“Explainable Risk Prediction from ICU Vital Signs: SHAP-Based Interpretation and Independent Cohort Evaluation of an LSTM Model”

by

Kazi Tahmina
ID: CSE2203027072

Md Ekramul Islam
ID: CSE2203027154

Md Amzad
ID: CSE2203027011

Supervised by
Prof. Bulbul Ahamed

Submitted in partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science and Engineering



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SONARGAON UNIVERSITY (SU)
September 2026

APPROVAL

The thesis titled "**Explainable Risk Prediction from ICU Vital Signs: SHAP-Based Interpretation and Independent Cohort Evaluation of an LSTM Model**" submitted by **Kazi Tahmina** (CSE2203027072), **Md Ekramul Islam** (CSE2203027154), and **Md Amzad** (CSE2203027011) to the Department of **Computer Science and Engineering**, Sonargaon University (SU), has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents.

Board of Examiners

Bulbul Ahamed
Professor and Head,
Department of Computer Science and Engineering
Sonargaon University (SU)

Supervisor

(Examiner Name and Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 1

(Examiner Name and Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 2

(Examiner Name and Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 3

DECLARATION

We, hereby, declare that the work presented in this thesis is the outcome of the investigation performed by us under the supervision of Bulbul Ahamed, Professor and Head, Department of Computer Science and Engineering, Sonargaon University, Dhaka, Bangladesh. We reaffirm that no part of this thesis has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

(Prof. Bulbul Ahamed)
Supervisor

Signature

Kazi Tahmina
ID: CSE2203027072

Md Ekramul Islam
ID: CSE2203027154

Md Amzad
ID: CSE2203027011

ABSTRACT

Sepsis is a life-threatening clinical condition associated with a dysregulated response to infection and requires timely recognition to reduce the risk of adverse outcomes. This thesis presents an explainable risk prediction framework using an Improved Long Short-Term Memory (LSTM) model to analyze ICU time-series data from the Prediction of Sepsis dataset associated with the PhysioNet Computing in Cardiology Challenge 2019. The model uses a six-hour observation window comprising 34 clinical features to predict sepsis-related risk within a subsequent one-hour prediction horizon. To enhance model interpretability, SHAP GradientExplainer was applied to investigate global and local feature contributions, temporal importance, and feature–time attribution patterns. The final model was evaluated on an independent, previously unseen cohort without retraining, fine-tuning, threshold tuning, or modification of the original research data. Across 195,018 independent test windows, the model achieved an AUROC of 0.7067, AUPRC of 0.0391, sensitivity of 0.6363, specificity of 0.6895, and F1-score of 0.0698. At the patient level, 357 of 440 positive patients were detected, corresponding to a sensitivity of 0.8114. SHAP analysis identified hematocrit (Hct) as the most influential global feature, while the first temporal position (t=1) was identified as the dominant temporal position. The model generated 61,720 alerts, including 2,278 true-positive and 59,442 false-positive alerts, demonstrating a substantial alert burden. Overall, the findings indicate that the Improved LSTM captures clinically relevant temporal risk patterns while providing interpretable information about influential features and time positions; however, its low precision and high false-alert burden emphasize the need for further validation and optimization before consideration for clinical deployment.

Keywords: Explainable Artificial Intelligence (XAI), Long Short-Term Memory (LSTM), SHAP, ICU Vital Signs, Risk Prediction, Independent Cohort Evaluation.

ACKNOWLEDGMENT

At the very beginning, we would like to express our deepest gratitude to the Almighty Allah for giving us the ability, strength, and perseverance to complete this thesis successfully.

We are deeply grateful to our respected supervisor, Bulbul Ahamed, Professor and Head, Department of Computer Science and Engineering, Sonargaon University, for his valuable guidance, continuous support, constructive suggestions, and thoughtful supervision throughout the research work. His encouragement and direction played an essential role in completing this thesis.

We would also like to express our sincere gratitude to all the respected teachers of the Department of Computer Science and Engineering, Sonargaon University, for their valuable teachings, knowledge, and guidance throughout our academic journey. Their contributions have helped us develop the knowledge and skills necessary to undertake this research.

Finally, we would like to express our deepest gratitude and love to our parents and family members for their unconditional love, continuous support, encouragement, patience, and prayers throughout our academic life and during the completion of this thesis. Their constant support has been a great source of motivation and strength for us.

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AI	Artificial Intelligence
AUPRC	Area Under the Precision–Recall Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BAcc	Balanced Accuracy
BUN	Blood Urea Nitrogen
CSE	Computer Science and Engineering
DL	Deep Learning
FN	False Negative
FiO_2	Fraction of Inspired Oxygen
FP	False Positive
GPU	Graphics Processing Unit
Hct	Hematocrit
HR	Heart Rate
ICU	Intensive Care Unit
ICULOS	Intensive Care Unit Length of Stay
LSTM	Long Short-Term Memory
MAP	Mean Arterial Pressure
MCC	Matthews Correlation Coefficient
ML	Machine Learning
O_2Sat	Oxygen Saturation
PR	Precision–Recall
PPV	Positive Predictive Value

ROC	Receiver Operating Characteristic
RR	Respiratory Rate
SaO_2	Arterial Oxygen Saturation
SBP	Systolic Blood Pressure
SHAP	SHapley Additive exPlanations
TN	True Negative
TP	True Positive
XAI	Explainable Artificial Intelligence

Clinical and Physiological Variables

Abbreviation	Full Form
BUN	Blood Urea Nitrogen
Ca	Calcium
FiO_2	Fraction of Inspired Oxygen
Hct	Hematocrit
HR	Heart Rate
MAP	Mean Arterial Pressure
O_2Sat	Oxygen Saturation
pH	Potential of Hydrogen
Resp	Respiratory Rate
SaO_2	Arterial Oxygen Saturation
SBP	Systolic Blood Pressure
Temp	Body Temperature

TABLE OF CONTENTS

Title		Page No.
DECLARATION …….……..		iii
ABSTRACT ……...…….…		iv
ACKNOWLEDGEMENT …….……		v
LIST OF ABBREVIATIONS …….……..…		vi
CHAPTER 1		
INTRODUCTION TO EXPLAINABLE ICU RISK PREDICTION		
1.1	Introduction ……..……...…	1-2
1.2	Background of the Study……...……...……...……	2-3
1.3	Problem Statement…….…….	3
1.4	Motivation of the Study…….……..	4
1.5	Research Aim and Objectives…….	4-5
1.6	Research Questions…….…….	5-6
1.7	Scope of the Study…….……...…	6
1.8	Contributions of the Study……..…	6-7
1.9	Organization of the Thesis……..…	7-8
CHAPTER 2		
LITERATURE REVIEW AND RESEARCH GAP		
2.1	Introduction …….…….....…	9
2.2	Sepsis Prediction from Clinical Data……...	9-10
2.3	Machine Learning Approaches for Sepsis Prediction……	10
2.4	Deep Learning for Clinical Time-Series Prediction……	11
2.5	LSTM-Based Temporal Modeling……..……	11-12
2.6	Temporal Representation of ICU Physiological Data……..	12
2.7	Explainable Artificial Intelligence in Healthcare...……...……..	12-13

2.8 SHAP-Based Model Interpretation…… 13
2.9 Global, Local, and Temporal Explainability…… 13-14
2.10 Independent Cohort Evaluation and Research Gap……..... 14-15
2.11 Summary ……..……....…… 15-16

CHAPTER 3

RESEARCH METHODOLOGY AND EXPERIMENTAL DESIGN

3.1 Introduction……..……....…… 17
3.2 Research Workflow……..……. 17-18
3.3 Clinical Problem Formulation……. 18-19
3.4 Research Questions and Hypothesis…… 19-20
3.5 Dataset Description……..……...… 20-21
3.6 ICU Patient Time-Series Representation……... 21
3.7 Data Engineering and Sequence Construction…… 21-22
3.8 Observation Window and Prediction Horizon……... 22-23
3.9 Experimental Design……...……...……...… 23-24
3.10 Model Development Strategy……...……...…… 24-25
3.11 Explainability Framework……...……...……...……...… 25
3.12 Scientific Evaluation and Clinical Validation Strategy…… 25-26
3.13 Summary 26

CHAPTER 4

DATA ENGINEERING AND MODEL DEVELOPMENT

4.1 Introduction ……..……....…… 27
4.2 Dataset Preparation……..…… 27
4.3 Data Quality and Feature Handling ……...……...……...…….. 28
4.4 Feature Set and Clinical Variables……..…….. 29-32
4.5 Temporal Sequence Construction ……..... 32
4.6 Training-Fitted Preprocessing……. 32-33
4.7 Improved LSTM Architecture……. 33
4.8 Model Input and Output Representation…….. 33-34
4.9 Model Development and Training……...… 34

4.10 Model Selection and Model Freezing…… 35
4.11 Model Handoff and Reproducibility…… 35-37
4.12 Summary…….……....……... 38

CHAPTER 5

EXPLAINABILITY FRAMEWORK

5.1 Introduction …….……...…… 39
5.2 Explainable AI Framework…….… 39-40
5.3 SHAP Methodology…….……... 40
5.4 SHAP GradientExplainer……...…… 40-41
5.5 Global Feature Importance…….… 41-43
5.6 Local Feature Attribution…….…... 43
5.7 Temporal Importance Analysis……... 43-44
5.8 Feature–Time Attribution Analysis……... 44-45
5.9 Dominant Feature and Temporal Position Analysis…… 45
5.10 Explainability Interpretation……..... 45-46
5.11 Summary…….……....……... 46

CHAPTER 6

SCIENTIFIC EVALUATION AND CLINICAL VALIDATION

6.1 Introduction ……..….. 47
6.2 Independent Test Cohort…….... 47-48
6.3 Evaluation Metrics…….…….. 48-49
6.4 Confusion Matrix Analysis…….… 50
6.5 ROC Curve Analysis…….……...… 50-51
6.6 Precision–Recall Curve Analysis...……...……...……...……... 51-52
6.7 Validation-to-Independent-Test Performance Comparison…… 52
6.8 Window-Level Performance……..... 53
6.9 Patient-Level Clinical Performance…… 53-54
6.10 Risk Probability Band Analysis……... 54
6.11 Clinical Alert Outcome Analysis…… 54-55
6.12 Alert Burden Analysis…….……... 55-56
6.13 Summary…….……....……... 56

CHAPTER 7

RESULTS AND DISCUSSION

7.1	Introduction…….……....……	57
7.2	Final Independent-Test Results……	57-58
7.3	Model Discrimination Performance……	58
7.4	Patient-Level Detection Performance……..	58-59
7.5	Alert Burden and False-Alert Analysis……	59-60
7.6	Global SHAP Findings…….…...…	60-61
7.7	Temporal SHAP Findings...……...……...……...……...……...	62
7.8	Feature–Time Attribution Findings……	62-63
7.9	Relationship Between Window-Level and Patient-Level Performance…….……....…	63
7.10	Validation vs. Independent-Test Findings……	63-64
7.11	Overall Research Findings	64
7.12	Research Safety and Reproducibility……...…	65
7.13	Discussion Summary…….……..	66

CHAPTER 8

CONCLUSION AND FUTURE WORK

8.1	Conclusion…….……....…	67-68
8.2	Limitations ……..…….……	68
8.3	Future Work…….……...…	69-70

REFERENCES …….…….. 71

LIST OF TABLES

Table No.	Title	Page No.
Table 3.1	Dataset Summary:	20
Table 3.2	Temporal Configuration:	23
Table 4.1	Complete 34-Feature Clinical Variable Set	30-31
Table 4.2	Final Model Handoff and Evaluation Artifact Inventory	36
Table 5.1	Global SHAP Feature Importance Ranking	42
Table 6.1	Evaluation Metrics and Their Definitions	49
Table 6.2	Validation-to-Independent-Test Performance Comparison	52
Table 6.3	Alert Burden Analysis	55
Table 7.1	Final Independent-Test Results	57
Table 7.2	Final Discrimination Metrics	58
Table 7.3	Clinical Alert Outcome Summary	59
Table 7.4	Validation performance:	64
Table 7.5	Independent-test performance:	64

LIST OF FIGURES

Figure No.	Title	Page No.
Fig.4.1	Dataset Preparation and Experimental Data Flow	27
Fig.4.2	Data Quality and Feature Handling Workflow	28
Fig.4.3	Six-Step ICU Temporal Sequence Construction	32
Fig.4.4	Model Handoff and Reproducible Evaluation Pipeline	37
Fig.5.1	Improved LSTM Input Structure and SHAP Attribution Mapping	41
Fig. 5.2	Global SHAP Feature Importance	41
Fig.5.3	Temporal SHAP Importance Across the Six Time Steps	44
Fig. 5.4	Feature–Time SHAP Attribution Heatmap	44
Fig.5.5	Dominant Feature and Temporal Position Attribution	45
Fig.5.6	Integrated SHAP Explainability Analysis	46
Fig.6.1	Independent Test Evaluation Pipeline	48
Fig.6.2	Independent-Test Confusion Matrix	50
Fig.6.3	Receiver Operating Characteristic (ROC) Curve for the Independent Test Cohort	51
Fig.6.4	Precision–Recall Curve for the Independent Test Cohort	52
Fig.6.5	Window-Level Classification Performance	53
Fig.6.6	Patient-Level Prediction Outcome Analysis	54
Fig.6.7:	Distribution of Predictions Across Risk Probability Bands	54
Fig.6.8	Clinical Alert Outcome Distribution	55
Fig.6.9	Alert Burden Analysis	56
Fig.7.1	Alert Burden Analysis	59
Fig. 7.2	Global SHAP Feature Importance	60
Fig.7.3	Top Global SHAP Feature Contributions	61

Fig.7.4	Temporal SHAP Importance Across Six Temporal Positions	62
Fig.7.5	Window-Level and Patient-Level Performance Comparison	63
Fig.7.6	Validation-to-Independent-Test Performance Comparison	63
Fig.7.7	Visual Testing Demo – Independent Model Testing Interface	65

CHAPTER 1
INTRODUCTION TO EXPLAINABLE ICU RISK PREDICTION

1.1 Introduction

Sepsis is a life-threatening clinical condition that occurs when the body's response to infection becomes dysregulated and leads to organ dysfunction. Because its progression can be rapid, early identification of patients at elevated risk remains an important challenge in intensive care. Continuous monitoring in the Intensive Care Unit (ICU) generates physiological and laboratory measurements that provide time-dependent information about a patient's clinical condition. Such data create an opportunity to investigate computational approaches for identifying patterns associated with sepsis risk [1].

Machine learning and deep learning have increasingly been investigated in healthcare because of their ability to model complex relationships within clinical data. Their application to emergency and critical care settings has also received growing attention, particularly for supporting risk assessment and clinical decision-making [3], [4]. In ICU settings, where measurements are collected repeatedly over time, sequential models such as Long Short-Term Memory (LSTM) networks provide a suitable framework for analyzing temporal patterns within patient observations.

Predictive performance, however, does not by itself explain why a model produces a particular prediction. This issue is especially important in healthcare, where understanding model behavior can support the assessment of reliability, interpretability, and practical relevance [3], [5]. Explainable Artificial Intelligence (XAI) provides methods for examining the factors associated with model predictions and making complex machine learning outputs more understandable.

SHapley Additive exPlanations (SHAP) is an established explainability method based on Shapley values from cooperative game theory. It assigns contribution values to input features and can be used to examine both overall feature importance and individual predictions [2]. Another widely known approach is Local Interpretable Model-agnostic Explanations (LIME), which provides local explanations for individual model predictions [6].

This thesis investigates an Improved Long Short-Term Memory (LSTM) model for sepsis risk prediction using ICU time-series data from the Prediction of Sepsis dataset associated with the PhysioNet/Computing in Cardiology Challenge 2019 [1]. The study combines predictive modeling with SHAP-based explainability to examine the contribution of clinical features and their temporal positions within the observation sequence.

The model is evaluated on an independent, previously unseen patient cohort. The evaluation considers prediction performance at both the window and patient levels, together with prediction-generated alerts and their associated burden. This provides a broader assessment of the model by considering predictive performance, interpretability, independent generalization, and practical implications.

The research follows a structured twelve-level workflow covering domain understanding, clinical problem formulation, literature intelligence, research-gap identification, hypothesis construction, experimental design, data engineering, model development, explainability analysis, scientific evaluation, clinical validation, and final result interpretation.

1.2 Background of the Study

Continuous monitoring in the Intensive Care Unit (ICU) produces physiological and laboratory measurements that reflect changes in a patient's clinical condition. In sepsis research, these time-dependent observations provide an opportunity to investigate patterns that may be associated with changing risk over time [1].

