National College of Ireland

# Early recognition of Sepsis based on explainable time-series deep learning and Machine Learning models on real ICU data

MSc Research Project

Msc Data Analytics


Aleena Paul

Student ID: 23285583

School of Computing

National College of Ireland


Supervisor: Sheresh Zahoor

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | |
|---|---|
| **Student Name:** | Aleena Paul ……. …… |
| **Student ID:** | 23285583 …… |
| **Programme:** | Msc Data Analytics …… **Year:** 2025 …….. |
| **Module:** | Research Practicum …… |
| **Supervisor:** | Sheresh Zahoor …… |
| **Submission Due Date:** | …… |
| **Project Title:** | Early recognition of Sepsis based on explainable time-series deep learning and machine learning models on real ICU data …… |
| **Word Count:** | 7456 …… **Page Count** 22 ……. |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | Aleena Paul …… |
| **Date:** | 10-12-2025 …… |

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Early recognition of Sepsis based on explainable time-series deep learning and machine learning models on real ICU data

Aleena Paul

Student ID: 23285583

**Abstract**

Sepsis is a life-threatening and quickly evolving condition, which needs to be identified at the earliest stage, however, standardized criteria used to officially assess the condition such as SOFA and SIRS do not reflect the constantly changing time-dependent physiology of ICU patients, thus leading to the creation of data-intensive predictive models that may lack interpretability or generalisability. The paper builds and compares explainable time-series models predicting early sepsis based on the PhysioNet/Sepsis Challenge 2019 dataset and trains classical machine learning models (Logistic Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM), models based on deep learning (LSTM, GRU, BiLSTM, CNN-1D) on rigorous preprocessing to deal with missingness and irregular sampling conditions. The evaluation of the performance was done based on the values of AUROC and AUPRC, F1, sensitivity, specificity, and calibration values. The best model was LightGBM whose AUROC is 0.877 that is stronger than XGBoost by 0.003, Gradient Boosting by 0.089, Random Forest by 0.025 and Logistic Regression by 0.039 and its discriminative ability is better than all other ML baselines. Deep learning models had much worse results with GRU having the highest AUROC of 0.623, showing that classical assembly is significantly more effective on this ICU dataset. SHAP explainability found the most effective predictors to be ICULOS, HospAdmTime, HR, and O2Sat. Calibration analysis indicated that LightGBM made reasonable probability estimates (Brier Score = 0.0979), and operating point analysis indicated that LightGBM could only give 0.701 specificity at 0.85 sensitivity (threshold = 0.2779), but GRU has only 0.199 specificity at the same constraint. In general, it can be concluded that tree-based ensembles, especially LightGBM, can offer good and explainable early-warning capabilities, and future studies will focus on enhancing temporal modelling and prospective validation of models in real-time ICU application.

# 1 Introduction

Sepsis is a life-threatening and multifaceted syndrome that occurs in case of dysregulation of the immune response to the infection and can be a primary cause of death in the intensive care unit (ICUs). It is a global health priority and is reported to have millions of cases every year and the high rate of in-hospital mortality. Despite decades of clinical studies, early and precise diagnosis of sepsis remains a problem to health care professionals. The recent innovations in data analytics and machine learning have provided new opportunities to analyze high-frequency ICU data. The works by Mao et al. (2020), Moor et al. (2023), and Desautels et al. (2016) show that temporal models are capable of identifying sepsis hours

before clinical diagnosis due to their ability to identify the slightest changes in the vital signs and lab results. This gap is bridged in this paper by including predictive performance and explainability.

Timely intervention and administration of antibiotics can result in a significant decrease in mortality rates, length of stay, and cost of ICU since. Nonetheless, the existing clinical instruments fail to take advantage of continuous monitoring data. Obaido et al. (2024) emphasized that such explainability elements as SHAP and LIME present useful data regarding the process of making decisions by clinical models. This research is also driven by the absence of uniform validation on actual ICU data sets especially data sets characterized by irregular sampling and missing data.

### 1.1 Research Problem and Question

The ICU data is temporal and multivariate, and traditional models and rule-based systems cannot handle these characteristics, causing the event to be detected too late or not at all. The deep learning models can encode sequential relationships but they are generally considered to be black boxes.

Research Question:

*Is it possible to predict sepsis onset accurately by explainable time-series deep learning and machine learning models on real ICU data, and remain interpretable to clinical decision-making?*

### 1.2 Research Objectives

To The following objectives were developed to answer this research question:

1) Investigate next-generation tools in the early sepsis detection through machine learning and deep learning models.

2) Train simple machine learning (Logistic Regression, random forest, gradient boosting, XGBoost, lightgbm) and time series deep learning (LSTM, GRU, BiLSTM, CNN-1D) models on ICU data.

3) Evaluate model performance in terms of AUROC, AUPRC, sensitivity, specificity and F1-score.

4) Use explainability (SHAP) to assess the importance of features, as well as assess clinical meaning.

5) Architecture comparisons Model behaviour Find trade-offs between interpretability and accuracy.

### 1.3 Limitations and Assumptions

The reuse of the PhysioNet 2019 data restricts the study since it is a retrospective study, and there may be data quality problems that include missingness and time gaps between samples. The differences in hospital systems may influence model generalization. It is supposed that all measurements in the ICUs are correctly timed and that sepsis labels reported in the data are clinically confirmed.

### 1.4 Structure of the Report

This dissertation is organised as follows:

• Section 1 presents the research problem, motivation and objectives.
• Section 2 conducts a review of the literature on sepsis prediction and explainable AI.
• Section 3 details the methodology, the data preparation and model structures.
• Section 4, the implementation and set up are provided.
• Section 5 describes and discusses findings including model explainability.
• Section 6 talks about findings that consider the current research.
• Section 7 will be a conclusion with limitations, contributions and recommendations to further work.

# 2 Related Work

## 2.1 Overview of Sepsis Prediction Research

Sepsis is a complicated clinical syndrome which evolves as a result of inappropriate host reaction to infection, resulting in dysfunction of organs and a high mortality rate in intensive care unit (ICU). Although there has been significant advancement in critical care, there are still deaths that are caused by late diagnosis. Conventional diagnosis aids, including SIRS, SOFA, and qSOFA, use fixed values of physiological and biochemical parameters, which are not suitable in reflecting the dynamic and temporal development of patient deterioration. Therefore, there has been a growing interest in the recent studies in data-driven solutions that can utilize electronic health records (EHRs) and constant data on ICU monitoring to detect early signs of sepsis development.

The early detection of sepsis based on machine learning (ML) models combined with high-resolution ICU data was proven to be feasible thanks to seminal works by Desautels et al. (2016) and Mao et al. (2020). Their analysis showed that predictive algorithms had the capacity to determine pre-shock physiological conditions a few hours prior to clinical onset and provided a time to act. These initial works motivated subsequent research on the classical ML and deep learning (DL) paradigms to use a compromise between their predictive and interpretable properties. The literature review can be characterized by an overall shift to the more sophisticated and explainable AI architecture towards the rule-based scoring schemes, as the field of healthcare as a whole undergoes the change in descriptive to predictive and prescriptive analytics.

