RESEARCH PAPER

Explainable AI-Driven ICU Mortality Prediction Using Temporal Deep Learning Models

K Vijay[1], Kodam Swathi[2], Ms N Bhargavi[3], Mr Farooqhusain[4]
[1]Assistant Professor, CSE, Siddhartha Institute of Technology & Sciences, Ghatkesar-500088
Email: vijaykoraveni.cse@siddhartha.co.in
[2]M.Tech, CSE, Siddhartha Institute of Technology & Sciences, Ghatkesar-500088
Email: 24TQ1D5806@siddhartha.org.in
[3]Assistant Professor, CSE, Siddhartha Institute of Technology & Sciences, Ghatkesar-500088
Email: bhargavi.cse@siddhartha.co.in
[4]Assistant Professor, CSE, Siddhartha Institute of Technology & Sciences, Ghatkesar-500088
Email: farooqhusain.cse@siddhartha.co.in

Received: 25th March, 2026 | Revised: 15 st Aprail, 2026 | Accept: 25th June, 2026 | Available Online: 7th Sep,

Abstract: Building upon these, this study presents a deep learning framework to predict ICU mortality using temporal clinical data with the help of explainable artificial intelligence via two salience techniques. Predicting the ICU mortality of critically ill patients is a significant healthcare challenge since, critically ill patients endure life-threatening organ failure and require continuous monitoring; hence, early identification of at-risk patients can greatly improve treatment decisions and help save lives. They analyze sequential patient observations to identify patterns that correlate with deterioration and the risk of death by processing preprocessed clinical datasets of physiological vital signs, lab test results, and patient demographic information (e.g. age). The inclusion of temporal patient observations leads to two changes in our prediction model: First, the input data should be converted to temporal sequences, where each sequence is a record of patient observations over a series of time windows, and Second, the architecture of deep learning should include (i) temporal learning layers, which are designed to learn temporal dependencies and (ii) fully-connected neural network layers to perform feature integration, and (iii) a binary output layer producing a probability score for the probability that the patient will die in the ICU. Combining interpretability and transparency, the proposed system develops Explainable Artificial Intelligence (XAI) SHAP (SHapley Additive exPlanations) techniques in which global explanations point out the clinical features that have the highest importance ranking for all patients along with local explanations with a high level of details that clarify how the model designed a prediction of mortality for each individual. The experimental results show that our prediction model maintains a level of predictive performance on the various evaluation metrics such as accuracy, confusion matrix, ROC-AUC, and precision, recall, and F1-score analysis, and the addition of SHAP-based explainability analysis further improves the system and makes it more useful in healthcare applications where clinicians need to interpret and understand the logic behind the AI-based recommendations. If further developed, AI driven mortality prediction systems can help clinicians identify patients at high risk of mortality early on their management in the intensive care units, and facilitate decision making in critical care practice.

Keywords: *ICU Mortality Prediction, Temporal Deep Learning, Explainable AI (XAI), SHAP, Clinical Decision Support, Sequential Data Analysis, Healthcare AI.*

How to cite this article: K Vijay1 Kodam Swathi2 Ms N Bhargavi3 Mr Farooqhusain4. Explainable AI-Driven ICU Mortality Prediction Using Temporal Deep Learning Models. Int J Drug Deliv Technol. 2026;16(80s):1400-1406. DOI: 10.25258/ijddt.16.80s.179.

1. INTRODUCTION

ICU is one of the most resource-hungry and life-critical environments in the hospital where patients need continuous monitoring and immediate clinical attention. One of the most formidable of these challenges is the decision-making of predicting patient outcomes, most notably in-hospital mortality. Early identification of those patients at risk of deterioration can facilitate early intervention and targeted treatment, efficient use of resources, and may improve survival [2, 4]. Yet the prediction task is particularly challenging for traditional clinical scoring systems [5] due to the dynamic and complex data of ICU patients, consisting of continuous streams of physiological signals (e.g., heart rate, respiratory rate, blood pressure, etc...) and laboratory measurements that are often offered at irregular sampling intervals.