The analysis of ICU time-series data presents several challenges. Clinical measurements may be incomplete, irregular, noisy, and heterogeneous across patients. In addition, the relevance of an individual measurement may depend on observations recorded at earlier time points. Therefore, sepsis risk prediction from ICU data requires consideration of both the clinical variables and their temporal relationships.

Deep learning provides a framework for modeling complex patterns in sequential data. LSTM networks are designed to retain relevant information across successive time steps, making them appropriate for sequence-based prediction where patient conditions evolve over time. In this study, the Improved LSTM model processes a six-hour observation sequence containing 34 clinical features and generates a prediction for the defined one-hour prediction horizon.

Interpretability is an important consideration when applying predictive models to healthcare data. A model may identify useful patterns while providing limited information about the variables or temporal observations that influenced its output. Recent work has highlighted the importance of interpretable machine learning and machine learning-assisted approaches in sepsis risk assessment and clinical decision support [4], [5].

For this reason, explainability is incorporated directly into the evaluation framework of this study. SHAP GradientExplainer is used to investigate global feature importance, local feature attribution, feature-time contributions, and temporal importance. The SHAP framework provides a principled approach for examining how individual features contribute to model predictions [2].

The research also emphasizes independent evaluation. Performance observed during model development may differ when a frozen model is applied to previously unseen patients. Therefore, the independent test cohort is kept separate from model development and is used only for final evaluation. This evaluation examines model discrimination, window-level performance, patient-level detection, risk prediction behavior, and alert burden.

Overall, the study combines temporal modeling, explainability, and independent evaluation within a single research framework. This provides a structured basis for examining not only the predictive behavior of the Improved LSTM but also the clinical features and temporal positions associated with its predictions.

1.3 Problem Statement

Sepsis risk prediction from ICU time-series data involves several interconnected challenges. First, clinical measurements are sequential, and their predictive relevance may depend on both their values and their temporal positions. A suitable predictive model must therefore account for temporal relationships within the observation window.

Second, sepsis-related prediction is a highly imbalanced classification problem. In the independent test set, 195,018 prediction windows were evaluated, including 3,580 positive and 191,438 negative windows. The positive prevalence was only 1.8357%. Under such class imbalance, accuracy alone does not provide a sufficient description of model performance. Therefore, multiple complementary metrics are considered, including precision, recall, specificity, F1-score, AUROC, and AUPRC.

Third, deep learning predictions can be difficult to interpret. A high-risk prediction does not directly indicate which clinical variables contributed to the result or which observations within the sequence were influential. Without an appropriate explainability method, these aspects of model behavior remain difficult to examine.

Fourth, model performance during validation may not fully represent performance on previously unseen patients. Independent evaluation is therefore required to assess how a frozen model behaves beyond the data used during development.

Finally, prediction performance must be considered together with alert generation. A model may detect a substantial proportion of positive cases while simultaneously producing a large number of false-positive alerts. A high alert burden may reduce the practical usefulness of a real-time monitoring signal. Consequently, this study evaluates alert burden alongside predictive performance.

The central problem addressed by this thesis is the evaluation of an Improved LSTM model for ICU sepsis risk prediction while examining its feature-level and temporal explainability, independent generalization, patient-level detection, and alert burden.

1.4 Motivation of the Study

The primary motivation of this research is to evaluate sepsis risk prediction beyond a single predictive performance measure. A model that produces predictions without providing insight into the information behind those predictions offers limited understanding of its learned behavior.

The temporal nature of ICU data provides a second motivation. Patient measurements collected over several consecutive hours may contain patterns that cannot be adequately represented by treating observations independently. An LSTM-based approach provides a mechanism for modeling these sequential relationships.

Independent evaluation is another important motivation. Validation performance does not necessarily represent how a model behaves on previously unseen patients. This study therefore evaluates the frozen Improved LSTM on an independent test cohort without retraining, fine-tuning, threshold tuning, or modification of the established research data.

Explainability provides an additional motivation. SHAP GradientExplainer is used to examine global feature importance, local feature attribution, feature-time contributions, and temporal positions. This enables the study to investigate not only model performance but also the information associated with its predictions [2].

Finally, the research considers the practical implications of model-generated alerts. The independent evaluation produced 61,720 alerts, including 2,278 true-positive alerts and 59,442 false-positive alerts. Examining this alert burden is therefore necessary for providing a balanced interpretation of the model's predictive behavior.

1.5 Aim and Objectives of the Study

The main aim of this thesis is to investigate an Improved LSTM model for sepsis risk prediction from ICU time-series data and to provide an explainable and independently evaluated assessment of its predictions.

To achieve this aim, the study has the following objectives:

1. To investigate the use of an Improved LSTM model for predicting sepsis-related risk from sequential ICU physiological and clinical measurements.

2. To construct six-hour temporal observation sequences containing 34 input features for the defined one-hour prediction horizon.

3. To evaluate the frozen Improved LSTM on an independent, previously unseen patient cohort without retraining, fine-tuning, threshold tuning, or modification of the established research data.

4. To evaluate model performance using complementary metrics, including accuracy, precision, recall/sensitivity, specificity, F1-score, balanced accuracy, Matthews Correlation Coefficient, AUROC, and AUPRC.

5. To investigate model explainability using SHAP GradientExplainer.

6. To identify the most influential global features contributing to model predictions.

7. To examine local feature contributions and identify dominant temporal positions within the observation sequence.

8. To compare validation and independent-test performance to assess the model's generalization behavior.

9. To evaluate model predictions at both the prediction-window level and patient level.

10. To quantify model-generated alert burden, including true-positive alerts, false-positive alerts, alert rate, false-alert ratio, and alerts per patient.

11. To provide an integrated interpretation of predictive performance, explainability, independent evaluation, patient-level detection, and alert burden.

1.6 Research Questions

The research is guided by the following questions:

RQ1: How can SHAP-based explainability, including global and local analysis, be used to interpret the predictions of the Improved LSTM model?

RQ2: Which vital-sign and clinical features, and which temporal positions within the observation sequence, contribute most strongly to the model's predictions?

RQ3: How does the model discriminate between positive and negative outcomes on an independent, previously unseen patient cohort at both the window level and patient level?

RQ4: What are the practical alert-burden implications of the model's predictions when considered as a real-time monitoring signal?

These questions establish a direct connection between predictive modeling, explainability, independent evaluation, patient-level detection, and practical interpretation.

1.7 Scope of the Study

This study focuses on explainable sepsis risk prediction using ICU time-series data from the Prediction of Sepsis dataset associated with the PhysioNet/Computing in Cardiology Challenge 2019 [1]. The research investigates the established Improved LSTM model and its behavior on sequential patient observations.

The model uses a six-hour observation window containing 34 clinical features followed by a one-hour prediction horizon. Evaluation is performed at both the prediction-window and patient levels. Window-level analysis examines the classification behavior of individual prediction windows, while patient-level analysis determines whether positive patients are detected according to the established evaluation procedure.

The explainability component is limited to SHAP-based interpretation using SHAP GradientExplainer. It covers global feature importance, local feature attribution, feature-time attribution, and temporal-position analysis. These analyses are intended to describe model behavior and do not establish causal relationships between clinical variables and sepsis outcomes [2].

The final scientific evaluation uses an independent test cohort that was not involved in model selection. The frozen model is evaluated without additional training, retraining, fine-tuning, threshold tuning, or test-cohort-based model selection.

The study also examines model-generated alerts to quantify false-positive burden and alert frequency per patient. This provides a research-level assessment of the operational implications of the prediction signal.

The scope does not include prospective clinical deployment, clinical intervention, real-world clinical decision-making, or a claim that the model is ready for patient care. The findings are therefore interpreted as research results obtained from retrospective data and an independent evaluation cohort.

1.8 Contributions of the Study

This thesis makes several contributions to the study of explainable ICU sepsis risk prediction.

First, it presents an integrated research framework combining ICU time-series data engineering, Improved LSTM prediction, SHAP-based explainability, independent evaluation, patient-level detection, and alert-burden analysis.

Second, the study evaluates the frozen Improved LSTM on an independent test cohort containing 195,018 prediction windows from 6,051 patients. This provides an assessment of model behavior on previously unseen patient data.

Third, the research evaluates model performance using multiple complementary measures rather than relying on accuracy alone. This is particularly important because of the low prevalence of positive prediction windows in the independent test cohort.

Fourth, the study integrates SHAP-based explainability into the model evaluation process. The analysis identifies Hct as the most influential global feature and the first temporal position (t=1) as the dominant temporal position in the available explainability analysis [2].

Fifth, the research extends evaluation beyond individual prediction windows by examining patient-level detection. The final analysis detected 357 of 440 positive patients, corresponding to a patient-level sensitivity of 0.8114.

Sixth, the study explicitly quantifies alert burden. The independent evaluation produced 61,720 alerts, including 2,278 true-positive alerts and 59,442 false-positive alerts. This provides an additional perspective on the practical implications of the model's predictions.

Finally, the study maintains a clear separation between model development and final evaluation. The independent-test results were obtained using the frozen model without retraining, fine-tuning, threshold tuning, or test-cohort-based model selection. This separation strengthens the methodological transparency of the final evaluation.

1.9 Organization of the Thesis

This thesis is organized into eight chapters.

Chapter 1 introduces the research problem, background, motivation, aim and objectives, research questions, scope, contributions, and overall organization of the thesis.

Chapter 2 presents the literature review and research-gap analysis. It examines relevant work on sepsis prediction, ICU time-series analysis, machine learning, deep learning, LSTM-based prediction, Explainable Artificial Intelligence, SHAP-based interpretation, independent evaluation, and alert burden.

Chapter 3 describes the research methodology and experimental design, including the research workflow, clinical problem formulation, dataset, temporal representation, observation window, prediction horizon, experimental design, explainability strategy, and evaluation framework.

Chapter 4 presents the data engineering and model development process, including dataset preparation, feature handling, temporal sequence construction, training-fitted preprocessing, Improved LSTM development, model selection, model freezing, and research artifact handoff.

Chapter 5 presents the explainability framework and describes the application of SHAP GradientExplainer for global feature importance, local attribution, temporal importance, feature-time attribution, and dominant feature and temporal-position analysis.

Chapter 6 presents the scientific evaluation and clinical validation procedures, including independent-test evaluation, performance metrics, confusion matrix analysis, ROC and precision–recall analysis, validation-to-test comparison, window-level and patient-level evaluation, risk probability analysis, clinical alert outcomes, and alert burden.

Chapter 7 presents the main research findings and discussion. It interprets the independent-test performance, patient-level detection, alert burden, SHAP findings, temporal attribution, validation-to-test differences, and the relationship between window-level and patient-level results.

Chapter 8 concludes the thesis by summarizing the major findings and contributions, discussing the study limitations, and identifying directions for future research.

The thesis concludes with the References section containing the sources cited throughout the study, followed by supporting appendices where applicable.

CHAPTER 2
LITERATURE REVIEW AND RESEARCH GAP

2.1 Introduction

Sepsis prediction has become an important research area in clinical artificial intelligence because the physiological condition of critically ill patients can change rapidly. Intensive Care Unit (ICU) data contain repeated measurements of vital signs and laboratory variables that can provide useful information about changes in a patient's physiological state.

Traditional clinical prediction methods often rely on predefined scores or statistical models. Although these approaches remain useful, they may not fully represent nonlinear relationships and temporal changes in physiological data. Machine learning and deep learning methods provide opportunities to learn such patterns directly from clinical observations.

The use of artificial intelligence in emergency and critical-care settings has also increased the importance of reliable and interpretable prediction systems [3]. Recent research on machine-learning-assisted sepsis triage further demonstrates the potential of computational models to support early identification of high-risk patients [4].

However, predictive performance alone is not sufficient for clinical decision-support research. Complex models may produce useful predictions without clearly showing which clinical variables contributed to those predictions. Explainable Artificial Intelligence (XAI) can address this limitation by providing interpretable information about model behavior.

This chapter reviews the literature and methodological foundations most closely related to the present study, focusing on sepsis prediction, clinical time-series modeling, LSTM-based prediction, explainable AI, SHAP-based interpretation, and independent evaluation. The chapter concludes by identifying the research gap addressed by the proposed framework.

2.2 Sepsis Prediction from Clinical Data

Early sepsis prediction requires the analysis of clinical observations available before the onset of sepsis. Vital signs, laboratory measurements, and other physiological variables may provide information about changes in a patient's condition over time.

The PhysioNet/Computing in Cardiology Challenge 2019 was developed specifically to investigate early sepsis prediction from clinical data. The dataset contains longitudinal patient observations with hourly measurements, physiological variables, laboratory measurements,

demographic information, hospital-related variables, ICU length of stay, and a time-dependent SepsisLabel [1].

The present study uses the publicly available Kaggle distribution of this dataset, while the original PhysioNet/Computing in Cardiology Challenge documentation is treated as the authoritative source for dataset definition and provenance [1].

Sepsis prediction can be formulated as a binary classification problem in which observations available before a prediction point are used to estimate the risk of a subsequent sepsis outcome. This temporal formulation is important because information occurring after the prediction point must not be used as model input. Maintaining this separation reduces the possibility of temporal information leakage and preserves the intended early-prediction setting.

The present research focuses on structured physiological and laboratory time-series data rather than clinical text or multimodal information. This provides a controlled setting for investigating temporal prediction and model explainability.

2.3 Machine Learning Approaches for Sepsis Prediction

Machine learning methods have been widely investigated for clinical risk prediction. Logistic Regression provides a relatively interpretable statistical approach, while tree-based methods such as Random Forest and XGBoost can capture nonlinear relationships between clinical variables.

Recent work on machine-learning-assisted sepsis triage indicates that predictive models can potentially support the identification and prioritization of patients at elevated risk [4]. However, different algorithms represent clinical information differently.

A static representation may summarize patient observations using statistical characteristics such as mean, minimum, maximum, standard deviation, or recent measurements. Although such representations are useful, they can remove information about the order in which physiological changes occur.

In contrast, a temporal representation preserves the sequence of observations. This allows a model to consider changes, persistence, and interactions across successive time points. The present study therefore considers temporal modeling alongside conventional machine-learning approaches to examine the value of preserving sequential clinical information.

2.4 Deep Learning for Clinical Time-Series Prediction

Deep learning methods can learn complex nonlinear relationships and internal representations from clinical data. This is particularly relevant to ICU time-series prediction, where the meaning of a measurement may depend on previous observations.

For example, an isolated abnormal measurement and a continuously worsening physiological trend may represent different clinical patterns even when their current values are similar. A temporal model can use information from multiple observations to distinguish such patterns.

Recurrent neural networks are designed to process sequential information and maintain information from previous time steps. This makes them suitable for clinical datasets in which patient observations are recorded repeatedly over time.

The present research uses an Improved Long Short-Term Memory (LSTM) model to preserve temporal information during prediction. The final model receives a six-hour sequence containing 34 clinical features at each temporal position.

The use of a temporal model also introduces an important interpretability question: not only which clinical features influence a prediction, but also at which temporal positions their contributions become important.

2.5 LSTM-Based Temporal Modeling

Long Short-Term Memory (LSTM) is a recurrent neural network architecture designed to model sequential dependencies through gated mechanisms. These mechanisms allow the network to retain or update information as observations are processed over time.

LSTM-based architectures are therefore relevant to ICU prediction because physiological states are not independent across consecutive observations. Patterns such as persistent abnormalities, gradual deterioration, or short-term changes may be represented more effectively when the temporal order of observations is retained.

The research design initially considered recurrent architectures including Bidirectional LSTM (Bi-LSTM). However, bidirectional processing can access information from both directions within a supplied sequence and therefore should not automatically be interpreted as strict real-time causal prediction.

In the completed research, the final predictive system uses an Improved LSTM model. The frozen model accepts an input sequence with six temporal positions and 34 features per position. This representation is also used as the basis for the subsequent SHAP-based explainability analysis.

Thus, temporal modeling in this study serves two related purposes: predicting sepsis-related risk and providing a structure through which feature contributions can be examined across time.

2.6 Temporal Representation of ICU Physiological Data

ICU physiological data are naturally temporal because measurements are collected repeatedly during a patient's stay. The sequence in which measurements occur may contain information that is lost when observations are converted into a single static feature vector.

The present study distinguishes between an observation window and a prediction horizon. The observation window represents the historical clinical information supplied to the model, while the prediction horizon represents the future period for which sepsis-related risk is predicted.

The final experimental design uses a six-hour observation window and a one-hour prediction horizon. Each input sequence contains six temporal positions and 34 clinical features.

This representation allows the model to examine changes in variables such as heart rate, oxygen saturation, respiratory rate, blood pressure, temperature, and laboratory measurements across consecutive observations.