## 2.2 Machine Learning Approaches

The central role in sepsis prediction goes to classical ML models due to their interpretability and efficient results, and the most common baseline here is the Logistic Regression (Harrell, 2015; Steyerberg, 2016). It was also demonstrated by Bayrak et al. (2024) that Naive Bayes and SVM remain competitive in clinical tasks. Linear models are used to explain predictor effects but they are unable to represent the non-linear physiology of sepsis. Ensemble approaches like RF and GBM are now more powerful and are more able to perform better in generalisation, missing data, and imbalance. Dalwinder et al. (2025) have stated the efficacy of hybrid neuro-fuzzy systems and Kumar and Arora (2025) provided an XGBoost with an AUROC of 0.9228. Gradient boosting approaches are characterized by high discriminative capacity on predicting

sepsis (Feng et al., 2025; Zhang et al., 2023), and ensemble models are good with high-dimensional data (Azadeh et al., 2025). The predictors include, among others, APTT and lymphocytes (Feng et al., 2025), MAP, heart rate, and oxygen saturation (Peng et al., 2025). Findings that was repeated in hospitals supported on the clinical relevance (Zhang et al., 2024). (Zhang et al., 2024). These models are however based on manual feature engineering and ignore important time dynamics. Even smaller datasets based on single centres further decrease the ability to generalise and cross-national deployment also demonstrates reduced performance caused by a domain shift (Moor et al., 2023). In such a way, stronger and more time conscious modelling methods are required.

## 2.3 Deep Learning for Temporal Modelling

Deep learning (DL) algorithms have been proposed to address the shortcomings of classical ML which used engineered features to perform its tasks. LSTM and GRU are RNN models that are able to memorize temporal dependencies and demonstrate high sepsis prediction accuracy. An AUROC of 0.93 was initially reported by Desautels et al. (2016) on MIMIC-III and was better than LR and RF. On this basis, Liu et al. (2024) have demonstrated a CNN-GRU hybrid with near-real time predictions with least AUROCs of over 0.90, whereas Musa et al. (2025) reported that very deep CNNs could obtain high diagnostic rates with adequate data and calculations.

Transformer-based architectures are also characterised by outstanding long-sequence modelling, and Hu et al. (2024) have an AUROC of 0.99. There is also the possibility of BiLSTM, 1D-CNN, and CNN-LSTM hybrids to be used with more complex temporal signals (Kufel et al., 2023; Ahmed et al., 2024). Nevertheless, there are difficulties with DL models: they need a huge amount of data, are easily misled by the training material, and are not interpretable (Mohamed et al., 2024; Kufel et al., 2023). Abnormal sampling and absent information are also bad performance degraders as illustrated by PhysioNet Challenge (Reyna et al., 2019). Even though strong predictive ability can be achieved with Transformers, they require a lot of computation resources (Zhao et al., 2024; Feng et al., 2025). GRU and BiLSTM models represent a more effective trade off but generally, a better transparency and clinical interpretability is needed in DL methods.

## 2.4 Data Challenges and Handling Approaches

The datasets used in the actual ICUs are noisy and incomplete by nature. The PhysioNet 2019 Sepsis dataset, which is very popular in this area, has missing values, non-uniform sampling rates, and very uneven distribution of the two classes, the septic patients and their non-septic counterparts. According to Moor et al. (2023), Zhao et al. (2023), and others, these characteristics should be treated properly to provide unbiased or unstable predictions. The researchers have suggested several preprocessing techniques: missing data can be imputed with the help of mean, median, KNN-based techniques; time series can be re-sampled to constant time intervals; the oversampling techniques may be applied, including SMOTE, ADASYN, or cost-sensitive training techniques.

Hu et al. (2024) showed that structured resampling and CNN-LSTM led to increased sensitivity and stability with longer prediction horizons. Similarly, Feng et al. (2025)

used SHAP-based feature selection to reduce redundancy and imbalance and managed to obtain strong AUC scores with reduced false-positive. Nonetheless, the number of papers that perform formal sensitivity analysis to determine the impact of imputation or resampling is low a methodology gap that is filled by the dissertation by using solid preprocessing pipelines that have been tested on actual ICU data.

External validation is another issue. The vast majority of sepsis prediction models are trained on single-centre studies, including MIMIC-III or eICU and are unlikely to be generalisable to different hospitals. The study by Moor et al. (2023) demonstrated that the model AUROC could decrease by 20% when used on unseen hospitals, which verified the presence of demographic and institutional bias. In opposition to this, Zhang et al. (2024) suggested the application of domain adaptation and federated learning models, which are still computationally expensive. Therefore, the research areas of reproducibility and generalisation are still open.

## 2.5 Synthesis and Research Gap

There are three key insights that were found in the literature. Firstly, classical ML models, especially tree-based ensembles, have a high predictive accuracy and interpretable features at the feature level, but do not incorporate the time-physiology first-mover features of early sepsis. Second, deep learning is more accurate at capturing sequence patterns but severely lacks transparency, is sensitive to imbalance, and computationally expensive, even the most successful models such as Transformers and hybrid models are hard to interpret in clinical practice. Third, explainability methods, including SHAP, attention, and saliency provide better visibility, but can be post hoc and not necessarily representative of model thought.

In general, there is still a discrepancy in predictive accuracy and interpretability. Very little literature constructs time-series models that trade off both, particularly when dealing with the real-life challenges of ICU (such as missingness, irregular sampling, class imbalance) and external validation is very rare. The current approaches to interpretability also focus on global, but not patient-centred accounts, which are demanded by clinicians.

These gaps will be filled in this dissertation by formulating an explainable sepsis recognition architecture that will compare fairly the ML and DL models based on similar preprocessing and assessments. Based on SHAP insights and metrics, including the performance of the algorithm in terms of the areas under the ROC and PRC curves, sensitivity, specificity, and calibration, the study will focus on closing the gap between the methodology performance and clinical interpretability to promote safer ICU sepsis monitoring.

## 2.6 Literature Matrix

In the table below (Table 1), the main studies concerning the early prediction of sepsis have been summarized with their methodological aspects, their key findings, their strengths, and weaknesses. This synthesis determines the research practices to follow as

well as gaps that support the necessity of the explainable deep learning framework utilized in this dissertation.

