



# **SAFE-ICU: Advancing AI-Driven Decision Support in the Intensive Care Units**

by

*Pradeep Singh*

*(PhD19208)*

Under the Supervision of

**Prof. Tavpritesh Sethi**

IIIT-Delhi

**&**

**Prof. Rakesh Lodha**

AIIMS-Delhi

**Department of Computational Biology**

**Indraprastha Institute of Information Technology**

**NEW DELHI- 110020**

**May, 2026**



# **SAFE-ICU: Advancing AI-Driven Decision Support in the Intensive Care Units**

by

*Pradeep Singh*

*(PhD19208)*

A Thesis

Submitted in partial fulfillment of the requirements for the degree of  
**Doctor of Philosophy**

**Department of Computational Biology**

**Indraprastha Institute of Information Technology**

**NEW DELHI- 110020**

**May, 2026**

# Certificate

This is to certify that the thesis titled “**SAFE-ICU: Advancing AI-Driven Decision Support in the Intensive Care Units**” being submitted by **Pradeep Singh** to the **Indraprastha Institute of Information Technology, Delhi**, for the award of the degree of **Doctor of Philosophy**, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards, fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or in whole to any other university or institute for the award of any degree/diploma.



**Prof. Tavpritesh Sethi**  
Professor, Computational Biology  
Founding Head,  
Center of Excellence in Healthcare  
IIT Delhi, 110020



**Prof. Rakesh Lodha**  
Professor, Deptt. Of Pediatrics  
Associate Dean (Academics),  
AIIMS Delhi, 110020

# Acknowledgements

As I reach the end of my Ph.D. journey, I find myself filled with deep gratitude for the many people who have shaped this path. This milestone is not only a reflection of my academic work but a culmination of countless moments of kindness, support, encouragement, and learning. Each interaction, whether big or small, has contributed to who I am today, and I carry with me immense appreciation for all who have guided and stood by me.

First and foremost, I would like to express my deepest gratitude to my advisor, **Prof. Tavpritesh Sethi**, whose mentorship, insight, and unwavering support have been instrumental in my growth as a researcher. His patience, belief in my potential, and persistent encouragement motivated me to challenge myself and strive for excellence. I am genuinely grateful for his time, his feedback, and his commitment to nurturing both my academic and personal development.

A special note of gratitude to my external advisor, **Dr. Rakesh Lodha** at AIIMS Delhi. His mentorship played a vital role in shaping my early research interests and career trajectory. His guidance, openness to collaboration, and insightful feedback helped me navigate complex challenges and grow both intellectually and personally. I also owe a debt of thanks to the **D5-ICU/MCB-PICU team at AIIMS**, especially **Mr. Anil Kumar Sharma** and **Mr. Varun Prakash**, for creating a supportive environment that allowed me to focus intensely on my research.

I extend my sincere thanks to **Prof. Gajendra P. S. Raghava (HoD CB)** and **Prof. Ranjan Bose (Director, IIIT-Delhi)** for welcoming me into the Ph.D. program and offering me the opportunity to be a part of this incredible institution. My heartfelt appreciation also goes to **Prof. Ganesh Bagler** and **Dr. Arul Murugun**, members of my annual review and comprehensive exam committees, for their valuable guidance and encouragement. I am equally thankful to **Dr. Vibhor Kumar**, **Dr. Sanat K. Biswas**, and **Prof. Raghava** for their impactful teaching during my coursework, which laid the foundation for much of my research.

I would like to acknowledge the ever-supportive administrative team at IIIT-Delhi, **Mrs. Priti Patel**, **Mrs. Shipra Jain**, **Mrs. Anshu Dureja**, and **Mr. Raju Biswas**, for always being there to resolve our academic queries with efficiency and warmth. My sincere thanks also go to **Mr. Imran Khan**, **Mrs. Sarika**, **Mr. Mohit**, **Mrs. Vanshika**, **Dr. Prerana** and **Mr. Kapil Dev Garg** for ensuring the timely processing of my stipend. I'm grateful to the IT support staff, especially **Mr. Bhawani**, for their constant help whenever I faced technical hurdles. IIIT-Delhi has been



more than an institution; it has been my home, and I'm thankful for its world-class infrastructure and the vibrant, supportive ecosystem it fosters.

I would like to express my heartfelt gratitude to **Full Prof. Frederic Precioso**, Full Professor at Université Côte d'Azur, Laboratoire I3S, and INRIA – Maasai Team, France, for his invaluable mentorship during my internship. His guidance, expertise, and encouragement played a pivotal role in shaping my research direction and expanding my perspective.

To my seniors at IIIT-Delhi, **Dr. Aditya Nagori**, **Dr. Raghav Awasthi**, and **Dr. Ridam Pal**, thank you for guiding me with wisdom and warmth. Your advice and camaraderie made this path easier to walk. A special thanks to **Aditya** and **Raghav**, whose mentorship felt like that of elder brothers, always approachable, always insightful.

I sincerely thank my PhD colleagues **Ayushi Gupta**, **Akshaya**, **Jasmine**, **Alok**, **Manas**, and **Aakansha**, who have made significant contributions to my research endeavors. Collaborating with you enriched my experience and reminded me of the beauty of shared passion and purpose. **Akshaya**, a heartfelt thank you for constantly checking in on me and offering your support through every phase of this journey.

To friends from other labs, **Vishakha Gautam**, **Mansi Goel**, **Gayatri Panda**, and **Sukriti Sacher**, thank you for the bond we shared over shared challenges and victories. Your friendship made this Ph.D. journey more joyful and memorable. Special thanks to **Vishakha** and **Mansi**, who have been my pillars of strength; your encouragement carried me forward during the most testing times.

I'm also grateful to my collaborators **Vanshika Vats**, **Harsh Bandhey**, **Deeptanshu**, and others, whose contributions helped bring many research ideas to fruition. Working with you added depth and dimension to my work.

A heartfelt thanks to colleagues from the department, **Dr. Sumeet Patiyal**, **Dr. Anjali Dhall**, **Dr. Neelam Sharma**, **Dr. Neteesh Pandey**, **Dr. Shreya Mishra**, **Dr. Indra Prakash Jha**, **Dr. Dilraj Kaur**, **Dr. Madhu Sharma**, **Aayushi Mittal**, **Sanjay Mohanty**, **Sakshi Arora**, **Ritu Tomar**, and **Nishant Kumar**, for your constant encouragement. Your kindness during difficult times helped me stay grounded and hopeful.

I am also profoundly grateful to my friends **Sandeep Choudhary**, **Asad Anwar**, **Mohit Pipal**, **Ritik Dubey**, **Arjun Rao**, **Arun Raveenderan**, **Pramod** and **Naman Bajpai** for being my constants. Your presence has brought laughter, comfort, and perspective throughout this journey. Your unwavering support during both the good times and the challenging ones has meant the world to me, and I genuinely value the bond we share.

To my family, my parents, **Ram Kishore** and **Pushap Devi**, and my siblings, **Kuldeep Singh**, **Akash Singh**, **Akansha Singh**, and **Sheetal Singh**, your unwavering support has been my

anchor. Your love, guidance, and sacrifices have shaped the person I am today. You believed in me when it mattered most, and your constant encouragement has been my greatest strength. I am deeply grateful for all that you've done to ensure my success. This achievement is as much yours as it is mine. Thank you for always being my rock. This journey would not have been possible without you.

And finally, to everyone whose name I may have unintentionally missed, thank you. Every conversation, piece of advice, and shared moment has played a meaningful role in this journey. I am genuinely grateful to have been part of such a supportive community, and I carry your contributions with me as I move forward.

*Pradeep Singh*

### **Declaration Regarding Use of AI in Thesis:**

I acknowledge that I am fully responsible for the entire content of my thesis, including any sections assisted by online tools, including Artificial Intelligence-based tools. I accept full accountability for any violations of ethical standards in publications arising from the use of such tools.

Signature of PhD Student: *Pradeep Singh*  
Name of PhD Student: PRADEEP SINGH  
Roll Number: PhD19208

# Abstract

Integrating artificial intelligence into healthcare presents an unprecedented opportunity to address critical challenges in Intensive Care Units. The objective of this thesis is to create, validate, and deploy AI models to increase safety through the development of early warning systems for drug-drug interactions and predicting clinical decompensation of patients.

Our first objective is creation of a data lake involving multimodal data, which combines vital signs, laboratory tests, treatment histories, patient files, and thermographic videos; not just images. The combination of structured and unstructured data contained in the datasets offers unmatched insight into intensive care unit patients, thereby enhancing the overall quality of research. The datasets used for prediction and diagnostics are pre-processed through various techniques like feature extraction, temporal alignment, etc.

The second Objective aims at the creation of actionable, predictive, and management tools for sophisticated critical care of ICU patients. For example, *ThermalShockNet*, which was created for pediatric ICU settings and utilizes thermal videos and patient vitals to identify subtle physiological changes indicative of hemodynamic shock. This model uses data from thermal frames and vital signs to make predictions. When combined using deep learning, these predictions can reveal the onset of shock before it appears in a patient. *ThermoGnose*, a pipeline developed specifically for the early prediction of hypothermia, uses non-linear features drawn from patient monitoring data to estimate hypothermia onset and provide vital lead time for clinical intervention. *SIgnose* predicts shock risk by analyzing vital sign trends. When combined, this variety of approaches highlight how AI can address complex challenges in the ICU.

The third Objective was to improve drug safety in the ICU. For this, the *ContraIndicator* tool was created. NLP processes the treatment chart and line listing of drugs to find probable adverse drug-drug interactions. The ContraIndicator insights assist clinicians in minimizing the chance of adverse reactions.

The end Objective was *SafeICU*, a modular and explainable AI-driven real-time platform for the pediatric ICU. It brings together early warning systems along with safety and infection surveillance modules, which include vital signs, thermal imaging, medication, and microbiology results. At SafeICU, they help flag unusual Shock Index values in many ICUs. The *ContraIndicator* is on the lookout for drug interactions through the examination of ICU prescriptions. *AMRtrack* monitors the emergence of antimicrobial resistance by processing

laboratory reports, mapping trend data, and informing infection control planning. When these modules are together, real-time alerts, risk scores, and infection monitoring are delivered into the ICU workflow.

**Overall**, this thesis provides a clinically oriented framework for using AI in ICUs. It shows how models that can be multimodal, explainable, and deployable work together in real time to assist clinical decision making in high-acuity clinical settings. These advances create a pathway to the secure and scalable exploitation of AI in critical care and, eventually, precision medicine in health care that are rich in data.

# INDEX

|                                                                       |       |
|-----------------------------------------------------------------------|-------|
| Certificate .....                                                     | i     |
| Acknowledgements .....                                                | ii    |
| Abstract .....                                                        | vi    |
| List of Figures .....                                                 | xviii |
| List of Tables .....                                                  | xxi   |
| List of Publications .....                                            | xxiii |
| Abbreviations .....                                                   | xxv   |
| Chapter-1 .....                                                       | 1     |
| Introduction .....                                                    | 1     |
| 1.1 Clinical Decision Support in the Intensive Care Environment ..... | 1     |
| 1.2 AI for Decision Support in the High-Risk ICU Settings .....       | 2     |
| 1.3 Types of Data in ICU .....                                        | 3     |
| 1. Physiological Time-Series Data .....                               | 3     |
| 2. Laboratory and Biochemical Data .....                              | 4     |
| 3. Medication Administration Records .....                            | 4     |
| 4. Unstructured Clinical Text .....                                   | 4     |
| 5. Medical Imaging .....                                              | 4     |
| 6. Microbiology Data .....                                            | 5     |
| 7. Multicenter Benchmark Datasets .....                               | 5     |

|                                                                                   |    |
|-----------------------------------------------------------------------------------|----|
| 1.4 Problem Statement .....                                                       | 5  |
| 1.5 Development of Models .....                                                   | 6  |
| 1.6 Validation of Models .....                                                    | 6  |
| 1.7 Deployment of Models .....                                                    | 7  |
| 1.8 Multi-centric Validation of Models .....                                      | 7  |
| 1.9 Resource Considerations and the Way Ahead for Resource-Limited Settings ..... | 8  |
| 1.10 Foundation Models in Clinical AI .....                                       | 9  |
| 1.11 Limitations and Challenges of AI in the ICU .....                            | 10 |
| A. Limitations of AI in Critical Care .....                                       | 10 |
| B. Failure of AI in ICU Settings .....                                            | 10 |
| C. Biases in Data and Algorithmic Fairness .....                                  | 11 |
| D. Human Factors and Clinical Integration .....                                   | 11 |
| E. Ethical and Regulatory Considerations .....                                    | 12 |
| 1.12 What Does the Thesis Address .....                                           | 12 |
| 1.13 Objectives and Structure of the Thesis .....                                 | 13 |
| Chapter-2 .....                                                                   | 16 |
| Background and Related Works .....                                                | 16 |
| 2.1 Introduction: Opportunities and Challenges for AI in ICUs .....               | 16 |
| 2.1.1 Clinical AI for Prediction and Monitoring .....                             | 17 |
| 2.1.2 Reinforcement Learning for Treatment Optimization .....                     | 18 |

|                                                                                       |    |
|---------------------------------------------------------------------------------------|----|
| 2.1.3 Usability and Adoption of AI in Clinical Workflows .....                        | 18 |
| 2.1.4 Generalizability, Multimodality, and Dataset Shift .....                        | 19 |
| 2.1.5 Drug–Drug Interaction (DDI) Detection and Prescribing Safety .....              | 19 |
| 2.1.6 Antimicrobial Resistance (AMR) Surveillance and Clinical Decision Support ..... | 20 |
| 2.2 Data-Centric Challenges in ICU AI .....                                           | 20 |
| 2.2.1 Data Heterogeneity and Interoperability .....                                   | 20 |
| 2.2.2 Bias and Underrepresentation in Clinical Data .....                             | 21 |
| 2.2.3 Temporal Data Drift and Informative Missingness .....                           | 21 |
| 2.2.4 Small Sample Size and Domain Scarcity in Pediatric ICUs .....                   | 22 |
| 2.3 Model-Centric Challenges .....                                                    | 23 |
| 2.3.1 Black-Box Behavior and Lack of Interpretability .....                           | 23 |
| 2.3.2 Model Drift and Updating in Clinical Environments .....                         | 23 |
| 2.3.3 Fragility and Adversarial Sensitivity .....                                     | 24 |
| 2.3.4 Overfitting and Lack of Generalization .....                                    | 24 |
| 2.3.5 Model Complexity vs. Clinical Utility .....                                     | 25 |
| 2.4 Validation and Evaluation Challenges .....                                        | 25 |
| 2.4.1 Single-Center Overfitting and Lack of External Validation .....                 | 26 |
| 2.4.2 Ambiguity in Clinical Ground Truths .....                                       | 26 |
| 2.4.3 Lack of Prospective and Real-Time Evaluation .....                              | 26 |
| 2.4.4 Evaluation Metrics Beyond Accuracy .....                                        | 26 |



|                                                                                        |    |
|----------------------------------------------------------------------------------------|----|
| 2.4.5 Interpretability and Clinician Validation .....                                  | 27 |
| 2.5 Deployment and Operationalization Challenges .....                                 | 28 |
| 2.5.1 Infrastructure and EHR Integration .....                                         | 28 |
| 2.5.2 Resource Constraints in Low-Resource Settings .....                              | 28 |
| 2.5.3 Resource Constraints in Low-Resource Settings .....                              | 28 |
| 2.5.4 Legal and Regulatory Challenges .....                                            | 29 |
| 2.5.5 Clinician Usability and Human-Centered Design .....                              | 29 |
| 2.6 Real-World Evaluation Metrics .....                                                | 29 |
| 2.6.1 Clinical Effectiveness and Patient Outcomes .....                                | 30 |
| 2.6.2 Economic Evaluation and Cost-Benefit Analysis .....                              | 30 |
| 2.6.3 Safety and Risk Management .....                                                 | 30 |
| 2.6.4 Capacity, Training, and Organizational Readiness .....                           | 30 |
| 2.6.5 Regulatory Evaluation Readiness .....                                            | 31 |
| 2.7 Research Gaps and Thesis Contributions .....                                       | 32 |
| 2.7.1 Identified Research Gaps .....                                                   | 32 |
| 2.7.2 SAFE-ICU: Contributions and Solutions .....                                      | 32 |
| Chapter-3 .....                                                                        | 34 |
| Construction of the SAFE-ICU Data Lake: Multimodal Integration and Preprocessing ..... | 34 |