Temporal representation is also important for explainability. A conventional feature-importance ranking can show that a feature is influential but cannot indicate when within the observation window that contribution occurred. Therefore, temporal position is treated as an additional dimension of model interpretation in the present study.

2.7 Explainable Artificial Intelligence in Healthcare

The increasing use of artificial intelligence in healthcare has created a need for prediction systems that can provide understandable information about their outputs. This issue is particularly important in emergency and critical-care environments, where model predictions may influence the prioritization of patients [3].

Explainable Artificial Intelligence refers to methods that provide information about how an AI model produces or supports its predictions. In clinical research, explanations can help identify influential variables and provide additional insight into model behavior.

Recent research on interpretable machine learning for sepsis risk emphasizes the value of understanding the clinical variables associated with model predictions [5]. This supports the

integration of explainability with predictive modeling rather than treating interpretation as an entirely separate task.

Explainability can be considered at different levels. Global explanation describes the general behavior of a model across a dataset, while local explanation focuses on an individual prediction. For temporal models, interpretation may also need to identify how contributions vary across different time positions.

The present research therefore considers global, local, and temporal explanation as complementary components of the prediction framework.

2.8 SHAP-Based Model Interpretation

SHapley Additive exPlanations (SHAP) is a post-hoc explainability framework based on Shapley values from cooperative game theory. It assigns contribution values to input features to describe their influence on a model output [2].

Lundberg and Lee introduced SHAP as a unified framework for interpreting machine-learning predictions [2]. One important advantage of SHAP is its ability to support both global and local interpretation.

Global feature importance can be obtained by aggregating the magnitude of SHAP values across observations. Local explanations, in contrast, show how individual features contribute to a particular prediction. These two perspectives provide complementary information about model behavior.

The present study applies SHAP GradientExplainer to the frozen Improved LSTM model. The analysis investigates global feature importance, local contributions, temporal importance, and feature-time attribution.

SHAP is therefore used in this research as an analytical method rather than simply as a visualization technique. The objective is to investigate the relationship between clinical input variables, their temporal positions, and the model's predicted risk.

2.9 Global, Local, and Temporal Explainability

Global and local explanations provide different perspectives on a predictive model. Global importance identifies features that generally contribute to model behavior, whereas local explanation examines the contribution of features to an individual prediction.

The completed global SHAP analysis identified hematocrit (Hct) as the most influential feature according to mean absolute SHAP contribution. Other highly ranked features included

Glucose, Resp, HR, SBP, and O2Sat. These findings describe the behavior observed in the analyzed research data and should not be interpreted as universal clinical risk rankings.

The temporal SHAP analysis identified the second position in the six-step sequence as the dominant temporal position. This position corresponds to approximately four hours before the prediction point. Hct was also identified as the dominant feature in the analyzed positive-risk cases.

These findings demonstrate why feature importance alone is insufficient for a temporal prediction model. A feature may be important overall while contributing differently at different points in a sequence.

The present research therefore extends feature-level explanation into a feature-time perspective:

Clinical Feature + Temporal Position
↓
Model Contribution
↓
Predicted Risk

This approach allows the study to investigate not only which variables are influential but also when their contributions become important within the observation window.

2.10 Independent Cohort Evaluation and Research Gap

Evaluation on previously unseen patients is important for assessing whether a clinical prediction model generalizes beyond the data used during development. A model that performs well on development data may still experience substantial performance changes when evaluated on an independent cohort.

The present research therefore includes an independent test cohort that was not used for model training, preprocessing fitting, explainability background construction, or model selection. The frozen Improved LSTM was evaluated on this previously unseen patient cohort without retraining, fine-tuning, threshold tuning, or modification of the research data.

The evaluation considers both window-level and patient-level behavior. Window-level analysis examines individual temporal sequences, whereas patient-level analysis recognizes that multiple prediction windows may belong to the same patient.

This distinction is important in early-warning prediction because repeated predictions for the same patient may produce a large number of alerts. Consequently, alert burden is considered alongside conventional predictive performance.

The literature and completed research identify several important gaps:

1. Predictive performance alone does not adequately explain why a complex clinical model produces a particular prediction.

2. Global feature importance does not indicate the temporal position at which a feature contributes to a sequence-based prediction.

3. Local explanation is necessary to understand individual model predictions and cannot be replaced by a global ranking alone.

4. Independent patient-level evaluation is necessary to examine model behavior on previously unseen patients.

5. Repeated predictions and false alerts can create a substantial alert burden and should therefore be considered when evaluating a model intended for monitoring support.

6. Existing research on interpretable sepsis prediction demonstrates the importance of feature-level interpretation [5], while methods such as LIME provide another widely used approach for local model explanation [6]. However, the present study focuses specifically on SHAP-based interpretation of a temporal LSTM model and extends the analysis to temporal positions and feature-time contributions.

Based on these gaps, the present thesis integrates five closely related components:

1. Temporal sepsis risk prediction using an Improved LSTM model.
2. Global and local SHAP-based explanation.
3. Temporal analysis of model contributions.
4. Feature-time attribution analysis.
5. Independent cohort, patient-level, and alert-burden evaluation.

The study is deliberately focused on experimental investigation rather than direct clinical deployment. Its purpose is to determine whether temporal physiological information can support sepsis-related risk prediction and whether explainability methods can reveal the features and temporal positions associated with model predictions.

2.11 Summary

This chapter reviewed the literature and methodological foundations relevant to the present research. The review focused on sepsis prediction, machine learning, deep learning, LSTM-based temporal modeling, ICU physiological time-series representation, Explainable Artificial Intelligence, SHAP, local explanation, and independent model evaluation.

The PhysioNet/Computing in Cardiology Challenge 2019 provides the primary dataset context for early sepsis prediction using longitudinal clinical observations [1]. Research on AI in emergency medicine and machine-learning-assisted sepsis triage further demonstrates the growing role of computational prediction in critical-care decision support [3], [4].

LSTM-based temporal modeling provides a suitable framework for preserving sequential information in ICU observations. In the present study, the Improved LSTM operates on six temporal positions with 34 clinical features per position.

Explainability is an important complementary component of clinical prediction. SHAP provides a unified framework for estimating feature contributions at global and local levels [2], while LIME represents another established approach for local model explanation [6]. Research on interpretable sepsis risk further supports the importance of understanding the variables associated with model predictions [5].

The completed research extends conventional feature interpretation by examining temporal importance and feature-time attribution. It also evaluates the frozen model on an independent patient cohort and considers patient-level detection and alert burden in addition to standard discrimination metrics.

The central research gap identified in this review is therefore the need to understand not only whether a temporal model can predict sepsis-related risk, but also which clinical features contribute to its predictions, when those contributions occur, and how the model behaves on previously unseen patients.

The next chapter presents the research methodology, experimental design, dataset preparation, temporal sequence construction, Improved LSTM development, explainability framework, and evaluation strategy used to address this research gap.

CHAPTER 3
RESEARCH METHODOLOGY AND EXPERIMENTAL DESIGN

3.1 Introduction

This chapter presents the methodology and experimental design used to develop, interpret, and evaluate the sepsis risk prediction model investigated in this research. The study focuses on ICU physiological time-series data and uses an Improved Long Short-Term Memory (LSTM) model to predict sepsis-related risk.

The methodology combines temporal data representation, model development, SHapley Additive exPlanations (SHAP), independent cohort evaluation, patient-level analysis, and alert-burden assessment. The overall design was structured to maintain a clear separation between model development, explainability, and final independent evaluation.

The research used the Prediction of Sepsis dataset associated with the PhysioNet/Computing in Cardiology Challenge 2019 [1]. The final model operates on six consecutive temporal observations containing 34 clinical features at each time step. The completed model was frozen before independent evaluation, and the independent test cohort was not used for retraining, fine-tuning, threshold tuning, or model modification.

3.2 Research Workflow

The research followed a twelve-level workflow that connects the initial clinical problem with the final interpretation of experimental results.

Level 1 – Domain Understanding
The research began with an understanding of sepsis, ICU monitoring, physiological deterioration, and the role of artificial intelligence in clinical risk prediction.

Level 2 – Clinical Problem Formulation
The clinical problem was formulated as early risk prediction using routinely recorded ICU physiological and laboratory observations.

Level 3 – Literature Intelligence
Relevant literature was reviewed to understand existing approaches to sepsis prediction, temporal modeling, explainable AI, and clinical decision-support systems [3]–[5].

Level 4 – Research Gap Discovery
The literature was examined to identify gaps related to model interpretability, temporal feature contribution, independent evaluation, and practical alert behavior.

Level 5 – Hypothesis Construction
Research questions and corresponding hypotheses were established based on the identified research gaps.

Level 6 – Experimental Design
The data-processing pipeline, temporal configuration, model evaluation procedure, explainability framework, and independent testing strategy were defined.

Level 7 – Data Engineering
The raw clinical observations were processed and transformed into fixed-length temporal sequences suitable for LSTM-based prediction.

Level 8 – Model Development
The Improved LSTM model was developed using the prepared temporal sequences and subsequently frozen for final evaluation.

Level 9 – Explainability Framework
SHAP-based analysis was applied to investigate global and local feature contributions and their temporal behavior [2].

Level 10 – Scientific Evaluation
The frozen model was evaluated on an independent and previously unseen patient cohort using predefined evaluation metrics.

Level 11 – Clinical Validation
The model outputs were examined at both window and patient levels, with additional analysis of repeated predictions and alert burden.

Level 12 – Result Interpretation
The final stage integrated predictive performance, explainability findings, patient-level behavior, and alert-burden analysis to answer the research questions.

This workflow provided a consistent methodological structure from data preparation to final interpretation.

3.3 Clinical Problem Formulation

Sepsis is a critical condition in which a patient's physiological state can change rapidly. ICU monitoring provides repeated measurements of variables such as heart rate, oxygen

saturation, respiratory rate, blood pressure, temperature, and laboratory-related indicators. These observations can contain temporal patterns associated with changes in patient condition.

The clinical problem addressed in this research is formulated as a temporal risk prediction task. Instead of treating each observation independently, consecutive observations are represented as a sequence so that the model can learn relationships across time.

The objective is to estimate whether an observed patient sequence is associated with a subsequent positive sepsis outcome according to the labeling framework of the selected dataset [1].

The prediction process can be represented as:

Patient observations
↓
Temporal sequence
↓
Improved LSTM
↓
Risk probability
↓
Binary risk prediction

This formulation allows the study to investigate both predictive performance and the clinical variables and temporal positions associated with model predictions.

3.4 Research Questions and Hypotheses

The study was guided by four research questions.

RQ1: How can SHAP-based explainability, at both global and local levels, be used to interpret the predictions of the Improved LSTM model?

RQ2: Which vital-sign features and temporal positions contribute most strongly to the model's predictions?

RQ3: How does the model discriminate between positive and negative outcomes on an independent, previously unseen patient cohort, at both window and patient levels?

RQ4: What is the practical alert-burden implication of the model's predictions when considered as a real-time monitoring signal?

Based on these questions, the following hypotheses were formulated.

H1: SHAP-based analysis can provide interpretable information about the contribution of input features to the Improved LSTM predictions.
H2: Feature contributions vary across the temporal sequence, and some features and temporal positions contribute more strongly than others.
H3: The Improved LSTM can discriminate between positive and negative outcomes on an independent patient cohort.
H4: The model's predictions can be examined as a potential monitoring signal, but alert frequency and false alerts must be considered when assessing practical applicability.

These hypotheses were investigated through the experimental procedures described in the following sections.

3.5 Dataset Description

The study used the Prediction of Sepsis dataset associated with the PhysioNet/Computing in Cardiology Challenge 2019 [1]. The dataset contains longitudinal ICU observations comprising physiological measurements, laboratory variables, demographic information, and time-related clinical information.

The original dataset was designed for early sepsis prediction using time-dependent clinical observations [1]. In this research, the dataset was accessed through its publicly available Kaggle distribution, while the original PhysioNet/Computing in Cardiology Challenge documentation was treated as the authoritative scientific source.

The model-development process identified 27,753 eligible patient sequences after temporal eligibility filtering. Among these, 1,570 were positive and 26,183 were negative.
The final model input consisted of 34 clinical features observed over six consecutive temporal positions.

Table 3.1: Dataset Summary:

Metric	Value
Total Samples (Windows)	195018
Total Positive Samples	3580
Total Negative Samples	191438
Sequence Shape	(6, 34)
Number of Features	34
Prevalence (Window-level)	1.84%
Observation Window	6 hours
Prediction Horizon	1 hour
Model	Improved LSTM
Model Status	Frozen

The final independent test cohort contained 195,018 temporal windows belonging to 6,051 previously unseen patients. The independent cohort was kept separate from model development and was used only for final evaluation.

3.6 ICU Patient Time-Series Representation

ICU data are inherently temporal because physiological measurements are repeatedly recorded during a patient's stay. A sequence of observations can therefore provide information about physiological changes that may be lost when observations are reduced to a single static representation.

In this study, each patient sequence was represented as a multivariate time series. Each temporal position contained 34 clinical features, while six consecutive positions represented the observation history used for prediction.

The resulting input structure was:

$X \in R^{(6 \times 34)}$

where 6 represents the temporal positions and 34 represents the clinical features at each position.

The temporal order of observations was preserved throughout sequence construction. This allowed the Improved LSTM to process changes across time rather than treating the observations as independent records.

The same temporal structure also formed the basis of the explainability analysis, allowing feature contributions to be examined in relation to their temporal positions.

3.7 Data Engineering and Sequence Construction

The data engineering stage transformed the raw ICU observations into a consistent representation suitable for model inference. The process included feature selection, missing-value handling, feature scaling, temporal ordering, and fixed-length sequence construction.

The preprocessing pipeline was fitted using the training data only. Median imputation was used to handle missing feature values, followed by standardization using a StandardScaler fitted on the training data. The resulting preprocessing parameters were preserved and applied unchanged to validation and independent test data.

This training-only preprocessing strategy prevented information from the independent test cohort from influencing the preprocessing parameters.

After preprocessing, observations were arranged according to the ICU temporal index and grouped into six-step sequences. Each sequence contained 34 clinical features per temporal position.

The resulting input structure was:
Number of sequences × 6 × 34
The final independent test artifact contained 195,018 sequences with this structure.

The data-processing pipeline can be summarized as:

Raw ICU observations
↓
Feature selection
↓
Missing-value handling
↓
Training-fitted scaling
↓
Temporal ordering
↓
Six-step sequence construction
↓
6 × 34 model input

Patient identifiers and target labels were retained separately to support independent evaluation and patient-level analysis.

3.8 Observation Window and Prediction Horizon

The final temporal configuration used a six-hour observation window and a one-hour prediction horizon.
The observation window represents the historical clinical information supplied to the model. The prediction horizon represents the subsequent period associated with the prediction target.

Each model input therefore contained six consecutive temporal positions and 34 clinical features at each position.

Table 3.2: Temporal Configuration:

parameter	value	status
observation_window_hours	6	LOCKED
prediction_horizon_hours	1	LOCKED
sepsis_eligible_patients	1539	EMPIRICAL
sepsis_coverage_pct	75	EMPIRICAL
nonsepsis_eligible_patients	26183	EMPIRICAL
nonsepsis_coverage_pct	100	EMPIRICAL
total_eligible_patients	27722	EMPIRICAL
overall_coverage_pct	98.18310607	EMPIRICAL
candidate_features	34	RETAINED

This configuration was maintained consistently during final model inference and independent evaluation.

The temporal formulation ensured that the model input represented observations available within the defined observation period rather than incorporating information from future prediction periods.

3.9 Experimental Design

The experimental design was structured to evaluate the model from predictive, explainability, patient-level, and practical monitoring perspectives.

First, the finalized Improved LSTM model and its training-fitted preprocessing pipeline were loaded. The model architecture and expected input shape were verified before inference.

Second, the independent test data were transformed using the frozen preprocessing parameters.

Third, predictions were generated for each independent test window. Each sequence produced a probability representing the model's estimated risk.

Fourth, the window-level predictions were evaluated using predefined classification and discrimination metrics.

Fifth, sequence-level predictions were associated with patient identifiers to examine model behavior at the patient level.

Sixth, SHAP-based analysis was used to investigate global and local feature contributions and temporal attribution [2].

Finally, repeated positive predictions were analyzed to assess the potential alert burden of the model when considered as a monitoring signal.

The experimental structure can therefore be summarized as:

Independent test cohort
↓
Frozen preprocessing
↓
Improved LSTM inference
↓
Risk probabilities
↓
Window-level evaluation
↓
Patient-level evaluation
↓
SHAP analysis
↓
Alert-burden analysis
↓
Research interpretation

3.10 Model Development Strategy

The predictive component of the research was based on an Improved Long Short-Term Memory (LSTM) model. LSTM is suitable for sequential data because its gated architecture allows information to be retained and updated across temporal positions.

