International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 08 | Aug 2024 www.irjet.net p-ISSN: 2395-0072

Detection of Septic Condition based on patient Laboratory Values using XGBoost Model

Komal Kumari Roy[1], Prof. Preeti Rai[2]

[1]Reseacrh Scholar, Department of CSE, Gyan Ganga Institute of Technology and sciences, Jabalpur, M.P.
[2]Professor, Departmet of CSE, Gyan Ganga Institute of Technology and Sciences, Jabalpur, M.P.

Abstract –Sepsis is a complex medical condition that can manifest in various ways and progress through different stages. Clinicians often categorize sepsis into different types or stages based on the severity and specific clinical characteristics. This paper review research that have already done by various researchers in the field as well as introduced a novel prediction model using XGBoost for septic patient's detection based on medical data. The research began by addressing the challenges of missing data in the available patient records. To counteract the potential negative impact of this missing information on prediction accuracy, a unique preprocessing and early prediction network is proposed. This model not only identified missing data patterns to enhance prediction accuracy but also extended its applicability to broader detection timeframes of the septic conditions of patients.

Keywords: Sepsis detection, Machine learning, Random forest, Boosting ensemble, early predictions.

1. 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 neural networks (DNNs) and machine learning for solving multivariate, complex, and nonlinear problems. Training such networks requires a significant volume of data. Meanwhile, the intensive care unit (ICU) patients are monitored consistently. This has generated an abundance of data, which allows for training ML and DNNs for event prediction or decision support in critical care cases [11].

Recent studies have incorporated such ML and DNN-based approaches using electronic health records (EHRs) for identifying the early stages of complex diseases [12], [13], [14]. In sepsis cases, attention has been placed on creating an accurate and swift prediction model as an extension of the clinical decision since the performance of the deep learning models has been found significantly higher than traditional scoring systems. In [15] and [16], authors have been systematically reviewed and evaluated studies employing machine learning for the prediction of sepsis in the ICU. Our strategy is to build a ML-based preprocessing block that has the ability to learn the underlying dependencies and correlations of the observed time series to estimate the missing values.

2. Related Work

Early detection and diagnosis of sepsis can increase a patient's chances of survival and improve long-term health. In this paper [17] we use shapley additive explanations

(shap) analysis to investigate variables that are frequently associated with the development of sepsis in patients and to examine different working models for classification. To develop our prediction model, we evaluate the contribution of different features to the prediction result in two periods, using data collected after the first and fifth hours of recognition. Our findings suggest that although there is much missing information in the early stages of presentation, this information may be useful for early prediction of sepsis. We also found a discrepancy between different features of the entry level, which should be taken into account when developing machine learning models.

Study [18] evaluated the performance of a combined machine learning model in predicting sepsis using six pre-existing vital signs. This study validates the tool and compares it with existing methods. The results show that the technology can predict sepsis up to 48 hours in advance. This method has high sensitivity and specificity and rarely loses data. Early diagnosis is critical to improving outcomes in patients with sepsis, and delaying treatment may put the patient at risk. Therefore, machine learning algorithms can play an important role in early detection and prediction of sepsis, thereby improving patient outcomes.

Many techniques have been used to identify sepsis before clinical outcomes are known, but machine learning techniques have taken the lead among other techniques and tools. More information is available to obtain more accurate results, and some researchers use their own information that is not readily publicly available. We [19] used two lists; set a and set b to train and test our models available on physionet.org. We use the entire set a and most of set b to train our algorithm, and the rest of set b to test. We achieved the necessary training process by applying forward filling and back filling to the data to fill in missing values and then applied the extreme gradient boosting (xgb) classifier to obtain sepsis diagnosis results. Diagnosis of sepsis is 92% accurate. The algorithm will be used to improve patient survival and save annual costs for sepsis diagnosis and treatment.