|                                                                                      |    |
|--------------------------------------------------------------------------------------|----|
| 3.1 Introduction .....                                                               | 34 |
| 3.2 Data Sources and Modalities .....                                                | 34 |
| 3.3 Data Ingestion and Storage Architecture .....                                    | 35 |
| 3.3.1 Vitals and Monitoring Data .....                                               | 36 |
| Step1: Data Transfer from Monitors to CMS (Local Intranet) .....                     | 36 |
| Step2: Pulling Vitals from CMS using the SAFE-ICU Client-Server                      |    |
| Pipeline .....                                                                       | 36 |
| 3.3.1.1 Vitals Data Preprocessing: Unsolicited CMS Data Stream .....                 | 37 |
| 3.3.2 Laboratory Test Results .....                                                  | 38 |
| 3.3.3 Prescription and Medication Data .....                                         | 39 |
| 3.3.4 Thermal Imaging Videos .....                                                   | 41 |
| 3.3.5 Microbiology Reports .....                                                     | 42 |
| Chapter-4 .....                                                                      | 45 |
| Predictive Modeling in Intensive Care .....                                          | 45 |
| 4.1 Early prediction of hypothermia in pediatric intensive care units using ML ..... | 45 |
| 4.1.1 Introduction .....                                                             | 45 |
| 4.1.2 Methods .....                                                                  | 47 |
| 4.1.2.1 Study data and pre-processing .....                                          | 47 |
| 4.1.2.1.1 Physiological vitals records .....                                         | 47 |
| 4.1.2.1.2 Treatment-charts .....                                                     | 47 |
| 4.1.2.2 Imputation .....                                                             | 47 |
| 4.1.2.3 Data Preprocessing and cohort-generation .....                               | 48 |

|                                                                                                                           |    |
|---------------------------------------------------------------------------------------------------------------------------|----|
| 4.1.2.3.1 Cohort construction .....                                                                                       | 48 |
| 4.1.2.3.2 Data pre-processing and scoring .....                                                                           | 48 |
| 4.1.2.4 Model Development and Evaluation .....                                                                            | 49 |
| 4.1.2.5 Shapley additive exPlanations value analysis for model<br>interpretability .....                                  | 51 |
| 4.1.3 Results .....                                                                                                       | 51 |
| 4.1.3.1 Data Characteristics .....                                                                                        | 51 |
| 4.1.3.2 Model performance on Time-series features .....                                                                   | 52 |
| 4.1.3.3 Model Interpretability Analysis .....                                                                             | 54 |
| 4.1.3.4 Prospective Validation of Model .....                                                                             | 58 |
| 4.1.4 Discussion .....                                                                                                    | 59 |
| 4.2 Early Prediction of Hemodynamic Shock in Pediatric Intensive Care Units With Deep<br>Learning on Thermal Videos ..... | 61 |
| 4.2.1 Introduction .....                                                                                                  | 61 |
| 4.2.2 Methods .....                                                                                                       | 63 |
| 4.2.2.1 Data Collection and Study Design .....                                                                            | 64 |
| 4.2.2.2 Classification Into Covered and Uncovered .....                                                                   | 66 |
| 4.2.2.3 Multiple Person Detection .....                                                                                   | 67 |
| 4.2.2.4 Segmentation and CPD Extraction From Abdomen and Feet ...                                                         | 67 |
| 4.2.2.5 LSTM Time Series Sequence Classification .....                                                                    | 68 |

|                                                                                                                                                    |    |
|----------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 4.2.2.6 Outcome Variable Binary Shock Index .....                                                                                                  | 69 |
| 4.2.2.7 Model Evaluation .....                                                                                                                     | 69 |
| 4.2.3. Results .....                                                                                                                               | 70 |
| 4.2.3.1 Patient Characteristics, Preprocessing, and Cohort Building ....                                                                           | 70 |
| 4.2.3.2 Segmentation of Abdomen and Feet Achieved a Total Dice Loss of<br>0.0391 Using ResUNet .....                                               | 70 |
| 4.2.3.3 The LSTM Model Was Found to be the Best Performing .....                                                                                   | 71 |
| 4.2.3.4 Evaluation of the LSTM Model at Different Lead Times Reveals<br>That CPD Increases the Model Performance .....                             | 71 |
| 4.2.4 Discussion .....                                                                                                                             | 73 |
| 4.3 Beyond Language Models: Clinically Validated SIgnose Outperforms LLMs in<br>Generalizable Early Shock Index Prediction in Intensive Care ..... | 77 |
| 4.3.1 Introduction .....                                                                                                                           | 77 |
| 4.3.2 Methods .....                                                                                                                                | 79 |
| 4.3.2.1 Dataset description & pre-processing .....                                                                                                 | 79 |
| 4.3.2.2 Cohort Construction .....                                                                                                                  | 80 |
| 4.3.2.3 Pre-processing and Scoring .....                                                                                                           | 80 |
| 4.3.2.4 Engineering Explainable Feature and Feature Selection .....                                                                                | 81 |
| 4.3.2.5 Model Development, Evaluation, and Interpretability .....                                                                                  | 81 |
| 4.3.2.6 SHAP value computation for model interpretability .....                                                                                    | 86 |
| 4.3.2.7 Model generalization & prospective validation .....                                                                                        | 86 |

|                                                                                                               |     |
|---------------------------------------------------------------------------------------------------------------|-----|
| 4.3.3 Results .....                                                                                           | 87  |
| 4.3.3.1 Data characteristics .....                                                                            | 87  |
| 4.3.3.2 Model performance on eICU data .....                                                                  | 91  |
| 4.3.3.3 Assessing LLM's predictive Performance on eICU Data .....                                             | 92  |
| 4.3.3.4 Model performance for different lead times, age groups, unit types,<br>and time since admission ..... | 93  |
| 4.3.3.5 Performance of external and prospective validation .....                                              | 94  |
| 4.3.3.6 Model Interpretability Analysis .....                                                                 | 98  |
| 4.3.3.7 Model validation across several diagnosis categories .....                                            | 99  |
| 4.3.3.8 Assessing LLM's predictive Performance on eICU Data .....                                             | 101 |
| 4.3.4 Discussion .....                                                                                        | 101 |
| Chapter-5 .....                                                                                               | 105 |
| Drug Safety and Interaction Analysis in the ICU .....                                                         | 105 |
| 5.1 Potential Drug-Drug Interactions in the Pediatric Intensive Care Unit .....                               | 105 |
| 5.1.1 Introduction .....                                                                                      | 105 |
| 5.1.2 Methods .....                                                                                           | 106 |
| 5.1.2.1 Study Setting .....                                                                                   | 106 |
| 5.1.2.2 Objectives and outcome measures .....                                                                 | 107 |
| 5.1.2.3 Variables collected, Procedure, and other measures .....                                              | 107 |
| 5.1.2.4 Population, Sample, and Patient Characteristics .....                                                 | 109 |

|                                                                                          |     |
|------------------------------------------------------------------------------------------|-----|
| 5.1.2.5 Statistical Analysis .....                                                       | 109 |
| 5.1.2.6 ContraIndicator .....                                                            | 110 |
| 5.1.3 Results .....                                                                      | 110 |
| 5.1.3.1 Baseline characteristics .....                                                   | 110 |
| 5.1.3.2 Frequency based on major, moderate, and minor pDDI .....                         | 113 |
| 5.1.3.3 Predictors of pDDI .....                                                         | 114 |
| 5.1.3.4 Predictors of length of stay .....                                               | 116 |
| 5.1.4 Discussion .....                                                                   | 117 |
| Chapter 6 .....                                                                          | 121 |
| SAFE-ICU: An Integrated AI-Enabled Decision Support Platform for Intensive Care Units .. | 121 |
| 6.1 Introduction .....                                                                   | 121 |
| 6.2 Motivation .....                                                                     | 121 |
| 6.3 System Architecture .....                                                            | 121 |
| 6.4 Module 1: Hemodynamic Shock Prediction .....                                         | 127 |
| 6.5 Module 2: Drug-Drug Interaction Monitoring .....                                     | 127 |
| 6.6 Module 3: AMR Surveillance .....                                                     | 129 |
| 6.7 SAFE-ICU User Interface .....                                                        | 134 |
| Chapter-7 .....                                                                          | 136 |
| Conclusion and Future Work .....                                                         | 136 |

|                                                           |     |
|-----------------------------------------------------------|-----|
| 7.1 Conclusion .....                                      | 136 |
| 7.2 Key Findings .....                                    | 137 |
| 7.2.1 Predictive Modeling of Clinical Deterioration ..... | 137 |
| 7.2.2 Drug Safety Analytics via Clinical NLP .....        | 138 |
| 7.2.3 AMR Surveillance Integration .....                  | 138 |
| 7.2.4 Summary of Key Contributions .....                  | 138 |
| 7.3 Limitations .....                                     | 139 |
| 7.4 Future Work .....                                     | 140 |
| 7.4.1 Clinical Deployment and Usability Testing .....     | 140 |
| 7.4.2 Polypharmacy Modeling with LLMs .....               | 140 |
| 7.4.3 Real-Time Resistance Forecasting .....              | 140 |
| 7.4.4 Multilingual Clinical NLP .....                     | 140 |
| 7.4.5 Fairness and Bias Audits .....                      | 140 |
| 7.4.6 Integration with Hospital Information Systems ..... | 140 |
| 7.5 Final Remarks .....                                   | 141 |
| References .....                                          | 142 |

# List of Figures

|                                                                                                                |    |
|----------------------------------------------------------------------------------------------------------------|----|
| 1.1 Total health expenditure per capita .....                                                                  | 1  |
| 1.2 Applications of Artificial Intelligence in Healthcare .....                                                | 14 |
| 3.1 Data Sources and Modalities in the SAFE-ICU Data Lake .....                                                | 35 |
| 3.2 Raw Format Illustration of ICU Vital Signs Data .....                                                      | 37 |
| 3.3 Cleaned and Structured Vitals Format from SAFE-ICU Pipeline .....                                          | 38 |
| 3.4 Raw Format Example of Laboratory Test .....                                                                | 39 |
| 3.5 Raw Format Example of Prescription and Medication Data .....                                               | 40 |
| 3.6 Structured Output of Prescription Data After NLP Processing .....                                          | 41 |
| 3.7 Thermal Video Frame Captured for ICU Patient Monitoring .....                                              | 41 |
| 3.8 Raw Format of Microbiology Reports from the ICU Laboratory System .....                                    | 43 |
| 4.1 Physiological variables for the prediction of Hypothermia in pediatric-ICU .....                           | 49 |
| 4.2 Overview of ThermoGnose pipeline .....                                                                     | 50 |
| 4.3 Length of stay is associated with the length of stay with hypothermia .....                                | 52 |
| 4.4 Model performance on Time-series features .....                                                            | 53 |
| 4.5 Detailed plots of individual top predictors for future Hypothermia .....                                   | 55 |
| 4.6 SHAP values of the top 20 Nonlinear features generated from the five-fold test set on pediatric data ..... | 58 |
| 4.7 Prospective Validation of Model .....                                                                      | 59 |
| 4.8 Shock prediction pipeline .....                                                                            | 63 |
| 4.9 Hemodynamic shock prediction with Deep Learning on Thermal Videos .....                                    | 66 |



|                                                                                                                                                                           |     |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 4.10 Quality assessment of models .....                                                                                                                                   | 71  |
| 4.11 Quality evaluation on the LSTM model .....                                                                                                                           | 72  |
| 4.12 Summary of the pipeline for generalizing the prediction of hemodynamic instability.....                                                                              | 78  |
| 4.13 Each 30-minute time window is labeled as shock-positive or shock-negative. A 420-minute observation window is used to predict shock status up to 8 hours ahead ..... | 81  |
| 4.14 Data Preparation and Cohort Formation .....                                                                                                                          | 88  |
| 4.15 Model training on TS-features .....                                                                                                                                  | 92  |
| 4.16 SIgnose with oversampling of the minority class, and random undersampling of the majority class .....                                                                | 92  |
| 4.17 Model performance on eICU data .....                                                                                                                                 | 93  |
| 4.18: SIgnose Performance .....                                                                                                                                           | 97  |
| 4.19 Model Interpretability Analysis .....                                                                                                                                | 99  |
| 5.1 Flow chart of potential drug–drug interactions frequency .....                                                                                                        | 111 |
| 5.2 Drug Pairs, Frequency based on Major pDDI and Moderate pDDI .....                                                                                                     | 114 |
| 6.1 SAFE-ICU Login Interface .....                                                                                                                                        | 122 |
| 6.2 SAFE-ICU Dashboard Interface .....                                                                                                                                    | 123 |
| 6.3 Patient Monitoring View in SAFE-ICU Dashboard .....                                                                                                                   | 126 |
| 6.4 Shock Risk Prediction Output in SAFE-ICU .....                                                                                                                        | 127 |
| 6.5 ContraIndicator Module in SAFE-ICU .....                                                                                                                              | 128 |
| 6.6 Drug Interaction Output in ContraIndicator (SAFE-ICU) .....                                                                                                           | 129 |
| 6.7 Login Interface of the Antimicrobial Stewardship Platform .....                                                                                                       | 130 |
| 6.8 Sample Type Distribution in Pediatric ICU AMR Dataset .....                                                                                                           | 131 |
| 6.9 Top 15 Isolated Organisms in Pediatric ICU AMR Dataset .....                                                                                                          | 132 |

|                                                   |     |
|---------------------------------------------------|-----|
| 6.10 Antibigram Query Interface in AMRTrack ..... | 133 |
| 6.11 Antibigram Visualization in AMRTrack .....   | 133 |

# List of Tables

|                                                                                                                                              |    |
|----------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.1 Summary table of model-centric challenges and SAFE-ICU solutions .....                                                                   | 25 |
| 2.2 Summary table validation challenges and SAFE-ICU solutions .....                                                                         | 27 |
| 2.3 Summary Table for deployment barriers and SAFE-ICU innovations .....                                                                     | 29 |
| 2.4 Summary table of real-world evaluation dimensions .....                                                                                  | 31 |
| 2.5 Addressing Existing Research Gaps Through SAFE-ICU .....                                                                                 | 33 |
| 3.1 Summary Table of Multimodal Clinical Data .....                                                                                          | 44 |
| 4.1 Characteristics of SafeICU-cohort observational window captured (30min-4 hours) earlier than Hypothermia .....                           | 51 |
| 4.2 Shows the results of the proposed models for predicting hypothermia/non-hypothermia based on physiological variables .....               | 53 |
| 4.3 Important non-linear features .....                                                                                                      | 54 |
| 4.4 Results of the prospective validation of a model for predicting hypothermia/non-hypothermia based on physiological variables .....       | 59 |
| 4.5 Cohort characteristics and statistical significance of control (non-shock) vs. affected (shock) classes .....                            | 70 |
| 4.6 Performance of the proposed model predicting the presence of shock/non-shock using automated CPD and heart rate .....                    | 72 |
| 4.7 Comparison between the results of machine learning techniques for hemodynamic shock prediction .....                                     | 75 |
| 4.8 Performance of the proposed model predicting the presence of Shock/Non-shock using automated CPD, heart rate, and respiratory rate ..... | 75 |
| 4.9 Non-linear time-series features Description .....                                                                                        | 83 |
| 4.10 Characteristics of the eICU cohort observational window captured 4 hours earlier than the abnormal shock index .....                    | 89 |
| 4.11 Characteristics of the MIMIC cohort .....                                                                                               | 90 |

|                                                                                                           |     |
|-----------------------------------------------------------------------------------------------------------|-----|
| 4.12 Characteristics of the safeICU cohort .....                                                          | 90  |
| 4.13 Association of mortality with the first three days of shock index status .....                       | 91  |
| 4.14 Association of LOS with the first three days of shock index status .....                             | 91  |
| 4.15 LLaMA Model Performance on eICU (5-Fold Cross-Validation, Mean $\pm$ SD) .....                       | 93  |
| 4.16 Impact of Fine-Tuning on Model Generalization using LLaMA .....                                      | 95  |
| 4.17 Model performance across several diagnosis categories .....                                          | 100 |
| 4.18 LLaMA Model Performance on eICU (5-Fold Cross-Validation, Mean $\pm$ SD) .....                       | 101 |
| 5.1 Comparison of DDI pairs across different databases .....                                              | 108 |
| 5.2 Patient characteristics based on exposure to pDDI .....                                               | 112 |
| 5.3 The frequency of drug interactions in PICU .....                                                      | 113 |
| 5.4 Analysis of Univariate Logistic regression for at least one major, one moderate, and all pDDI .....   | 115 |
| 5.5 Analysis of Multivariate Logistic regression for at least one major, one moderate, and all pDDI ..... | 116 |
| 5.6 Analysis of Univariate Linear regression for length of stay .....                                     | 117 |
| 6.1 SAFE-ICU is a web-based application developed using the following technology stack ...                | 122 |
| 6.2 Key Characteristics of the Hemodynamic Shock Prediction Module .....                                  | 127 |
| 6.3 Key Characteristics of the Drug–Drug Interaction Monitoring Module .....                              | 128 |
| 6.4 Key Characteristics of the AMR Module .....                                                           | 130 |
| 7.1 Summary of Key Contributions .....                                                                    | 138 |

