start the paper writing.

The summary of the proposed framework is illustrated in Figure below. It consists of four main blocks, which we call preprocessing, Feature Extraction, training and the prediction block.

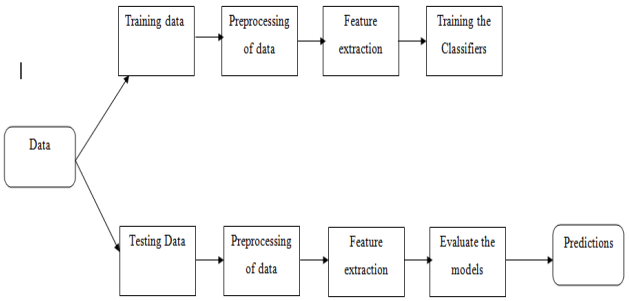

Figure 1: Proposed System

The dataset contained a substantial amount of missing values which were bound to have an adverse effect on the performance of the model unless treated properly. This issue was addressed by carrying forward the last observation to the subsequent missing values and constant imputation. One of the most common methods of dealing with missing data is complete case method, referred to as list wise deletion as well. In this technique, all the cases with any missing data are dropped. In this method of analysis, a missing follow-up value is replaced by (imputed as) that subject's previously observed value, that is, the last observation is carried forward. While performing the last observation carried forward in this paper, it was ensured that the data of one patient is not passed on to the next patient because it would make no sense. Each and every patient is independent in terms of their vital signs and laboratory value measurements. Unless there is any specific reason to do so, i.e. use one patient value for the other, which might be a case for disease which is related to age, gender, demographic area, keeping the patient data as separate entities is the most logical approach to handle the missing values. Even after replacement of the missing values in the aforementioned method, some values may still be non-numerical values which might appear problematic to the model. Therefore, single imputation was resorted to after performing the LOCF method.

The most common method of single imputation is constant replacement. This procedure requires a single value to be computed and then subsequently imputed for (i.e., to replace) a missing value. Despite being a simple method, it does not factor correlations between data and might introduce bias in the data itself. In this thesis, zero imputation was used. The question whether zero is a plausible value for the feature being replaced is also of concern. It can be noted here that the missingness itself could be classified as a new feature in and of itself in appropriate cases. Upon performing the data cleaning, the end tidal carbon-dioxide (EtCO2) feature was found to contain no values at all. So, its variance was zero and it contained no information. Any feature with little to no variance adds little to the predicting power of a machine learning model. As a consequence, this feature was dropped from the data. Among the remaining features, 19 features were then selected based on the analysis of variance test. The next step was to address the huge imbalance in the dataset. As is the case with any medical data where anomaly or disease detection is the goal, the number of ailing encounter is significantly lower than that of normal people. The disproportion between the minority and majority classes can be handled in a number of ways, one of which is cost-sensitive learning. The main theme behind cost-sensitive learning is to give weight to a class based on its proportion. That is, if a specific class has lower number of cases in it, then it is given more weight to compensate for it. This way, the model is not biased towards any of the classes.

Proposed algorithm is as follow:

Step 1: Read the dataset for sepsis detection.
Step 2: Preprocess the dataset.
Step 3: Apply feature engineering to select the required features.
Step 4: Split the dataset into train_set and test_set. In the train_set and test_set select data in the ration 80:20.
Step 5: Train the model using the XGBoost Classifiers.
Step 6: Train the model on the base classifiers.
Step 7: Evaluate the model by calculating "Accuracy" of Classes with 0 and 1.
Step 8: Print the Accuracy. Step 10: Exit.

IV. RESULT

The data is collected from Kaggle websites. We will use machine learning models to predict patients likely to become septic based on vital signs and laboratory values. Collected patient vitals and labs fetched on a recurring time interval and passed through the data transformation pipeline and model for Sepsis prediction. Provider manually enters values in a Sepsis application. After that the model performance will be calculated on following parameters. F1_score - measure provides a way to combine both precision and recall into a single measure that captures both properties. Recall - calculated as the number of true positives divided by the total number of true positives and false negatives; good for unbalanced data. Precision - quantifies the number of correct positive predictions made; good for unbalanced data.

Following steps will be done for the implementation work. Fetch the data. Load the data using Pandas. Load semi-colon separated data from disk. Explore the data Based on the attributes from the vital signs and laboratory values, below are target attributes to indicate Sepsis.

