Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240534

An Overview of Explainable AI Studies in the Prediction of Sepsis Onset and Sepsis Mortality

Andria NICOLAOU [a,b,1], Charithea STYLIANIDES [b], Waqar A. SULAIMAN [a], Zinonas ANTONIOU [c], Lakis PALAZIS [d], Anna VAVLITOU [d], Constantinos S. PATTICHIS [a,b] and Andreas S. PANAYIDES [b]
[a] *Department of Computer Science, University of Cyprus, Nicosia, Cyprus*
[b] *CYENS Centre of Excellence, Nicosia, Cyprus*
[c] *3aHealth, Nicosia, Cyprus*
[d] *State Health Services Organization, Cyprus*
ORCiD ID: Andria Nicolaou https://orcid.org/0000-0002-4863-6505,
Charithea Stylianides https://orcid.org/0009-0002-3568-3449,
Zinonas Antoniou https://orcid.org/0000-0002-5148-5197,
Constantinos S. Pattichis https://orcid.org/0000-0003-1271-8151,
Andreas S. Panayides https://orcid.org/0000-0001-9829-7946

Abstract. Explainable artificial intelligence (AI) focuses on developing models and algorithms that provide transparent and interpretable insights into decision-making processes. By elucidating the reasoning behind AI-driven diagnoses and treatment recommendations, explainability can gain the trust of healthcare experts and assist them in difficult diagnostic tasks. Sepsis is characterized as a serious condition that happens when the immune system of the body has an extreme response to an infection, causing tissue and organ damage and leading to death. Physicians face challenges in diagnosing and treating sepsis due to its complex pathogenesis. This work aims to provide an overview of the recent studies that propose explainable AI models in the prediction of sepsis onset and sepsis mortality using intensive care data. The general findings showed that explainable AI can provide the most significant features guiding the decision-making process of the model. Future research will investigate explainability through argumentation theory using intensive care data focused on sepsis patients.

Keywords. Explainable AI, ICU, prediction, sepsis, mortality, review

1. Introduction

The integration of artificial intelligence (AI) in healthcare has lighted up the way of new possibilities and challenges [1],[2]. As complex machine learning (ML) and deep learning (DL) models are developed to analyze vast amounts of medical data, there is a growing imperative to understand the decisions made by these algorithms. Explainable AI is emerging as vital to bridge the gap between the black-box nature of advanced ML and DL algorithms and the necessity for transparent and interpretable decision-making.

[1] Corresponding Author: Andria Nicolaou; E-mail: a.nicolaou@cyens.org.cy.

Explainability in healthcare refers to the capability to provide clear and understandable explanations in diagnosis, treatment, and prognosis, considering medical knowledge and ethical standards, and as a result, gain the trust of physicians and patients [2].

Sepsis is defined as a life-threatening organ dysfunction caused by infection [3],[4]. It impacts over 30 million individuals each year globally and stands as a leading cause of mortality among critically ill patients worldwide [4], affecting a large proportion of intensive care unit (ICU) patients [3]. The ICU is a specialized area equipped with high staffing levels, advanced and continuous monitoring, and organ support to improve patient outcomes [3]. However, successful intensive care extends beyond the confines of the ICU and necessitates a holistic approach, including early warning systems using advanced AI technology that can assess critical illnesses, and predict mortality [5].

The objective of this work was to review explainable AI studies in the prediction of sepsis onset and sepsis mortality using intensive care data. This paper is organized as follows. The study selection and analysis are described in Section 2. The results of the analyzed studies are summarized in Section 3. The discussion is presented in Section 4, and finally, the concluding remarks are addressed in Section 5.

2. Methods

2.1. Study selection

The openly available web search engine Google Scholar was the main database used to identify studies presented in this work. The search strategy was focused on the explainable AI models in the prediction of sepsis onset and sepsis mortality. The terms "explainable AI model", "interpretable model", "sepsis onset" and "sepsis mortality" were used to search the database. The searches were limited by selecting a custom year range of 2021 to 2024, and results were ranked by relevancy. Only 13 of the 7920 publications found through these searches satisfied the review's inclusion criteria and were thoroughly analyzed by one author (AN).