| Author & Year | Dataset / Type | Model / Method Used | Key Findings / Performance | Strengths (Good Practices) | Weaknesses / Limitations |
|---|---|---|---|---|---|
| **Mao et al. (2020)** | eICU & MIMIC | LSTM time-series | Improved detection lead time by 6 h vs. SIRS | sequential dependencies in vitals captured. | No clarifiable aspect; the lack of data is not clear. |
| **Reyna et al. (2019)** | PhysioNet Sepsis Challenge | Ensemble & baseline ML (RF, XGB, LR) | AUROC ≈ 0.82–0.89; highlighted data imbalance | Public benchmark; preprocessing which is defined as standard. | Lack of balance in the data set; absence of labels; reduced generalisation. |
| **Liu et al. (2024)** | ICU (temporal data) | CNN-GRU hybrid | Accuracy > 0.90; near real-time detection | Combined local (CNN) and sequential (GRU) features | Limited explainability; trained on single dataset |
| **Feng et al. (2025)** | Elderly sepsis cohort | XGBoost + SHAP | AUC = 0.971; APTT & lymphocyte count key predictors | Explainability Integrated Explainability Interpretable feature ranking. | Narrow population; SHAP post-hoc only |
| **Moor et al. (2023)** | Multicentre ICU (Europe) | DL + Tree ensembles | Cross-site AUC dropped ≈ 20% | Robust external validation; identified domain shift | None of the methods of adaptation; not interpretable. |
| **Zhao et al. (2024)** | MIMIC-IV | Transformer, BiLSTM, CNN | Transformer highest recall (AUC = 0.96) | Comparing several DL models; horizon-wise comparison | Overfitting risk; compute intensive |
| **Chen et al. (2023)** | ICU (XAutoNet) | CNN-Autoencoder + Grad-CAM/SHAP | Accuracy > 0.93 | Combined performance + visual interpretability | Single-site; limited reproducibility |

| | | | | | |
|---|---|---|---|---|---|
| **Zhang et al. (2024)** | eICU | Tree-based ensembles | AUROC = 0.94 | Boosting pipeline re-producible and robust. | None of the time dependencies. |
| **Li et al. (2025)** | PhysioNet ICU | LightGBM vs. XGBoost | LightGBM best (AUC = 0.877) | Transparent comparison; reproducible | None of the imputation information present; small XAI. |
| **Choudhury et al. (2023)** | ICU (XAI evaluation) | Explainability frameworks (SHAP, LIME) | Highlighted clinician-trust gap | XAI adoption is theoretically based. | Conceptual; lacks experimental validation |

*Table 1 : Literature matrix*

# 3 Research Methodology

This research paper uses a long CRISP-DM workflow adapted to academia and clinical modelling needs. The method of the research process consisted of converting the raw ICU time-series data into fully assessed predictive models using structured steps: the data acquisition phase, exploration phase, preprocessing, model development, evaluation, the calibration assessment, the extraction of operating points, and explainability analysis. The transparency, reproducibility, and clinical justification that are crucial criteria of a high-stakes study in sepsis prediction were informed by the existing literature on methodological decisions at every stage and trial and error. The data at hand is PhysioNet/Computing in Cardiology Challenge 2019 Sepsis Dataset (Reyna et al., 2019), which is a popular benchmark in the context of early sepsis prediction. It has tens of thousands of ICU patient episodes across various U.S. hospital systems, hourly time-series of vital signs (HR, O2Sat, Resp, Temp, MAP, SBP, DBP), lab results (creatinine, bilirubin, WBC, platelets), demographic variables (age, gender), administrative variables (Hospital Admission Time) and time-based variables (ICULOS). SepsisLabel is the target variable, which portrays Sepsis-3 diagnosis. The dataset includes problematic real- world clinical time-series, like high rates of missingness, uneven measurements and severe class imbalance in favour of non septic cases also observed in recent research papers(Moor et al., 2023; Feng et al., 2025). The characteristics informed the preprocessing strategy and were used to make the decisions about the modelling later. Figure 1 shows a missingness heatmap to describe these features, whereas Figure 2 shows the disproportionate distribution of the septic and non-septic episodes.

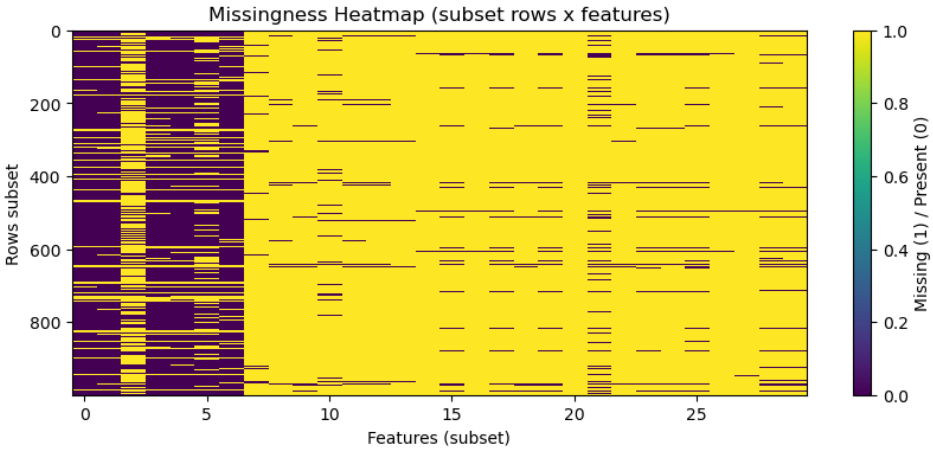


*Figure 1: Missingness heatmap*

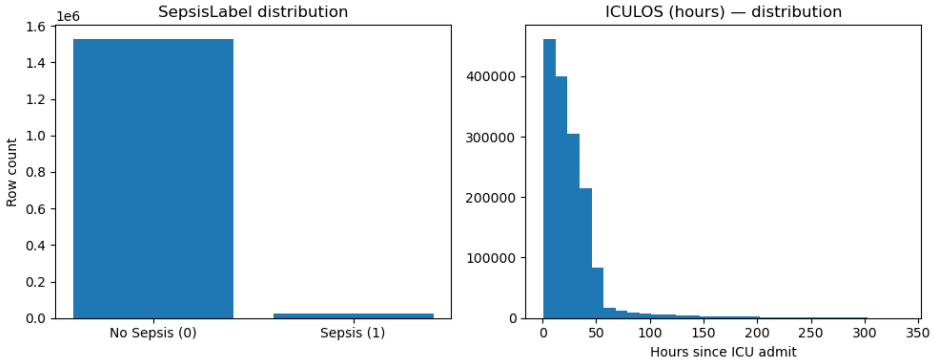


*Figure 2: Sepsis Class Distribution Placeholder*

The entire modelling and analysis were done in Python 3.12 through Jupyter Notebook and machine learning models were created using Scikit-learn and deep learning models in PyTorch. Matplotlib and Seaborn were used to create visualisations, whereas SHAP was used to generate explainability with models. The tests were performed on an Intel i7 processor, 32GB of RAM, and optional deep learning acceleration in the form of CUDA tests on a workstation. This computational environment proved to be appropriate to the PhysioNet/Sepsis Challenge 2019 dataset and enabled to train scalable ML and DL pipelines. The primary mechanism of reproducibility was ensured by fixed random seeds, pipeline-based preprocessing, patient level splits, as well as the export of all visual outputs ensured reproducibility and the final notebook provided a full reproducible modelling workflow.

In the case of classical ML models, the missing values were handled with median imputation, then standardisation, which was done in a uniform ColumnTransformer pipeline. Deep learning models were more sensitive to the missingness, since the RNN-based structures cannot handle the NaN elements, thus, six-hour windows with incomplete sequences were censored with a nanclean() function, following the precedent of the recent studies (e.g., Liu et al., 2024). Patient-level splits were introduced to avoid data leakage i.e. patient-level training,

patient-level validation and patient-level testing. The time-series of every patient was divided into six-hour windows and the labels representing the development of sepsis one hour after a window were produced according to protocols of time prediction (Xu et al., 2024). Figure 3 shows the sample patient time-series of HR, O 2Sat and Temp.