XGBoost classifier result is shown below:

```
from sklearn.metrics import confusion_matrix,accuracy_score
cm = confusion_matrix(y_test, y_pred)
acc1 = accuracy_score(y_test,y_pred)
print('accuracy of XGBoost model is=', acc1)
cm
```

```
accuracy of XGBoost model is= 0.9381627874948354
array([[6698,   34],
       [ 415,  114]])
```

Figure 2: XGBoost classifier

Accuracy comparison is shown below:

Classifier	Accuracy (in %)
SVM	89.90
KNN	92.42
Logistic Regression	92.75
Random Forest	93.14
Naive Bayes	86.25
MLP	89.32
SGD	92.59
XGBoost Model	**93.81**

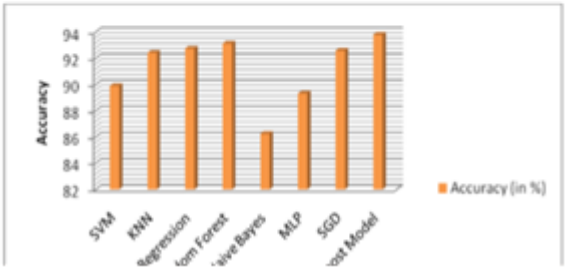

Figure 3: Accuracy comparison

V. CONCLUSION

In this paper, we presented and studied a novel prediction model with the XGBoost model for septic patients. Our initial findings reveal that the available data contained a considerable amount of missing values. Thus, to mitigate the negative effect of the missing information on the performance of the prediction model, we proposed a novel prepossessing and early prediction network. Our model not only discovers the missing patterns to improve the prediction results, but also can cooperate with broader detection windows. We concluded that capturing the uncertainty in the time series is specifically important in the medical settings to mitigate the propagation of error for the purpose of prediction. Our proposed method showed superior results, and it was shown to be applicable for any applications involving infrequently recorded health records. To our knowledge, this is the latest study to demonstrate a sepsis prediction algorithm over incrementally longer time windows with a significant performance involving adversarial training.

REFERENCES

[1] A. Rhodes, L. E. Evans, W. Alhazzani, M. M. Levy, M. Antonelli, R. Ferrer, A. Kumar, J. E. Sevransky, C. L. Sprung, and M. E. Nunnally, ‘‘Surviving sepsis campaign: International guidelines for management of sepsis and septic shock: 2016,’’ Intensive Care Med., vol. 43, no. 3, pp. 304–377, 2017.

[2] D. C. Angus, W. T. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, and M. R. Pinsky, ‘‘Epidemiology of severe sepsis in the United States: Analysis of incidence, outcome, and associated costs of care,’’ Crit. Care Med., vol. 29, no. 7, pp. 1303–1310, Jul. 2001.

[3] C. J. Paoli, M. A. Reynolds, M. Sinha, M. Gitlin, and E. Crouser, ‘‘Epidemiology and costs of sepsis in the United States—An analysis based on timing of diagnosis and severity level,’’ Crit. Care Med., vol. 46, no. 12, p. 1889, 2018.

[4] V. X. Liu, V. Fielding-Singh, J. D. Greene, J. M. Baker, T. J. Iwashyna, J. Bhattacharya, and G. J. Escobar, ‘‘The timing of early antibiotics and hospital mortality in sepsis,’’ Amer. J. Respiratory Crit. Care Med., vol. 196, no. 7, pp. 856–863, Oct. 2017.

[5] C. W. Seymour, F. Gesten, H. C. Prescott, M. E. Friedrich, T. J. Iwashyna, G. S. Phillips, S. Lemeshow, T. Osborn, K. M. Terry, and M. M. Levy, ‘‘Time to treatment and mortality during mandated emergency care for sepsis,’’ New England J. Med., vol. 376, no. 23, pp. 2235–2244, 2017.

[6] A. Kumar, D. Roberts, K. E. Wood, B. Light, J. E. Parrillo, S. Sharma, R. Suppes, D. Feinstein, S. Zanotti, and L. Taiberg, ‘‘Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock,’’ Crit. Care Med., vol. 34, no. 6, pp. 1589–1596, 2006.