Over the last decade, Artificial Intelligence (AI), and more specifically, deep learning has proved to be a strong competitor for healthcare analytics [1, 11]. "Temporal deep learning models learn from sequential data rather than treating each data point independently as machine learning models typically would." From the longitudinal patient records, they can automatically extract complex nonlinear time-dependent patterns that characterize the underlying physiology evolving toward clinical end points such as death [14, 16]. It renders them highly ideal for use cases such as sepsis prediction and ICU mortality risk stratification [3, 13, 15].

Although they can make highly accurate predictions, the opaque nature of many deep learning models is a major obstacle to their use in the clinic. One of the reasons for clinicians not wanting to trust and implement the recommendations made by AI is the lack of knowledge of the underlying rationale [2, 6]. In order to overcome this gap, the Explainable AI (XAI) community has emerged, stressing the need for models to be accurate as well as interpretable. Methodologies such as SHapley Additive exPlanations (SHAP) offer a principled approach for quantifying the contribution of each input feature toward a model prediction, yielding both global interpretation (which features across the population are most important) and local explanations (why a specific patient received a specific prediction) [1, 15].

In this study, we focus on the pressing requirement for reliable and interpretable prediction tools available in ICU settings. Here, we present a new framework that combines a temporal deep learning model for mortality risk prediction with SHAP-based explainability. Abstractcontext: The system analyzes sequential clinical data (vital signs, laboratory results, and demographics) to predict patient outcomes. Our framework aims to provide trust in AI-enabled decision-support tools by offering both high predictive performance and parsimonious, clinically-actionable explanations, thus promoting the use of AI to augment clinical workflows in critical care.

2. RELATED WORK

Machine learning and deep learning have been increasingly used in the prediction of negative outcomes in ICU patients. This section examines the body of work in the fields of mortality and sepsis prediction, the use of temporal models, and finally, the essential incorporation of explainability.

Predicting ICU Outcomes using AI: Many Works are available on predictive modeling in ICU Stylianides et al. A recent paper [4] has summarized the advances of AI in ICU with a limited focus on sepsis prediction and pointed out that these techniques may change healthcare. Similarly, Huang et al. ABSTRACT [11] reviewed different early prediction methodologies for in-hospital ICU mortality and suggested that even first-day data can be used for rapid risk stratification and treatment to improve the outcomes. The clinical relevance and sophistication of AI applications in this field are thus highlighted by these reviews.

Sequential Nature of ICU Data: Because of how the information in an ICU is sequenced over time, the field has progressed from static models to deep learning models capable of processing sequential data directly. At the moment, recurrent neural networks (RNNs) and long short-term memory (LSTM) networks are the building blocks in this space. Temporal Dynamics Background- Cooper [13] shows the effectiveness of capturing temporal dynamics via LSTM networks on real-time wearable sensor data for sepsis onset prediction. Badhon et al. also explore more sophisticated architectures. Abstract: To enhance the interpretability and actionability of deep learning predictions related to Acute Kidney Injury, [5] proposed an approach called Temporal Concept Tracing. Furthermore, Escudero-Arnanz et al. Explainable spatio-temporal GCNNs designed for explanation of multivariate time series data can model patient data over time with increasing accuracy,[14] representing just a taste of the advancement of the field towards more multilayer and abstract patient modeling. The work by Li et al. There is even new focus coming from a neighbouring field; [16] on interpretableAI for ICU mortality prediction had us writing this for models that "think as a doctor";