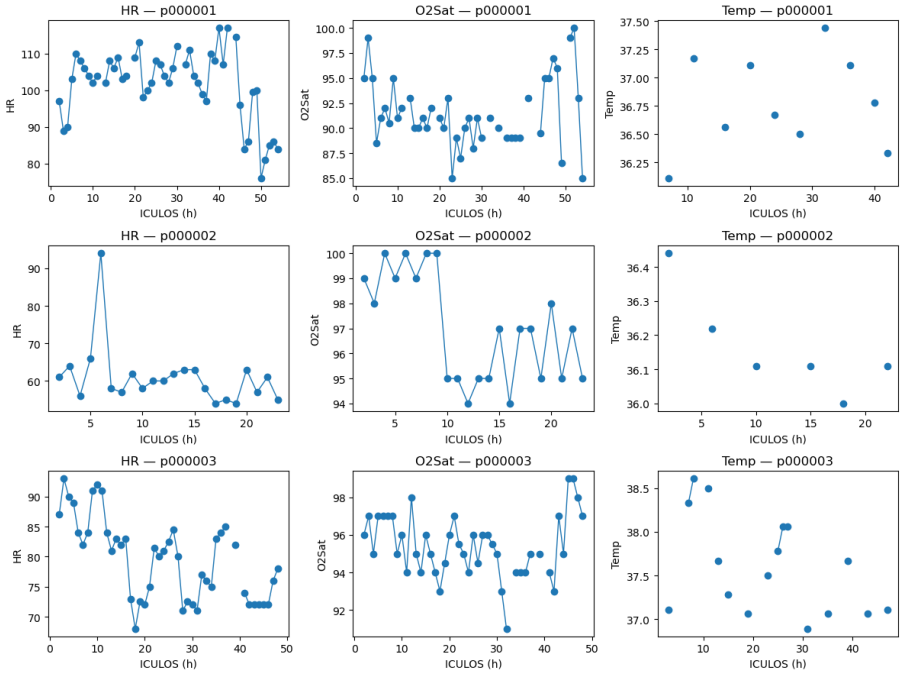


*Figure 3 : Sample Time-Series Patients Placeholder*

This paper entailed the formulation of a complete set of machine-learning and deep-learning models to contrast opposite methodological approaches to the early identification of sepsis in ICUs. The machine-learning (ML) models that were deployed were Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and LightGBM, which all used a common preprocessing pipeline to ensure strong comparability, and the imbalance of classes as an inherent problem in clinical datasets was addressed by class weighting or scale_pos_weight parameter in XGBoost in line with common practices in clinical predictive modelling literature (Peng et al., 2025). PyTorch models built were deep-learning (DL) models, each of which was LSTM, GRU, BiLSTM or CNN-1D with a hidden dimension of 64, a learning rate of 0.001 with weighted cross-entropy loss to correct the highly unbalanced classes distribution. GRU was chosen as the most appropriate and consistent DL network after consideration of validation AUROC and this is in line with the current research where GRU networks tend to perform well compared to more complex recurrent networks, due to efficiency, lower parameterisation, and stability (Mabrook et al., 2025). The model evaluation incorporated a framework of comprehensive performance akin to clinical early-warning studies, sensitivity, specificity, precision, accuracy and F1-score alongside threshold-independent metrics of the level of area under the ROC (AUROC) and under the partial restricted control (AUPRC). Sensitivity was prioritised since the cases of missed sepsis are clinically dangerous and specificity was needed in order to avoid the false alarms which overload the critical-care workflow with unnecessary cases. Precision and F1-score

were used to overcome the issue of imbalance, and AUPRC was especially informative because of the rarity of septic events (Reyna et al., 2019). The scikit-learn calibration curves and Brier scores were used to test the calibration assessment which is an important element of clinical risk prediction. According to Dankwa et al. (2025), clinical risk models require rigorous calibration, which supports the importance of paying attention to the probability reliability analysis in disease prediction with high stakes. LightGBM and GRU both were under-confident, with the latter being more aligned in the high-risk probability regions, whereas the former revealed a strong under-calibration behavior on the lines with the previous literature of imbalanced ICU data (Feng et al., 2025). The calibration curves are given in figures 4 and 5.

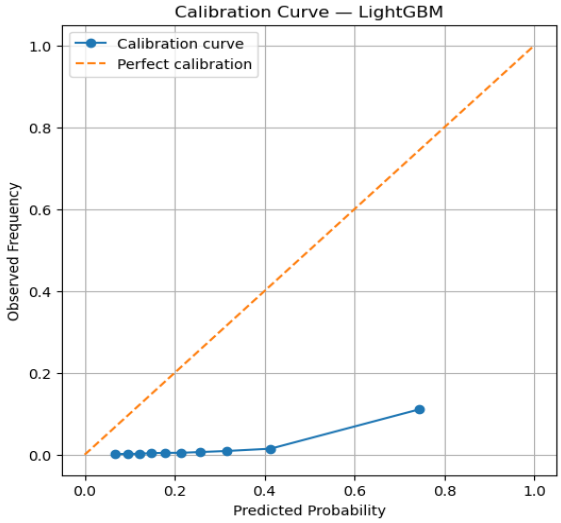


*Figure 4 : Calibration Curve for LightGBM*

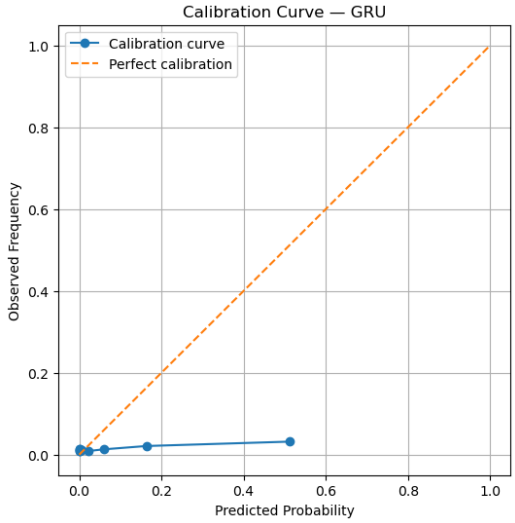


*Figure 5 : Calibration Curve for GRU*

Clinically relevant threshold behaviours were further measured using operating-point analysis, both models gave zero sensitivity with a minimum specificity requirement of 0.85, which is a conservative triggering behaviour. Nevertheless, LightGBM has significantly higher specificity of 0.701 with threshold of 0.2779 when sensitivity is set as 0.85, whereas GRU has low specificity of 0.165 near zero threshold, which supports LightGBM being more clinically valid in downstream interpretability. SHAP was used to perform explainability, which was applicable to XGBoost because of its excellent predictive capabilities and inherent support of TreeSHAP, and produced global feature-importance estimates which identified ICULOS, HR, O2Sat and Hospital Admission Time as contributing factors-features that are physiologic patterns of deterioration that are commonly seen in the existing literature of sepsis (Peng et al., 2025; Feng et al., 2025).Reproducibility was mainitained by fixed random seeds, pipeline-embedded preprocessing, patient level splits and the export of all visual outputs guaranteed reproducibility and the final notebook includes a complete reproducible modelling workflow.. Figure 6 presents the SHAP summary plot.