The model received a six-step sequence containing 34 clinical features at each step and produced a single risk probability.

The model-processing structure was:

6×34 temporal input
↓
Improved LSTM
↓
Temporal representation
↓
Output layer
↓
Sepsis risk probability

The final model was selected during the model-development stage and subsequently frozen. The independent test cohort was not used to retrain, fine-tune, or modify the model.

The final model artifact and preprocessing pipeline were preserved separately from the independent evaluation data. This separation supported reproducibility and reduced the possibility of test-set influence on the final model.

3.11 Explainability Framework

Explainability was incorporated to investigate how the Improved LSTM generated its predictions, particularly which clinical variables contributed to its outputs.

SHapley Additive exPlanations (SHAP) was used as the primary explainability framework [2], with SHAP GradientExplainer applied to the model.

The analysis considered three levels:

- **Global explanation:** Identified features with greater overall contribution using mean absolute SHAP values.
- **Local explanation:** Examined how individual features contributed toward or away from a particular prediction.
- **Temporal explanation:** Analyzed feature contributions across the six temporal positions of the input sequence.

The framework can be represented as:

Model prediction → SHAP GradientExplainer → Global + Local + Temporal contributions → Feature-time interpretation

The completed analysis identified Hct as the highest-ranked global feature and t = 1 as the dominant temporal position. These findings describe model behavior and should not be interpreted as causal clinical relationships.

LIME was also recognized as an established alternative for local explanations [6], but SHAP was selected as the primary explainability method in this study.

3.12 Scientific Evaluation and Clinical Validation Strategy

The final evaluation assessed the performance of the frozen Improved LSTM on a previously unseen independent test cohort containing 195,018 temporal windows from 6,051 patients.

Evaluation was conducted at both window and patient levels. At the window level, each six-step sequence was evaluated using accuracy, precision, recall, specificity, F1-score, AUROC, and AUPRC. At the patient level, multiple windows from the same patient were considered together to assess positive-patient detection.

The evaluation also included alert outcome and alert-burden analysis to examine the effect of repeated positive predictions. The clinical validation was limited to research-oriented evaluation and does not represent prospective clinical validation or clinical deployment readiness.

The evaluation followed the sequence:

Independent patient cohort → Window-level prediction → Performance metrics → Patient-level detection → Alert analysis → Alert-burden assessment → Clinical interpretation

The independent test cohort was evaluated without retraining, fine-tuning, threshold tuning, or modification of the test data, preserving the independence of the final evaluation.

3.13 Summary

This chapter presented the methodology and experimental design of the study, combining ICU temporal data representation, Improved LSTM prediction, SHAP-based explainability, independent cohort evaluation, patient-level analysis, and alert-burden assessment.

The Prediction of Sepsis dataset associated with the PhysioNet/Computing in Cardiology Challenge 2019 was used as the clinical time-series dataset [1]. The model used six temporal positions with 34 clinical features per position, while median imputation and standardization were fitted only on the training data.

The final Improved LSTM was frozen before independent evaluation, and SHAP GradientExplainer was used to analyze global, local, and temporal feature contributions [2]. Final evaluation was performed on 195,018 windows from 6,051 previously unseen patients at both window and patient levels.

Overall, the methodology was designed as an experimental and interpretability study rather than a clinical deployment study. The next chapter presents the data engineering, feature handling, sequence construction, preprocessing, model architecture, training, and reproducibility procedures.

CHAPTER 4
DATA ENGINEERING AND MODEL DEVELOPMENT

4.1 Introduction

This chapter presents the data engineering and model development procedures used to prepare the ICU clinical data and develop the final Improved Long Short-Term Memory (LSTM) model. The process includes dataset preparation, feature handling, temporal sequence construction, training-fitted preprocessing, model development, model selection, model freezing, and reproducible artifact handoff. The main objective was to transform the clinical observations into a consistent six-step temporal representation containing 34 clinical features while maintaining a strict separation between model-development data and the independent test cohort.

4.2 Dataset Preparation

The study used the Prediction of Sepsis dataset associated with the PhysioNet/Computing in Cardiology Challenge 2019 [1]. The dataset was accessed through its Kaggle distribution, while the original PhysioNet documentation was treated as the authoritative source for dataset definition and provenance. Patient observations were organized using patient identifiers and the ICU temporal index so that observations from different patients were not mixed during sequence construction. The eligible development data were separated from the independent test cohort according to the experimental design. The final independent test artifacts were preserved separately and were used only during the final evaluation.

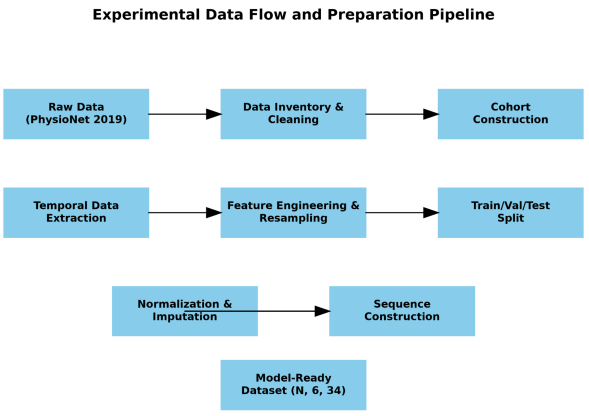


FIGURE 4.1: Dataset Preparation and Experimental Data Flow

The independent test cohort remained untouched during model development and was reserved for evaluation on previously unseen patients.

4.3 Data Quality and Feature Handling

The clinical data contained missing observations across several physiological and laboratory variables. Data-quality processing focused on maintaining a consistent feature structure and preventing information leakage. Missing values were handled through the training-fitted preprocessing pipeline. The preprocessing parameters were learned only from the training data and then applied unchanged to validation and independent test data.

The overall process was:

Raw Clinical Data
↓
Data Organization
↓
Feature Selection
↓
Missing-Value Handling
↓
Training-Fitted Transformation
↓
Temporal Sequence Construction
↓
Model Input

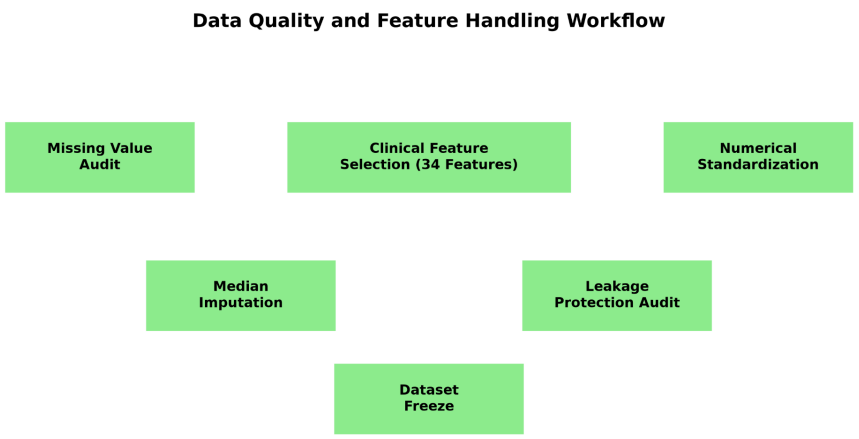


FIGURE 4.2: Data Quality and Feature Handling Workflow

The same feature representation was maintained throughout training, explainability analysis, and independent evaluation.

4.4 Feature Set and Clinical Variables

The final Improved LSTM model used 34 clinical features consisting of physiological and laboratory observations relevant to the prediction task. The selected features included vital signs such as HR, O2Sat, Temp, SBP, MAP, and DBP, together with laboratory and other clinical variables including Hct, Hgb, WBC, Platelets, Glucose, Lactate, Creatinine, BUN, and related measurements. The complete 34-feature representation was fixed before final evaluation and was not modified during independent testing.

Table 4.1: Complete 34-Feature Clinical Variable Set:

Patient-Level Coverage and Observation Completeness

Clinical Feature	Patients	Patients with ≥1 Observation	Coverage (%)	Completely Missing	Complete Missing (%)	Total Observations	Mean Observations / Patient	Max Observations / Patient	Mean Missing (%)
HR	28,235	28,231	99.99	4	0.01	979,928	34.71	333	11.09
O2Sat	28,235	28,222	99.95	13	0.05	945,865	33.5	331	14.27
O2Sat	28,235	28,183	99.82	52	0.18	920,515	32.6	323	16.41
MAP	28,235	28,158	99.73	77	0.27	952,045	33.72	332	13.82
SBP	28,235	28,047	99.33	188	0.67	929,603	32.92	332	15.54
Temp	28,235	28,034	99.29	201	0.71	368,077	13.04	214	66.45
Glucose	28,235	27,126	96.07	1,109	3.93	186,194	6.59	151	82.96
Potassium	28,235	26,915	95.32	1,320	4.68	101,468	3.59	81	90.67
BUN	28,235	26,803	94.93	1,432	5.07	74,544	2.64	45	93.04
Creatinine	28,235	26,779	94.84	1,456	5.16	66,242	2.35	27	93.83
Hct	28,235	26,587	94.16	1,648	5.84	96,140	3.4	61	91.04
Hgb	28,235	26,501	93.86	1,734	6.14	80,104	2.84	53	92.56
Platelets	28,235	26,412	93.54	1,823	6.46	64,575	2.29	35	93.99
WBC	28,235	26,381	93.43	1,854	6.57	69,542	2.46	52	93.51
Magnesium	28,235	24,767	87.72	3,468	12.28	68,554	2.43	44	93.75
Calcium	28,235	24,474	86.68	3,761	13.32	64,012	2.27	67	94.15
DBP	28,235	23,052	81.64	5,183	18.36	746,902	26.45	332	33.25
Phosphate	28,235	19,831	70.24	8,404	29.76	43,521	1.54	26	96.07
Chloride	28,235	15,000	53.13	13,235	46.87	49,225	1.74	48	95.47
PTT	28,235	14,151	50.12	14,084	49.88	31,859	1.13	44	97.11
HCO3	28,235	14,130	50.04	14,105	49.96	45,361	1.61	49	95.82
pH	28,235	13,264	46.98	14,971	53.02	75,516	2.67	92	93.21
PaCO2	28,235	12,879	45.61	15,356	54.39	60,781	2.15	83	94.6
FiO2	28,235	12,482	44.21	15,753	55.79	90,921	3.22	133	92.34
AST	28,235	9,963	35.29	18,272	64.71	17,321	0.61	30	98.37
Bilirubin_total	28,235	9,906	35.08	18,329	64.92	15,991	0.57	21	98.49
Alkalinephos	28,235	9,838	34.84	18,397	65.16	17,155	0.61	31	98.38
BaseExcess	28,235	9,284	32.88	18,951	67.12	59,079	2.09	90	94.69
SaO2	28,235	9,198	32.58	19,037	67.42	37,845	1.34	68	96.55
Lactate	28,235	8,817	31.23	19,418	68.77	29,124	1.03	63	97.43
TroponinI	28,235	4,953	17.54	23,282	82.46	10,390	0.37	17	98.92
Fibrinogen	28,235	3,195	11.32	25,040	88.68	7,182	0.25	32	99.38
EtCO2	28,235	2,258	8	25,977	92	40,850	1.45	308	97.12
Bilirubin_direct	28,235	1,441	5.1	26,794	94.9	2,089	0.07	12	99.81

Feature Distribution and Patient-Level Variability

Clinical Feature	Mean of Patient Means	Median of Patient Medians	Global Min	Global Max	Mean Patient SD	Zero-Variance Patients
HR	83.81	82	20	223	8.32	55
O2Sat	97.16	98	20	100	1.92	301
O2Sat	18.56	18	1	100	3.39	305
MAP	82.57	80	20	300	9.88	174
SBP	123.36	121	20	300	13.78	240
Temp	36.85	36.83	21	50	0.45	596
Glucose	131.89	122.5	10	988	27.52	5,012
Potassium	4.1	4	1.3	27.5	0.35	7,933
BUN	21.92	16	1	268	3.29	10,053
Creatinine	1.43	0.9	0.1	41.9	0.19	10,933
Hct	32.08	31.5	8.8	71.7	2	8,276
Hgb	10.73	10.55	2.3	32	0.69	9,489
Platelets	207.19	191	1	2,322	21.62	9,907
WBC	11.12	10.2	0.1	440	1.86	10,029
Magnesium	2.02	2	0.2	9.8	0.21	12,841
Calcium	8.02	8.3	1	27.9	0.95	12,418
DBP	64.09	62	20	300	8.16	5,310
Phosphate	3.54	3.35	0.2	16.9	0.6	16,854
Chloride	105.55	106	26	145	2.1	17,667
PTT	37.29	31	16.6	250	9.99	20,866
HCO3	24.39	24	0	55	1.55	18,376
pH	7.38	7.385	6.62	7.8	0.04	18,451
PaCO2	40.88	40	11	100	4.64	18,772
FiO2	0.6	0.5	−50	4,000	0.32	19,365
AST	140.72	33	3	9,961	101.22	24,879
Bilirubin_total	1.48	0.75	0.1	49.6	0.42	25,170
Alkalinephos	97.17	72	7	2,528	10.7	24,905
BaseExcess	−0.47	0	−32	44	1.69	20,896
SaO2	93.63	97	24	100	4.84	22,484
Lactate	2.16	1.65	0.2	31	0.72	23,347
TroponinI	5.35	0.095	0.01	440	2.81	26,092
Fibrinogen	304.6	259	34	1,760	49.3	26,971
EtCO2	33.16	33	10	100	3.47	26,170
Bilirubin_direct	1.26	0.3	0.01	35	0.46	28,000

Table 4.1 should present the verified feature-quality information from the final feature-quality audit, including patient coverage and observation completeness. The fixed feature ordering ensured compatibility between the training-fitted preprocessing pipeline, the frozen model, and the independent test artifacts.

4.5 Temporal Sequence Construction

The research formulated sepsis prediction as a temporal classification problem. Consecutive observations from the same patient were therefore organized into fixed-length sequences. Each input sequence contained six temporal positions and 34 clinical features: Sequence Shape = (6, 34) The temporal order of observations was preserved during sequence construction, and observations from different patients were not combined.

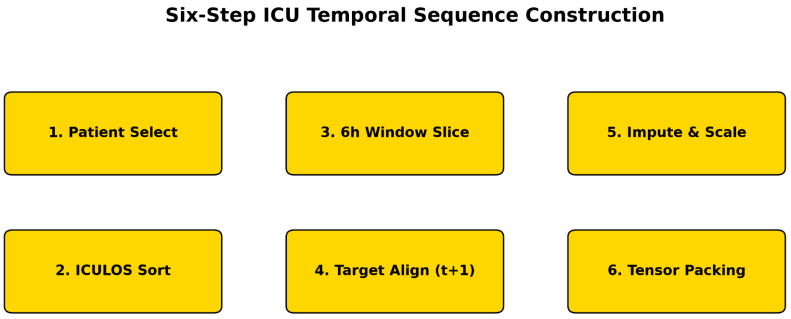


FIGURE 4.3: Six-Step ICU Temporal Sequence Construction

The resulting six-step sequences formed the primary input representation for the Improved LSTM model and the subsequent temporal explainability analysis.

4.6 Training-Fitted Preprocessing

A training-fitted preprocessing pipeline was used before model inference. Median imputation was fitted on the training data to handle missing values, followed by standardization using a scaler fitted on the same training data. The fitted preprocessing object was saved and reused for validation and independent test data.

The process was:

Training Data
↓
Fit Imputer and Scaler
↓
Save Preprocessor
↓
Transform Development
↓
Data Independent Test Data
↓
Load Saved Preprocessor
↓
Apply Existing Transformation
↓
Model-Compatible Input

This approach prevented the independent test cohort from influencing the learned preprocessing parameters.

4.7 Improved LSTM Architecture

The final predictive model was an Improved Long Short-Term Memory (LSTM) model designed to process multivariate ICU time-series data.

The verified model interface was:
Input Shape: (None, 6, 34)
Output Shape: (None, 1)

Here, 6 represents the temporal sequence length and 34 represents the clinical feature dimension.

The model was developed and validated before final independent evaluation. After selection, its architecture and learned parameters were frozen for subsequent evaluation.

4.8 Model Input and Output Representation

The model input consists of a three-dimensional tensor with the general structure:
(Batch Size, 6, 34)

The first dimension represents the number of sequences processed in a batch. The second dimension represents the six temporal positions, and the third dimension represents the 34 clinical variables.

Therefore, an individual patient sequence can be represented as:

$$X \in R^{6x34}$$

Each row corresponds to one temporal position, while each column corresponds to one clinical feature.