XAI in Critical Care: The recent integrating of XAI now-rank as recent area of research [56] that tackled on the drawback of the opaqueness of compositional designs. Not surprisingly, the focus of our work aligns with the findings of a systematic review highlighted by Athukorala and Ilmini [2], stressing the importance of interpretable models for sepsis and ICU mortality prediction. SHAP is a common way to explain models in many modern studies. For example, Begum et al. SHAP was applied in an XAI-enabled multimodal deep learning framework for early sepsis prediction, according to [1]. Yilmaz Başer et al. Those [15] went even further using SHAP to provide explanations for a meta-ensemble prediction model for ICU mortality and demonstrate its practical applicability. Similarly, Kim et al. Methods: The explainable machine learning technique, SHAP, was used for early and timely prediction of cardiac arrest.[3] Nicolaou et al. Both works added to the growing body of literature on XAI in sepsis prediction, with [6] providing an overview of XAI studies in sepsis prediction and strengthening the focus on interpretable AI in critical care.

And some researchers have looked at the latter (more specialized ones) individually as innovative approaches. Cheng et al. An explainable causal AI model was proposed [7] to recover hidden physiological dynamics beyond traditional correlation-based predictions. Motivating by a need to overcome data scarcity challenges in some clinical settings, Pandey [12] investigated the issue of mortality prediction using small datasets with XAI. Pioneering but more peripheral than our work are those by Tadikonda [9] on privacy-preserving interpretable transformer models, and by Gowda [10] on dynamic AI models for real-time monitoring, which push promising approaches for deployable, trustworthy systems.

In summary, while the existing literature has made great strides towards AI-based prediction of ICU mortality, most of these studies either purport to maximize prediction performance or interpretation. In its direct embodiment of a strong temporal deep learning model, validated global Explainable Aritifical Intelligence (XAI) methodology (SHAP) providing accurate and plausibly clinically actionable information from a structured preprocessing pipeline, our proposed framework has the components necessary to get closer than more generalized methods to addressing the trustworthiness AI needs to find its place in temporal critical care decision making.

3. METHODS AND PROCEDURES

The overall workflow of our proposed ICU mortality prediction framework is illustrated in Fig. 1. Figure 1 A novel pipeline overview — preprocessing of dataset, temporal modelling of the pooled information to mortality prediction and explainability

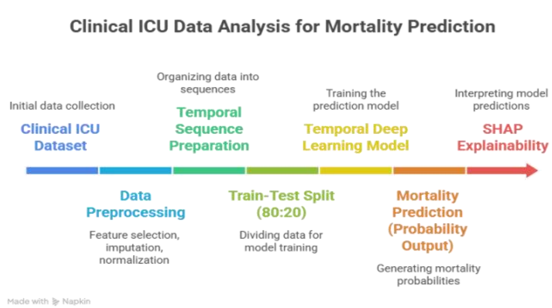

Fig 1. Proposed Methodology

A. Dataset

Clinical dataset — In this paper we used a clinical cohort synthesized for mortality prediction where physiological signals and laboratory results recorded during the hospitalization of a patient in the intensive care unit (ICU)[10]. Predicting mortality accurately in the ICU is an important healthcare task as it has been largely shown that clinical decision-making and overall patient survival can be improved through earlier identification of these high-risk patients in their treatment course. About: The data of millions of database includes observations of patients over their entire stay in the ICU. The dataset contains observations of a patient status at a time interval, where each observation contains physiological parameters, lab test results, and demographical attributes. Taken together, these features describe the patients' time-varying physiological status and contain information for prediction tasks. The dataset includes many clinical variables including Vital signs, Biochemical Measurements, and Demographic Characteristics. In the ICU, vital signs like heart rate, blood pressure, respiratory rate, body temperature, and oxygen saturation are continuously measured and monitored and are major indicators of patient stability. Additionally, Laboratory testing as blood chemistry, metabolic markers and hematological parameters may provide detailed insight into the deeper internal physiological status of the patient, indicate underlying pathological mechanisms via abnormalities driven by disease progression or organ dysfunction.