2.2. Study analysis

The selected studies were analyzed focusing on the proposed explainable AI methods. Most of the researchers obtained data from the Medical Information Mart for Intensive Care (MIMIC) III and IV databases that are large, de-identified, and publicly available collections of medical records of patients admitted to the Beth Israel Deaconess Medical Centre in Boston, Massachusetts [6]-[18]. Some studies used datasets from the Emory University Hospital in Atlanta [8],[13], the University of California San Diego [13], and the private Historical Database of Ruijin Hospital in Shanghai, China [10]. Then, the data were preprocessed, and classification analysis was performed. ML and DL algorithms were trained and evaluated.

According to the current literature review, explainable AI methods were applied to identify the significant features impacting the model's prediction. They were post-hoc as they performed explainability after the development of the model, including model-specific and model-agnostic methods [2]. Model-specific methods were the gradient-weighted class activation mapping (GradCAM) [6] that generates visual explanations from deep networks, the sensitivity analysis [8],[13] that investigates the significance of each model input on the model's output, and heatmaps [12] which are graphical

representations of data and are utilized to interpret deep networks. On the contrary, model-agnostic methods were the feature importance (FI) [7],[8],[11],[14],[18]] that assigns a score to the input features based on how significant they are in the model's prediction, Shapley additive explanations (SHAP) [6],[7],[10],[11],[14]-[18] that interpret the model's prediction, generating both local and global explanations, and local interpretable model-agnostic explanations (LIME) [8],[16] which explain a specific prediction of the model.

3. Results

The main findings of the selected and analyzed explainable AI studies in the prediction of sepsis onset and sepsis mortality are presented in Table 1. It is shown that ensemble models (e.g., random forest, extreme gradient boosting) were the best-performing ML algorithms for both prediction tasks. The model-agnostic methods such as FI, and SHAP were the most frequently used to explain the models.

Regarding the significant features of the analyzed studies (see Table 1), heart rate and temperature were strong indicators of sepsis onset. On the other hand, age, urine output, and blood urea nitrogen significantly affected sepsis mortality. It is also observed that some features impacted both prediction tasks such as respiratory rate and Glasgow coma scale. It's worth mentioning that most of these features matched the measurements of existing guidelines such as sequential organ failure assessment (SOFA) score and acute physiology and chronic health evaluation (APACHE) II score [4].

Table 1. An overview of explainable AI studies in the prediction of sepsis onset and sepsis mortality

Author [ref.]	Data	Classification	Explainability	Results	Significant features
Prediction of sepsis onset					
Chakraborty *et al.* [6]	MIMIC III	DL	GradCAM, SHAP	ACC=0.93	temp, SIRS, HCT, respiratory rate
Jiang *et al.* [7]	MIMIC IV	ML (XGB)	FI, SHAP	AUC=0.75	PaO_2, temp, heart rate
Zhang *et al.* [8]	MIMIC IV	ML (ET)	LIME	AUC=0.76 ACC=0.71	temp, systolic blood pressure
Chen and Hernández [9]	MIMIC III, EUH	ML (RF)	FI, Sensitivity analysis	AUC=0.85 ACC=0.82	temp, FiO_2, pulse oximetry, respiratory rate
Chen *et al.* [10]	MIMIC III, RH	ML (LGBM), DL	SHAP	AUC=0.98 ACC=0.91	age, antibiotics, fluid balance, ventilation
Nesaragi and Patidar [11]	MIMIC III	ML (LGBM)	FI, SHAP	AUC=0.86	PaO_2-FiO_2, creatinine, temp
Rosnati and Fortuin [12]	MIMIC III	DL	Heatmaps	AUC=0.66	temp, heart rate, pulse oximetry
Shashikumar *et al.* [13]	MIMIC III, UCSD, EUH	DL	Sensitivity analysis	AUC=0.90	heart rate, temp, GCS
Prediction of sepsis mortality					