# List of Publications

## Publications and Preprints

Singh, P., Nagori, A., Lodha, R., & Sethi, T. Early prediction of hypothermia in pediatric intensive care units using machine learning. *Frontiers in Physiology*, 13, 921884. <https://doi.org/10.3389/fphys.2022.921884>

Singh, P., Vats, V., Nagori, A. et al. Early prediction of hemodynamic shock in pediatric intensive care units with deep learning on thermal videos. *Frontiers in Physiology*, 13, 872411. <https://doi.org/10.3389/fphys.2022.862411>

Singh, P., Devadiga, A., Pal, R. et al. ContraIndicator: A Natural Language Processing-Based Approach to Potential Drug–Drug Interaction Detection in Pediatric Intensive Care. *Clin Drug Investig* 46, 331–341 (2026). <https://doi.org/10.1007/s40261-026-01529-z>

Nagori, A., Singh, P., et al. A Prospectively Validated Generalizable Model for Outcome Prognostication Using Shock Index in Intensive Care Units. (In Review)

## Other Publications and Preprints

Ansari, M. S., Singh, P., et al. Spectral pattern of cough sounds as an aid to diagnosis in children with common respiratory diseases. *International Journal of Contemporary Pediatrics*. <https://doi.org/10.18203/2349-3291.ijcp20250770>

Awasthi, R., Pal, R., Singh, P., et al. (2020). CovidNLP: A Web Application for Distilling Systemic Implications of COVID-19 Pandemic with Natural Language Processing. *medRxiv*. <https://doi.org/10.1101/2020.04.25.20079129>

Kaur, A., Prajapati, S., Singh, P., et al. (2022). Heart rate assessment in a pediatric ICU with non-contact infrared thermography and machine learning. *medRxiv*. <https://doi.org/10.1101/2022.11.18.2228244>

Akshaya Devadiga, Pradeep Singh, Jhuma Sankar, Rakesh Lodha, & Tavpritesh Sethi(2026). Generalizing intensive care AI across time scales in resource-limited settings. *medrxiv*. <https://doi.org/10.64898/2026.04.23.26351588>

Kainth D, Gupta A, Singh P, Prakash S, Thukral A, Deorari A, et al. A machine learning model for prediction of early-onset neonatal sepsis in low-income and middle-income countries: development and validation study. *BMJ Paediatrics Open*. 2026;10:e003561. <https://doi.org/10.1136/bmjpo-2025-003561>

Ridam Pal, Shaswat Patel, Akshala Bhatnagar, Hardik Garg, Pradeep Singh, et al. ShockModes: A Multimodal Model for Prognosticating Intensive Care Outcomes from Physician Notes and Vitals. (In Review)

Singh, P., Devadiga, A., et al. Rethinking critical care with artificial intelligence: A review of opportunities, challenges, and the road ahead. (In Review)

# Abbreviations

|                   |                                                                       |
|-------------------|-----------------------------------------------------------------------|
| <b>ABP</b>        | Arterial Blood Pressure                                               |
| <b>AI</b>         | Artificial Intelligence                                               |
| <b>AIIMS</b>      | All India Institute of Medical Sciences                               |
| <b>AKI</b>        | Acute Kidney Injury                                                   |
| <b>AMR</b>        | Antimicrobial Resistance                                              |
| <b>ASHA</b>       | Accredited Social Health Activist                                     |
| <b>AST</b>        | Antibiotic Susceptibility Testing                                     |
| <b>AUC</b>        | Area Under the Curve                                                  |
| <b>AUC-ROC</b>    | Area Under the Receiver Operating Characteristic Curve                |
| <b>AUPRC</b>      | Under the Precision Recall Curve                                      |
| <b>BERT</b>       | Bidirectional Encoder Representations from Transformers               |
| <b>CDSS</b>       | Clinical Decision Support Systems                                     |
| <b>CMS</b>        | Central Monitoring System                                             |
| <b>CNNs</b>       | Convolutional Neural Networks                                         |
| <b>CONSORT-AI</b> | Consolidated Standards of Reporting Trials in Artificial Intelligence |
| <b>CPD</b>        | Center-to-Peripheral Difference                                       |
| <b>CPD</b>        | Centre-to-Peripheral Difference                                       |
| <b>CPOE</b>       | Computerized Physician Order Entry                                    |
| <b>CRP</b>        | C-reactive Protein                                                    |
| <b>CT</b>         | Computed Tomography                                                   |

|                |                                                           |
|----------------|-----------------------------------------------------------|
| <b>CV</b>      | Computer Vision                                           |
| <b>DDI</b>     | Drug-Drug Interaction                                     |
| <b>Dia-bp</b>  | Diastolic Blood Pressure                                  |
| <b>DL</b>      | Deep Learning                                             |
| <b>ECG</b>     | ElectroCardioGram                                         |
| <b>EHRs</b>    | Electronic Health Records                                 |
| <b>eICU</b>    | Electronic Intensive Care Unit                            |
| <b>EQUATOR</b> | Enhancing the QUALity and Transparency Of Health Research |
| <b>EVM</b>     | Eulerian Video Magnification                              |
| <b>FDA</b>     | Food and Drug Administration                              |
| <b>FHIR</b>    | Fast Healthcare Interoperability Resources                |
| <b>FPS</b>     | frames per second                                         |
| <b>GBM</b>     | Gradient Boosting Machine                                 |
| <b>GDP</b>     | Gross Domestic Product                                    |
| <b>GDPR</b>    | General Data Protection Regulation                        |
| <b>GNA</b>     | Generative Adversarial Network                            |
| <b>GPT</b>     | Generative Pre-trained Transformer                        |
| <b>GPU</b>     | Graphical Processing Unit                                 |
| <b>GRU-D</b>   | Gated Recurrent Unit with Decay                           |
| <b>GRU</b>     | Gated Recurrent Unit                                      |
| <b>HIPAA</b>   | Health Insurance Portability and Accountability Act       |
| <b>HL7</b>     | Health Level 7                                            |
| <b>HR</b>      | Heart Rate                                                |
| <b>ICD</b>     | International Classification of Diseases                  |



|              |                                                |
|--------------|------------------------------------------------|
| <b>ICU</b>   | Intensive Care Unit                            |
| <b>IoU</b>   | Intersection over Union                        |
| <b>IQR</b>   | Interquartile Range                            |
| <b>LAN</b>   | Local Area Network                             |
| <b>LIME</b>  | Local Interpretable Model-agnostic Explanation |
| <b>LIS</b>   | Laboratory Information Systems                 |
| <b>LlaMa</b> | Large Language Model Meta AI                   |
| <b>LLM</b>   | Large Language Model                           |
| <b>LMIC</b>  | Low- and Middle-Income Countries               |
| <b>LOS</b>   | Length of stay                                 |
| <b>LSTM</b>  | Long Short-Term Memory                         |
| <b>MAR</b>   | Medication Administration Records              |
| <b>MAR</b>   | Missing at Random                              |
| <b>MCAR</b>  | Missing Completely at Random                   |
| <b>MIMIC</b> | Medical Information Mart for Intensive Care    |
| <b>ML</b>    | Machine Learning                               |
| <b>MLOPs</b> | Machine Learning Operations                    |
| <b>MNAR</b>  | Missing Not at Random                          |
| <b>NER</b>   | Named Entity Recognition                       |
| <b>NLP</b>   | Natural Language Processing                    |
| <b>ONNX</b>  | Open Neural Network Exchange                   |
| <b>OVISS</b> | Optimal Vasopressin Initiation in Septic Shock |
| <b>PACS</b>  | Picture Archiving and Communication Systems    |

|                        |                                           |
|------------------------|-------------------------------------------|
| <b>PDDI</b>            | Potential Drug-Drug Interaction           |
| <b>PICU</b>            | Pediatric Intensive care Unit             |
| <b>PICUs</b>           | Pediatric ICUs                            |
| <b>PPG</b>             | Photoplethysmography                      |
| <b>PPI</b>             | Peripheral Perfusion Index                |
| <b>PPV</b>             | Positive Predictive Value                 |
| <b>Q-CRT</b>           | Quantified Capillary Refill Rime          |
| <b>RAM</b>             | RandomAccess Memory                       |
| <b>RL</b>              | Reinforcement Learning                    |
| <b>RNN</b>             | Recurrent Neural Network                  |
| <b>RR</b>              | Respiratory Rate                          |
| <b>SAFE-ICU</b>        | Sepsis Advanced Forecasting Engine        |
| <b>SaMD</b>            | Software as a Medical Device              |
| <b>SD</b>              | Standard Deviation                        |
| <b>SHAP</b>            | SHapley Additive exPlanations             |
| <b>SIPA</b>            | Shock-index pediatric age-adjusted        |
| <b>SMOTE</b>           | Synthetic Minority Oversampling Technique |
| <b>SOP</b>             | Standard Operating Procedure              |
| <b>SOP</b>             | Standard Operating Protocol               |
| <b>SpO<sub>2</sub></b> | Oxygen Saturation                         |
| <b>StructTS</b>        | Structural Time-Series                    |
| <b>SVM</b>             | Support Vector Machine                    |
| <b>SVM</b>             | Support Vector Machines                   |
| <b>SVR</b>             | Systemic Vascular Resistance              |

|                    |                                                                                                                            |
|--------------------|----------------------------------------------------------------------------------------------------------------------------|
| <b>Sys-bp</b>      | Systolic Blood Pressure                                                                                                    |
| <b>TCF</b>         | Twin-Core Fiber                                                                                                            |
| <b>TCN</b>         | Temporal Convolutional Networks                                                                                            |
| <b>THE</b>         | Total Health Expenditure                                                                                                   |
| <b>TREWScore</b>   | Targeted Real-time Early Warning Score                                                                                     |
| <b>TRIPOD-AI</b>   | Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis in Artificial Intelligence |
| <b>TS features</b> | Time-series Features                                                                                                       |
| <b>UMLS</b>        | Unified Medical Language System                                                                                            |
| <b>VAEs</b>        | Variational Autoencoders                                                                                                   |
| <b>VIF</b>         | Variation Inflation Factor                                                                                                 |
| <b>WSN</b>         | Wireless Sensor Networks                                                                                                   |



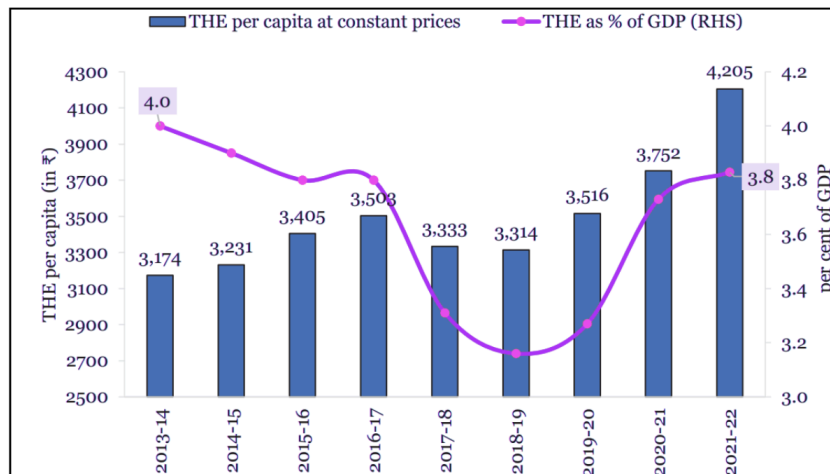
# Chapter-1

## Introduction

### 1.1 Clinical Decision Support in the Intensive Care Environment

An Intensive Care Unit (ICU) is a clinical area where decisions made in real time can mean life or death for children. Pediatric ICUs (PICUs) take care of children who require sophisticated interventions and continuous monitoring due to rapid deterioration in their physiological condition. During such high-stakes situations, large amounts of patient data are recorded in Electronic Health Records (EHRs) by clinicians. This data includes the vitals, lab results, medications, clinical notes, and others. However, the sheer volume and complexity of this data can overwhelm decision-making, increasing the risk of delayed or suboptimal care. In addition, the cost of healthcare is increasing at a rate higher than that of the gross domestic product (GDP). U.S. health care spending grew 7.5 percent in 2023, reaching \$4.9 trillion or \$14,570 per person ([National Health Expenditures Data: Historical, 2024](#)). In comparison, India's total health expenditure stood at only 3.8 percent of its GDP in 2021–22, with per capita expenditure ([Figure 1.1](#)), indicating the resource constraints and disparities in health system investments. Notably, 87.3 percent of India's Total Health Expenditure (THE) is directed toward current health expenditure, while only 12.7 percent is allocated to capital expenditure ([Maithani & Prasad, 2025](#)).

**Total Health Expenditure (THE) per capita and as a share of GDP**



Source: National Health Accounts, Ministry of Health and Family Welfare ( MoHFW)

*Figure 1.1 **Total health expenditure per capita:** Trends in THE per capita at constant prices and as a percentage of GDP in India from 2013 to 2022. The chart highlights a gradual increase in per capita expenditure over time, with fluctuations in THE as a share of GDP, reaching 3.8% in 2021–22*

Due to the high complexity of data and limited resources in healthcare, there is a need for Clinical Decision Support Systems (CDSS). These systems can support clinicians by generating sophisticated, multidimensional data, resulting in better clinical decisions that are more accurate, timely, and consistent. Especially in resource-limited environments like India, CDSS provides a viable solution to improve the delivery of care, lessen the burden on providers, and enhance patient outcomes.

AI can help CDSS to automate data interpretation (AI interprets data) and help in risk stratification. AI can also help us in prediction. Models powered by AI can study huge amounts of past and live patient data to spot trends, anticipate medical decline, and suggest proper interventions. In simple terms, by including AI in the CDSS, it will not only produce accurate diagnoses and decision-making but also be scalable for improving healthcare quality in various climates.

## **1.2 AI for Decision Support in the High-Risk ICU Settings**

AI is changing critical care medicine for the second time. AI is modifying decision support in real-time in high-risk environments, such as ICUs. These circumstances yield multi-modal, high-frequency data, from physiological waveforms and laboratory values to clinical notes and imaging, that is well-suited for machine learning. For example, AI systems (especially deep learning models) can synthesize these complex data streams to provide timely insights and predict deterioration to guide therapy and ultimately improve patient outcomes.

Over recent years, the usage of AI for critical care has witnessed a huge advance. In the 1980s and 1990s, early decision support systems such as MYCIN and INTERNIST were essentially static - they were based on rule-based logic codified by clinical experts ([Shortliffe, 1976](#)). While pioneering in intent, these systems were inflexible, not scalable, and incapable of evolving in changing clinical settings. In the 2000s, there was a move towards statistical and machine learning (ML) methods as EHRs and bedside monitors became more widespread. We trained classical algorithms, logistic regression, support vector machines (SVMs), decision trees, and random forests on structured datasets to predict mortality, ICU readmissions, and sepsis ([Obermeyer, Z., & Emanuel, E. J., 2016](#)). These are more scalable and interpretable models, especially with feature importance.

The decade of the 2010s deep learning (DL) witnessed the introduction of recurrent neural networks (RNNs), especially Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, for temporally ordered ICU data. These architectures were influential in predicting acute clinical events, such as shock, sepsis, or respiratory failure, that arise in clinical settings several hours in advance ([Choi et al., 2016](#); [Ghassemi et al., 2015](#); [Tomašev et al., 2019](#)).

At the same time, convolutional neural networks (CNNs) started being used for automated analysis of chest x-rays, ultrasound, and thermography alongside clinical investigations.

More recently, transformer-based models, initially developed for natural language processing (e.g., BERT, GPT), are being adapted for tabular, time-series, and multimodal biomedical data ([Acosta et al., 2022](#); [Rasmy et al., 2021](#)). These foundation models enable contextual reasoning across multiple modalities, allowing for richer predictions that combine patient vitals, medications, and clinical narratives.