Implement additional patient demographics such as age for patients and demographics such as sex, to contextualize the impact this has on patient outcomes. Each patient record (one line per patient) is associated with the binary outcome variable (mortality during ICU stay) The target variable is survived or died in the hospital The first cause of missing data in this dataset is the natural clinical environment in which the data are collected, in that patients are not constantly tested, and that some values may not be present in the particular measurements. Thus, datasets require preprocessing for handling the missing values and uniformity. The data is then divided in 80% for training and 20% of the test from the model having the proper learning model and estimation of the model. Every sample includes measurements from different physiological systems, lab records, demographics — and a morta generator label.

B. Data Preprocessing

Prior to training with the deep learning model, the dataset is processed in several steps to ensure data quality and consistency. They often include missing values, noise and heterogeneous feature distributions which can be a challenge and should be pre-processed correctly for optimal model performance.

Preprocessing begins with feature selection, in which irrelevant attributes (e.g., patient identifiers, site or administrative metadata) are removed. We can get rid of those features that are not supportive to the predictive modeling in the training process and only act as a source of noise over the data set and hence removing it will result in increasing the computational efficiency and the accuracy of the model. Then Incompleted data undergoes vaue imputation. Directly observed clinical measurements are collected in an irregular fashion at a subset of time-steps, and therefore statistical imputation techniques are used to impute values for missing measurements. So that it is regularized and able to accept it for deep learning algorithms and thus reduce bias.

Feature engineering: Making the dataset more predictive. Which leverages clinical variables with more relatedness that are better characterising the health status of a patient by creating features from them. By capturing relationships between variables, feature engineering allows the model to learn much more complex patterns ultimately. After this, we normalize it to get that every feature has a comparable scale. Clinical variables have varied ranges and therefore there is a need tonormalize them so that large magnitude features would not govern thelearning process. Which make it easy to converge and stability when training the model.

This step is a preparatory step of preprocessing since the data from an ICU is temporal in nature. Observations of patients are organized to sequential format where each sequence indicates measured observations over time. This enables the model to take into account temporal dependencies and trends that may be predictive of patient deterioration. Next we split our dataset into train and test subset in 80:20 ratio. The preprocessing steps discussed above, ensure that the dataset is clean, structured and optimized to be fed into temporal deep learning models.

C. Model

A deep learning architecture specifically designed for sequential data analysis is utilized for the prediction of ICU mortality. Temporal deep learning models differ from traditional machine learning models in that they are better suited for capturing time-dependent relationships in clinical data, as classical machine learning models typically treat observations as independent. The nature of ICU patient data is sequential due to the temporal evolution of physiological parameters. The temporal nature of these observations is subsequently used by our proposed model to analyze sequences of individual patient observations, in turn identifying features that lead to decompensation or recovery.

The architecture of the model includes an input layer taking sequential clinical data followed by temporal learning layers to learn the time-step dependencies. LSTMs keep information of past observations while they process new ones, making it possible for the model to learn long-term relationships in patient health states. These temporal features are fed into fully connected layers that combine the learned representations and output the final prediction. The output layer generates a probability score that indicates the likelihood of the patient dying in the ICU.

D. Proposed Model

We introduce a framework that combines temporal deep learning with explainable-AI methods to ensure the ICU-mortality predictions are precise and interpretable. Our framework starts with processing the preprocessed dataset and generating patient observations to temporal sequences to adapt the model inputs. The deep learning model processes these during training and identifies patterns related to patient decline. Using optimization algorithms, the model constantly updates parameters based on the prediction error for the able to classify better. After many iterations of this training, the model learns to differentiate between survival and death cases from subtle clinical patterns.

After training, the model generates probability scores that represent the likelihood of a patient with a specific set of features surviving or succumbing to death These outputs with probabilities transfer the risk into a quantitative number (i.e., a percentage) that aids in clinical decision making. Relief is an explainable artificial intelligence (XAI) framework that can be understood based on the SHAP (SHapley Additive exPlanations) technique to minimize such limitations. It provides global and local interpretability because it measures the contribution of every feature to the prediction of the model. Global explanations find the clinical features that have the greatest effect on prediction in the global dataset, while local explanations explain the prediction for a single patient and provide the most relevant features for that specific prediction. This improves clinical utility and reliability.