Li *et al.* [14]	MIMIC IV	ML (XGB)	FI, SHAP	AUC=0.85 ACC=0.77	urine output, age, supplemental oxygen therapy
Hu *et al.* [15]	MIMIC IV	ML (XGB)	SHAP	AUC=0.88 ACC=0.90	GCS, BUN, respiratory rate, urine output, age
Hu *et al.* [16]	MIMIC IV	ML (RF), DL	SHAP, LIME	ACC=0.85	GCS, urine output, BUN
Ke *et al.* [17]	MIMIC IV	ML (XGB)	SHAP	AUC=0.87	acute physiology score III, age
Jiang *et al.* [18]	MIMIC III	ML (LGBM)	FI, SHAP	AUC=0.73	BUN, age, albumin, glucose

MIMIC: Medical Information Mart for Intensive Care, EUH: Emory University Hospital, RH: Ruijin Hospital, UCSD: University of California San Diego, ML: Machine Learning, DL: Deep Learning, SHAP: Shapley Additive Explanations, LIME: Local Interpretable Model-Agnostic Explanations, FI: Feature Importance, GradCAM: Gradient-weighted Class Activation Mapping, RF: Random Forest, XGB: Extreme Gradient Boosting, LGBM: Light Gradient Boosting Machine, ET: Extremely Randomized Trees, ACC: Accuracy, AUC: Area Under receiver operating characteristic Curve, SIRS: Systematic Inflammatory Response Syndrome, HCT: Hematocrit, PaO_2: Partial pressure of oxygen, FiO_2: Fraction of inspired oxygen, GCS: Glasgow Coma Scale, BUN: Blood Urea Nitrogen.

4. Discussion

The objective of this study was to review explainable AI studies in the context of the ICU. The main results indicated that explainability methods can provide the most significant features guiding the decision-making process of the AI models.

It's worth mentioning that explainability through argumentation theory focusing on intensive care data has not been investigated yet to the best of the authors' knowledge. However, other studies proposed explainable AI models in healthcare using argumentation theory. More specifically, Prentzas *et al.* [19] developed an argumentation framework for explainable ML called ArgEML, based on a novel approach that integrates sub-symbolic methods with logical methods of argumentation to provide explainable solutions to learning problems, applied to gynecological cancer prognosis.

Another explainable AI model was implemented to assess Multiple Sclerosis (MS) disease using clinical data and brain MRI lesion texture features [20]. Different ML models were employed to classify the MS subjects between benign and progressive form of the disease. Argumentation-based reasoning was then performed using the extracted rules from the ML models, achieving high precision in providing understandable and interpretable information for the progression of the disease.

5. Conclusions

Explainability facilitates unprecedented opportunities in the use of AI for the prediction of sepsis onset and sepsis mortality as it can provide significant insights to medical experts, enhancing AI-models trust and transparency, while removing bias. Future work will investigate explainability through argumentation theory using intensive care data.

Acknowledgment

This work was funded by the European Union Recovery and Resilience Facility of the NextGenerationEU instrument, through the Research and Innovation Foundation of Cyprus (Project: CODEVELOP-ICT-HEALTH/0322/0071, Hospital Transformation through Artificial Intelligence – HOSPAITAL).