Despite this progress, several limitations impede clinical adoption. First, most AI models are trained for narrow, single-site tasks, which restricts their generalizability across different patient populations, ICUs, and countries ([Moor et al., 2023](#); [Zhuang et al., 2023](#)). Second, deep learning models often act as "black boxes" with limited interpretability, which hinders clinical trust and transparency. Third, most studies rely on retrospective datasets, such as Medical Information Mart for Intensive Care (MIMIC) or Electronic Intensive Care Unit (eICU), with few models prospectively validated in live hospital environments ([Topol, 2019](#)). Also, before AI is deployed in full-scale use, algorithmic bias, fairness, and human–AI interaction will require the solution of ethical issues.

As AI is deployed beyond research and into the clinic, there is a new need for explainability, validation in the real world, fairness auditing, and clinician-in-the-loop systems. The AI-CDSS of the future should streamline into clinical work procedures, be adaptable across institutions, and have interpretability for widespread uptake in the ICU.

## 1.3 Types of Data in ICU

The Intensive Care Unit (ICU) is one of the most data-intensive environments in modern medicine, characterized by rapid intervention and high-stakes clinical decision-making. It creates huge amounts of data streams that include many modes and happen at a high frequency, allowing for AI-based decision support. The ICU data is heterogeneous, big, and of coarse temporal granularity, thus rendering it particularly suitable for predictive analytics and machine learning applications.

### 1. Physiological Time-Series Data

ICU monitors record vital signs, including heart rate (HR), respiratory rate (RR), arterial blood pressure (ABP), oxygen saturation (SpO<sub>2</sub>), and core temperature, at a frequency of minutes or seconds. These constant streams form the foundation of real-time monitoring of clinical processes and the development of Early Warning Systems. Deep learning models have adopted the use of recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and temporal convolutional networks (TCNs) to predict clinical decompensation, which includes sepsis and shock ([Shashikumar et al., 2017](#); [Nemati et al., 2018](#)). This has also been shown to

work in resource-limited settings through the SAFE-ICU initiative, which demonstrated feasibility with low-cost thermographic data and time-series monitoring for early shock detection and tele-diagnosis (Sethi et al, ICEGOV).

## **2. Laboratory and Biochemical Data**

Another source of structured data within the ICU comes from routine blood work, including complete blood counts, metabolic panels, blood gas analyses, and infection markers. Evaluating organ function, diagnosing sepsis, and informing treatment protocols all rely on these values. When combined with time series data, these algorithms boost the accuracy of severity scores such as APACHE-II, SOFA, and NEWS2. Recent works have used ensemble learning to combine labs and vitals for high-fidelity mortality risk prediction and multiorgan dysfunction prediction ([Alaa et al., 2019](#)).

## **3. Medication Administration Records**

Software systems that capture physicians' orders in a structured format, often embedded in an electronic health record (EHR), are called computerized physician order entry (CPOE) systems. CPOE systems and MARs have complete records for all drug administrations when and in what dose, by what route. Drug safety surveillance relies heavily on such data useful for detecting drug-drug interactions. The new module presented in this thesis is called ContraIndicator, which makes use of named entity recognition (NER) for extracting drugs and mapping their pairs with existing curated pharmacological interaction databases. Recent studies have incorporated drug administration trends, as well as pharmacokinetic data, to produce real-time toxicity warnings ([Hanf et al., 2022](#)).

## **4. Unstructured Clinical Text**

A significant portion of ICU data is stored in unstructured form, including physician progress notes, discharge summaries, and radiology reports. Natural Language Processing (NLP) models, particularly transformer-based architectures such as ClinicalBERT and BioGPT, have demonstrated strong performance in extracting structured clinical information, including diagnoses, symptoms, and medications, from free text ([Alsentzer et al., 2019](#); [Luo et al., 2022](#)). This thesis employed the MED7 model to parse ICU prescriptions and support downstream pharmacovigilance tasks.

## **5. Medical Imaging**

Medical imaging plays a crucial role in ICU diagnostics, particularly in evaluating respiratory and neurological conditions. Standard modalities include chest radiographs and computed tomography (CT) scans. In recent years, thermal imaging has gained attention as a non-invasive technique for assessing peripheral perfusion and detecting early signs of shock or hypothermia,



particularly in pediatric populations. Research has shown that thermal gradients, such as center-to-peripheral temperature differences (CPD), can be informative biomarkers of hemodynamic instability ([Nagori et al., 2022](#)). Proposed on CNNs and LSTM-based pipeline to analyze the features from thermal video frames and the vital sign trend for an early warning system, showing neonatal thermal video containing rich features and multi-model fusion.

## 6. Microbiology Data

The culture results and profiles of antibiotics can microbiological data can be very useful to see the infection with the help of the infection control committee in an ICU setup. As AMR is becoming a bigger and bigger problem, there is a need to track these trends. Many systems have been created to streamline the extraction and visualization of AST reports to guide empiric therapy decisions and support antimicrobial stewardship. Studies that use hospital-wide antibiograms show feasibility for the integration of resistance monitoring into clinical workflows to enhance real-time decision support ([Topol, 2020](#); [Tacconelli et al., 2018](#)).

## 7. Multicenter Benchmark Datasets

Datasets publicly available for intensive care units such as **MIMIC-III**, **MIMIC-IV**, and the **eICU Collaborative Research Database** have become valuable resources for critical care research. The databases contain de-identified, high-resolution, time-stamped clinical data from diverse populations across multiple institutions, including vitals, laboratory results, medication administration, and notes([Johnson et al., 2016](#); [Pollard et al., 2018](#)). Using them widely helps to reproduce, validate externally, and benchmark the AI models. Many papers have used these datasets to design and deploy prediction models for sepsis and mortality risk prediction along with early warning systems, thus allowing researchers to examine the generalizability of their models beyond local settings ([Harutyunyan et al., 2019](#); [Shickel et al., 2019](#)).

## 1.4 Problem Statement

In spite of the enormous increase in digital health data and advances in artificial intelligence (AI), pediatric intensive care units (PICUs) still lack integrated real-time systems capable of translating data into clinically useful actions. Current decision support tools primarily rely on rules or narrow domains and have limited interpretability, poor generalisability across institutions, and little capacity to harness the complete multimodal ICU data available (especially time-series vitals, thermal images, treatment, and microbiology data).

The core research problem addressed in this thesis is:

*To develop, validate and deploy AI models for capturing early warning signals for patient deterioration and enabling treatment decision support pediatric ICU leveraging multimodal data, with a specific focus on antimicrobial resistance.*

## 1.5 Development of Models

The development of AI models in critical care begins with high-value clinical targets and timely decisions. As a consequence, timely decisions can have an important impact on outcomes. For example, early prediction of sepsis, hypoxia, or drug-related adverse events, and often also physiological data collection, will form the basis for these AI models, which will be designed to help clinicians. The early models primarily used classical machine learning techniques like logistic regression, decision trees, and random forests. Although interpretable, these methods required heavy manual feature engineering and were sensitive to missing data or imbalanced data ([Obermeyer et al., 2016](#)).

The transition to deep learning marked a paradigm shift. RNNs, LSTM, and GRU have performed successfully in modeling temporal patterns present in ICU data. These architectures were used successfully in the early detection of sepsis and respiratory failure and circulatory collapse using raw vital sign streams ([Choi et al., 2016](#); [Tomašev et al., 2019](#)).

A major milestone in this field was the creation of TREWScore, an early warning tool for septic shock that used multivariate temporal features and probabilistic modelling to give real-time risk scores with high discriminative power ([Henry et al., 2015](#)).

Newer multimodal architectures and transformer-based ClinicalBERT and BEHRT models capable of integrating structured data with unstructured clinical narratives have come up ([Rasmy et al., 2021](#), [Acosta et al., 2022](#)). We've improved the contextual comprehension and scaling of various healthcare data sets.

## 1.6 Validation of Models

Validating models will ensure that they are clinically relevant and fair. Using k-fold cross-validation or train-test validation for internal validation is still common, but no longer enough to be ready for deployment. Contemporary model validation encompasses.

- **Temporal validation:** Assesses the ability of a model to predict future unseen data. It is important in ICUs where patient characteristics and diseases change continuously.
- **Subgroup Validation:** Model efficacy testing in subgroup of patients of different age groups, gender, ethnicity etc. reveals disparities and biases in the predictive performance ([Khor et al., 2023](#)).
- **Calibration Assessment:** “How closely do predicted losses or gains match observations?” Not all accurate models are reliable, which may lead to adverse outcomes ([Van Calster et al., 2019](#)).

Explainability is now a critical component of validation. Tools such as SHAP LIME and Integrated Gradients can give clinicians an account of how input features are weighted by the models ([Lundberg et al., 2020](#)).

## 1.7 Deployment of Models

Deployment aids in the transfer of knowledge from research to clinical utility. Using trained AI systems in a hospital involves a high-level integration into hospital workflows, most often through EHR systems or custom dashboards. Key considerations include.

- **Real-time Responsiveness:** Real-time responsiveness means that the model should produce predictions fast enough to inform an immediate decision like in Intensive Care Units which will have crucial implications.
- **User Interface Design:** AI outputs must be conveyed in a form that clinicians can easily act on and interpret (e.g., visual explanations and confidence intervals).
- **Compliance and Regulation:** Deployable models must comply with the device regulations that are often characterised by clearance from regulators like the US FDA. In addition, it is necessary to comply with data privacy and security laws, such as HIPAA and GDPR ([Neupane et al., 2025](#)).

Examples which have been able to establish credibility are Google Health’s sepsis prediction engine and Stanford’s embedding of AI-CDSS (Clinical Decision Support System), both of which enabled enhanced functioning ([Johnson et al., 2023](#)).

## 1.8 Multi-centric Validation of Models

AI systems must perform well beyond the environment of their initial deployment if they are to be trustworthy. To ensure that a model works well on different datasets and to avoid it interpreting something wrong, validation must take place multi-centrally.

This method became popular with publicly available ICU datasets like:

- **MIMIC-III and MIMIC-IV:** are critical care data sets of patients who were admitted in the ICU at Beth Israel Deaconess Medical Center ([Johnson et al., 2016](#)).
- **eICU-CRD:** A real-world multi-hospital dataset that helps in generalization (Pollard et al., 2018) ([Pollard et al., 2018](#)).

Models like **VentNet** and **DeepSOFA** are assessed for scaling and fairness using cross-dataset, multi-institutional validation strategies. **VentNet**, developed by ([Shashikumar et al., 2021](#)), was

prospectively validated on ICU data from 2 health systems, which were UC San Diego and Massachusetts General Hospital. It showed excellent predictive performance ( $AUC > 0.88$ ) even in COVID-19 cohorts, and it did so without the need for local retraining. Likewise, **DeepSOFA**, a deep learning model to predict illness severity over time, was validated across multiple hospital systems to ensure it is robust in heterogeneous ICUs ([Shickel et al., 2018](#)). This shows a general shift toward multi-center validation for trustworthy and fair clinical deployment in low-resource or international settings, where retraining might not be feasible and data is scarce or hard to come by.

## 1.9 Resource Considerations and the Way Ahead for Resource-Limited Settings

Artificial Intelligence (AI) has the potential to positively impact outcomes in critical care settings. These benefits could be particularly pronounced in low- and middle-income countries (LMICs), where clinician-to-patient ratios are often inadequate, and access to advanced diagnostics is limited. Yet, implementation in resource-limited settings faces several challenges:

1. **Infrastructure Limitations:** Many ICUs, especially in low-resource settings, face a lot of infrastructural issues. They hardly have an uninterrupted power supply, high-performance computers, internet, etc. Thus, it limits the deployment of cloud-based AI or GPU-dependent AI like ChatGPT. Using models that can run offline, which can be deployed at the edge, resulting in formats such as Open Neural Network Exchange (ONNX) or TensorRT, can offer a practical alternative ([Xu et al., 2025](#)).
2. **Data Quality and Standardization:** Training and transfer of AI can be hindered by the heterogeneity of EHR systems, inconsistent coding practices, and limited annotation capacity (data quality and standardization). Standardized data collection frameworks, such as HL7, FHIR, should invest in ([Rajkomar et al., 2018](#)). International data registries can also help with model portability.
3. **Human Resource Constraints:** Having a human resource constraint means that the clinicians should be trained in interpreting AI outputs. Also, the culture of the institution towards automation can affect the adoption. Medical staff local involvement ensures the CDSS is relevant, usable, and trusted.
4. **Cost-Effectiveness and Sustainability:** There should be a justification of their value in low-resource settings. Models for early triage or infection control may have high impact and low computational demands, making them ideal for deployment in tier-2 or rural hospitals ([Topol, 2019](#)).

## 1.10 Foundation Models in Clinical AI

Foundation models, which are large-scale, pre-trained neural networks designed to generalize across multiple downstream tasks, represent a significant leap forward in clinical AI. These include language models (e.g., GPT, LLaMA, BioGPT) and vision-language and multimodal transformers that can process various types of data (e.g., text, image, time series).

Advantages in Healthcare:

- **Transferability:** Foundation models like **Med-BERT**, trained on massive biomedical and clinical corpora (e.g., MIMIC, PubMed, UMLS), can be fine-tuned with **minimal labeled data** to deliver strong performance on downstream tasks, such as diagnosis prediction, note summarization, or adverse event detection. For example, [Rasmy et al. \(2020\)](#) demonstrated that pre-training on 28 million patient records and fine-tuning for as few as 300–500 samples could yield improvements of over 20% in AUC, achieving results comparable to those of models trained on ten times more data. The impact of foundation models on efficient and versatile low-resource learning in a wide range of clinical applications.
- **Multimodal Fusion:** Models like Med-PaLM (and a multimodal derivative, Med-PaLM M) as well as vision-language transformers such as BEiT-3 can reason over mixed clinical data modalities, including structured inputs, free-text, medical imaging, and even genomics, better capturing the complexity seen in ICUs ([Singhal et al., 2023](#); [Wang et al., 2022](#)). For example, scientists evaluated Med-PALM on a benchmark called MultiMedBench, which has 14 medical tasks like writing radiology reports, answering medical questions, and interpreting genomic variants. Med-PALM consistently outperformed specialist models. Also, in 40% of comparisons for interpreting chest X-rays, clinicians preferred Med-PALM ([Singhal et al., 2023](#)). In the same way ([Wang et al., 2022](#)) describe BEiT-3, a unified vision-language transformer, which has reached state of the art across vision and vision-language benchmarks, exhibiting practical general-purpose reasoning with (image, text) as input. We demonstrated the multimodal foundation models (MFMs) potential for important ICU-relevant tasks, such as prognosis, decompensation detection and multimodal interpretation, all of which would typically be supported by much richer contexts.
- **Context-Aware Reasoning:** The models which are trained on large-scale biomedical in the context of medicine i.e., PubMed, MIMIC, UMLS encode a wide-ranging medical context, which makes them adapt on a few-shot basis or even zero-shot basis to novel clinical tasks, thereby improving their interpretability, and generalization in decision-making. **Med-PaLM 2** received a score of 86.5% on MedQA and produced expert-level output. Similarly, **GPT-4** attained a score of 86.7% on the USMLE, indicating they both possess a great deal of knowledge within the medical domain ([Qiu et](#)

[al., 2024](#); [Singhal et al., 2023](#)). The Med-PaLM M was also fine-tuned on MultiMedBench multimodal benchmark and showed a good capacity for clinical reasoning over structured data, images, and genomics ([Tu et al., 2023](#)). The rich context-aware inference given by these models is more than the task specific systems can provide.

## 1.11 Limitations and Challenges of AI in the ICU

### A. Limitations of AI in Critical Care

While the progress has been considerable, there are limitations to the usage of AI in ICUs:

- **Data Quality and Heterogeneity:** ICU datasets are notoriously inconsistent and heterogeneous, mainly due to variations in institutional protocols, monitoring frequency, data resolution, and EHR architecture. The different methods used to measure and assess patients in ICUs make predictive models less reliable when applied to other ICUs. These may fail to generalize because an ICU's predictive model was trained there. If the model doesn't account for the heterogeneity across sites, it will learn site-specific artifacts and not actual clinically meaningful patterns ([Pei et al., 2023](#)).
- **Temporal Drift and Non-Stationarity:** Clinical AI models are trained on past data. Over time, it is possible that the model performance goes down. If there is a shift in the population of patients, disease prevalence, clinical practices, etc. the non-performance happens due to the change in the target variable over time. This is known as drift or concept shift or non-stationarity. Models could become outdated without continuous reassessment and changes to the methods used in points of dispensation. There are many works in the literature of machine learning to identify drift and adapt the model to ensure stable performance ([Lu et al., 2018](#)).
- **Label Noise and Ground Truth Uncertainty:** Clinical outcomes (e.g. sepsis or shock) may not be truly predictive as these labels are subject to a lot of interpretation and varying coding practices, generating substantial label noise. Predictors that are trained on imperfect ground truth can learn spurious or non-clinical associations that impede prediction reliability. A scoping review investigating label noise in medical deep learning showed that inter-observer variability, coding mistakes, and ambiguities in labeling protocols often hamper the performance of models ([Wei et al., 2024](#)). Further research has demonstrated that medical image models that are trained with noisy labels experience a measurable decline in calibration and accuracy ([Karimi et al., 2020](#)).