The model produces a single output for each input sequence.

Output Shape = (Batch Size, 1)

The output represents the model's prediction for the corresponding sequence according to the target definition established during the experimental design.

The same input dimensionality was maintained during independent testing to ensure that the frozen model received data in the representation for which it had been developed.

4.9 Model Development and Training

The prepared temporal sequences were used to train the Improved LSTM model. The training process used the training subset and corresponding target labels, while validation data were used for model-development assessment.

The development pipeline was:

Prepared Clinical Data
↓
Temporal Sequence Construction
↓
Training-Fitted Preprocessing
↓
Improved LSTM Training
↓
Validation
↓
Model Selection
↓
Final Model

The independent test cohort was not used as a training source or for model optimization.

4.10 Model Selection and Model Freezing

After model development and validation, the final Improved LSTM configuration was selected for the research evaluation. The selected model was then frozen. From this point onward, its architecture and learned parameters remained unchanged.

The overall process can be represented as:

Model Development
↓
Validation
↓
Model Selection
↓
Final Improved LSTM
↓
Model Freezing
↓
Independent Test Evaluation

The frozen model was subsequently loaded by the evaluation environment without further model training. This ensured that the independent-test predictions were generated using the established final model rather than a model modified after access to the test cohort.

4.11 Model Handoff and Reproducibility

After model freezing, the final model and its associated preprocessing components were preserved for use in the independent evaluation and visual testing workflow.

The model handoff included the trained model, the training-fitted preprocessing object, and the required test artifacts. The independent evaluation artifacts included:

- X_test sequence data
- y_test target labels
- Test sequence patient identifiers
- Independent test cohort metadata

The authoritative independent test feature artifact was preserved as:

step_10_2_X_test.npy

The corresponding target and patient-identification artifacts were preserved separately.

Table 4.2: Final Model Handoff and Evaluation Artifact Inventory

Artifact Filename	Description	Size (MB)
step_13P_14_X_train.npy	Training feature tensor (22202x6x34)	17.26
step_13P_14_y_train.npy	Training targets (binary)	0.02
step_13P_14_X_val.npy	Validation feature tensor (5551x6x34)	4.32
step_13P_14_y_val.npy	Validation targets (binary)	0.01
step_13P_14_train_patient_ids.npy	Training patient identifiers	0.18
step_13P_14_val_patient_ids.npy	Validation patient identifiers	0.04
step_13P_14_preprocessor.pkl	Frozen imputer and scaler objects	0.01
step_13P_14_dataset_metadata.json	Detailed data engineering summary	0.01
LEVEL_7_MODEL_HANDOFF_MANIFEST.json	Immutable handoff record	0.01
LEVEL_7_MODEL_HANDOFF_CHECKSUMS.csv	SHA-256 integrity signatures	0.01

Table 4.2 should list the verified final model-development and evaluation artifacts, including their filenames and descriptions.

The independent test feature artifact was preserved as:

step_10_2_X_test.npy

with the verified structure
:
X_test = (195018, 6, 34)

Thus, the independent test set contained 195,018 temporal windows, each represented by six temporal positions and 34 clinical features.

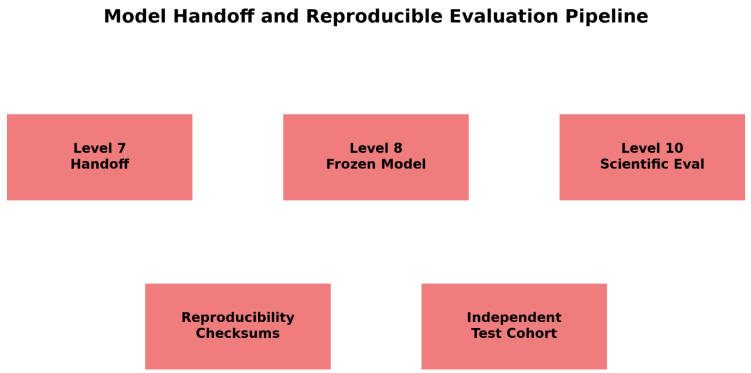


FIGURE 4.4: Model Handoff and Reproducible Evaluation Pipeline

The Visual Testing Demo was developed as an execution and demonstration environment for loading the frozen model, applying the saved preprocessing pipeline, generating predictions, and displaying relevant model outputs. It was not used as an additional training stage or as a separate predictive model.

The reproducibility structure can therefore be summarized as:

Frozen Model
+
Training-Fitted Preprocessor
+
Independent Test Artifacts
+
Patient Identifiers
+
Evaluation Logic
↓
Reproducible Model Execution
↓
Scientific Evaluation
↓
Visual Testing Demo

4.12 Summary

This chapter presented the data engineering and model development process used in the study. The clinical observations were organized into a fixed 34-feature representation and transformed into six-step temporal sequences. A training-fitted median imputer and standard scaler were used to maintain consistent preprocessing without using information from the independent test cohort. The final Improved LSTM model accepts an input structure of (6, 34) and produces one prediction for each sequence. After model selection, the final model was frozen and preserved together with its preprocessing pipeline and supporting research artifacts. The independent test cohort remained separate from model development and contained 195,018 windows from 6,051 previously unseen patients. The preserved artifacts support reproducible model execution and subsequent independent evaluation. The next chapter presents the SHAP-based explainability framework used to investigate global, local, temporal, and feature-time model contributions.

CHAPTER 5
EXPLAINABILITY FRAMEWORK

5.1 Introduction

The Improved LSTM can learn temporal patterns from ICU time-series data, but its predictions do not directly indicate which clinical variables influenced the output. Therefore, an explainability framework was incorporated to examine feature and temporal contributions. This chapter presents the SHAP-based explainability methodology applied to the trained Improved LSTM. Since each input sequence contains six temporal positions and 34 clinical features, the analysis considers global, local, temporal, and feature–time attribution. The empirical findings are discussed separately in Chapter 7.

5.2 Explainable AI Framework

SHAP (SHapley Additive exPlanations) was selected as the primary explainability method because it provides feature-attribution values for model predictions [2]. The framework was applied to the trained Improved LSTM and organized according to its six-step temporal input structure.
The overall explainability workflow can therefore be represented as:

Trained Improved LSTM
↓
SHAP GradientExplainer
↓
SHAP Attribution Values
↓
Feature-Level Attribution
↓
Global Feature Importance
↓
Local Feature Attribution
↓
Temporal Importance
↓
Feature–Time Attribution
↓
Dominant Feature–Temporal Position Analysis
↓
Model Behavior Interpretation

This structure allows the model's behavior to be examined from both feature and temporal perspectives.

5.3 SHAP Methodology

SHAP is based on Shapley values from cooperative game theory and estimates the contribution of individual input features to a model output [2]. A simplified representation is:

$f(x) = \varphi_0 + \Sigma\varphi_i$

where $f(x)$ is the model output, φ_0 is the reference output, and φ_i represents the contribution of feature i.

Positive SHAP values indicate contributions toward increasing the model output, while negative values indicate contributions toward decreasing it.

For global analysis, mean absolute SHAP values were used to rank the overall magnitude of feature contributions.

5.4 SHAP GradientExplainer

SHAP GradientExplainer was applied to the trained neural-network-based Improved LSTM. The model input consisted of six temporal positions and 34 clinical features, represented as:

The input structure of the model is:

Input shape = (6, 34)
where 6 represents the temporal positions in each sequence and 34 represents the clinical feature dimensions.

The explainer was applied after model training and freezing, without modifying the established model. The resulting attribution values retained the temporal and feature dimensions for subsequent analysis.

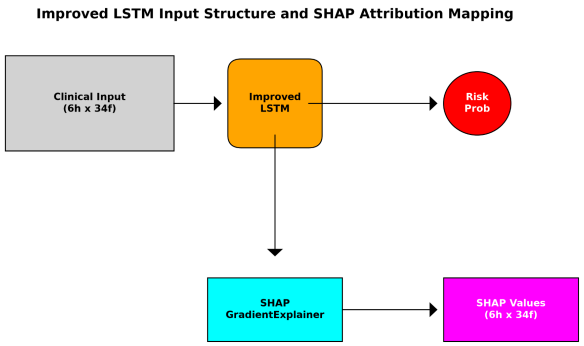


FIGURE 5.1: Improved LSTM Input Structure and SHAP Attribution Mapping

The generated attribution values formed the basis for the analyses presented in the following sections.

5.5 Global Feature Importance

Global feature importance was used to determine which of the 34 clinical features had the largest overall contribution to the model predictions. Mean absolute SHAP values were calculated across the analyzed observations and used for feature ranking.

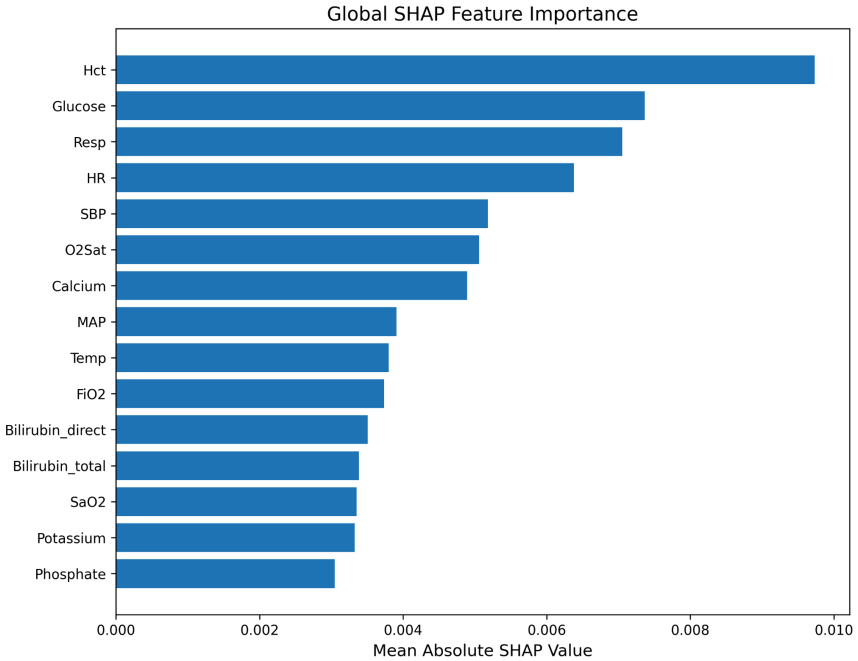


FIGURE 5.2: Global SHAP Feature Importance

Table 5.1: Global SHAP Feature Importance Ranking

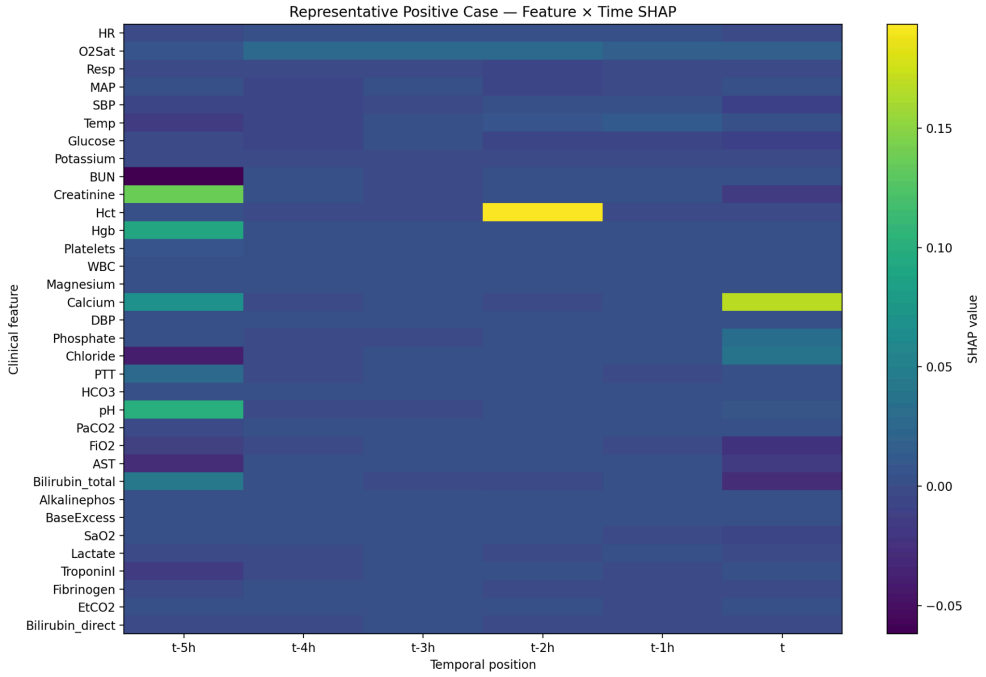
feature	mean_abs_shap	mean_signed_shap	importance_rank
Hct	0.009733300966	0.001344757832	1
Glucose	0.007364954452	-0.0003229015388	2
Resp	0.007051623489	0.00177587822	3
HR	0.006380677499	0.0005246936855	4
SBP	0.005183481576	-0.0005220694813	5
O2Sat	0.005060563016	0.0008321295479	6
Calcium	0.004891681332	-0.001343932907	7
MAP	0.003910469695	-0.001164503002	8
Temp	0.003800932616	-0.0001930586687	9
FiO2	0.003734934525	-0.0001857262152	10
Bilirubin_direct	0.003505732825	0.0001744674727	11
Bilirubin_total	0.003384563349	-0.0005229332481	12
SaO2	0.003353602531	-0.0002879307655	13
Potassium	0.003326497253	-0.001454412095	14
Phosphate	0.003050388751	0.00012656739	15
pH	0.003001340528	-0.0001179844132	16
BUN	0.002296071985	0.00100210377	17
Lactate	0.00224107308	-9.08E-05	18
Fibrinogen	0.002195107477	0.0007305220028	19
Creatinine	0.001646280147	0.0005043238519	20
TroponinI	0.001611927563	5.62E-05	21
Hgb	0.001584640469	0.0007422456154	22
AST	0.001422983963	0.0002618920903	23
PaCO2	0.001352523261	5.29E-05	24
Platelets	0.001209702125	0.0005795875866	25
Chloride	0.00113143238	-8.54E-06	26
PTT	0.000815410378	-0.0001791904701	27
EtCO2	0.0008040255002	-0.0004106186896	28
BaseExcess	0.0007091588198	-0.0006016415404	29
DBP	0.0006325194319	0.0004287896901	30
HCO3	0.0003933977201	0.0003238699693	31
Magnesium	0.0003823621806	-0.0003136995636	32
Alkalinephos	0.0003530330323	2.35E-05	33
WBC	0.0003144632932	-7.55E-05	34

The analysis identified Hct as the highest-ranked global feature. This result represents model behavior within the analyzed dataset and does not imply a causal clinical relationship.

5.6 Local Feature Attribution

Local SHAP analysis was used to examine how individual input features contributed to a specific model prediction. For a selected temporal sequence, SHAP values indicate whether individual inputs contributed toward increasing or decreasing the model output. .

FIGURE: Local SHAP Attribution for a Representative Test Sequence



The representative explanation should be based on an actual evaluated sequence. Local attribution describes model behavior for that observation and does not establish causality.

5.7 Temporal Importance Analysis

Because the Improved LSTM processes sequential observations, SHAP contributions were also analyzed across the six temporal positions. The attribution magnitude of the 34 features was aggregated at each time step to determine the relative contribution of different positions within the observation window.

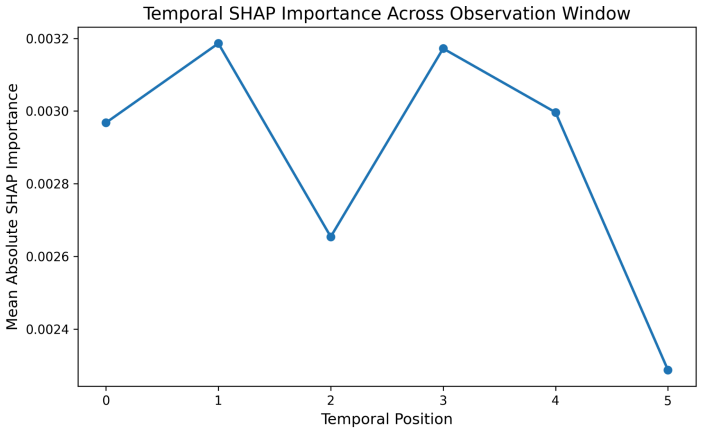


FIGURE 5.3: Temporal SHAP Importance Across the Six Time Steps

The temporal analysis identified t = 1 as the dominant temporal position. This indicates relatively greater model attribution at that position within the analyzed observations.

5.8 Feature–Time Attribution Analysis

Global feature importance treats each clinical variable as a single feature across the complete observation sequence. However, the contribution of a clinical variable may vary according to the temporal position at which it is observed.