References

[1] Panayides AS, et al. AI in Medical Imaging Informatics: Current Challenges and Future Directions. IEEE J Biomed Heal Informatics. 2020;24(7):1837–57.
[2] Prentzas N, Kakas A, Pattichis CS. Explainable AI applications in the Medical Domain: a systematic review. arXiv Prepr arXiv230805411. 2023;1–19.
[3] Jackson M, Cairns T. Care of the critically ill patient. Surg (United Kingdom). 2021;39(1):29–36, doi: 10.1016/j.mpsur.2020.11.002
[4] Huang M, Cai S, Su J. The pathogenesis of sepsis and potential therapeutic targets. Int J Mol Sci. 2019;20(21).
[5] Stylianides C, et al. AI-based solutions for predicting sepsis in ICUs. In: 2023 IEEE EMBS Spec Top Conf Data Sci Eng Heal Med Biol. 2024;163–4.
[6] Chakraborty S, Kumar K, Reddy BP, Meena T, Roy S. An Explainable AI based Clinical Assistance Model for Identifying Patients with the Onset of Sepsis. In: 2023 IEEE 24th Int Conf Inf Reuse Integr Data Sci. 2023; p. 297–302.
[7] Jiang Z, et al. Interpretable machine-learning model for real-time, clustered risk factor analysis of sepsis and septic death in critical care. Comput Methods Programs Biomed. 2023;241, doi: 10.1016/j.cmpb.2023.107772
[8] Zhang TY, Zhong M, Cheng YZ, Zhang MW. An interpretable machine learning model for real-time sepsis prediction based on basic physiological indicators. Eur Rev Med Pharmacol Sci. 2023;27(10):4348–56.
[9] Chen M, Hernández A. Towards an Explainable Model for Sepsis Detection Based on Sensitivity Analysis. Irbm. 2022;43(1):75–86, doi: 10.1016/j.irbm.2021.05.006
[10] Chen Q, et al. Transferability and interpretability of the sepsis prediction models in the intensive care unit. BMC Med Inform Decis Mak. 2022;22(1):1–10, doi: 10.1186/s12911-022-02090-3
[11] Nesaragi N, Patidar S. An Explainable Machine Learning Model for Early Prediction of Sepsis Using ICU Data. Infections and Sepsis Development. IntechOpen; 2021, doi: 10.5772/intechopen.98957
[12] Rosnati M, Fortuin V. MGP-AttTCN: An interpretable machine learning model for the prediction of sepsis. PLoS One. 2021;16(5):1–21, doi: 10.1371/journal.pone.0251248
[13] Shashikumar SP, et al. DeepAISE – An interpretable and recurrent neural survival model for early prediction of sepsis. Artif Intell Med. 2021;113, doi: 10.1016/j.artmed.2021.102036
[14] Li S, et al. Developing an Interpretable Machine Learning Model to Predict in-Hospital Mortality in Sepsis Patients: A Retrospective Temporal Validation Study. J Clin Med. 2023;12(3).
[15] Hu C, et al. Interpretable Machine Learning for Early Prediction of Prognosis in Sepsis: A Discovery and Validation Study. Infect Dis Ther. 2022;11(3):1117–32.
[16] Hu C, Li L, Li Y, Wang F, Hu B, Peng Z. Explainable Machine-Learning Model for Prediction of In-Hospital Mortality in Septic Patients Requiring Intensive Care Unit Readmission. Infect Dis Ther. 2022;11(4):1695–713.
[17] Ke X, et al. Interpretable Machine Learning to Optimize Early In-Hospital Mortality Prediction for Elderly Patients with Sepsis: A Discovery Study. Comput Math Methods Med. 2022;2022.
[18] Jiang Z, et al. An explainable machine learning algorithm for risk factor analysis of in-hospital mortality in sepsis survivors with ICU readmission. Comput Methods Programs Biomed. 2021;204, doi: 10.1016/j.cmpb.2021.106040
[19] Prentzas N, et al. Argumentation-based Explainable Machine Learning (ArgEML): a Real-life Use Case on Gynecological Cancer. In: CEUR Workshop Proceedings. 2022. p. 1–13.
[20] Nicolaou A, Pantzaris M, Loizou CP, Kakas A, Pattichis CS. An Explainable AI model in the assessment of Multiple Sclerosis using clinical data and Brain MRI lesion texture features. In: 2023 IEEE EMBS Int Conf Biomed Heal Informatics; 2023 Oct 15-18; Pittsburgh, PA. USA.