### B. Failure of AI in ICU Settings

The ICU has seen the failures and underperforming trials of notable AI tools:

- **AI for diabetic retinopathy screening**, though highly accurate in controlled laboratory evaluation, faced real-world limitations during deployment in Thailand, including poor image quality from low-light settings and inconsistent capture techniques, which led to a significant percentage of ungradable scans in clinical practice ([Ruamviboonsuk et al., 2022](#)).
- **AI for Oncology**: a well known model was criticized for unsafe or irrelevant recommendations in real-world settings, primarily due to limited training data and a lack of transparency ([Strickland, 2019](#)).

Often, the failure stems from a lack of context in the study, low involvement of the clinician, or over-relying on the performance of the benchmark without operational testing.

### C. Biases in Data and Algorithmic Fairness

Bias remains a core challenge:

- **Demographic bias**: Refers to the unequal performance of predictive models depending upon ethnicity, gender, or age group when training data is unbalanced. For example, a tool that predicts who will need health care issues was shown to underestimate the health risks of Black patients ([Obermeyer et al., 2019](#)).
- **Label bias**: In simplest terms, what this means is that the use of a marker other than the gold standard can introduce bias in the model. Moreover, when the surrogate marker has systemic inequities, such as in coding or access to the healthcare system, it can result in significant biases that do not reflect the true performance in the field.

To lessen bias, we can check stratified model evaluations, fairness-aware learning, and inclusive data practices.

### D. Human Factors and Clinical Integration

Human factors significantly influence model adoption:

- **Trust and usability**: Predictions from a black box without an explanation can undermine faith. Studies show that explainability is vital to adoption, especially if predictions conflict with clinician intuition ([Tonekaboni et al., 2019](#)).
- **Workflow integration**: AI models that disrupt rather than complement existing ICU workflows face resistance. It is critical to embed models within routine systems such as EHR alerts, dashboards, or mobile apps.
- **Training and awareness**: If clinicians do not get trained about AI tools and best practices, the tools may be misused or overused.



## E. Ethical and Regulatory Considerations

Bringing AI into the high-risk domain is fraught with ethical and legal challenges.

- **Informed consent and autonomy:** AI systems, if used to predict or triage a patient, call into question what it means to consent to treatment, especially when the model is working in the background ([Gerke et al., 2020](#)).
- **Regulatory oversight:** Few AI tools have received regulatory approval from the likes of U.S. agencies like the FDA. Models that adapt to incoming data over time (e.g., continual learning) pose regulatory challenges
- **Liability:** Legal liability is hazy in determining if the clinician or the system designer is responsible.

### 1.12 What Does the Thesis Address

This thesis fills critical gaps in the use of artificial intelligence to deliver time-sensitive clinical decision support in high-risk pediatric ICUs. While AI has demonstrated promise in retrospective prediction and post-hoc analysis, its real-time, actionable deployment in the ICU, especially in resource-constrained settings, remains limited. The work presented here seeks to move beyond proof-of-concept models toward robust, interpretable, and clinically integrable systems.

The central focus of this thesis is on developing multimodal, explainable, and deployable AI solutions for early detection, therapeutic guidance, and safety surveillance in pediatric critical care. Through the integration of high-frequency time-series signals, thermal imaging, unstructured clinical texts, and microbiological reports, the thesis builds novel machine learning pipelines capable of anticipating clinical deterioration and adverse events with actionable lead times.

Key challenges addressed include:

- Temporal prediction using multimodal data: Leveraging non-linear physiological patterns and novel sensing modalities (e.g., thermal videos) to predict events like hypothermia or shock.
- Data heterogeneity and generalizability: Addressing model robustness across datasets through symbolic, non-linear, and foundation model-based approaches, and validating performance across diverse cohorts.
- Drug safety in polypharmacy scenarios: Automating the detection of potential drug-drug interactions using NLP pipelines applied to real-world ICU prescription data.



- Infection control through AMR surveillance: Mining temporal antibiogram trends from AST reports to support empirical therapy planning in ICUs

Ultimately, this thesis culminates in the development and integration of the SAFE-ICU platform. This modular, real-time clinical AI system delivers predictive insights, risk stratification, DDI alerts, and antimicrobial resistance trends directly to bedside providers. Unlike most prior work, SAFE-ICU is not a single-task algorithm but a scalable infrastructure for clinically grounded, context-aware AI, supporting end-to-end translation from research models to real-time decision support.

This thesis thus addresses the translational frontier of AI in healthcare, where technical performance must intersect with clinical relevance, usability, and ethical responsibility. It aims to provide a replicable and deployable blueprint for the responsible implementation of AI in pediatric ICUs, with implications for broader applications in global critical care.

### **1.13 Objectives and Structure of the Thesis**

This dissertation elucidates the scope and applicability of leveraging multimodal data and advanced analytics for the early detection of clinical deterioration in intensive care units. It explains the development of novel predictive models, ThermoGnose, ThermalShockNet, and SIgnose, which use continuous vitals and thermal imageries to predict hypothermia, hemodynamic shock, and abnormal Shock Index with a key focus on accuracy, interpretability, and greater lead-time utility. It assumes that our current classifier will be able to generalise well across patients of diverse geographies, settings, and stages of disease. To justify this assumption, we will use AIIMS data, MIMIC-III data ( USA dataset), and eICU data ( USA dataset) to test SIgnose and address cross-institutional robustness. The thesis presents tools called ContraIndicator and AMRtrack, which can extract medications given in the ICU in real-time and warn about likely drug interactions, and an automated method for AMR surveillance through AST reports, respectively. Through its deployable clinical interface, SAFE-ICU, which is user-centered, the modules are integrated to provide real-time risk forecasting, prescription safety alerts, and AMR analytics.

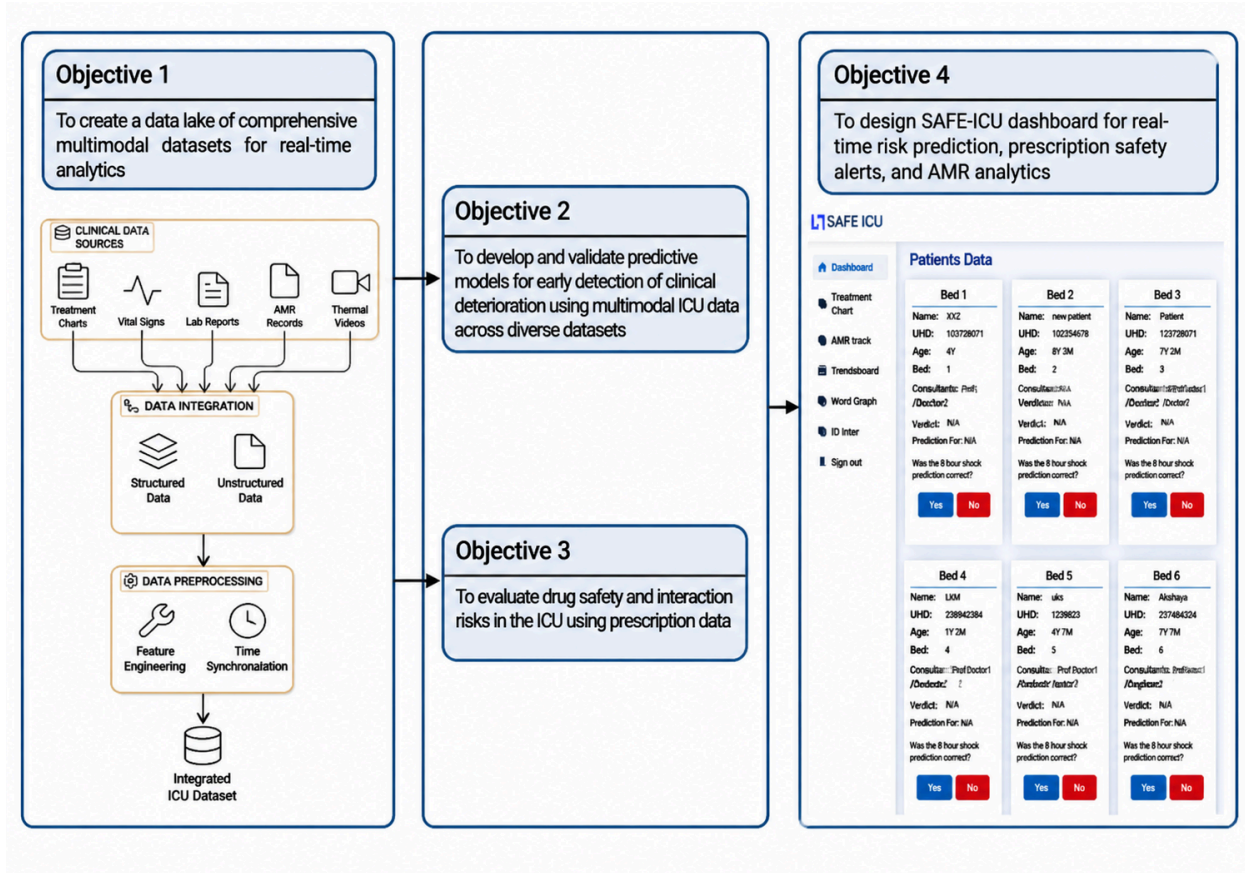


Figure 1.2: The primary objective of this thesis is to design, develop, and evaluate an integrated AI-powered decision support platform, SAFE-ICU, for use in Pediatric Intensive Care Units. This platform addresses critical clinical safety domains, including deterioration prediction, pharmacovigilance, and antimicrobial stewardship.

Chapter 2 of this dissertation provides a background and an overview of related works. It reviews the current landscape of AI in critical care, with emphasis on time-series modeling, drug safety analytics, natural language processing in clinical settings, and antimicrobial resistance surveillance. It also surveys literature on generalizability challenges and transformer-based methods for clinical data.

Chapter 3 outlines the construction of the SAFE-ICU data lake and data warehouse by integrating multimodal clinical data streams from a pediatric ICU setting. This includes raw and processed data from bedside monitors (vitals), laboratory test systems, microbiology reports, prescription charts, and thermal imaging videos. The chapter details the ingestion pipelines, data cleaning procedures, timestamp synchronization, and standardized storage strategies. It emphasizes preprocessing workflows for both structured and unstructured inputs, forming the foundation for the downstream AI modeling and real-time analytics in subsequent chapters.

Chapter 4 presents the development and evaluation of three key models. ThermoGnose (hypothermia prediction), ThermalShockNet (shock forecasting via thermal imaging and vitals), and Signose (abnormal Shock Index classification across datasets). This chapter details dataset

preparation, temporal feature engineering, LLM-based experiments, and cross-dataset validation of predictive performance.

Chapter 5 describes the drug safety and interaction analysis in the ICU for real-time potential DDI (pDDI) detection using the MED7 NER model. It includes statistical analyses of pediatric ICU prescriptions, interaction severity stratification, and a discussion of integration with DDI knowledge bases.

Chapter 6 explains the full platform integration of the predictive, drug safety, and AMR modules into the deployable SAFE-ICU dashboard. This chapter also covers the AMR visualization system built from AST reports and discusses system-level design considerations for usability and scalability.

Chapter 7 summarizes the thesis contributions and limitations. It also proposes future extensions, including LLM-driven polypharmacy modeling, multilingual adaptation of the NLP pipeline, and clinical deployment pathways for SAFE-ICU in real-world hospital environments.

# Chapter-2

## Background and Related Work

### 2.1 Introduction: Opportunities and Challenges for AI in ICUs

Artificial Intelligence has emerged as a transformative force in healthcare, particularly within high-acuity environments such as ICUs. These settings generate vast quantities of multimodal data from continuous physiological signals and laboratory diagnostics to EHRs, medical imaging, and free-text clinical notes. This data-rich context has catalyzed the development of advanced AI systems designed to predict adverse outcomes, optimize treatment regimens, and support real-time clinical decision-making.

Several foundational models have demonstrated the efficacy of AI in this domain. For example, ClinicalBERT and BEHRT ([Rasmy et al., 2021](#)) achieved state-of-the-art performance in mortality prediction and sepsis detection by fusing structured EHR data with unstructured clinical narratives. Likewise, models such as BEHRT+ and Med-BERT have shown the effectiveness of contextual pretraining on longitudinal patient records, especially in handling temporally misaligned data.

Despite these advances, real-world clinical adoption remains limited. A key barrier is the disconnect between retrospective performance and deployment in dynamic hospital settings. ([Acosta et al., 2022](#)) proposed interpretable transformer-based models for ICU forecasting but emphasized the challenge of integrating these tools into live workflows. A similar gap was evident in the work by Sendak et al. ([2020](#)), who underscored the organizational and workflow barriers that prevent AI systems from scaling beyond pilot programs.

One of the few deployed systems, the TREWS, studied by Katharine E. et al., demonstrated that **provider trust, timing of alerts, and UI design** are critical for clinician adoption. These human factors are just as crucial as model accuracy. Similarly, [Lekadir et al. \(2023\)](#) provided a framework for measuring trustworthiness in clinical AI, emphasizing the need for transparency, robustness to shift, and participatory design with end-users.

In terms of intervention-focused AI, the Optimal Vasopressin Initiation in Septic Shock (OVISS) study ([Kalimoultou et al., 2023](#)) stands out. It applied reinforcement learning (RL) to optimize vasopressor initiation in septic shock, demonstrating clinically meaningful improvements. Liu et al. ([2021](#)) similarly applied RL to dose titration tasks, highlighting the promise of AI in dynamic treatment planning, especially under uncertainty.

Studies have also addressed the challenge of generalization. [Burdick et al. \(2020\)](#) conducted a multicenter evaluation of AI-based sepsis prediction, revealing substantial performance variability across sites and EHR systems. In response, [Guo et al. \(2022\)](#) proposed domain adaptation techniques that adjust for dataset-specific shifts, a method now commonly used to enhance model portability.

The importance of multimodal data fusion has been addressed in several recent publications. [Soenksen et al. \(2022\)](#) proposed a design framework for multimodal clinical AI, integrating vital signs, laboratory results, text, and imaging to enhance predictive robustness. [Lijian et al. \(2024\)](#) introduced MedViLaM, which demonstrates strong generalizability and explainability across diverse medical tasks by integrating clinical text and imaging using a single set of model weights.

Finally, the issue of equity and fairness in AI has received growing attention. [Obermeyer et al. \(2019\)](#) exposed racial bias in commercial health algorithms, prompting a wave of research into subgroup validation and fairness-aware training. Recent research has demonstrated that ICU AI systems can exhibit significant performance disparities across various socioeconomic and demographic groups, including age, race, ethnicity, and insurance status, highlighting the ethical imperative for comprehensive fairness audits. Expert consensus and emerging studies underscore the importance of making subgroup performance validation an integral part of conditional model deployment, particularly in high-stakes ICU environments ([Josephson et al., 2023](#); [van Schaik et al., 2022](#)).

To bridge these challenges, technical, translational, and ethical, this thesis introduces SAFE-ICU: a real-time, modular, open-source AI platform for ICU monitoring and early intervention. SAFE-ICU builds upon the lessons from these pivotal studies by emphasizing explainability, generalizability, equity, and clinician-centered usability. SAFE-ICU represents a significant advancement toward truly translational and trustworthy clinical AI.

### **2.1.1 Clinical AI for Prediction and Monitoring**

The use of ML in ICUs has expanded dramatically over the past decade, propelled by access to large EHR datasets such as MIMIC-III, eICU-CRD, and institutional registries. These data have been used to develop early warning systems for sepsis, shock, hypoxia, acute kidney injury, and mortality.

[Tomašev et al. \(2019\)](#) demonstrated one of the earliest and most cited applications of deep learning in critical care by forecasting acute kidney injury 48 hours in advance using longitudinal EHR data. Their model, trained on over 700,000 encounters, set a benchmark for deployment-ready predictive systems.

[Rasmy et al. \(2021\)](#) introduced ClinicalBERT and BEHRT, which fused structured EHR inputs and unstructured clinical notes using transformer architectures. These models significantly outperformed traditional embeddings in mortality and comorbidity prediction tasks, highlighting the importance of contextual awareness.

[Zhang et al. \(2022\)](#) proposed a self-supervised contrastive pretraining method to improve ICU time-series representation. Their method enhanced performance across multiple downstream tasks even with minimal labels, an insight that SAFE-ICU builds upon through modular design and transferability.