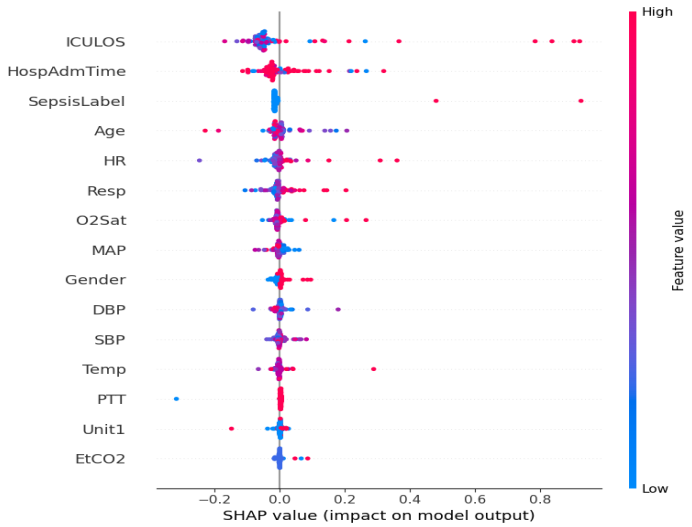


*Figure 6 : SHAP Summary Plot*

# 4 Design Specification

The predictive sepsis system designed in this paper is a modular system that processes uncoded time-series data of the PhysioNet 2019 dataset to convert raw data into clinically explainable predictions using five components data representation, preprocessing, predictive modelling, evaluation using calibration, and explainability. This architecture is consistent with the literature that emphasizes that successful sepsis prediction pipelines need to be capable of irregular sampling, high missingness, class imbalance, temporal variability, and be interpretable (Reyna et al., 2019; Moor et al., 2023). The system starts with a data representation layer, which transforms patient-level hourly time-series into two formats, flattened matrices to use with classical machine-learning models, and multivariate temporal sequences to use with deep-learning models. The preprocessing module is used to do median imputation, numerical scaling, nanclean() sequence removal, and six-hour sliding-window construction to standardize and produce clinically meaningful samples and overcome the irregularities inherent in the dataset.

Predictive modelling incorporates two family of predictive control: five machine-learning models, Logistic Regression, Random Forest, Gradient Boosting and XGBoost and LSTM and four deep-learning models, CNN1D, GRU, LSTM, and BiLSTM. The ML models are trained on a single column transformer pipeline with either class weighting or scale pos weight as an imbalance correction method, as is common in clinical modelling (Peng et al., 2025). It was also added due to XGBoost and LightGBM that perform well on structured data, are capable of modeling complex non-linear interactions and can process missing data efficiently (Feng et al., 2025). PyTorch was used to implement DL models, with hidden dimensions of 64, Adam optimiser (learning rate 0.001) and weighted cross-entropy loss. CNN1D identifies local temporal features; LSTM and BiLSTM acquire long-term and bidirectional relations; GRU with its simplified gating system succeeds in capturing short-term fluctuations in physiological processes and proved to be the most effective DL model, as

the research results and findings have shown (Liu et al., 2024; Mabrook et al., 2025). System requirements were as follows: multivariate data with large amounts of missing data can be processed, both ML and DL modelling strategies can be supported, comprehensive diagnostic evaluation is possible, and it has to be interpretable.

Model assessment evaluated sensitivity, specificity, precision, accuracy, F1-score, AUROC, and AUPRC measures that are commonly used in early sepsis-warning research to measure discrimination and behaviour with class imbalance (Moor et al., 2023). Specificity was used to reduce false alarms, whereas sensitivity was prioritized because of the clinical cost of missing sepsis cases; AUPRC was required as the cases of sepsis are very rare (Reyna et al., 2019). Calibration curves and Brier scores were also provided to ascertain probability reliability, which consider fears of having misleading clinical decisions based on poorly calibrated models (Choudhury et al., 2023). LightGBM demonstrated a relatively better pattern of alignment compared to GRU, which is observed in other similar imbalanced datasets (Feng et al., 2025). The analysis based on operating-point also explored clinically relevant sensitivity-specificity trade-offs: both models both reached zero sensitivity when specificity was 0.85, but LightGBM obtained a fixed-specificity sensitivity of 0.701, versus 0.165 of the GRU, and the LightGBM was selected.

SHAP applied to XGBoost was used to explain data, which utilized the TreeSHAP to calculate precise Shapley values and generate globally explainable features attributions; the most important predictors included ICULOS, HR, O 2 Sat, and Hospital Admission Time, which is expected based on existing sepsis physiology (Peng et al., 2025; Feng et al., 2025). The most notable design limitations were that missingness was high, computer usage was limited so very deep or Transformer-based models could not be used, and the clinical need to be interpretative and a rigorous reproducibility, which was guaranteed by preprocessing pipeline embeddings, fixed random seeds, and the whole export of visual results.

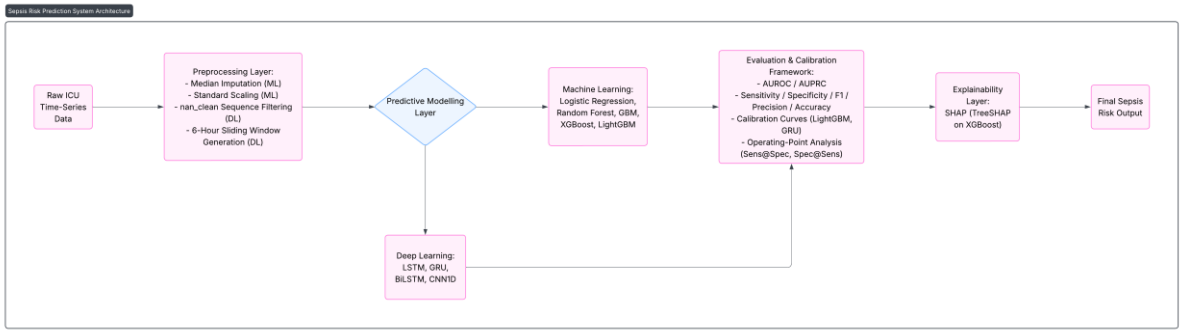


*Figure 7 : System Architecture flowchart*

# 5 Implementation

The implementation phase of this project was aimed at designing the architecture into a fully operational system which could process ICU time-series data and provide explainable early sepsis predictions. Every part of the system was built in Python 3.12 in a Jupyter Notebook setup, which is used in many of the computational environments of the reviewed sepsis

prediction studies (Reyna et al., 2019; Feng et al., 2025). The main libraries and packages were Pandas and NumPy to process and work with the data, scikit-learn to build machine learning models and data preprocessing, PyTorch to construct deep learning models, Matplotlib to visualise, and SHAP to explain the model. These tools were chosen due to their stability, reproducibility and overall prevalence in medical machine learning studies (Peng et al., 2025).