The article [20] focuses on predicting sepsis in its early stages: machine learning algorithms. Sepsis is a serious form of the body's response to an infection. When the body's response goes wrong, it can lead to tissue damage, organ failure, and even death. Sepsis can be difficult to diagnose because it develops quickly and can be confused with other diseases. Early detection of sepsis can save lives; late detection can be fatal and consume significant hospital resources. Therefore, in this study, sepsis patient identification is based on machine learning algorithms to provide better performance. This project was created in visual studio and uses python for accessibility. The aim is to use physiological data to predict early sepsis. Patient data is analyzed and validated using algorithms from different models such as gradient boosting, logistic regression, support vector machines, and decision tree classifiers to predict clinical outcomes. Svm and logistic regression are more effective in high-level environments. Boosting algorithms provide machine learning models with multiple capabilities to improve prediction accuracy. Boosting algorithm is one of the most commonly used types of algorithms.

The research study in [21] will use preliminary data (using decision trees, randomization) from the mimic-iii database to create a prediction model for the first 15 hours of sepsis onset. Forest, Adaboost, gradient boosting trees and various sensor layers. The model was validated using 10-fold cross-validation. The roc-auc score is often used to evaluate the performance of the prediction model. In the model comparison, an additional set of predictive models was created 10 hours before the onset of sepsis, using the same algorithm to compare their performance with the previously generated predictive model. Model comparisons showed that gradient boosted tree had the best roc-auc score for the 15-hour and 10-hour prediction models before the onset of sepsis; the 15-hour forecast pattern was 0.777 and the 10-hour forecast pattern was 0.769. Results can be further refined using more data and derived from augmented trees algorithms.

Early and accurate diagnosis of sepsis is important because delayed treatment may increase mortality. The aim of the study was to create a classification that could predict sepsis 6 hours before the diagnosis of the disease. This is done through the patient's emr, vital signs, and demographic information. This work introduces various interactions and proposes a new filling algorithm called hybrid filling. The key features that make the interpreter predictable are explained, making it easier for medical staff to interpret the examples. Six models including random forest, logistic regression, optical gradient boosting technology, cloud gradient boosting, neural network, and short time were studied to classify patients. The parameters obtained are unprecedented and useful in terms of duration and accuracy of sepsis.

This article [23] proposes a method to predict the occurrence of sepsis 6 hours in advance using various machine learning and deep learning models and suggests learning strategies for the same rate. The medical data mart for intensive care iii (mimic3) dataset is used to test traditional machine learning methods such as random forests (rf), XGBoost, and deep learning techniques such as neural networks and autoencoders with XGBoost.

In [24], the authors proposed a good learning technique to predict missing data in the data. Our model includes a convolutional neural network (gan) that uses a short-term temporal (lstm) network as a generator and distinguishes text events in the classroom. Deep lstm networks are also used for prediction purposes. The prediction network is trained using gan conditional results and evaluated in blind testing to examine the effectiveness of the proposed model.

Here, we show that the proposed framework can identify not only long-term temporal but also missing patterns. We present performance results and compare them with other well-known methods. The method achieved receiver operating characteristic (auroc) of 94.49%, 93.74%, and 94.01% for 4-hour, 8-hour, and 12-hour prediction of sepsis, respectively. Here, it is shown that advances in detection and prediction promise to provide an effective means for early detection of sepsis in high-risk patients.

The aim of this study [25] is to develop and implement a machine learning (ml)-based technology that can predict the severity of shock and sepsis and evaluate its impact on healthcare and patients. This study is a type of integrated retrospective algorithm derivation and validation with pre- and post-intervention evaluation. For non-icu cases, algorithms are derived and used for specific times. The system used in this study was derived and implemented using electronic health records (ehrs), which are initially silent but alert medical personnel to the prediction of sepsis. Patients selected to train the classification system must have an icd and updated codes for acute or septic shock. Additionally, patients must have good blood results and proof of high blood pressure (sbp) or lactate levels at the time of contacting the hospital. The accuracy, sensitivity, and specificity of the classification algorithm are 93.84%, 93.22%, and 95.25%, respectively. In terms of reporting, diagnostic criteria have made small but important contributions to iv use and clinical trials.