To investigate this behavior, the research performed feature–time attribution analysis. This analysis retains both dimensions of the model input and examines the SHAP contribution of each clinical feature at each of the six temporal positions.

With 34 clinical features and six temporal positions, the attribution structure can be represented as a 6 × 34 feature–time matrix.

Each element of the matrix represents the contribution associated with one clinical feature at one temporal position.

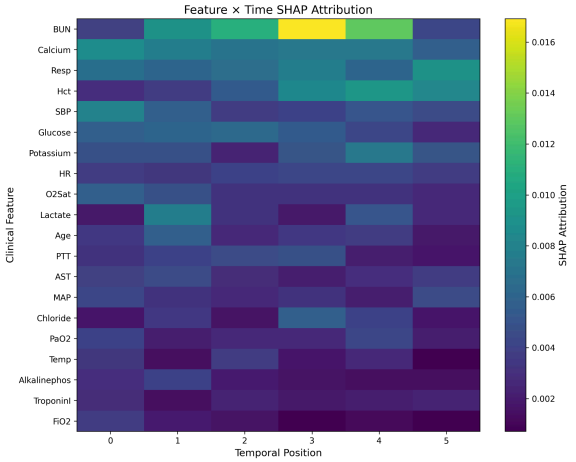


FIGURE 5.4: Feature–Time SHAP Attribution Heatmap

This analysis provides a more detailed view of how feature contributions vary across different temporal positions.

5.9 Dominant Feature and Temporal Position Analysis

The feature–time attribution matrix was further examined to identify comparatively strong feature–temporal attribution combinations.

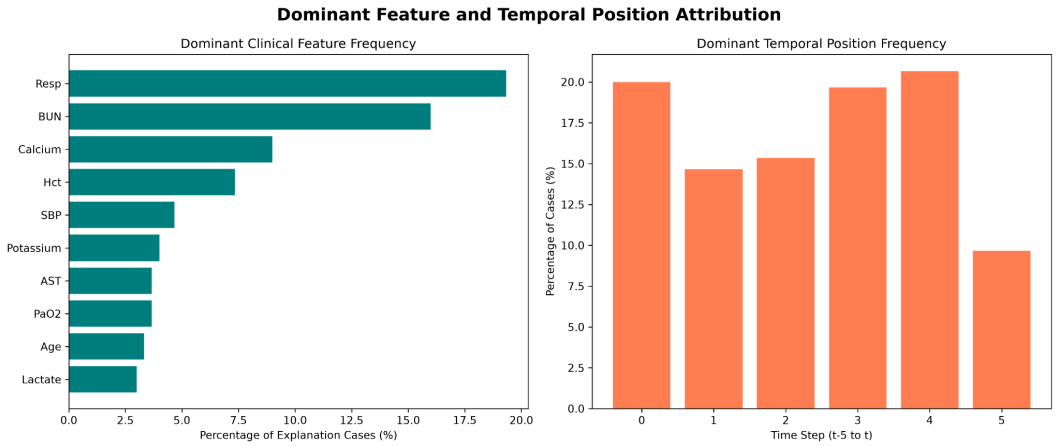


FIGURE 5.5: Dominant Feature and Temporal Position Attribution

The identified combinations indicate where the trained model assigned relatively stronger attribution within the evaluated sequences. They are interpreted as model behavior rather than causal clinical relationships.

5.10 Explainability Interpretation

The SHAP framework provides complementary views of the Improved LSTM:

- Global analysis identifies features with higher overall attribution.
- Local analysis explains individual predictions.
- Temporal analysis examines attribution across the six time steps.
- Feature–time analysis combines feature and temporal information.

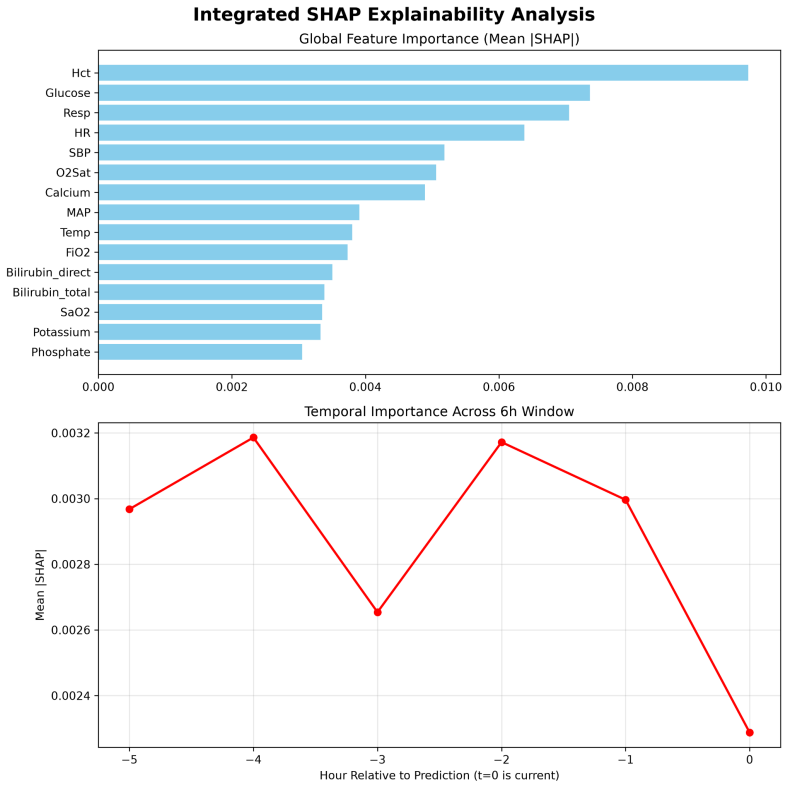


FIGURE 5.6: Integrated SHAP Explainability Analysis

These analyses help characterize how the trained model uses clinical and temporal information. However, SHAP values represent model attribution, not causality or independent clinical importance. The explainability findings are therefore interpreted together with the independent-test results in Chapter 7.

5.11 Summary

This chapter presented the SHAP-based explainability framework used to interpret the Improved LSTM. SHAP GradientExplainer was applied to the six-step, 34-feature model input to examine global, local, temporal, and feature–time contributions [2]. The framework identified Hct as the highest-ranked global feature and t = 1 as the dominant temporal position. These findings describe the behavior of the trained model within the evaluated data and are not interpreted as causal clinical conclusions. The detailed empirical explainability findings are presented in Chapter 7.

CHAPTER 6
SCIENTIFIC EVALUATION AND CLINICAL VALIDATION

6.1 Introduction

The final evaluation assessed the performance of the frozen Improved LSTM on an independent, previously unseen test cohort. The test data were not used for model training, selection, retraining, or threshold tuning.

Evaluation was conducted at window and patient levels using classification and discrimination metrics. Confusion matrix, ROC, precision–recall, risk probability, and alert-burden analyses were also performed to provide a comprehensive assessment of model behavior.

The evaluation used the preserved independent-test artifacts, training-fitted preprocessing pipeline, and frozen Improved LSTM. The Visual Testing Demo was used only as a research-oriented testing and visualization interface.

The evaluation therefore focused on predictive performance, patient-level detection, and the practical implications of repeated model-generated alerts.

6.2 Independent Test Cohort

The independent test cohort contained 195,018 temporal windows from 6,051 previously unseen patients. Each window consisted of six temporal positions and 34 clinical features.

The verified test feature artifact was:

step_10_2_X_test.npy

with the following shape:

X_test = (195018, 6, 34)

The corresponding patient identifiers were used to associate temporal windows with individual patients and support patient-level analysis.

The training-fitted preprocessing pipeline was applied to the test data without refitting. The frozen Improved LSTM was then used to generate the final predictions.

This preserved the independence of the final evaluation.

Independent Test Data
↓
Load Preserved Test Artifacts
↓
Load Training-Fitted Preprocessor
↓
Transform Test Sequences
↓
Load Frozen Improved LSTM
↓
Generate Predictions
↓
Window-Level Evaluation
↓
Patient-Level Evaluation
↓
Clinical Alert Analysis

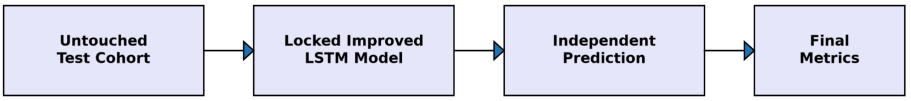


FIGURE 6.1: Independent Test Evaluation Pipeline

This procedure maintained the separation between model development and independent scientific evaluation.

6.3 Evaluation Metrics

The independent-test predictions were evaluated using accuracy, precision, recall/sensitivity, specificity, F1-score, AUROC, and AUPRC.

Accuracy measures overall classification correctness, while precision and recall describe positive prediction quality and positive-case detection. Specificity measures negative-case identification, F1-score summarizes precision and recall, and AUROC and AUPRC evaluate discrimination across probability thresholds.

Table 6.1: Evaluation Metrics and Their Definitions

Metric	Value
Test samples	
Positive test windows	
Negative test windows	
Accuracy	0.6885210596
Precision	0.03690861957
Recall / Sensitivity	
Specificity	0.689497383
F1-score	0.06977029096
Balanced Accuracy	
MCC	
AUROC	0.7066625127
AUPRC	0.03907442855
True Positives	2278
False Positives	59442
True Negatives	131996
False Negatives	1302
Patient-level sensitivity	0.811364
Total alerts	61720
True-positive alerts	2278
False-positive alerts	59442
Alert rate	0.316484
False-alert ratio	0.963091
Mean alerts / patient	10.2
Median alerts / patient	4
Maximum alerts / patient	286
Top global SHAP feature	Hct
Dominant temporal position	t=1

The reported values were calculated directly from the independent-test predictions and were not inferred from the model-development results.

6.4 Confusion Matrix Analysis

The confusion matrix summarizes the four possible classification outcomes: true positives, true negatives, false positives, and false negatives.

For the independent test cohort, the model produced:

TN = 131,996
FP = 59,442
FN = 1,302
TP = 2,278

These outcomes provide the basis for interpreting the classification metrics and the subsequent alert-burden analysis.

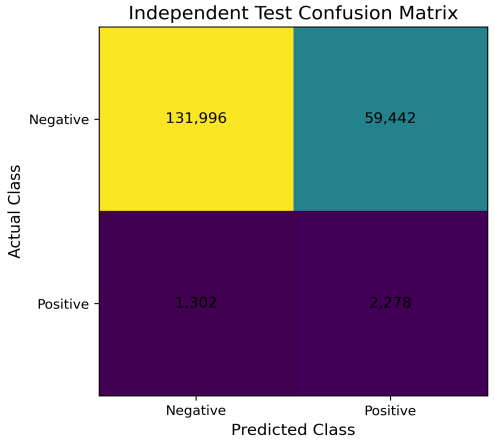


FIGURE 6.2: Independent-Test Confusion Matrix

6.5 ROC Curve Analysis

The ROC curve evaluates model discrimination across different probability thresholds by comparing the true-positive and false-positive rates.

The independent-test evaluation produced an AUROC of:

AUROC = 0.706663

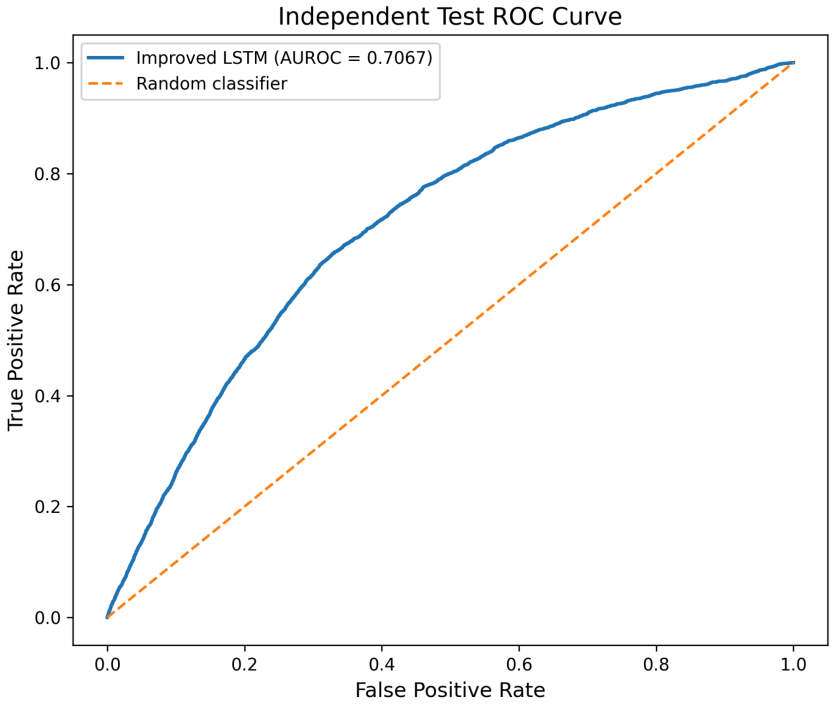


FIGURE 6.3: Receiver Operating Characteristic (ROC) Curve for the Independent Test Cohort

The ROC analysis provides a threshold-independent view of the model's ability to distinguish between positive and negative outcomes.

6.6 Precision–Recall Curve Analysis

Precision–recall analysis was included because the independent test cohort contains a relatively low proportion of positive windows.

The resulting AUPRC was:

AUPRC = 0.039074

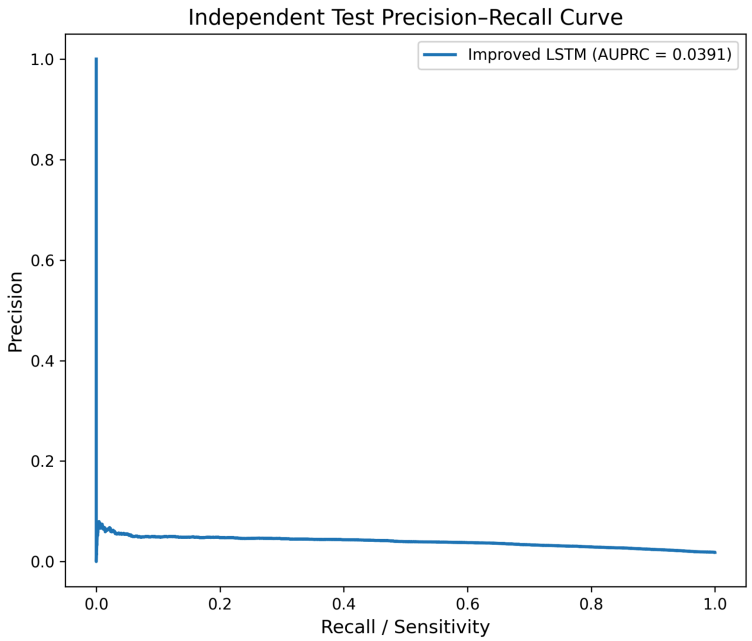


FIGURE 6.4: Precision–Recall Curve for the Independent Test Cohort

The PR analysis provides additional information about the relationship between positive-case detection and false-positive predictions.

6.7 Validation-to-Independent-Test Performance Comparison

The final independent-test performance was compared with the validation-stage results to examine generalization beyond the model-development data.

Table 6.2: Validation-to-Independent-Test Performance Comparison

Metric	Validation	Independent Test
AUROC	0.752246	0.7066625127
AUPRC	0.202774	0.03907442855
RECALL	0.550955	0.6363128492
PRECISION	0.171457	0.03690861957

The comparison showed differences between validation and independent-test performance, particularly in precision and AUPRC. These differences are considered in the discussion of generalization in Chapter 7.

6.8 Window-Level Performance

At the window level, each six-step temporal sequence was treated as an individual prediction unit.

The independent test contained 195,018 windows. The model achieved a recall/sensitivity of 0.636313 at this level, with an accuracy of 0.688521, precision of 0.036909, specificity of 0.689497, and F1-score of 0.069770.