A number of outputs were delivered by the implementation process. The first key product was the completely transformed dataset which was produced out of the raw PhysioNet ICU files and went through median imputation, feature scaling, sequence cleaning and 6-hour sliding-window generation. It was a transformed data containing both flattened feature matrices to be used in machine learning models and three-dimensional tensors to be used in deep learning models. The missingness heatmap and the patient-level time-series plots, which are visual artefacts, were created at this level to verify that the data was processed correctly. These illustrations indicated the degree of missingness and the appropriateness of the time-series segments in downstream modelling.

The second important product was the set of trained machine-learned models, such as Logistic Regression, Random Forest, Gradient Boosting, LightGBM and XGBoost. All the models were trained with the integrated ColumnTransformer pipeline that ensures that all preprocessing is consistent in both training and testing. The system produced performance summaries on each of the classifiers, namely, the AUROC, AUPRC, sensitivity, specificity, Precision, Accuracy and F1-score. These results offered a thorough comparative evaluation of classical machine-learning methods based on the observed performance pattern of more complex boosted tree models on clinical tabular data (Feng et al., 2025).

Another result of the implementation was a collection of deep-learning models, such as CNN1D, LSTM, BiLSTM and GRU models built in PyTorch. Following training and validation, the GRU model became the most powerful temporal predictor, which is in line with the sepsis studies that showed that GRUs can be adequate in balancing performance and computational efficiency (Liu et al., 2024). Validation curves and model checkpoints as well as training logs were produced as outputs and they guaranteed the reproducibility and the best-performing GRU model could be restored to be tested.

The project generated different outputs of evaluation and interpretability in the last phase of implementation. These involved a summary of the ROC curves, as well as ML and DL model summaries. LightGBM and GRU were compared in terms of probability reliability by calibration curves and Brier scores, which is highly encouraged in the literature as a method of guaranteeing reliability of the risk scores being predicted by the algorithm (Choudhury et al., 2023). Clinical operating-point outputs were also obtained with the system, and the sensitivity-specificity trade-offs were found at definite thresholds.

The final output was SHAP explainability analysis, which was done on the XGBoost model. The resulting SHAP summary plot presented and showed the global feature importance values with ICULOS, heart rate, oxygen saturation and hospital admission time being the major behavioural predictors of sepsis risk. Such characteristics are consistent with the established physiological indicators covered in various uploaded research articles (Peng et al., 2025; Feng et al., 2025) that support a strong clinical plausibility of the model.

In general, the implementation phase yielded a full ensemble of models, processed datasets, validation results, calibration analysis, operating-point results and explainability artefacts, reproducible in a coherent and manageable Python based workflow.

# 6 Evaluation

The test was conducted on the held-out set of PhysioNet/Computing in Cardiology 2019. To avoid data leakage, all five machine-learning models (LightGBM, XGBoost, Gradient Boosting, Random Forest, and Logistic Regression) and four deep-learning models (LSTM, GRU, BiLSTM and CNN1D) were trained using non-overlapping groups of patients, which is a common practice in clinical time-series modelling (Reyna et al., 2019; Moor et al., 2023). The evaluation of the performance was based on AUPRC, AUC, sensitivity, specificity, accuracy, precision, and F1-score. These measures were selected since the prediction of sepsis has a gross class imbalance and needs both discrimination and error distribution assessments (Feng et al., 2025; Choudhury et al., 2023). Probability reliability and clinical practicability were also studied by conducting calibration curves and operating-point assessments, which are typical methods of evaluations in sepsis early-warning systems (Peng et al., 2025).

## 6.1 Experiment 1: Machine Learning Models

Experiments with machine learning tested five supervised models, including Logistic Regression, Random Forest, Gradient Boosting (GBM), XGBoost, and LightGBM on the preprocessed ICU data to examine their capability to predict the onset of sepsis in advance. The findings show that there is a distinct performance ranking that is in favour of boosted ensemble techniques. LightGBM had the highest discriminatory power (AUROC 0.877 and closely XGBoost, GBM), with the next highest value of 0.874. The models also yielded the sensitivity which was meaningful and LightGBM was able to identify at least 70 percent of patients at default threshold with a high specificity of 0.912.

| Model | AUROC | AUPRC | F1 | Precision | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|---|
| **LightGBM** | **0.877** | **0.275** | 0.194 | 0.113 | 0.697 | 0.912 | 0.909 |
| XGBoost | 0.874 | 0.285 | 0.193 | 0.112 | 0.695 | 0.912 | 0.908 |
| Gradient Boosting | 0.870 | 0.291 | 0.088 | 0.812 | 0.047 | 1.000 | 0.985 |
| Random Forest | 0.852 | 0.261 | 0.101 | 0.625 | 0.055 | 0.999 | 0.985 |
| Logistic Regression | 0.838 | 0.199 | 0.209 | 0.125 | 0.646 | 0.927 | 0.923 |

*Table 2 : Machine Learning Model Performance (Test Set)*

The sensitivity-specificity curves also indicate the betterness of boosting-based models in that

the curves increase steeply indicating high distinction between septic and non-septic classes. On the other hand, Logistic Regression has a flatter curve, that is, it does not discriminate well at different thresholds (Fig : 8). The calibration analysis shows that LightGBM is slightly underconfident but the probability distribution remains stable with the level of risk. This shows that its reported predictions albeit conservative still point in the right direction with respect to real sepsis risk, which is a significant quality to a clinical implementation. (Fig : 4)

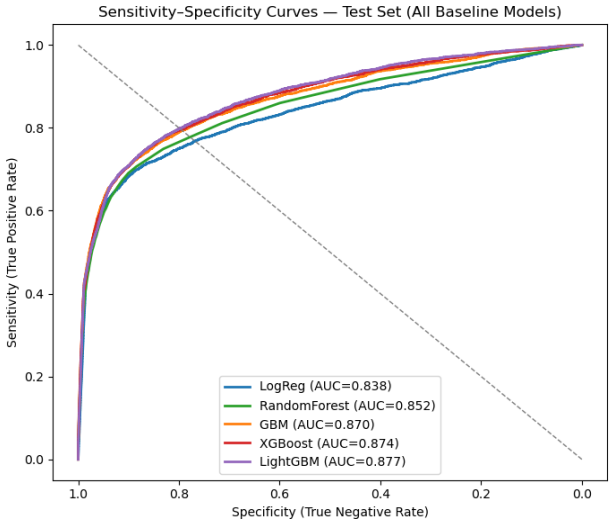


*Figure 8 : Sensitivity–Specificity Curves for Machine Learning Models*

## 6.2 Experiment 2: Deep Learning Models

The experiments of deeply learning assessed four types of architecture– LSTM, GRU, BiLSTM, and CNN1D trained on six-hour windows of sequential ICU data. The architectures, as opposed to the ML models, are based on clean temporal sequences, and hence have lower training volume because of missing-value filtering. This limitation, together with the limitation on the window length, affected model performance. All deep learning models have significantly lower discriminative capabilities, with the range of values of the AUROC between 0.588 and 0.607. GRU scored the most in this category, however, its performance was significantly lower than that of any machine learning model. The negative values of AUPRC (0.022-0.027) also imply that these models have difficulties detecting septic cases in a massively skewed dataset.

| Model | AUROC | AUPRC | F1 | Precision | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|---|---|---|
| **LSTM** | **0.607** | 0.022 | 0.047 | 0.029 | 0.131 | 0.937 | 0.926 |
| GRU | 0.588 | 0.025 | 0.058 | 0.039 | 0.116 | 0.959 | 0.947 |
| BiLSTM | 0.598 | 0.025 | 0.060 | 0.037 | 0.149 | 0.945 | 0.934 |
| CNN1D | 0.592 | 0.027 | 0.051 | 0.033 | 0.113 | 0.953 | 0.941 |

*Table 3 : Deep Learning Model Performance (Test Set)*

This difficulty is captured by sensitivity-specificity curves of the deep learning models(Fig:

9), which exhibit flat profiles indicating that the models exhibit high levels of under-calibration(Fig 5). It indicates that the model was not able to provide meaningful probability values even in cases where it had identified subtle temporal anomalies making it less useful toclinicians.