[Acosta et al. \(2022\)](#) extended this direction by presenting a multimodal transformer model that integrated structured data, free-text, and medical imaging for holistic clinical forecasting. Their emphasis on interpretability and real-time compatibility directly informs SAFE-ICU's approach to integrating vitals, prescriptions, thermal imaging, and microbiology into a deployable platform.

### **2.1.2 Reinforcement Learning for Treatment Optimization**

Beyond predictive tasks, RL has been explored for guiding dynamic treatment decisions. The OVISS study ([Kalimoultou et al., 2023](#)) by Kalimoultou et al. applied RL to determine the optimal vasopressor initiation timing in septic shock, achieving superior simulation outcomes compared to clinician-guided baselines.

[Zhang et al. \(2024\)](#) similarly used deep Q-networks to personalize norepinephrine titration in ICU patients. Their model responded dynamically to evolving patient states, indicating that policy learning could enhance treatment precision in complex care environments.

These studies represent a shift from passive risk prediction to actionable decision-support, an approach that SAFE-ICU incorporates through real-time inference modules designed to close the loop between prediction and action.

### **2.1.3 Integration of AI in Clinical Workflows**

The TREWS early warning system ([Katharine E. et al., 2023](#)) remains a landmark deployment of real-time sepsis detection in practice. Their study on adoption behaviors revealed that clinicians' perceptions of system accuracy, timing, and workflow fit were crucial for sustained use.

[Austrian et al. \(2018\)](#) conducted a pragmatic retrospective evaluation showing that real-time AI alerts reduced time-to-antibiotic and ICU length of stay. Complementing this, a cluster-randomized controlled trial of a deterioration alerting system confirmed improved outcomes, thus providing rare evidence for clinical impact beyond accuracy metrics.

Such findings inform SAFE-ICU’s clinician-in-the-loop design philosophy, emphasizing transparency, usability, and modularity to promote adoption in diverse ICU settings.

#### **2.1.4 Opportunity for building generalizable models**

Recent studies have demonstrated that deep learning models for sepsis prediction often experience significant performance degradation when transferred across clinical sites due to differences in data distributions and institutional protocols.

[Moor et al. \(2023\)](#) conducted a retrospective, multi-site international validation of a deep learning model, finding that even well-performing models trained on one site generalize poorly to others. Similarly, [Rockenschaub et al. \(2023\)](#) highlighted the challenges of transitioning from single-hospital to multi-center applications, emphasizing the need for enhancing generalizability in ICU adverse event prediction through better design and training across diverse datasets. These findings reinforce the rationale behind frameworks like SAFE-ICU, which train models on datasets such as MIMIC, eICU, and custom-labeled institutional data to improve robustness in real-world settings.

[Acosta et al. \(2022\)](#) proposed a unified framework for multimodal clinical AI, integrating structured signals, imaging, and unstructured narratives to improve robustness and interpretability. SAFE-ICU mirrors this approach with its multi-sensor integration and dashboard design aligned to ICU workflows.

[Henry et al. \(2015\)](#) followed up their TREWScore model with probabilistic temporal modeling and human-in-the-loop feedback, concluding that explainability and human alignment enhance both trust and clinical impact principles deeply embedded in the SAFE-ICU ecosystem.

#### **2.1.5 Adverse Drug Reactions and Prescribing Safety**

Polypharmacy is ubiquitous in ICUs and increases the risk of adverse drug events due to harmful DDIs. Deep learning approaches, such as DeepDDI ([Ryu et al., 2018](#)), model DDIs using pharmacological features and achieve high predictive accuracy. [Tan et al. \(2024\)](#) further developed EHR-based models that combined patient context with drug ontology data to highlight high-risk medication combinations.

In practice, DDI alert systems have struggled with alert fatigue due to a lack of specificity. [Vo et al. \(2022\)](#) emphasized the need for explainable and context-sensitive DDI warnings, especially in critical care.

In this thesis, the chapter on **ContraIndicator** module in SAFE-ICU is an attempt to address these challenges by integrating NER-extracted medication data with rule-based severity mappings to deliver targeted, real-time prescribing alerts.



### 2.1.6 Antimicrobial Resistance (AMR) Surveillance and Clinical Decision Support

Timely empirical therapy in ICU sepsis care is increasingly undermined by antimicrobial resistance. [Rawson et al. \(2024\)](#) identified AMR-aware AI as a frontier in personalized treatment of infectious diseases. [Lim et al. \(2020\)](#) created one of the first AMR surveillance dashboards operational in LMICs, visualizing pathogen-drug resistance patterns for therapy guidance.

[Hardan et al. \(2024\)](#) developed a deep learning model that integrates structured data, time-series vital signs, and clinical notes from MIMIC-IV to predict antimicrobial resistance in near real-time. The model achieves high accuracy across various pathogens and demonstrates the feasibility of embedded, offline-compatible resistance prediction, eliminating the need for a centralized data repository. [Truong et al. \(2021\)](#) provide guidance affirming the benefits of unit-specific antibiogram reporting, particularly in ICUs, emphasizing better therapy alignment and stewardship outcomes.

In this thesis, the **AMRTrack** module in SAFE-ICU extends these ideas by offering clinicians interactive, time-stamped visualizations of resistance trends linked to empirical therapy recommendations. It is designed for real-time use in high-burden, low-resource environments.

## 2.2 Data-Centric Challenges in ICU AI

### 2.2.1 Data Heterogeneity and Interoperability

- One of the most pervasive challenges in applying AI to ICU data is the heterogeneity of sources, structures, and semantics inherent in clinical data. Intensive care units generate a voluminous, multimodal data stream that includes structured vital signs, laboratory results, unstructured clinician notes, drug prescriptions, microbiology reports, and medical images. Each data stream is acquired at different frequencies and adheres to inconsistent documentation standards, even within a single hospital system ([Johnson et al., 2016](#); [Goldstein et al., 2017](#)). For example, while high-frequency vitals may be recorded every second, laboratory reports are updated intermittently, and clinical notes are often typed retrospectively.
- These inconsistencies become more pronounced when comparing datasets across institutions or geographies. The MIMIC-III and eICU Collaborative Research Database ([Pollard et al., 2018](#)) exemplify this divergence: although both are large, de-identified ICU datasets from the United States, they differ significantly in variable naming conventions, measurement units, and temporal resolution. Moreover, even though interoperability frameworks such as HL7 and FHIR exist, their adoption is sporadic, and mappings between systems often require manual harmonization.



- In this context, the SAFE-ICU platform addresses these limitations through a robust data engineering pipeline that standardizes input from multiple modalities and harmonizes semantically equivalent variables. A dedicated preprocessing module converts real-world EHR fields into a unified schema that feeds the AI models, ensuring consistency across training, validation, and deployment environments.

### 2.2.2 Bias and Underrepresentation in Clinical Data

Bias in medical AI emerges from both the structure of healthcare systems and the choices made during data curation and model development. Algorithms trained on data that underrepresent specific populations, such as pediatric patients, elderly individuals, or racial minorities, are prone to systemic inaccuracies when applied to these groups. This is especially critical in intensive care, where physiological baselines, treatment protocols, and responses to medication vary across age, gender, and ethnicity.

A landmark study by [Obermeyer et al. \(2019\)](#) highlighted this issue by revealing that a widely used commercial algorithm designed to allocate healthcare services underestimated the needs of Black patients. The algorithm used historical healthcare costs as a proxy for health needs, implicitly encoding systemic disparities into its predictions. This form of proxy bias can lead to skewed resource allocation and exacerbate healthcare disparities.

The SAFE-ICU platform incorporates mechanisms to assess and mitigate such biases. Subgroup-specific performance metrics are computed for all predictive models, including those focused on pediatric patients and low-resource settings. Additionally, fairness-aware validation pipelines are employed to detect disparate performance across demographic strata, and re-weighting or re-sampling strategies are applied where disparities are identified.

### 2.2.3 Temporal Data Drift and Informative Missingness

In dynamic clinical environments such as ICUs, data distributions are not stationary. This phenomenon, known as data drift, arises from shifts in patient demographics, evolving clinical practices, changes in diagnostic criteria, or even upgrades in monitoring infrastructure and hospital workflows ([Subbaswamy & Saria, 2020](#); [Finlayson et al., 2021](#)). Models trained on a static historical dataset may perform poorly when exposed to new data from a different temporal window or healthcare setting. For instance, a predictive model trained during a pre-pandemic era may fail to generalize to data collected during COVID-19 due to systemic changes in care protocols, medication usage, and disease profiles.

Compounding this problem is the issue of missing data, which is not always random in ICUs. Traditional statistical models often assume that data are Missing Completely at Random (MCAR) or Missing at Random (MAR), allowing standard imputation techniques to be used. However, ICU data are frequently Missing Not at Random (MNAR). For example, a lab test may

be absent not due to a random error but because a physician made a conscious decision not to order it, perhaps because the patient appeared clinically stable. In such cases, the missingness itself is an informative signal. Failing to consider this aspect can lead to biased imputation and downstream model inaccuracies.

Recent approaches address these issues using time-aware imputation techniques and generative modeling. Models like GRU-D (Gated Recurrent Unit with Decay) incorporate masking and decay terms to model how long it has been since the last observation and how the absence of data influences predictions ([Che et al., 2018](#)). In addition, variational autoencoders and GAN-based frameworks have been applied to learn the latent distribution of time-series data, thereby enabling more realistic imputations and simulations.

Kalman Filters, particularly in their recursive Bayesian formulation, provide an efficient method for imputing missing values in linear time series while accounting for measurement noise and underlying temporal dynamics ([Welch & Bishop, 2006](#)). SAFE-ICU integrates Kalman-based smoothing for high-resolution signals, such as heart rate and temperature, providing clinically plausible imputations that align with physiological continuity.

The SAFE-ICU platform incorporates these time-aware imputation methods in its preprocessing pipelines and leverages informative missingness in its feature engineering step. This ensures that temporal gaps in ICU monitoring, far from being discarded, are used to enhance model performance and interpretability.

#### **2.2.4 Small Sample Size and Domain Scarcity in Pediatric ICUs**

Another critical bottleneck in building robust AI models for pediatric critical care is the **limited volume of high-quality, labeled data**. Compared to adult ICUs, pediatric and neonatal intensive care units (PICUs/NICUs) manage smaller patient populations, each with highly variable physiological baselines, age-dependent pharmacodynamics, and rapidly changing clinical profiles. This data sparsity limits the statistical power of classical models and increases the risk of overfitting in complex neural networks.

Traditional techniques such as **data augmentation** and **bootstrapping** are often inadequate in these contexts. Emerging methods, such as **transfer learning** where knowledge from adult ICU datasets is adapted to pediatric contexts, and **synthetic data generation** using time-series Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), are proving to be promising. For instance, recent studies have utilized recurrent GANs to generate realistic heart rate and respiratory rate signals, thereby enhancing the diversity of training sets while maintaining biological plausibility.

The SAFE-ICU platform leverages transfer learning from larger datasets (e.g., MIMIC-III) while tuning and validating models specifically on pediatric ICU data from India.

## 2.3 Model-Centric Challenges

As AI in critical care moves from theory to implementation, attention must shift from "can the model predict accurately?" to "can the model be trusted, interpreted, updated, and sustained in a live clinical context?" While deep learning models have demonstrated superior performance on benchmark datasets, the real-world ICU setting introduces several model-related challenges, including opacity, fragility, drift, and generalizability.

### 2.3.1 Black-Box Behavior and Lack of Interpretability

Deep models, especially CNNs, RNNs, and transformers, have become standard in ICU prediction tasks due to their ability to learn from raw, high-dimensional data. However, their **opaque decision boundaries** hinder interpretability, raising a barrier to clinical acceptance.

- **Clinical Risk:** In high-stakes environments, such as ICUs, clinicians require transparent reasoning. Black-box predictions without causal pathways or rationale are ethically and operationally problematic ([Rudin, 2019](#)).
- **Interpretability Tools:**
  - SHAP ([Lundberg & Lee, 2017](#)): Assigns feature-level contributions to individual predictions using Shapley values from cooperative game theory.
  - LIME: Perturbs input data to approximate local decision boundaries using interpretable models ([Ribeiro et al., 2016](#)).
  - Attention Maps: Used in transformer-based models and CNNs to visualize feature importance over time or space.

While these tools provide post hoc explanations, they often lack robustness under distributional shift and may mislead when the underlying model logic changes subtly.

### 2.3.2 Model Drift and Updating in Clinical Environments

Even well-performing models decay over time due to **concept drift** (change in the relationship between inputs and outputs) or **covariate drift** (change in the distribution of input features). For example, changes in patient demographics, treatment protocols, or monitoring devices can subtly degrade performance ([Widmer & Kubat, 1996](#); [Gama et al., 2014](#)).

#### Types of Drift:

- Gradual Drift: Slow changes in patient population or clinical guidelines.
- Sudden Drift: Shocks like COVID-19, which instantly change data distributions.

- Recurring Drift: Seasonal variations in illness (e.g., influenza).

#### **Drift Detection and Adaptation Techniques:**

- Statistical Monitoring: ADWIN, Page-Hinkley tests, and Kolmogorov–Smirnov tests for detecting shifts.
- Model-based Adaptation:
  - Ensemble Learning: Use of multiple models trained on different temporal windows.
  - Online Learning: Continuously updates model weights as new data arrives.
  - Meta-Learning: Learns how to adapt across different environments and datasets.

#### **2.3.3 Fragility and Adversarial Sensitivity**

Recent studies have shown that even state-of-the-art medical models are vulnerable to adversarial perturbations, often imperceptible input changes that can dramatically alter output predictions ([Finlayson et al., 2019](#)).

- Sensor Sensitivity: ICU sensors are prone to artifacts (e.g., loose leads, motion noise). A model sensitive to these artifacts may produce erratic or unsafe outputs.
- Robustness Evaluation:
  - Saliency Maps and Occlusion Tests assess which features dominate predictions.
  - Noise-Injection Testing and Sensitivity Analysis help determine how input fluctuations affect outcomes.

#### **2.3.4 Overfitting and Lack of Generalization**

High-capacity models often tend to overfit to the unique characteristics of the training data, especially when trained on single-center or small datasets. This overfitting results in inflated internal validation metrics that do not accurately reflect the model's generalizability. Consequently, such models exhibit poor performance when applied to external datasets and carry a significant risk of failure in real-world clinical settings.

Strategies for mitigation include:

- Data Augmentation (e.g., time-warping, window slicing).
- Regularization techniques (e.g., dropout, weight decay).

- Cross-institution training and testing (as done in SIngnose across SAFE-ICU, MIMIC-III, and eICU).
- Transfer Learning and Fine-Tuning on small but local datasets.

### 2.3.5 Model Complexity vs. Clinical Utility

There is an inherent trade-off between model performance and clinical usability. Although complex models may achieve higher AUCs and predictive accuracy, they often demand greater computational resources, larger volumes of training data, and specialized infrastructure. These requirements can hinder their practical deployment, particularly in resource-constrained clinical settings.

Clinicians, in contrast, often prefer simpler models that offer reasonable accuracy with high interpretability.

*Table 2.1 Summary table of model-centric challenges and SAFE-ICU solutions.*

| Challenge                         | Implications                            | SAFE-ICU Mitigation                                       |
|-----------------------------------|-----------------------------------------|-----------------------------------------------------------|
| Lack of Interpretability          | Clinician mistrust, regulatory barriers | SHAP explanations, rule-based logic, feature-based models |
| Drift and Concept Shift           | Performance decay, unsafe outputs       | Drift monitoring, retraining thresholds, ensemble models  |
| Fragility/Adversarial Inputs      | Patient safety risk, erratic behavior   | Noise filtering, SHAP variance analysis, Kalman smoothing |
| Overfitting to Local Data         | Failure in external settings            | Multicenter validation, regularization, transfer learning |
| Model-Clinical Usability Tradeoff | Low adoption of high-performance tools  | Modular hybrid design, user-driven model selection        |

## 2.4 Validation and Evaluation Challenges

The credibility of AI models in critical care hinges not only on their performance metrics in retrospective datasets but also on how well they generalize, align with clinical ground truths, and perform under real-world operational constraints. The pathway from lab-grade evaluation to live deployment is fraught with barriers that must be methodically addressed. This section outlines these limitations and explains how SAFE-ICU addresses them systematically.