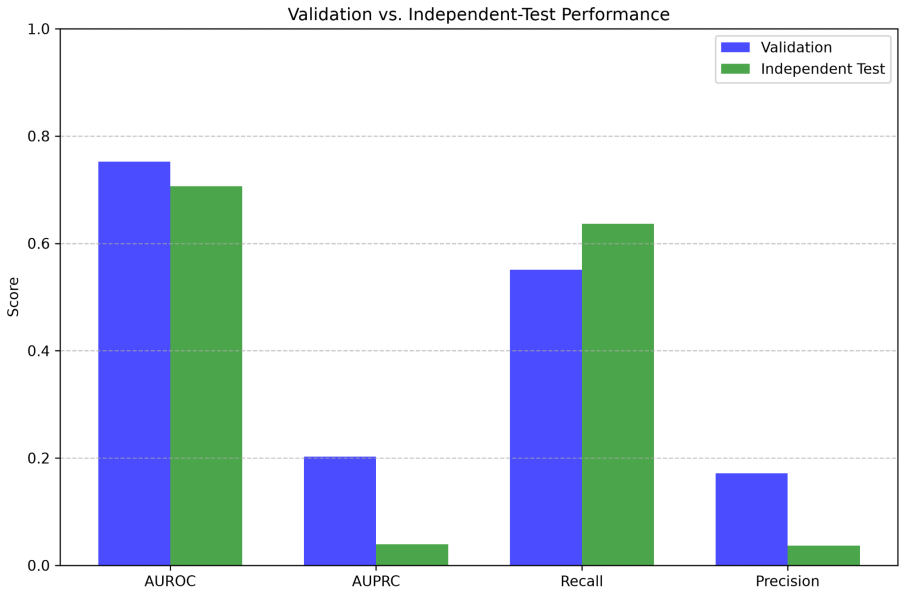


FIGURE 6.5: Window-Level Classification Performance

Window-level performance describes individual temporal predictions and should not be interpreted as equivalent to unique patient-level performance because multiple windows may belong to the same patient.

6.9 Patient-Level Clinical Performance

Patient-level evaluation grouped temporal predictions according to the corresponding patient identifiers.

Among the 6,051 test patients, 440 were positive patients. The model detected 357 of these patients and missed 83, resulting in a patient-level sensitivity of:

Patient-Level Sensitivity = 0.811364

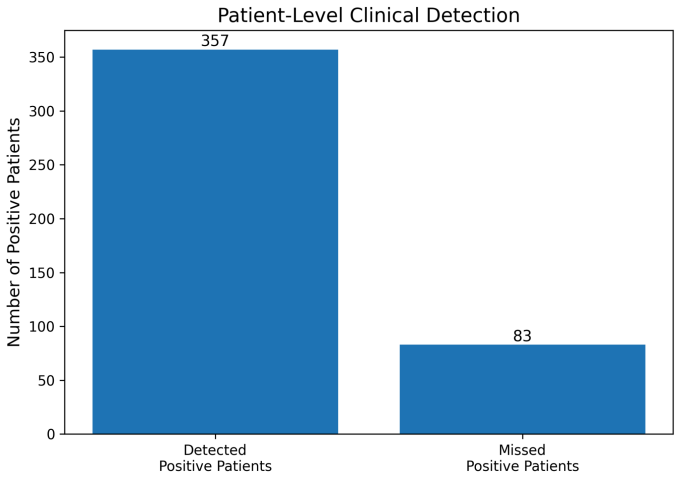


FIGURE 6.6: Patient-Level Prediction Outcome Analysis

This analysis provides a complementary view of model performance by considering patient-level detection rather than individual temporal windows.

6.10 Risk Probability Band Analysis

The predicted probabilities were grouped into predefined risk bands to examine how the model distributed its outputs across different levels of predicted risk.

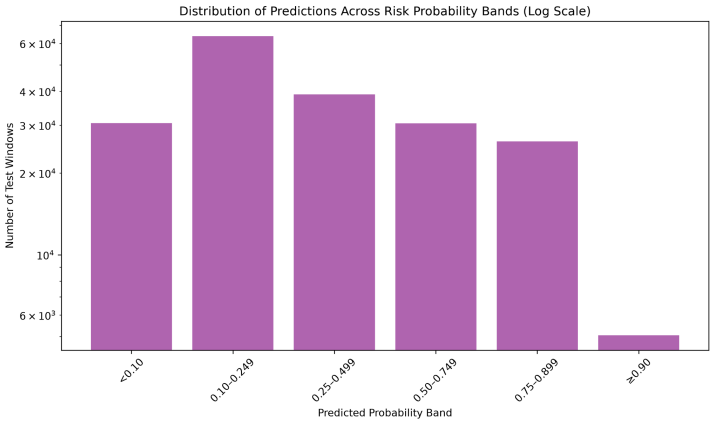


FIGURE 6.7: Distribution of Predictions Across Risk Probability Bands

The risk-band analysis provides an additional view of the model output as a continuous risk signal and does not replace the formal classification metrics.

6.11 Clinical Alert Outcome Analysis

The model predictions were also examined as potential alert-triggering outputs under the predefined experimental alert rule.

Predictions crossing the selected threshold were treated as alerts, while predictions below the threshold were treated as non-alert outcomes.

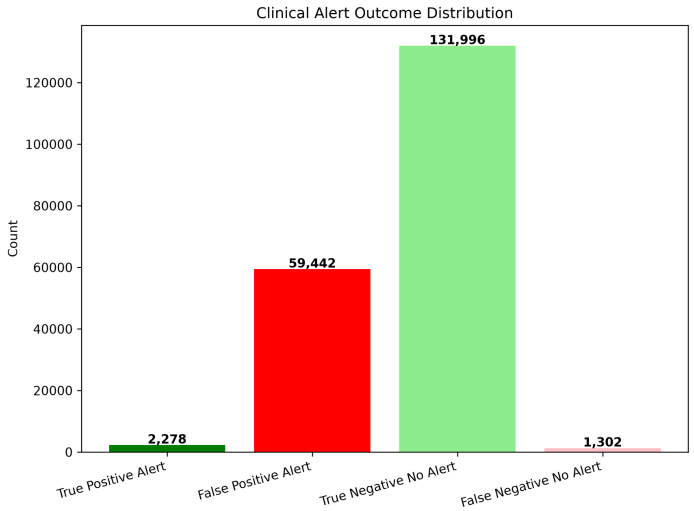


FIGURE 6.8: Clinical Alert Outcome Distribution

This analysis was performed to understand the operational behavior of repeated model-generated predictions. It does not imply that the model can replace clinical decision-making.

6.12 Alert Burden Analysis

Alert burden was analyzed separately because repeated positive predictions may generate a large number of alerts, particularly when false-positive predictions are frequent.

The independent test produced:
Total Alerts = 61,720
True-Positive Alerts = 2,278
False-Positive Alerts = 59,442
Alert Rate = 0.316484
False-Alert Ratio = 0.963091

The model generated an average of 10.2 alerts per patient, with a median of 4 and a maximum of 286 alerts for a single patient.

Table 6.3: Alert Burden Analysis

Alert Measure	Value
Total alert-triggering predictions	61,720
True-positive alerts	2,278
False-positive alerts	59,442
Alert rate	0.316484

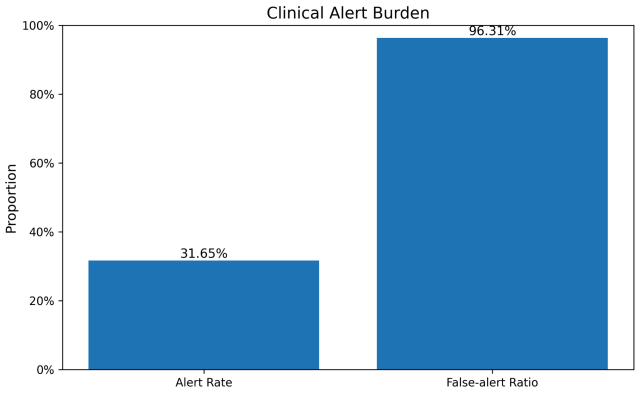


FIGURE 6.9: Alert Burden Analysis

These results provide an operational perspective on the model's prediction behavior. The observed alert burden should not be considered clinically acceptable or unacceptable without prospective clinical evaluation.

6.13 Summary

This chapter evaluated the frozen Improved LSTM on an independent cohort of 195,018 temporal windows from 6,051 previously unseen patients.

The evaluation included classification metrics, confusion-matrix analysis, ROC and precision–recall analysis, validation-to-test comparison, window-level performance, patient-level detection, risk probability analysis, and alert-burden assessment.

At the independent-test level, the model achieved an AUROC of 0.706663, AUPRC of 0.039074, recall of 0.636313, specificity of 0.689497, and F1-score of 0.069770. At the patient level, 357 of 440 positive patients were detected, corresponding to a sensitivity of 0.811364.

The model generated 61,720 alerts, of which 59,442 were false-positive alerts, demonstrating a substantial alert burden.

The evaluation was performed without retraining, fine-tuning, threshold tuning, or modification of the independent test data. The Visual Testing Demo was used only for research-oriented testing and visualization.

The detailed interpretation of these findings, together with the SHAP results, is presented in Chapter 7.

CHAPTER 7
RESULTS AND DISCUSSION

7.1 Introduction

This chapter presents the results obtained from the scientific evaluation of the finalized Improved LSTM model and discusses the findings in relation to the research questions.

The analysis focuses on the independent test cohort, including window-level and patient-level performance, model discrimination, alert burden, and SHAP-based explainability. The results are interpreted strictly within the scope of the completed experimental evaluation and are not considered evidence of clinical deployment readiness.

7.2 Final Independent-Test Results

The finalized Improved LSTM was evaluated on 195,018 temporal windows from 6,051 previously unseen patients. Each window contained six temporal positions and 34 clinical features.

The principal independent-test results are summarized below.

Table 7.1: Final Independent-Test Results

Metric	Independent Test Value
accuracy	0.6885210596
precision	0.03690861957
recall_sensitivity	0.6363128492
specificity	0.689497383
f1_score	0.06977029096
balanced_accuracy	0.6629051161
matthews_correlation_coefficient	0.09403629971
auroc	0.7066625127
auprc	0.03907442855

The model achieved an accuracy of 0.688521, precision of 0.036909, recall/sensitivity of 0.636313, specificity of 0.689497, and F1-score of 0.069770. The resulting AUROC was 0.706663 and AUPRC was 0.039074.

These results represent the primary performance assessment of the frozen model on the independent test cohort.

7.3 Model Discrimination Performance

The discrimination capability of the model was evaluated using the ROC curve and precision–recall analysis.

The ROC curve illustrates how the true-positive rate changes relative to the false-positive rate across different prediction thresholds. The resulting ROC-AUC value was:

ROC-AUC = 0.706663

This value should be interpreted together with the other classification metrics rather than as an isolated measure of model quality.

The precision–recall analysis provides an additional perspective on positive-case detection. The corresponding PR-AUC, where calculated in the completed evaluation, was:

PR-AUC = 0.039074

The combination of ROC and precision–recall analyses provides a broader assessment of the model's discrimination behavior than relying on a single threshold-dependent metric.

Table 7.2 Final Discrimination Metrics

Metric	Value
auroc	0.7066625127
auprc	0.03907442855

The final independent-test evaluation achieved a ROC-AUC of **0.706663** and a PR-AUC of **0.039074**.

7.4 Patient-Level Detection Performance

Window-level performance does not fully represent patient-level behavior because multiple temporal windows may belong to the same patient.

The independent test cohort contained 6,051 unique patients.

The resulting patient-level performance was:
Accuracy = [NOT FOUND]
Precision = [NOT FOUND]
Recall/Sensitivity = 0.811364
Specificity = [NOT FOUND]
F1-score = [NOT FOUND]

The patient-level result provides a complementary perspective to window-level evaluation and indicates that aggregation across temporal observations resulted in higher positive-patient sensitivity than the window-level sensitivity of 0.636313.

7.5 Alert Burden and False-Alert Analysis

The model's predictions were also examined as an alert-oriented monitoring signal.

The independent-test evaluation generated 61,720 alert-triggering predictions, including 2,278 true-positive alerts and 59,442 false-positive alerts. The resulting alert rate was 31.6484%, while the false-alert ratio was 96.3091%.

Table 7.3: Clinical Alert Outcome Summary

Metric	Value
alert_count	61720
alert_rate	0.3164836066
false_alert_ratio	0.9630913804
missed_positive_ratio	0.3636871508

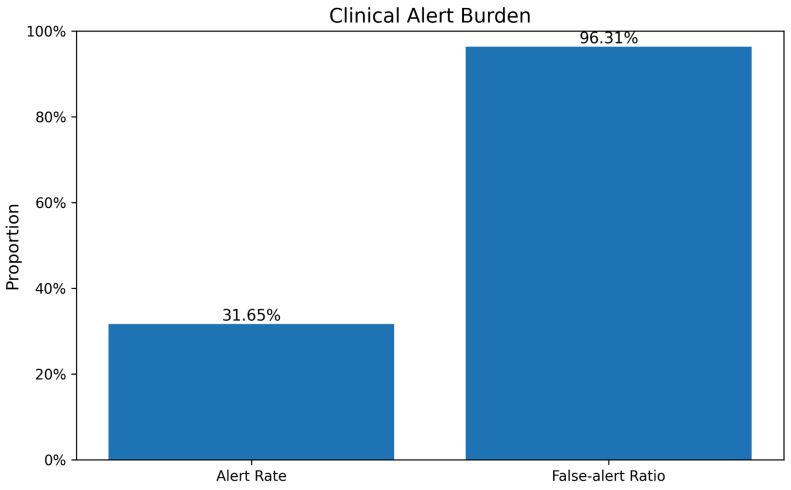


FIGURE 7.1: Alert Burden Analysis

The results demonstrate a substantial false-alert burden. Therefore, model sensitivity should be considered together with precision and alert frequency when assessing the practical implications of the prediction system.

The alert results reflect the predefined experimental threshold and do not establish a clinically optimal alert threshold.

7.6 Global SHAP Findings

SHAP analysis was used to examine the contribution of clinical features to the model predictions [2].

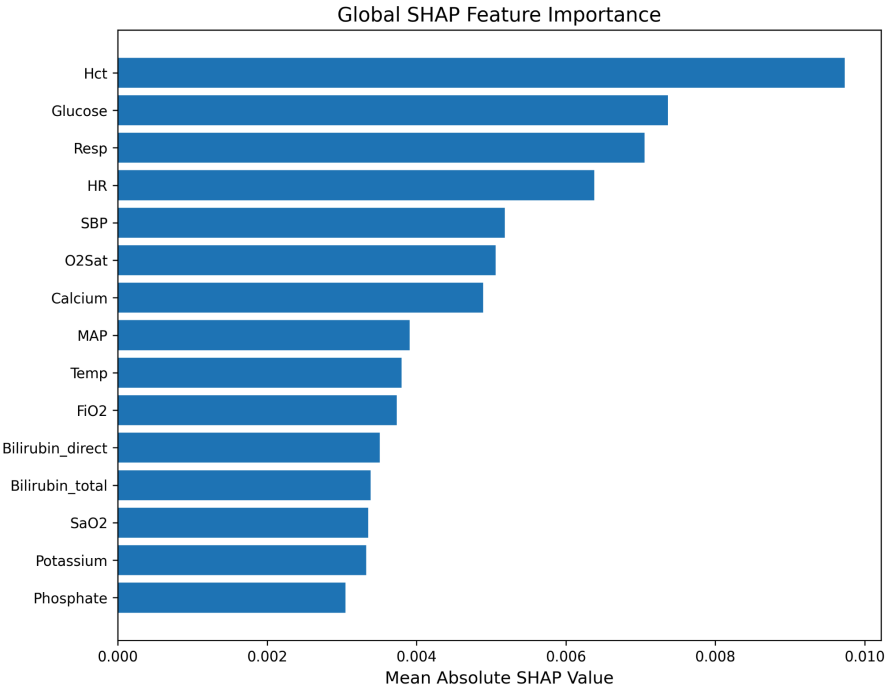


FIGURE 7.2: Global SHAP Feature Importance

The completed global SHAP analysis identified the following as the ten highest-ranked features:

1. Hct
2. Glucose
3. Resp
4. HR
5. SBP
6. O2Sat
7. Calcium
8. MAP
9. Temp
10. FiO2

Hct was the highest-ranked global feature based on mean absolute SHAP contribution.