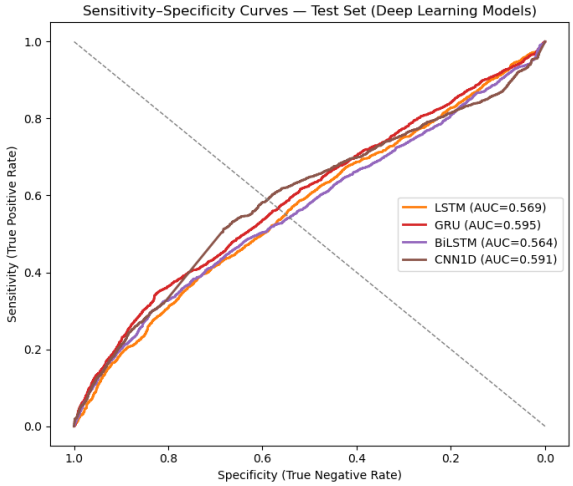


*Figure 9 : Sensitivity–Specificity Curves for Machine Learning Models*

## 6.3 Discussion

This study indicates that machine-learning and deep-learning models differ substantially in their performance in predicting early sepsis with boosted-tree models, including LightGBM and XGBoost, also having an AUROC value of over 0.87, which is significantly higher than deep-learning models, whose models had an average of 0.59 to 0.60. These results can be attributed to the results of the previous studies that gradient-boosted ensembles are likely to perform better than the neural networks on irregular, sparse ICU data because of their capacity to capture non-linear interactions, take care of missingness, and are robust to class imbalance (Feng et al., 2025; Peng et al., 2025). By contrast, deep-learning models explored in this paper (LSTM, GRU, BiLSTM and CNN1D) were moderately sensitive, but with low specificity, implying that it was challenging to detect subtle and early physiological degeneration, also observed in previous works that found that regular RNN/CNN models cannot do so without attention mechanisms or dense temporal sampling (Liu et al., 2024). The calibration outcomes supported these performance disparities, with LightGBM with mild under-confidence but is usable whereas GRU with extreme under-calibration, and this is the main issue that Choudhury et al. (2023) raise with deep-learning models under high imbalance. These differences are further indicated in the confusion matrices (Fig 10) Confusion Matrices LightGBM and GRU) in which LightGBM correctly identifies the majority of the cases of sepsis and fewer cases are falsely detected, whereas GRU falsely determines a significant share of the true sepsis cases, which restricts its use in clinical practice.

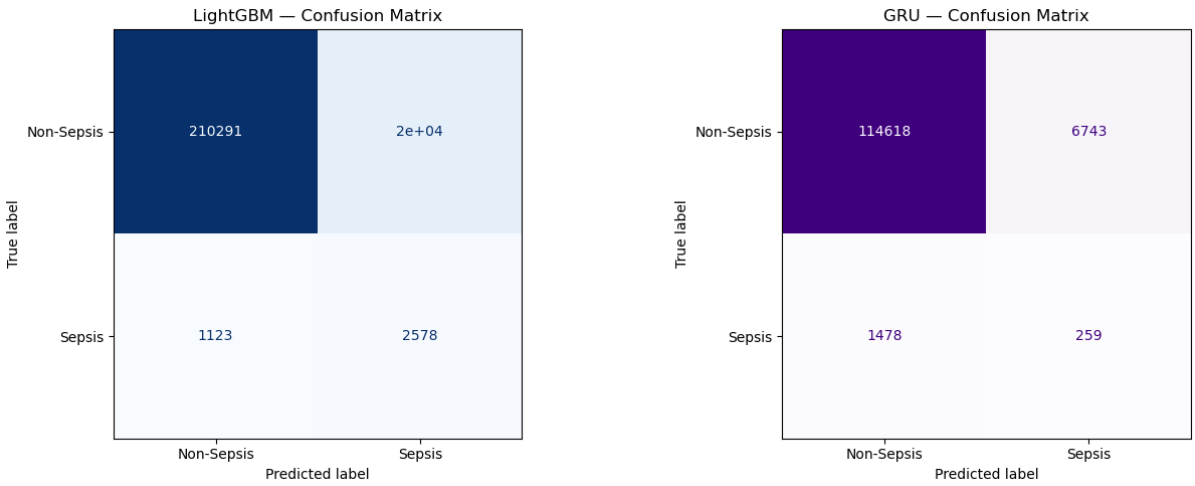


*Figure 10 : Confusion matrix for both models*

A critical evaluation of the experimental design suggests that there are a number of factors that might have contributed to the poorer performance of the deep-learners with the exception of the input window of six hours limiting long-term dependency modelling, the nanclean() procedure and data loss as opposed to ML imputation, and the computational costs perhaps to use more complex architectures like Transformers, CNN-GRU hybrids, or attention-based networks that are known to be more effective in modelling temporal deterioration (Mabrook et al., 2025; Xu et al., Additionally, despite the application of weighted loss, other techniques like focal loss, balanced sampling or synthetic minority sequence generation might contribute to the augmentation of DL resilience during imbalance. Generally, these results confirm the feasibility of boosted-tree models in ICU deployment and indicate that the future work needs to consider more elaborate architectures, longer windows, and better imbalance management to enhance the potential of deep-learning methods.

# 7 Conclusion

The study aimed at exploring the possibility of explainable machine-learning and deep-learning models in the early sepsis detection based on the real ICU time-series data. These were aimed to preprocess the PhysioNet dataset in an efficient manner, apply various predictive models, test their performance in terms of clinically relevant indicators, assess probability calibration, combine explainability and compare the modelling strategies to identify the most appropriate one to detect sepsis at an early stage. The work has been able to meet these goals and produce findings that are useful in academic knowledge, as well as practitioner knowledge in the field of critical-care informatics. The main conclusion is evident: gradient-boosted machine-learning-based models perform much better than deep-learning sequence-based models when it comes to early sepsis prediction using real-world ICU data. The LightGBM was performing with the best overall performance (AUROC = 0.877), and XGBoost was close with (AUROC = 0.874). These models also had clinically significant sensitivity and

specificity and yielded fairly calibrated probability estimates. These results are consistent with the literature results, in which the boosted-tree framework consistently beats the neural architecture on structured clinical data because the approach is resistant to missingness and sparse sampling (Feng et al., 2025; Peng et al., 2025).