### 2.4.1 Single-Center Overfitting and Lack of External Validation

AI models in ICU research are often trained and evaluated on data from a single center, frequently on well-curated public datasets such as MIMIC-III. While such datasets provide a foundation, they do not capture the variability in patient demographics, sensor resolution, EHR practices, or therapeutic regimens seen across institutions and regions ([Johnson et al., 2016](#); [Harutyunyan et al., 2019](#)). Consequently:

- Models may overfit to the idiosyncrasies of one dataset.
- Reported performance (e.g., AUROC, precision, recall) may not replicate in other environments.
- Generalization failure can undermine safety and clinician trust when deployed.

### 2.4.2 Ambiguity in Clinical Ground Truths

Establishing ground truth in critical care is notoriously challenging, as labels for conditions such as sepsis, shock, or hypothermia are often unreliable. These labels may be subjective, relying on clinical notes or physician discretion; retrospectively assigned, typically based on ICD-10 codes; or temporally misaligned, with significant delays between the actual onset of the condition and its documentation.

For instance, a shock diagnosis may occur hours after the physiological onset, introducing **label leakage** or temporal bias. Furthermore, definitions evolve, e.g., Sepsis-1 vs. Sepsis-3 ([Singer et al., 2016](#)), leading to inconsistencies across datasets. These factors contribute to label noise, which degrades the performance of supervised models, while uncertainty in event timing further distorts the reliability of lead-time-based prediction models.

### 2.4.3 Lack of Prospective and Real-Time Evaluation

The vast majority of AI models in healthcare are evaluated offline using test-train splits of retrospective data, which fail to capture the complexities of real-world deployment. Such evaluations do not simulate scenarios involving delayed or missing inputs in real-time, human-AI interaction and override behavior, or the effects of alert fatigue and clinician desensitization. One of the few exceptions is Sendak et al. ([2020](#)), who evaluated a real-time sepsis predictor and found that live performance often lagged behind retrospective estimates due to clinician distrust and environmental noise.

### 2.4.4 Evaluation Metrics Beyond Accuracy

Commonly reported metrics, such as AUROC, AUPRC, and F1-score, are valuable but insufficient for evaluating clinical utility. They fail to account for:

- **Timeliness:** How early is the risk predicted before it is clinically recognized?
- **Actionability:** Can clinicians take any action based on the prediction?
- **Outcome Influence:** Does the model reduce mortality, ICU length of stay, or intervention delay?

Emerging evaluation strategies include:

- **Lead Time Evaluation:** Number of hours between model flag and actual deterioration.
- **Decision Curve Analysis:** Weighing true positives against false positives at different threshold levels.
- **Clinical Utility Metrics:** Days in ICU, adverse event rates, and rate of unnecessary interventions.

## 2.4.5 Interpretability and Clinician Validation

Validation must go beyond numerical performance to include **interpretability validation**:

- Do the features driving predictions align with clinical intuition?
- Can clinicians anticipate or confirm model decisions based on their expertise and experience?

*Table 2.2 Summary table validation challenges and SAFE-ICU solutions*

| Challenge                        | Impact                                     | SAFE-ICU Mitigation                                                            |
|----------------------------------|--------------------------------------------|--------------------------------------------------------------------------------|
| Single-center overfitting        | Poor generalization                        | Cross-institution training/validation: SAFE-ICU, MIMIC, eICU                   |
| Ground truth ambiguity           | Label noise, leakage                       | Physiological thresholds, consensus-based annotation                           |
| Lack of real-time validation     | Alert fatigue, lower live performance      | Live dashboard deployment, clinician feedback integration                      |
| Limited clinical utility metrics | Disconnect between prediction and outcomes | Lead-time tracking, decision curve analysis, and post-alert patient monitoring |
| Interpretability validation      | Misalignment with clinical heuristics      | SHAP visualizations, clinician feature review studies                          |

## **2.5 Deployment and Operationalization Challenges**

The deployment of AI models in critical care extends far beyond technical performance; it requires seamless integration with clinical workflows, adherence to regulatory compliance, readiness of infrastructure, and sustained monitoring. Despite academic enthusiasm, the chasm between prototype and practice remains wide. This section explores the barriers to operationalization and the structural innovations adopted in the SAFE-ICU platform.

### **2.5.1 Infrastructure and EHR Integration**

Most healthcare AI prototypes remain research tools, disconnected from hospital infrastructure. Deployment requires real-time integration with EHR systems (e.g., Cerner, Epic), Laboratory Information Systems (LIS), Picture Archiving and Communication Systems (PACS), and bedside monitors. Standards such as HL7 and FHIR aim to facilitate interoperability, but they are not universally adopted or consistently implemented.

Key technical and organizational challenges include:

- Lack of standardization in data schemas and timestamps.
- Hospital IT departments' reluctance to deploy third-party tools due to cybersecurity risks.
- Difficulty in ensuring uptime and latency control for real-time use cases.

### **2.5.2 Resource Constraints in Low-Resource Settings**

Many AI tools are developed with the assumption of access to high-performance GPUs, reliable network connectivity, and dedicated machine learning operations (MLOps) teams. However, these assumptions often do not hold in LMICs or even in rural hospitals within high-income nations. In such settings, servers may lack sufficient random access memory (RAM) or computational power, internet connectivity may be intermittent, rendering cloud-dependent systems unreliable, and clinicians may have limited bandwidth to monitor, interpret, or manage AI-generated alerts.

### **2.5.3 Post-Deployment Maintenance (MLOps)**

AI models must be continuously monitored post-deployment. This includes tracking input drift, updating model weights, logging failures, and revalidating outputs, collectively referred to as MLOps. However, healthcare lacks the same level of DevOps maturity found in commercial software development.

Without structured MLOps:



- Models may silently degrade in performance due to **concept drift**.
- Bugs or infrastructure outages can compromise **patient safety**.
- Regulatory non-compliance becomes a risk as model behavior changes over time.

#### 2.5.4 Legal and Regulatory Challenges

AI systems in healthcare are increasingly subject to regulatory oversight. In the U.S., the FDA has released draft guidelines for "Software as a Medical Device" (SaMD), emphasizing explainability, transparency, and lifecycle risk management throughout the device's development and use. In the EU, the MDR framework requires CE certification for software impacting clinical decisions.

Challenges include:

- Vague definitions of “decision support” vs. “decision making.”
- Regulatory Grey Zones for Continuously Learning Systems.
- Lack of standard reporting for model validation and updates.

#### 2.5.5 Clinician Usability and Human-Centered Design

Even technically sound models fail if clinicians reject them due to poor UI/UX, interruptive alerts, or lack of perceived value. Human factors such as trust, cognitive load, and decision accountability must be integrated into system design.

*Table 2.3 Summary Table for deployment barriers and SAFE-ICU innovations*

| Barrier                  | Impact                                        | SAFE-ICU Innovation                                               |
|--------------------------|-----------------------------------------------|-------------------------------------------------------------------|
| Resource limits in LMICs | Unsustainable computation or cloud dependency | Edge-optimized models, on-premise deployments, low-latency design |
| Lack of MLOps pipelines  | Silent degradation, audit risk                | Continuous monitoring, drift alerts, and retraining support       |
| Regulatory ambiguity     | Delays in approval or risk of non-compliance  | CONSORT-AI/TRIPOD-AI alignment, explainability features           |
| Clinician resistance     | Low adoption, override of predictions         | Co-designed UI, role-based alerts, and real-time annotation tools |

## 2.6 Real-World Evaluation Metrics

Even the most technically sophisticated AI models must ultimately be judged by their clinical utility, economic value, safety, and adoption in real-world settings. Traditional metrics, such as AUROC or F1-score, provide only a partial picture of effectiveness. In this section, we expand the evaluation lens to include patient outcomes, cost-effectiveness, human factors, and regulatory readiness, key aspects for translating AI into sustainable clinical tools.

### 2.6.1 Clinical Effectiveness and Patient Outcomes

Most ICU-focused AI literature reports internal validation metrics, such as accuracy, AUROC, precision-recall, or calibration. However, these metrics are insufficient to assess **the real clinical benefit**, especially in high-stakes domains such as intensive care. What truly matters is:

- **Lead time:** How far in advance can a condition like shock or sepsis be predicted?
- **Intervention impact:** Did the model enable earlier or better intervention?
- **Patient-level outcomes:** Did it reduce ICU length of stay, mortality, or readmission?

Studies such as [Topol \(2019\)](#) and [Sendak et al. \(2020\)](#) advocate for the inclusion of **clinically meaningful endpoints** in evaluations.

### 2.6.2 Economic Evaluation and Cost-Benefit Analysis

Despite AI's potential to reduce costs through early interventions and efficient resource use, very few studies offer robust economic evaluations. Methods such as:

- Cost-utility analysis (using QALYs or DALYs),
- Decision curve analysis,
- Budget impact models

### 2.6.3 Safety and Risk Management

AI deployment in healthcare must proactively account for potential failure modes, such as false positives that contribute to alarm fatigue or false negatives that may delay critical interventions. Risk assessment should be conducted along multiple axes, including error propagation; particularly in modular systems where one component's mistake can cascade through the pipeline; clinical override behavior, which reflects how frequently clinicians reject or ignore AI alerts; and the broader legal and ethical consequences of relying on automated decision support in high-stakes environments.

### 2.6.4 Capacity, Training, and Organizational Readiness

The success of an AI deployment is tied to organizational capacity: are clinicians trained to use the tool? Is there support staff for maintenance? Are workflows adapted to act on AI insights?

Studies show that failure to address sociotechnical readiness often leads to non-adoption or even adverse events, despite the presence of technically valid models ([Greenhalgh et al., 2017](#); [Vayena et al., 2018](#)).

### 2.6.5 Regulatory Evaluation Readiness

Institutions and developers must document models using standardized reporting frameworks such as:

- Consolidated Standards of Reporting Trials in Artificial Intelligence (CONSORT-AI) for clinical trials involving AI
- Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis in Artificial Intelligence (TRIPOD-AI) for predictive model development and validation

*Table 2.4 Summary table of real-world evaluation dimensions*

| Evaluation Dimension    | Traditional Metric     | Real-World Metric                           | SAFE-ICU Feature                                |
|-------------------------|------------------------|---------------------------------------------|-------------------------------------------------|
| Clinical Performance    | AUROC, F1-score        | Lead time, reduced mortality, ICU-free days | Predictive horizons, alert-to-action tracking   |
| Economic Value          | N/A                    | Cost reduction, QALYs, resource utilization | Modular deployment, log collection              |
| Safety                  | Precision/Recall       | Harm risk, override rates, alert fatigue    | Override logging, SHAP-based interpretability   |
| Capacity & Training     | Not addressed          | Training completion, alert adherence        | Interactive training suite, clinician co-design |
| Regulatory Preparedness | Informal documentation | CONSORT-AI, SPIRIT-AI, IRB readiness        | Structured model documentation, ethical audits  |

## **2.7 Research Gaps and Thesis Contributions**

The rapid proliferation of AI tools in healthcare has led to a wide array of predictive, diagnostic, and surveillance models. However, a closer examination of the existing literature and deployed technologies reveals critical gaps that limit the translational impact of these systems in ICUs, particularly in pediatric settings. This section synthesizes the key limitations across the reviewed domains and positions the SAFE-ICU platform as a response to those gaps.

### **2.7.1 Identified Research Gaps**

- 1. Lack of Generalizable Models**

Most AI systems in ICU settings are trained on single-center data, which limits their transferability. Differences in patient demographics, clinical practices, and data capture protocols across hospitals exacerbate this.

- 2. Siloed, Single-Task Models**

Existing systems are typically focused on narrow problems (e.g., predicting sepsis or identifying DDIs) and lack interoperability across tasks.

- 3. Limited Multimodal Integration**

Few AI systems leverage the full spectrum of ICU data, structured vitals, unstructured notes, imaging, and microbiology reports, despite their combined diagnostic power.

- 4. Poor Interpretability of Deep Models**

Clinician trust is hindered by black-box predictions with little explanation, especially in deep learning frameworks used for time-series or image-based prediction.

- 5. Inadequate Real-Time Usability**

Most systems are retrospective and not embedded into live workflows, making them impractical for daily clinical decision support.

- 6. Underuse of LLMs for Structured and Multimodal ICU Data**

While NLP has been transformative in processing unstructured data, large language models (LLMs) remain underexplored for structured time-series encoding.

- 7. Sparse Evaluation of Economic Impact**

The cost-effectiveness, resource utilization, and economic feasibility of AI tools are rarely addressed, particularly in low-resource settings.

### **2.7.2 SAFE-ICU: Contributions and Solutions**

This thesis contributes to the advancement of trustworthy, clinically relevant AI by developing the SAFE-ICU platform, a modular, open-source, and real-time deployable solution designed for bedside integration in Pediatric ICUs. It emphasizes translational impact, bridging the gap

between research prototypes and real-world clinical utility. By ensuring accessibility and availability in low-resource settings, incorporating explainability mechanisms such as SHAP and rule-based alerts, and validating across diverse datasets, the platform promotes both generalizability and clinician trust. The system is designed with a plug-and-play architecture that enables modular deployment, allowing hospitals to select and customize components according to their specific needs. Moreover, by integrating multimodal data streams (vitals, prescriptions, imaging, microbiology) and aligning predictions with clinical workflows, SAFE-ICU demonstrates a significant step toward scalable and sustainable AI-driven critical care. Specific innovations and translational contributions are summarized below:

*Table 2.5 Addressing Existing Research Gaps Through SAFE-ICU*

| Research Gap                       | SAFE-ICU Solution                                                                                                                  |
|------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|
| Generalization across institutions | Multi-institutional training and validation on MIMIC-III, eICU, and SAFE-ICU; cross-dataset AUC and lead-time analysis             |
| Task interoperability              | Modular design integrating ThermoGnose (hypothermia), SInose (shock index), ContraIndicator (drug safety), and AMR surveillance    |
| Multimodal data fusion             | Integration of time-series vitals, typed prescriptions, thermal imaging, and microbiology AST reports                              |
| Interpretability                   | SHAP-based attribution in ThermoGnose and SInose; alerts of DDI in ContraIndicator; interpretable dashboards                       |
| Real-time deployment               | Lightweight RESTful APIs for real-time prediction; clinician-facing interfaces with live updates and feedback logging              |
| Structured data modeling with LLMs | Prompt-based LLaMA embeddings used to encode vitals and predict shock, benchmarked against traditional ML and deep learning models |
| Economic and resource feasibility  | Supports local edge deployment without cloud dependence; module-wise activation to minimize hardware and staffing burdens          |

# Chapter- 3

## Construction of the SAFE-ICU Data Lake and Data Warehouse: Multimodal Integration and Preprocessing

### 3.1 Introduction

A robust, comprehensive, and interoperable data foundation is essential for building effective clinical AI systems. The SAFE-ICU platform relies on an integrated multimodal data lake curated from real-world ICU environments, with a particular focus on pediatric critical care in resource-constrained settings. This chapter presents the systematic construction of the SAFE-ICU data lake, detailing the acquisition, integration, preprocessing, and quality assurance workflows that underpin all subsequent predictive and decision-support modules.

The objective was to create a unified representation of ICU patients by harmonizing structured data (e.g., vital signs, labs, prescriptions) and unstructured content (e.g., clinical notes, thermal imaging) into a machine-readable format. The resulting data infrastructure supports a diverse array of downstream tasks, including hemodynamic shock prediction, drug interaction analysis, AMR surveillance, and live ICU monitoring.

### 3.2 Data Sources and Modalities

The SAFE-ICU data lake aggregates data from multiple clinical subsystems within the ICU (Figure 3.1), each contributing a different aspect of the patient's clinical trajectory:

- I. **Vitals and Monitoring Data:** Collected continuously from bedside monitors, including heart rate, respiratory rate, temperature, SpO<sub>2</sub>, and blood pressure. Sampling frequencies range from 1 Hz to 0.1 Hz, depending on the parameter and monitoring equipment.
- II. **Laboratory Results:** Periodically recorded hematological and biochemical parameters, including white blood cell count, creatinine, lactate, C-reactive protein (CRP), hemoglobin, and others relevant for sepsis and organ dysfunction diagnosis.
- III. **Prescription and Medication Data:** Extracted from the treatment chart and digital prescription systems. Includes drug name, dosage, frequency, route of administration, and timing information. These are processed via NLP for NER, summarization, and contextual encoding.

- IV. **Thermal Imaging Videos:** Non-contact thermal recordings acquired using FLIR-based devices placed near ICU beds. These videos are primarily used in the *ThermalShockNet* module for predicting shock risk.
- V. **Microbiology Reports:** Structured outputs of AST, including organism identified, sample type, drug tested, and resistance phenotype.