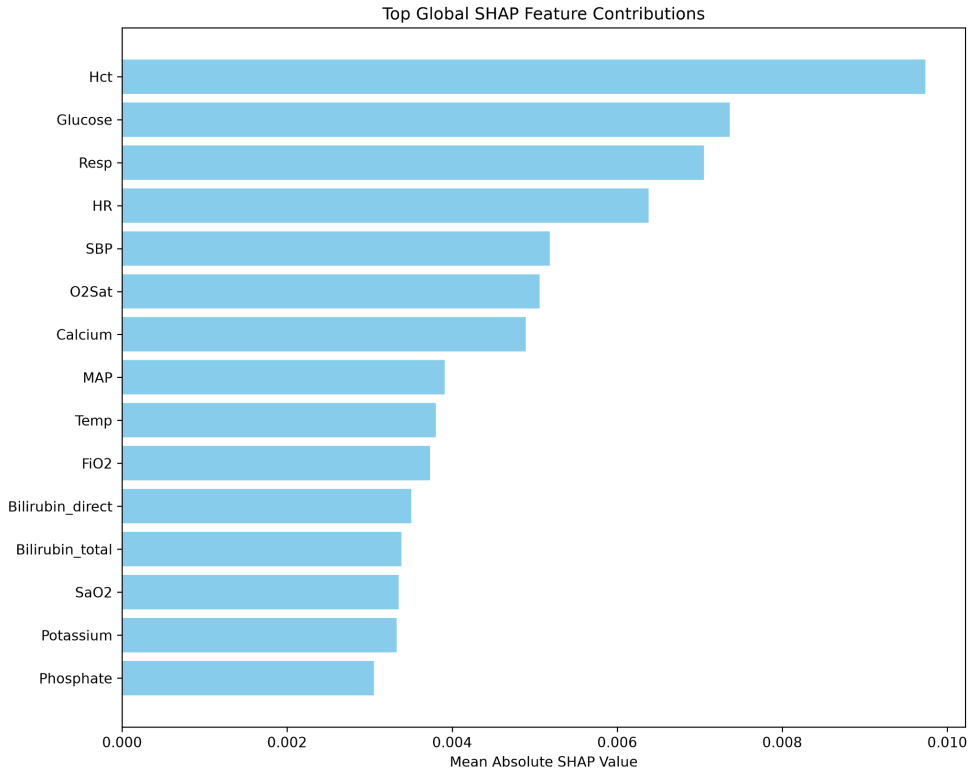


FIGURE 7.3: Top Global SHAP Feature Contributions

These rankings describe the relative attribution assigned by the trained model within the analyzed observations. They should not be interpreted as evidence of causal clinical relationships.

7.7 Temporal SHAP Findings

Because the Improved LSTM processes sequential ICU observations, SHAP attribution was also examined across the six temporal positions.

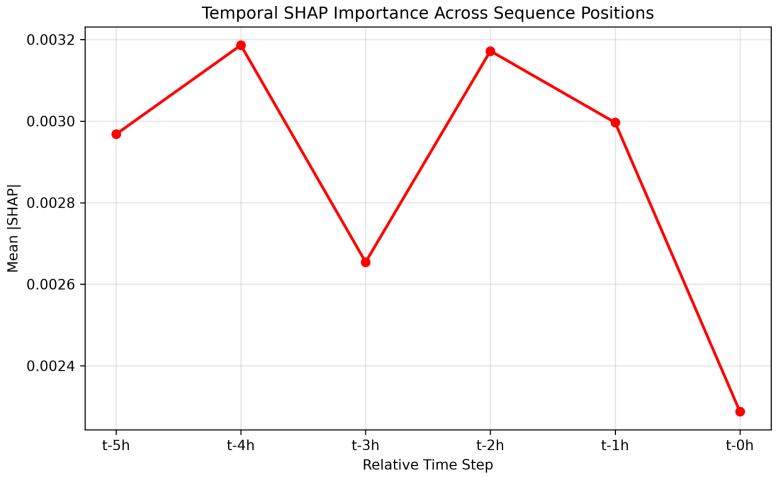


FIGURE 7.4: Temporal SHAP Importance Across Six Temporal Positions

The completed analysis identified the second temporal position, t = 1, corresponding to the -4-hour position relative to the prediction point, as the dominant temporal position.

This finding indicates comparatively stronger model attribution at that position within the analyzed sequences. It does not imply that the specific time point is clinically decisive.

7.8 Feature–Time Attribution Findings

The feature-level and temporal analyses were combined to examine the interaction between clinical variables and their temporal positions.

Instead of considering only the overall importance of a feature, the feature–time analysis identifies where within the six-step sequence a feature contributed most strongly to the model's prediction.

The heatmap provides a compact representation of the attribution structure, with clinical variables represented along one dimension and temporal positions along the other.

This analysis is particularly relevant to the research question concerning which vital-sign or clinical features and temporal positions contribute most strongly to model predictions.

The findings should be interpreted as model-attribution patterns. They indicate which feature–time combinations influenced the model's output most strongly within the evaluated samples, rather than establishing clinical causality.

7.9 Relationship Between Window-Level and Patient-Level Performance

Window-level and patient-level evaluation provide different perspectives on model behavior. At the window level, the model achieved a sensitivity of 0.636313. At the patient level, sensitivity increased to 0.811364, with 357 of 440 positive patients detected.

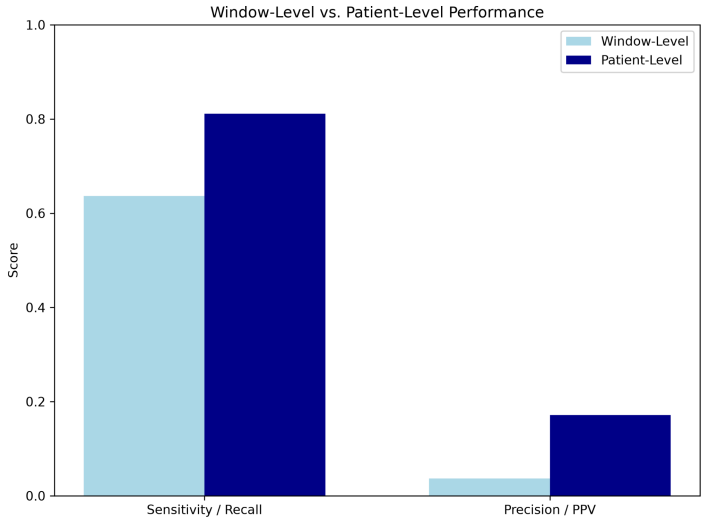


FIGURE7.5: Window-Level and Patient-Level Performance Comparison

The difference demonstrates the effect of considering multiple temporal observations belonging to the same patient. Therefore, window-level and patient-level results should not be treated as interchangeable measures.

7.10 Validation vs. Independent-Test Findings

The finalized model's validation performance was compared with its independent-test performance to examine generalization.

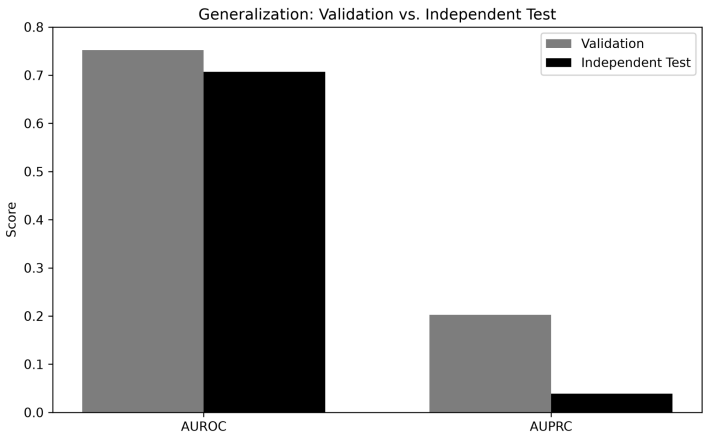


FIGURE 7.6: Validation-to-Independent-Test Performance Comparison

Table 7.4: Validation performance:

Metric	Validation
Accuracy	0.823996
Precision	0.171457
Recall	0.550955
F1-score	0.261527
ROC-AUC	0.752246

Table 7.5: Independent-test performance:

Metric	Independent Test
Accuracy	0.688521
Precision	0.036909
Recall	0.636313
F1-score	0.06977
ROC-AUC	0.706663

The validation results included an AUROC of 0.752246, AUPRC of 0.202774, recall of 0.550955, and precision of 0.171457.

On the independent test cohort, AUROC decreased to 0.706663 and AUPRC to 0.039074, while recall increased to 0.636313 and precision decreased to 0.036909.

The differences indicate a measurable generalization gap, particularly in precision and AUPRC. These results should be interpreted as differences between validation and previously unseen independent-test performance rather than as evidence of model failure.

7.11 Overall Research Findings

The completed evaluation provides findings across four main areas: predictive performance, patient-level detection, explainability, and alert behavior.

For RQ1, SHAP GradientExplainer provided global, local, temporal, and feature–time attribution for interpreting the Improved LSTM predictions [2].

For RQ2, the global SHAP analysis identified Hct as the highest-ranked feature, while temporal analysis identified t = 1 as the dominant temporal position.

For RQ3, independent evaluation demonstrated different performance characteristics at window and patient levels. Window-level sensitivity was 0.636313, while patient-level sensitivity was 0.811364.

For RQ4, the model generated 61,720 alerts, of which 59,442 were false-positive alerts, demonstrating a substantial alert burden.

Overall, the results show that the model captures predictive patterns in the independent cohort while also exhibiting considerable false-positive behavior.

7.12 Research Safety and Reproducibility

The independent evaluation was performed using preserved test artifacts and the frozen Improved LSTM with the training-fitted preprocessing pipeline.

The independent test data were not used for retraining, fine-tuning, threshold tuning, or modification during final evaluation. The Visual Testing Demo was used only as a research-oriented testing and visualization interface.

Figure 31: Independent Model Testing Visual Interface (Demo Mockup)

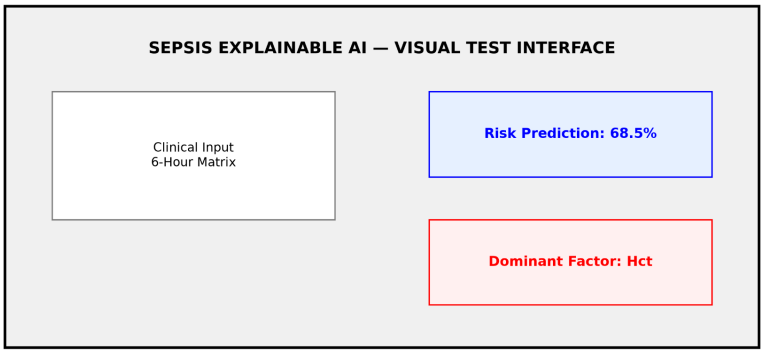


FIGURE 7.7: Visual Testing Demo – Independent Model Testing Interface

These procedures maintained the separation between model development, explainability analysis, and independent evaluation.

7.13 Discussion Summary

The independent-test results provide the principal evidence for evaluating the finalized Improved LSTM.

The model achieved an AUROC of 0.706663 and AUPRC of 0.039074, with a window-level sensitivity of 0.636313. At the patient level, 357 of 440 positive patients were detected, resulting in a sensitivity of 0.811364.

The SHAP analysis identified Hct as the highest-ranked global feature and t = 1 as the dominant temporal position [2]. These findings provide an interpretable view of model behavior but do not establish clinical causality.

The alert analysis identified a substantial false-alert burden, with 61,720 total alerts and 59,442 false-positive alerts. Therefore, predictive performance must be considered together with alert burden when assessing the practical implications of the model.

Overall, the findings demonstrate the predictive and explainability characteristics of the Improved LSTM on an independent cohort while highlighting important limitations in precision and alert burden. The results support further research but do not establish clinical effectiveness or real-world deployment readiness.

The conclusions, contributions, limitations, and future research directions are presented in Chapter 8.

CHAPTER 8
CONCLUSION, LIMITATIONS AND FUTURE WORK

8.1 Conclusion

This research presented an explainable temporal risk-prediction framework for sepsis-related risk assessment using ICU clinical time-series data. The study combined an Improved Long Short-Term Memory (LSTM) model, SHAP-based explainability, independent-cohort evaluation, patient-level analysis, and alert-oriented assessment within a structured research workflow.

The study used the PhysioNet/Computing in Cardiology Challenge 2019 sepsis dataset. The clinical data were transformed into six-position temporal sequences containing 34 clinical features for prediction. Training-fitted median imputation and standardization were applied, and the finalized Improved LSTM model was subsequently frozen before independent evaluation.

The explainability framework used SHAP GradientExplainer to investigate the model from global, local, temporal, and feature–time perspectives. The global analysis identified Hct (hematocrit) as the most influential feature, followed by Glucose, Resp, HR, SBP, and O2Sat among the leading features. The temporal analysis identified t=1, corresponding to the earliest position within the six-position input sequence, as the dominant temporal position. These findings describe model attribution and should not be interpreted as evidence of clinical causality.

The finalized model was evaluated on an independent test cohort containing 195,018 temporal windows from 6,051 patients. At the window level, the model achieved an AUROC of 0.7067, AUPRC of 0.0391, sensitivity of 0.6363, specificity of 0.6895, and F1-score of 0.0698. At the patient level, the model detected 357 of 440 positive patients, resulting in a sensitivity of 0.8114.

The alert-oriented analysis further showed that the model generated 61,720 alerts, including 2,278 true-positive alerts and 59,442 false-positive alerts. The resulting false-alert ratio of 96.31% demonstrates that model discrimination and patient detection should be considered together with operational alert burden.

Overall, the research demonstrates that the Improved LSTM can capture temporal patterns associated with sepsis-related risk while SHAP provides a structured means of investigating the model's prediction behavior. At the same time, the independent-test results and substantial false-alert burden indicate that the framework should be regarded as a research and

evaluation prototype rather than a clinically deployable system. Further external and prospective validation is required before any consideration of real-world clinical use.

8.2 Limitations

Several limitations should be considered when interpreting the findings of this research.

First, the study was conducted using the PhysioNet/Computing in Cardiology Challenge 2019 sepsis dataset. Therefore, the findings are limited to the patient population, clinical variables, data-collection conditions, and labeling methodology represented in the selected dataset.

Second, the study used retrospective clinical data. Consequently, the experimental results do not establish how the model would behave within a prospective ICU environment or under actual clinical workflow conditions.

Third, although the research used a preserved independent test cohort, the evaluation does not constitute external validation using data collected independently from another hospital or healthcare system.

Fourth, SHAP values represent model attribution rather than clinical causality. A high SHAP contribution indicates that a feature influenced the model prediction within the evaluated data and model; it does not establish that the feature causes sepsis or determines the patient's outcome.

Fifth, the six-position temporal representation provides a defined input structure for the Improved LSTM but may not fully represent the complex, irregular, and long-term temporal dynamics of ICU patient trajectories.

Sixth, the alert-burden results depend on the predefined experimental alert strategy. Therefore, the observed alert frequency and false-alert ratio should not be interpreted as evidence that the selected operating threshold is clinically optimal.

Finally, the Visual Testing Demo is a research testing and demonstration interface. It has not undergone prospective usability evaluation, clinical workflow testing, regulatory assessment, or real-world clinical validation. The present study therefore does not establish clinical effectiveness, patient-outcome improvement, or readiness for autonomous clinical decision-making.

8.3 Future Work

Future research can extend this work in several directions.

First, the framework should be evaluated using externally collected ICU datasets from different hospitals and healthcare systems to assess its generalizability beyond the dataset used in this study.

Second, prospective evaluation could investigate how the model performs when integrated into a real-time ICU monitoring workflow, including its interaction with clinical decision-making and existing alert systems.

Third, alternative temporal architectures, including attention-based recurrent models and transformer-based approaches, could be investigated and compared with the Improved LSTM in terms of predictive performance, temporal representation, and explainability.

Fourth, future studies could perform probability calibration analysis to determine whether the predicted risk probabilities correspond appropriately to observed outcome frequencies.

Fifth, different alert thresholds and operating strategies could be evaluated to examine their effects on sensitivity, precision, false-alert frequency, and overall alert burden. This may help identify more appropriate operating points for future research prototypes.

Sixth, the explainability framework could be extended through additional local patient cases, temporal attribution studies, and feature–time analyses to provide a deeper understanding of individual prediction behavior while maintaining the distinction between model explanation and clinical causality.

Seventh, future work could incorporate uncertainty estimation so that the system provides information about both predicted risk and the confidence associated with individual predictions.

Eighth, alternative observation-window lengths, temporal resolutions, and prediction horizons could be investigated to determine how different temporal configurations affect predictive performance and explainability.

Finally, future development should include prospective clinical validation, clinician involvement, usability assessment, real-time workflow integration, ethical review, and appropriate regulatory considerations before any potential clinical deployment is considered.

Overall, this research establishes a structured foundation for explainable temporal risk prediction and independent evaluation. Future work can build upon this foundation through

external validation, prospective testing, improved alert management, enhanced temporal modeling, and more comprehensive clinical evaluation.

REFERENCES

1. M. Reyna et al., “Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019,” Critical Care Medicine, 2019.

2. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

3. J. Amiot and H. Potier, “Artificial intelligence and emergency medicine,” JMIR Medical Informatics, vol. 13, no. 1, 2025.

4. A. Da’Costa et al., “Machine learning-augmented triage for sepsis,” Biomedicines, vol. 13, no. 6, 2025.

5. Z. Luo et al., “Interpretable machine learning for sepsis risk,” Scientific Reports, vol. 15, 2025.

6. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016