3. Proposed Work

Early detection of sepsis can be treated using antibiotics and completely curable and save millions of people. In this scenario the early detection or prediction of sepsis using Machine Learning is a hot area for research.Sepsis is activated by the immune system present in your body that works all the time in order to prevent the infection from entering. During this stage, the enormous number of synthetic substances discharged into the blood causes broad irritation. For the patient the practicality of detecting sepsis disease occurrence in development is an important factor in the result. The primary goal of this research is to build, train and test a model using data that is available in the form of electronic clinical health data and predicts outcome of class labels as sepsis or no-sepsis for unseen health records using machine learning model. The secondary goal is to compare the accuracy of the various models. The proposed system is developed with XGBoost machine learning model.

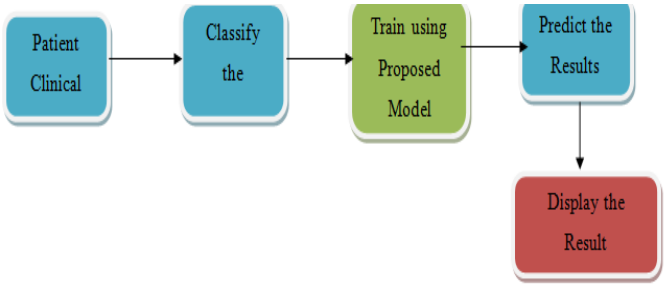

Figure 1: Proposed architecture

The first step is to get the data from dataset. Import the important libraries. Load the data as a Dataframe using Pandas. The data is stored in the form of .CSV files. The data is being downloaded from kaggle.com. Proposed system uses kaggle dataset. Parameters are defined for identification of sepsis. Then data visualization and exploratory data analysis are performed to get clear picture about various attributes and available data. Statistical description with values is carried out. Following calculus measures are carried out:

- Null values are ignored in the summary
- Highest number of entries for a column ~ 36k likely to consider if it helps indicate the prediction
- Lowest number of entries for a column ~1k likely to drop because there are too many missing entries to train a model on Average (mean), minimum (min), maximum (max) are self-explanatory.
- Standard deviation (std) how dispersed the values are normal (Gaussian) distribution follows 68-95-99.7 rule
    - 68% of values are within 1 std
    - 95% of values are within 2 std
    - 99.7% of values are within 3 std
- 1st (25%), median (50%), 3rd (75%) quartiles or percentiles, for example:
    - 25% of the patients had a temp lower than 36.3°C
    - 75% of the patients had a resp higher than 20.5 breaths per minute
- min/max
    - if min == 0 and max == 1, likely discrete categorical data.

4. Result Analysis

The Evaluations of various classifier algorithms according to accuracy are displayed below:

Table 1: Performance Evaluation.

Method	Testing Accuracy (%)
SVC	70.34
NB	18.8
KNN	55.74
SGD	19.15
LR	20.01
MLP	51.10
XGBoost	**95.01**

It is observed that proposed XGBoost classifier gives the better results in terms of accuracy.

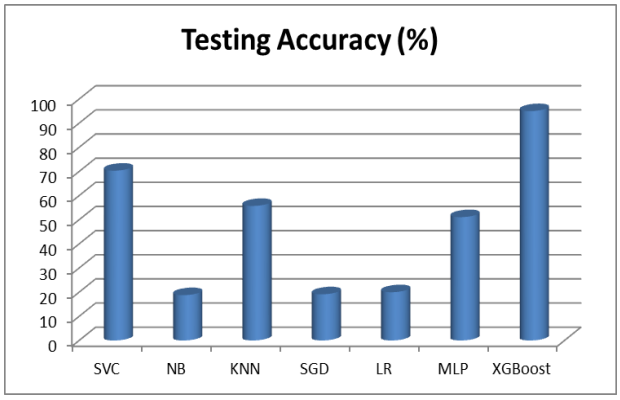

Figure 2: Accuracy Chart.