[7] M. M. Levy, M. P. Fink, J. C. Marshall, E. Abraham, D. Angus, D. Cook, J. Cohen, S. M. Opal, J.-L. Vincent, and G. Ramsay, ‘‘2001 SCCM/ESICM/ACCP/ATS/SIS international sepsis definitions conference,’’ Intensive Care Med., vol. 29, no. 4, pp. 530–538, Apr. 2003.

[8] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, and C. M. Coopersmith, “The third

international consensus definitions for sepsis and septic shock (sepsis-3),'' Jama, vol. 315, no. 8, pp. 801– 810, 2016.

[9] S. Lambden, P. F. Laterre, M. M. Levy, and B. Francois, ''The SOFA score—Development, utility and challenges of accurate assessment in clinical trials,'' Crit. Care, vol. 23, no. 1, pp. 1– 9, Dec. 2019.

[10] C. P. Subbe, A. Slater, D. Menon, and L. Gemmell, ''Validation of physiological scoring systems in the accident and emergency department,'' Emergency Med. J., vol. 23, no. 11, pp. 841–845, Nov. 2006.

[11] A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford, ''Machine learning and decision support in critical care,'' Proc. IEEE, vol. 104, no. 2, pp. 444–466, Feb. 2016.

[12] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, ''Deep learning for health informatics,'' IEEE J. Biomed. Health Informat. vol. 21, no. 1, pp. 4– 21, Jan. 2017.

[13] S.-L. Wang, F. Wu, and B.-H. Wang, ''Prediction of severe sepsis using SVM model,'' in Advances in Computational Biology. New York, NY, USA: Springer, 2010, pp. 75–81.

[14] J. S. Calvert, D. A. Price, U. K. Chettipally, C. W. Barton, M. D. Feldman, J. L. Hoffman, M. Jay, and R. Das, ''A computational approach to early sepsis detection,'' Comput. Biol. Med., vol. 74, pp. 69–73, Jul. 2016.

[15] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J. D. Chiche, C. M. Coopersmith, and R. S. Hotchkiss, "The third international consensus definitions for sepsis and septic shock (Sepsis-3)," Jama, 315(8), 2016, pp.801-810.

[16] C. Rhee, R. Dantes, L. Epstein, D. J. Murphy, C. W. Seymour, T. J. Iwashyna, S. S. Kadri, D. C. Angus, R. L. Danner, A. E. Fiore, and J. A. Jernigan, "Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014," Jama, 318(13), 2017, pp.1241- 1249.

[17] S. K. Gadre, M. Shah, E. Mireles-Cabodevila, B. Patel, and A. Duggal, "Epidemiology and Predictors of 30-Day Readmission in Patients with Sepsis," Chest, 155(3), 2019, pp.483490.

[18] A. Keeley, P. Hine and E. Nsutebu, "The recognition and management of sepsis and septic shock: a guide for non-intensivists," Postgraduate Medical Journal. 93, 2017, pp.626-634.

[19] P. E. Marik and A. M. Taeb, "SIRS, qSOFA and new sepsis definition," Journal of thoracic disease, 9(4), 2017, p.943.

[20] J. M. Williams, J. H. Greenslade, J. V. McKenzie, K. Chu, A. F. Brown, and J. Lipman, "Systemic inflammatory response syndrome, quick sequential organ function assessment, and organ dysfunction: insights from a prospective database of ED patients with infection," Chest, 151(3), 2017, pp.586-596.

[21] C. J. Paoli, M. A. Reynolds, M. Sinha, M. Gitlin, and E. Crouser, "Epidemiology and costs of sepsis in the United States—An analysis based on timing of diagnosis and severity Level," Critical care medicine, 46(12), 2018, p.1889.

[22] M. Reyna, C. Josef, R. Jeter, S. Shashikumar, M. Westover, S. Nemati, G. Clifford, and A. Sharma, "Early prediction of sepsis from clinical data: the Physionet/Computing in Cardiology Challenge 2019," Critical Care Medicine, 2019.

[23] D. W. Shimabukuro, C. W. Barton, M. D. Feldman, S. J. Mataraso, and R. Das, "Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial," BMJ open respiratory research, 4(1), 2017, p.e000234.

[24] 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, and K. Pavan, "A Machine Learning Algorithm to Predict Severe Sepsis and Septic Shock: Development, Implementation, and Impact on Clinical Practice," Critical care medicine, 47(11), 2019, pp.1485.