RESULTS AND DISCUSSION

The experimental evaluation was done to gain an in-depth insight into the predictive power of the proposed deep learning framework for the ICU mortality prediction task. We trained the model on the clinical dataset prepared in Section 4.1, and we provided performance evaluation via multiple metrics to ensure reliability and comprehensiveness. The relationships of the features with each outcome of interest (here, patient deterioration and mortality) were learned iteratively by the model. The model identifying signals of patient risk included physiological identifiers (e.g., abnormal vital signs), lab measures, and other demographics. The performance evaluation results of the trained temporal deep learning model indicate that it can effectively assess high risk patients and predict mortality outcomes accurately.

Training Performance Analysis

Analysis of training performance assesses how well the deep learning model learns from the dataset within the process of training depicted in Fig 2, 3. With epoch-wise trainings, the model was able to decrease prediction loss with each iteration, enabling the model to differentiate between survival and mortality cases more effectively. As the training loss quickly fell, it showed that the model was picking up on good signals from the dataset. Meanwhile, the performance on training was also better because the model started to "learn" the internal representation of variables related to the clinical case. If the training stable converged, then it shows that the preprocessing pipeline and model architecture were a good fit for the dataset.

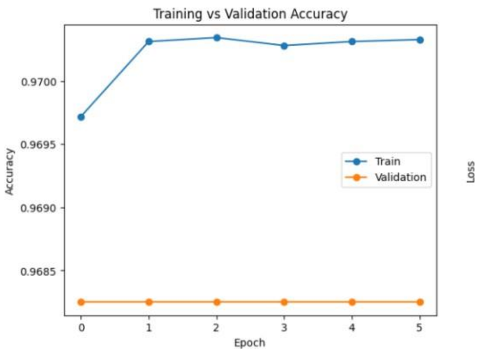

Fig 2: Training Vs Validation Accuracy

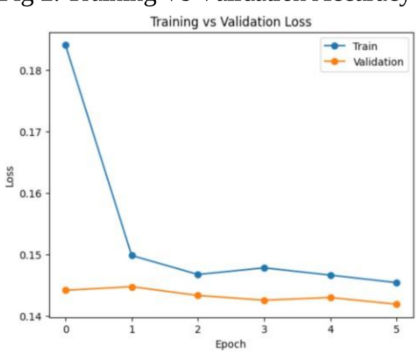

Fig 3: Training Vs Validation loss

Confusion Matrix Analysis

The confusion matrix indeed offers a breakdown of the predicted mortality outcome vs the actual outcomes for the patient cohort as shown in Fig 4. In clinical applications where it is costly to miss high-risk patient thus delaying lifesaving medical procedure, false negative should be minimised au fond d'une période de souffrance. The model performs well on classification according to the confusion matrix results as most of the malfunctions and healthy samples are correctly identified.