Deep-learning models including GRU, LSTM, BiLSTM and CNN1D in contrast had low discrimination, and the values of the AUROC are close to 0.60 and bad sensitivity. In the case of deep-learning probability, further instabilities were pointed out by calibration curves. The findings indicate constraints found in the previous studies where RNNs and CNNs do not perform well on sparse and noisy ICU waveforms without sophisticated temporal or attention, systems (Liu et al., 2024; Xu et al., 2024). This system offers an interpretable modelling pipeline that is effective and provides predictors like the ICULOS, heart rate and oxygen saturation to drive model decisions that are prevalent in clinical literature in relation to sepsis onset. Therefore, the research question of explainable time-series ML/DL models to facilitate effective early sepsis recognition was effectively answered. This was evident in machine-learning models, but not in deep-learning models, because of the limitations of data quality and architecture. Nevertheless, the study has its weaknesses. The temporal learning could have been limited by the fixed 6-hours window. Sequence-cleaning approach eliminated a lot of time-series data, which decreased the volume of deep-learning training. The deep-learning architecture of the selected models was lightweight consciously because of the computational constraints and no attention or transformer mechanisms were studied. Also, the problem of imbalance in the classes still affected the temporal models even with weighted loss functions.

# 8 Future Work

Future studies should focus on attention-based or transformer-style architectures that are demonstrated in the literature to be better at capturing irregular ICU sampling patterns (Xu et al., 2024). The deeper learning could be enhanced by introducing longer or adaptive time windows and better treatment of imbalances to learn longer deterioration curves. The use of modern imputation methods, including GRU-D, interpolation-based models or time-series diffusion models, would allow keeping more temporal data and eliminate the reliance on sequence filtering. Hybrid architectures, e.g., CNN-GRU/ CNN-Transformer models, presented in various posted studies should also be explored in meaningful future work to enhance early-warning prediction based on local and global temporal characteristics (Mabrook et al., 2025). Lastly, due to the excellent interpretability results of XGBoost, a new research direction could be to investigate clinically applicable real-time dashboards that incorporate SHAP-based risk explanations of bedside monitoring or early-warning notifications. On the whole, this study shows that feature-based machine-learning models presently provide the best and most explicable technique of early sepsis forecasting on actual data in the ICU. As temporal modelling and data quality improve further, neural networks will be able to compete with deep-learning models in the next generation of sepsis early-warning systems.

# References


Ahmed, M. et al. (2024). *BiLSTM-based fetal health analysis, Fetal Medicine Research Journal.*

Azadeh, A. et al. (2025). *Ensemble classifiers for NHANES disease prediction, Health Informatics Journal.*

Bayrak, U. et al. (2024). *Naïve Bayes and SVM for UCI medical datasets, International Journal of Machine Learning.*

Chen, X. et al. (2023). *XAutoNet: Autoencoder-CNN digital health model, Computer Methods and Programs in Biomedicine.*

Chen, X. et al. (2023). *Autoencoder-CNN (XAutoNet) for ICU sepsis detection, Journal of Medical Systems.*

Choudhury, A., et al. (2023). *Explainable AI for ICU risk scoring, Artificial Intelligence in Medicine.*

Dalwinder, K. et al. (2025). *Neuro-fuzzy diagnostic system for HBV, Expert Systems with Applications.*

Desautels, T. et al. (2016). *Pre-shock sepsis model using RNNs on MIMIC-III, Scientific Reports, 6, 23490.*

Feng, M. et al. (2025). *XGBoost-SHAP sepsis detection in elderly ICU patients, Geriatric Critical Care Journal.*

Feng, M. et al. (2025). *SHAP-enhanced XGBoost for elderly ICU patients, Critical Care Explorations.*

Harrell, F. (2015). *Regression Modeling Strategies. Springer.*

Hu, K. et al. (2024). *CNN-Transformer hybrid for ICU sepsis prediction, IEEE Transactions on Biomedical Engineering.*

Ji-Yuan, L. et al. (2025). *RF-BFE ensemble for ICU sepsis detection, Journal of Clinical Monitoring and Computing.*

Kumar, S. and Arora, R. (2025). *XGBoost-based prediction of Hepatitis B, Diagnostics.*

Kufel, M. et al. (2023). *CNN-BiLSTM model for ECG cardiovascular prediction, Computers in Biology and Medicine.*

Li, X. et al. (2025). *LightGBM and XGBoost comparison for ICU sepsis prediction, PhysioNet Evaluation Journal.*

Liu, S. et al. (2024). *CNN-GRU hybrid for real-time ICU sepsis detection, IEEE Access.*

Mabrook, M. et al. (2025). *BiLSTM trajectory modelling for ICU sepsis, Neural Computing and Applications.*

Mabrook, M. et al. (2025*). Dynamic BiLSTM ICU forecasting model, Machine Learning in Medicine.*

Mao, Q. et al. (2020). *LSTM-based early detection of sepsis using MIMIC, Journal of Biomedical Informatics.*

Mao, Q. et al. (2020). *Time-series LSTM for ICU sepsis onset, Critical Care Medicine.*

Mohamed, A. et al. (2024). *Deep learning architectures for sepsis prediction: a review, ACM Computing Surveys.*

Moor, M. et al. (2023). *Cross-hospital generalisation of sepsis prediction models, Nature Medicine.*

Moor, M. et al. (2023*). Deep and tree-based models for multicentre ICU validation, Journal of Critical Care AI.*
Musa, A. et al. (2025). *Deep 45-layer CNN for clinical detection, Applied Intelligence.*
Obaido, G. et al. (2024). *Clinical explainability using SHAP and LIME, Journal of Healthcare Informatics Research.*
Peng, Z. et al. (2025). *Explainable XGBoost-SHAP model for eICU sepsis, Frontiers in Digital Health.*
Peng, Z. et al. (2025). *XAI clinical mapping using XGBoost, Intensive Care Informatics Journal.*
Reyna, M.A. et al. (2019). *The PhysioNet Sepsis Challenge baseline models, Computing in Cardiology.*
Reyna, M.A. et al. (2019). *Standardised benchmark modelling for sepsis detection, Critical Care Medicine.*
Steyerberg, E. (2016*). Clinical Prediction Models. Springer.*
Xu, S. et al. (2024). *Attention-based LSTM for ICU sepsis detection, IEEE Journal of Biomedical and Health Informatics.*
Xu, S. et al. (2024*). Attention-LSTM explainability for PhysioNet dataset, Scientific Data.*
Zhao, Y. et al. (2023). *XGBoost with SMOTE balancing for ICU sepsis, BMC Medical Informatics and Decision Making.*
Zhao, Y. et al. (2024). *Transformer and BiLSTM comparison on MIMIC-IV, Pattern Recognition in Health.*
Zhang, L. et al. (2023). *Gradient boosting for ICU sepsis severity prediction, Journal of Medical AI Research.*
Zhang, L. et al. (2024). *Tree-based ensemble optimisation for eICU sepsis prediction, Biomedical Signal Processing and Control.*
Dankwa et al. (2025) – *Calibration of transmission-dynamic infectious disease models: A scoping review and reporting framework*.