5. Conclusion

In summary, this thesis introduced a novel prediction model using XGBoost for septic patients. The research began by addressing the challenge of missing data in the available patient records. To counteract the potential negative impact of this missing information on prediction accuracy, a unique preprocessing and early prediction network was proposed. This model not only identified missing data patterns to enhance prediction accuracy but also extended its applicability to broader detection timeframes. The key take away from this study is the recognition of the crucial role of capturing uncertainty in time series data, particularly in medical contexts. This approach helps mitigate error propagation, ultimately improving prediction outcomes.

The results of this research demonstrated the superiority of the proposed method, which can be applied to various applications involving infrequently recorded health records. Notably, this study represents the latest advancement in sepsis prediction algorithms, offering enhanced performance through adversarial training over progressively longer time windows. Overall, this work contributes to the field of medical prediction models, offering a valuable tool for sepsis prediction and underscoring the importance of addressing missing data in healthcare analytics.

6. 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. Ravì, 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] L. M. Fleuren, T. L. T. Klausch, C. L. Zwager, L. J. Schoonmade, T. Guo, L. F. Roggeveen, E. L. Swart, A. R. J. Girbes, P. Thoral, A. Ercole, M. Hoogendoorn, and P. W. G. Elbers, ‘‘Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy,’’ Intensive Care Med., vol. 46, no. 3, pp. 383–400, Mar. 2020.

[16] M. Moor, B. Rieck, M. Horn, C. R. Jutzeler, and K. Borgwardt, ‘‘Early prediction of sepsis in the ICU using machine learning: A systematic review,’’ Frontiers Med., vol. 8, May 2021, Art. no. 607952.

[17] E. Shakeri, E. A. Mohammed, Z. Shakeri H.A. and B. Far, "Exploring Features Contributing to the Early Prediction of Sepsis Using Machine Learning," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, 2021, pp. 2472-2475, doi: 10.1109/EMBC46164.2021.9630317.

[18] M. R. Kumar, N. Natteshan, J. Avanija, K. R. Madhavi, N. S. Charan and V. Kushal, "SMOTE-TOMEK: A Hybrid Sampling-Based Ensemble Learning Approach for Sepsis Prediction," 2023 2nd International Conference on Edge Computing and Applications (ICECAA), Namakkal, India, 2023, pp. 724-729, doi: 10.1109/ICECAA58104.2023.10212208.

[19] A. Ullah, H. Qayyum, M. K. Khan and F. Ahmad, "Sepsis Detection Using Extreme Gradient Boost (XGB): A Supervised Learning Approach," 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), Karachi, Pakistan, 2021, pp. 1-6, doi: 10.1109/MAJICC53071.2021.9526260.

[20] N. Shanthi and A. A, "A Novel Machine Learning Approach to predict Sepsis at an Early Stage," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-7, doi: 10.1109/ICCCI54379.2022.9741000.

[21] T. X. Ying and A. Abu-Samah, "Early Prediction of Sepsis for ICU Patients using Gradient Boosted Tree," 2022 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), Shah Alam, Malaysia, 2022, pp. 78-83, doi: 10.1109/I2CACIS54679.2022.9815467.

[22] A. Shankar, M. Diwan, S. Singh, H. Nahrpurawala and T. Bhowmick, "Early Prediction of Sepsis using Machine Learning," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 837-842, doi: 10.1109/Confluence51648.2021.9377090.

[23] N. Shah, J. Bhatia, N. Vasavat, R. Desai and P. Sonawane, "Early Sepsis Detection using Machine Learning and Neural Networks," 2021 2nd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2021, pp. 1-6, doi: 10.1109/GCAT52182.2021.9587778.

[24] M. Apalak and K. Kiasaleh, "Improving Sepsis Prediction Performance Using Conditional Recurrent Adversarial Networks," in IEEE Access, vol. 10, pp. 134466-134476, 2022, doi: 10.1109/ACCESS.2022.3230324.

[25] B. Y. Al-Mualemi and L. Lu, "A Deep Learning-Based Sepsis Estimation Scheme," in IEEE Access, vol. 9, pp. 5442-5452, 2021, doi: 10.1109/ACCESS.2020.3043732.