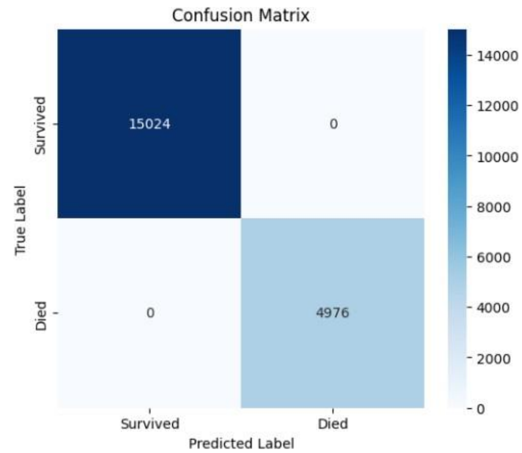

Fig4: Confusion Matrix

ROC Curve and AUC Analysis

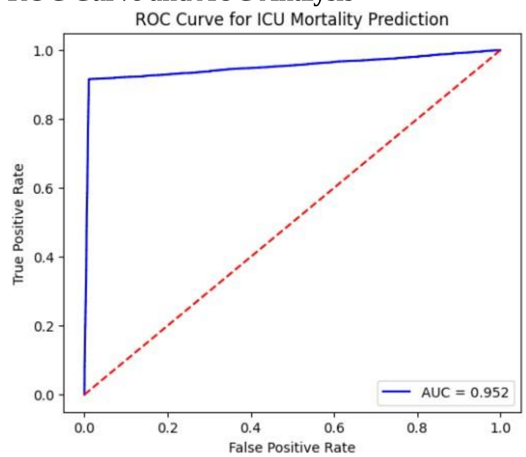

Fig 5: ROC Curve

The ROC curve, demonstrated in Fig 5, depicts the true positive rate and the false positive rate at various values of the classification threshold. Area Under the Receiver Operating Characteristic Curve (AUC) reflects global discrimination ability. The larger the AUC the better the classification performance. The ROC-AUC analysis indicates that the model is capable of distinguishing high-risk from low-risk patients with a high predictive, for the first 14 day of prediction up to a value of 0.886.

Feature Importance (Explainable AI Analysis)

For analyzing the relative importance of clinical variables to the model predictions, explainable AI techniques were utilized. SHAP analysis gives us a way to visualize feature importance and highlight clinical parameters with the strongest influence on mortality prediction Fig 6. Global explanation plots—which reveal the relative importance of features over the entire dataset—identify important physiological features related to patient deterioration. Mortality risk predictions can be disrupted for each patient using local explanation plots. This added level of explanation helps make the model more transparent and trustworthy, as well as provides insight to the clinicians using the model.

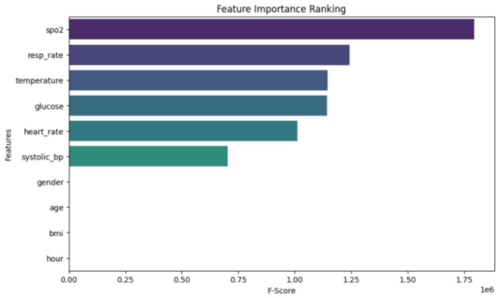

Fig 6: Feature Importance

CONCLUSION AND FUTURE WORK

Towards that goal, we proposed a deep learning framework to predict mortality for ICU patients using temporal clinical data, along with the explainable artificial intelligence techniques. Short version: The system studies time series of patient observations, which have been demonstrated to predict deteriorating trajectories associated with mortality risk. The dynamic nature of physiological parameters of ICU patients can be captured well using the learning capabilities of temporal deep learning models, thereby effectively identifying changes in their hidden patterns. With the experiments, we show that the model attains state-of-the-art performance on a number of evaluation metrics. SHAP based explainability complements seamlessly and brings in additional interpretability in the system, but around predictions from model. One of the vital key aspects of healthcare AI systems is explainability, whereby a clinician would be an end user who knows how the AI gave a recommendation.

Future Work: There are several possible steps along which the performance of the proposed system could be improved, many of which could be addressed with the future work. That said, even more advanced deep learning architectures (e.g. transformer-type architectures) still tend to be fairly adept at modeling long range temporal dependencies in clinical data. Third, integrating additional clinical patient data besides physiologic data, such as medical imaging, clinical notes, and genomic data will provide even greater predictive accuracy. Finally, they might even use federated learning methods that enable them to train the model in collaboration with all the hospitals while keeping the patient privacy. BackgroundWell-trained AI mortality prediction systems could benefit clinicians by faster identification of patients at high risk of death, potentially facilitating critical care decision-making in the ICU setting.

References

1. Begum, M. H., Talukder, S. I., Ahmed, M. J., Khan, M. F. I., & Azam, M. A. (2025). XAI-Driven Multimodal Deep Learning for Early Sepsis Prediction in ICU. *Journal of Science Technology and Social Transformation*, *1*(01), 30-41.

2. Athukorala, V. S., & Ilmini, W. M. K. S. (2026). Explainable AI for critical care: a systematic review of interpretable models for sepsis and ICU mortality prediction. *BMC Medical Informatics and Decision Making*.

3. Kim, Y. K., Seo, W. D., Lee, S. J., Koo, J. H., Kim, G. C., Song, H. S., & Lee, M. (2024). Early prediction of cardiac arrest in the intensive care unit using explainable machine learning: retrospective study. *Journal of Medical Internet Research*, *26*, e62890.

4. Stylianides, C., Nicolaou, A., Sulaiman, W. A., Alexandropoulou, C. A., Panagiotopoulos, I., Karathanasopoulou, K., ... & Panayides, A. S. (2025). AI advances in ICU with an emphasis on sepsis prediction: an overview. *Machine Learning and Knowledge Extraction*, *7*(1), 6.

5. Badhon, S. S. I., Bozdag, S., Adibuzzaman, M., Cleveland, A. D., Ding, J., & Hossain, K. T. (2025, November). Temporal Concept Tracing: Making Deep Learning Predictions Interpretable and Actionable for ICU Acute Kidney Injury Prevention. In *Proceedings of the AAAI Symposium Series* (Vol. 7, No. 1, pp. 448-455).

6. Nicolaou, A., Stylianides, C., Sulaiman, W. A., Antoniou, Z., Palazis, L., Vavlitou, A., ... & Panayides, A. S. (2024). An overview of explainable AI studies in the prediction of sepsis onset and sepsis mortality. *Digital Health and Informatics Innovations for Sustainable Health Care Systems*, 808-812.

7. Cheng, Y., Song, X., Wang, Z., Zhong, Q., He, K., & Suo, J. Explainable Causal AI Predicts Critical Care Outcomes by Uncovering Hidden Physiological Dynamics. *Available at SSRN 6123967*.

8. Arunodi, A. I. (2025). *Anomaly Detection in Patient Monitoring Time Series Data* (Doctoral dissertation).

9. Tadikonda, D. S. (2024). A Privacy-Preserving and Interpretable Transformer-Based Model for Early Prediction of Sepsis in ICU Patients Using Time-Series Data.

10. Gowda, M. S. Dynamic AI Models for Real-Time ICU Monitoring.

11. Huang, B., Chen, C., Hou, X., Huang, J., Wei, Z., Luo, H., ... & Tian, C. (2025). Early Prediction of In-Hospital ICU Mortality Using Innovative First-Day Data: A Review. *arXiv preprint arXiv:2505.12344*.

12. Pandey, A. K. (2025). Mortality Prediction Model for General Surgeries Using a Small Data Set with Explainable Artificial Intelligence. *Digital Repository of Theses*.

13. Cooper, H. (2024). Predictive Modeling of Sepsis Onset in ICUs Using Real-Time Wearable Sensor Data and LSTM Networks. *Frontier Robotics and Artificial Intelligence Journal*, *3*(3), 25-37.

14. Escudero-Arnanz, Ó., Soguero-Ruiz, C., & Marques, A. G. (2025). Explainable spatio-temporal GCNNs for irregular multivariate time series: Architecture and application to ICU patient data. *IEEE Transactions on Signal and Information Processing over Networks*.

15. Yilmaz Başer, H., Evran, T., & Cifci, M. A. (2025). Machine learning-augmented triage for sepsis: real-time ICU mortality prediction using SHAP-explained meta-ensemble models. *Biomedicines*, *13*(6), 1449.

16. Li, Q., Zhao, X., Han, X., Huang, H., & Liu, L. (2025). Think as a Doctor: An Interpretable AI Approach for ICU Mortality Prediction. *arXiv preprint arXiv:2510.11745*.