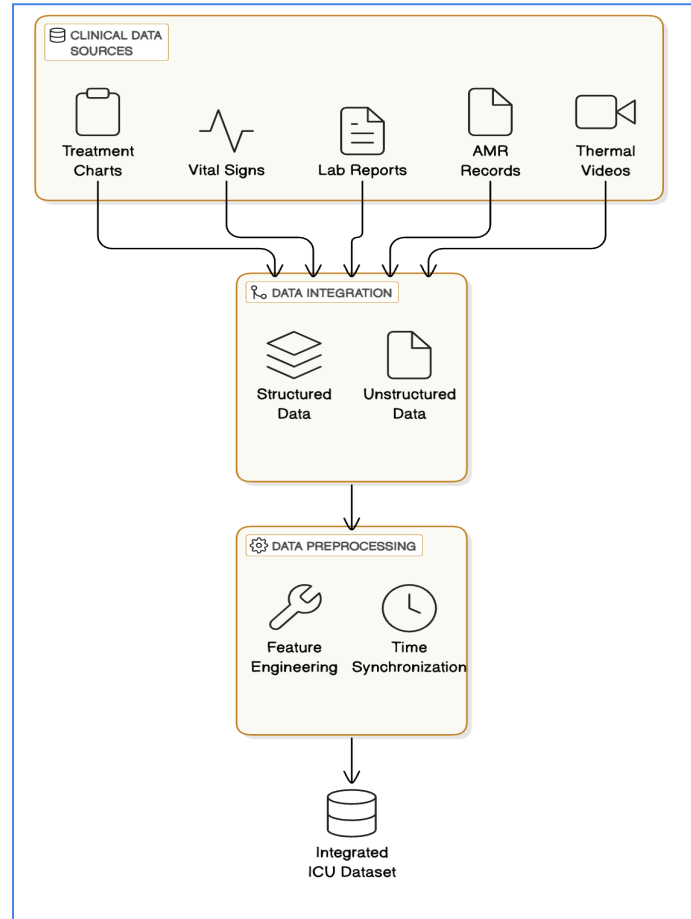


Figure 3.1 Data Sources and Modalities in the SAFE-ICU Data Lake.

### 3.3 Data Ingestion and Storage Architecture

The SAFE-ICU data lake is built to robustly ingest, normalize, and store heterogeneous data modalities originating from disparate ICU information systems. To accommodate the real-world structure and semantics of clinical data, each modality undergoes a customized ingestion pipeline that ensures fidelity, temporal alignment, and semantic harmonization. Below is a breakdown of each data stream and its associated ingestion pipeline.

### 3.3.1. Vitals and Monitoring Data

In the SAFE-ICU system, physiological vitals such as HR, RR, SpO<sub>2</sub>, systolic/diastolic blood pressure (NIBP), and temperature (T1, T2) are continuously monitored and streamed from ICU bedside devices. These monitors are integrated through a Central Monitoring System (CMS) that acts as the intermediate data collection layer within the ICU's network infrastructure.

#### Step 1: Data Transfer from Monitors to CMS (Local Intranet)

Each ICU bed is equipped with a high-frequency patient monitor that continuously captures and streams vitals. The monitors are connected over a local area network (LAN) to a centralized CMS server that collects and displays this information in real time for clinicians.

#### Step 2: Pulling Vitals from CMS using the SAFE-ICU Client-Server Pipeline

To access and organize this data for modeling, a custom **client-server pipeline** was developed:

- The client periodically *pulls both unsolicited at 15s and solicited files* from the CMS

#### Daily Binary Dump from CMS

- Each day, the CMS logs unsolicited HL7-encoded messages into a flat binary file that aggregates readings from all ICU beds.

**Raw Format:** vitals data appears as high-frequency time-stamped rows where each row corresponds to a measurement from sensors (e.g., heart rate, respiratory rate, temperature, blood pressure).



## Raw Format Illustration

```

Date 22:34:59.369
MSH|^~&M|MindrayGateway|||ORU^R01|2|P|2.3.1|
PID|Patients id|PAL^|20130101|M|||^^^|
PV1|I|^P|ICU6563232235625660|||P|20190621000000
OBR||Mindray Monitor||20190706224146|
OBX|NM|52^Height|106.0|||F
OBX|NM|51^Weight|11.8|||F
OBX|CE|2302^BloodType|0^|||F
OBX|CE|2303^PACE_Switch|0^Off|||F
OBX|NM|101^HR|2101|97|||F
OBX|NM|102^PVCs|2101|-100|||F
OBX|NM|106^ST-II|2101|-100.0|||F
OBX|CE|2038^|101|0^|||F
OBX|CE|2039^|101|0^|||F
OBX|CE|2038^|102|0^|||F
OBX|CE|2039^|102|0^|||F
OBX|CE|2038^|106|0^|||F
OBX|CE|2039^|106|0^|||F
OBX|NM|151^RR|2102|35|||F
OBX|CE|2038^|151|0^|||F
OBX|CE|2039^|151|0^|||F
OBX|NM|888^SpO2 Desat|2103|-100|||F
OBX|NM|160^SpO2|2103|93|||F
OBX|CE|2038^|888|0^|||F
OBX|CE|2039^|888|0^|||F
OBX|CE|2038^|160|0^|||F
OBX|CE|2039^|160|0^|||F
OBX|NM|200^T1|2104|34.90|||F
OBX|NM|201^T2|2104|35.00|||F
OBX|NM|202^Td|2104|0.10000|||F
OBX|CE|2038^|200|0^|||F
OBX|CE|2039^|200|0^|||F
OBX|CE|2038^|201|0^|||F
OBX|CE|2039^|201|0^|||F
OBX|CE|2038^|202|0^|||F
OBX|CE|2039^|202|0^|||F
OBX|NM|500^ART-Sys|2116|119|||F
OBX|NM|501^ART-Mean|2116|74|||F
OBX|NM|502^ART-Dia|2116|51|||F
OBX|CE|2038^|500|0^|||F
OBX|CE|2039^|500|0^|||F
OBX|CE|2038^|501|0^|||F
OBX|CE|2039^|501|0^|||F
OBX|CE|2038^|502|0^|||F
OBX|CE|2039^|502|0^|||F
OBX|NM|600^PR|2115|90|||F
OBX|CE|2038^|600|0^|||F
OBX|CE|2039^|600|0^|||F
OBX|NM|171^Dia|2105|61|||F|APERIODIC|20190630082938
OBX|NM|172^Mean|2105|87|||F|APERIODIC|20190630082938
OBX|NM|170^Sys|2105|107|||F|APERIODIC|20190630082938
OBX|CE|4||2419^Recorder Out of Paper|||||F|TECH_ALM|
OBX|CE|4||2422^Recorder Queue Full|||||F|TECH_ALM|
Date 22:34:59.582

```

Figure 3.2 Raw Format Illustration of ICU Vital Signs Data.

### 3.3.1.1 Vitals Data Preprocessing: Unsolicited CMS Data Stream

To transform the unsolicited bedside monitor data into a usable form, a custom R-based preprocessing pipeline was developed to extract vital signs from unstructured HL7-like logs generated by the CMS. These raw data streams, received in 15-second intervals across all ICU beds, are parsed to identify key metadata, including patient identifiers, timestamps, and vital sign measurements (e.g., heart rate, respiratory rate, SpO<sub>2</sub>, temperature, and blood pressure). The script applies a series of regex-based filters to isolate relevant OBX segments, dynamically

generates structured data frames for each patient, and removes artifacts, missing values, or formatting inconsistencies. The final output is a cleaned, timestamp-aligned “.csv” file per patient, ready for integration into the SAFE-ICU data lake, where it can be used for modeling, visualization, or real-time monitoring tasks.

## Cleaned Vital Illustration

|    | time_start | PID | PN1  | Height | Weight | HR   | PVCs   | ST/BE | PR   | SpO2.Desat | SpO2 | awRR | ECG2 | PIC2 | ART.Sys | ART.Mean | ART.Dia | Dis  | Mean   | Sys    | ST/BE  | PR           | time_end     | datestamp    | timestamp |
|----|------------|-----|------|--------|--------|------|--------|-------|------|------------|------|------|------|------|---------|----------|---------|------|--------|--------|--------|--------------|--------------|--------------|-----------|
| 1  | 900012.882 | NA  | 61.0 | 5.8    | 127    | -100 | -130.0 | 30    | -100 | 96         | -100 | -100 | -100 | 83   | 59      | 47       | 52      | 61   | 86     | -99    | -99    | 00:00:12.885 | 00:00:12.885 |              |           |
| 2  | 900012.885 |     | 3    | 0.0    | 6.0    | 142  | -100   | -99   | 21   | -100       | -100 | -99  | -99  | -99  | -99     | -99      | -99     | 62   | 79     | 111    | -100.0 | -100         | 00:00:12.888 | 00:00:12.888 |           |
| 3  | 900013.074 |     | 6    | 60.0   | 4.0    | 131  | -100   | -99   | 101  | -100       | 100  | -99  | -99  | -99  | 84      | 46       | 37      | 51   | 58     | 77     | -100.0 | 101          | 00:00:13.082 | 00:00:13.077 |           |
| 4  | 900013.262 | 8   | 0.0  | 13.1   | 136    | -100 | -99    | 20    | -100 | 99         | -99  | -99  | -99  | -99  | -99     | -99      | -100    | -100 | -100.0 | -100.0 | 89     | 00:00:13.449 | 00:00:13.266 |              |           |
| 5  | 900013.448 | 4   | 0.0  | 28.0   | 138    | -100 | -99    | 39    | -100 | 98         | -99  | -99  | -99  | 106  | 77      | 62       | -100    | -100 | -100.0 | -100.0 | 108    | 00:00:13.683 | 00:00:13.453 |              |           |
| 6  | 900013.683 | 2   | 0.0  | 0.0    | -99    | -99  | -99    | -99   | -99  | -99        | -99  | -99  | -99  | -99  | -99     | -99      | -99     | -99  | -99    | -99    | -99    | 00:00:13.886 | 00:00:13.686 |              |           |
| 7  | 900013.886 | 5   | 0.0  | 14.0   | 95     | 0    | -99    | 21    | -100 | 95         | -99  | -99  | -99  | 96   | 66      | 51       | 43      | 56   | 82     | -100.0 | 95     | 00:00:14.089 | 00:00:13.886 |              |           |
| 8  | 900014.089 | 7   | 0.0  | 3.0    | 140    | -100 | -99    | 32    | -100 | 94         | -99  | -99  | -99  | -99  | -99     | -99      | 89      | 100  | 121    | -100.0 | 141    | 00:00:27.679 | 00:00:14.089 |              |           |
| 9  | 900017.679 | NA  | 61.0 | 5.8    | 135    | -100 | -130.0 | 27    | -100 | 96         | -100 | -100 | -100 | 84   | 61      | 49       | 52      | 61   | 86     | -99    | -99    | 00:00:27.682 | 00:00:27.677 |              |           |
| 10 | 900017.882 | 3   | 0.0  | 6.0    | 140    | -100 | -99    | 21    | -100 | -100       | -99  | -99  | -99  | -99  | -99     | -99      | -99     | 62   | 79     | 111    | -100.0 | -100         | 00:00:28.070 | 00:00:27.882 |           |
| 11 | 900018.070 | 6   | 60.0 | 4.0    | 131    | -100 | -99    | 130   | -100 | 98         | -99  | -99  | -99  | 85   | 47      | 38       | 51      | 58   | 77     | -100.0 | 100    | 00:00:28.257 | 00:00:28.077 |              |           |
| 12 | 900018.257 | 8   | 0.0  | 13.1   | 137    | -100 | -99    | 16    | -100 | 99         | -99  | -99  | -99  | -99  | -99     | -99      | -100    | -100 | -100.0 | -100.0 | 107    | 00:00:28.490 | 00:00:28.257 |              |           |
| 13 | 900018.490 | 4   | 0.0  | 28.0   | 137    | -100 | -99    | 26    | -100 | 98         | -99  | -99  | -99  | 96   | 70      | 57       | -100    | -100 | -100.0 | -100.0 | 107    | 00:00:28.679 | 00:00:28.490 |              |           |
| 14 | 900018.679 | 2   | 0.0  | 0.0    | -99    | -99  | -99    | -99   | -99  | -99        | -99  | -99  | -99  | -99  | -99     | -99      | -99     | -99  | -99    | -99    | -99    | 00:00:28.882 | 00:00:28.677 |              |           |
| 15 | 900018.882 | 5   | 0.0  | 14.0   | 96     | 0    | -99    | 22    | -100 | 95         | -99  | -99  | -99  | 96   | 65      | 49       | 43      | 56   | 82     | -100.0 | 95     | 00:00:29.085 | 00:00:28.882 |              |           |
| 16 | 900019.085 | 7   | 0.0  | 3.0    | 140    | -100 | -99    | 29    | -100 | 94         | -99  | -99  | -99  | -99  | -99     | -99      | 89      | 100  | 121    | -100.0 | 141    | 00:00:42.889 | 00:00:29.085 |              |           |
| 17 | 900042.888 | NA  | 61.0 | 5.8    | 130    | -100 | -130.0 | 28    | -100 | 96         | -100 | -100 | -100 | 83   | 59      | 47       | 52      | 61   | 86     | -99    | -99    | 00:00:42.877 | 00:00:42.888 |              |           |
| 18 | 900042.877 | 3   | 0.0  | 6.0    | 139    | -100 | -99    | 25    | -100 | -100       | -99  | -99  | -99  | -99  | -99     | -99      | -99     | 62   | 79     | 111    | -100.0 | -100         | 00:00:43.084 | 00:00:42.877 |           |
| 19 | 900043.064 | 6   | 60.0 | 4.0    | 130    | -100 | -99    | 130   | -100 | 100        | -99  | -99  | -99  | 84   | 46      | 37       | 51      | 58   | 77     | -100.0 | 101    | 00:00:43.267 | 00:00:43.064 |              |           |
| 20 | 900043.267 | 8   | 0.0  | 13.1   | 137    | -100 | -99    | 19    | -100 | 99         | -99  | -99  | -99  | -99  | -99     | -99      | -100    | -100 | -100.0 | -100.0 | 107    | 00:00:43.454 | 00:00:43.267 |              |           |
| 21 | 900043.454 | 4   | 0.0  | 28.0   | 138    | -100 | -99    | 41    | -100 | 98         | -99  | -99  | -99  | 98   | 72      | 59       | -100    | -100 | -100.0 | -100.0 | 109    | 00:00:43.688 | 00:00:43.454 |              |           |
| 22 | 900043.688 | 2   | 0.0  | 0.0    | -99    | -99  | -99    | -99   | -99  | -99        | -99  | -99  | -99  | -99  | -99     | -99      | -99     | -99  | -99    | -99    | -99    | 00:00:43.876 | 00:00:43.688 |              |           |
| 23 | 900043.876 | 5   | 0.0  | 14.0   | 95     | 0    | -99    | 22    | -100 | 95         | -99  | -99  | -99  | 97   | 65      | 49       | 43      | 56   | 82     | -100.0 | 95     | 00:00:44.089 | 00:00:43.876 |              |           |
| 24 | 900044.089 | 7   | 0.0  | 3.0    | 142    | -100 | -99    | 27    | -100 | 92         | -99  | -99  | -99  | -99  | -99     | -99      | 89      | 100  | 121    | -100.0 | 142    | 00:00:57.888 | 00:00:44.089 |              |           |
| 25 | 900057.888 | NA  | 61.0 | 5.8    | 127    | -100 | -130.0 | 30    | -100 | 96         | -100 | -100 | -100 | 85   | 60      | 48       | 52      | 61   | 86     | -99    | -99    | 00:00:57.885 | 00:00:57.888 |              |           |
| 26 | 900057.888 | 3   | 0.0  | 6.0    | 140    | -100 | -99    | 29    | -100 | -100       | -99  | -99  | -99  | -99  | -99     | -99      | -99     | 62   | 79     | 111    | -100.0 | -100         | 00:00:58.088 | 00:00:57.888 |           |
| 27 | 900058.068 | 6   | 60.0 | 4.0    | 131    | -100 | -99    | 70    | -100 | 100        | -99  | -99  | -99  | 83   | 46      | 37       | 51      | 58   | 77     | -100.0 | 101    | 00:00:58.292 | 00:00:58.068 |              |           |
| 28 | 900058.292 | 8   | 0.0  | 13.1   | 136    | -100 | -99    | 24    | -100 | 100        | -99  | -99  | -99  | -99  | -99     | -99      | -100    | -100 | -100.0 | -100.0 | 108    | 00:00:58.484 | 00:00:58.292 |              |           |
| 29 | 900058.484 | 4   | 0.0  | 28.0   | 137    | -100 | -99    | 37    | -100 | 98         | -99  | -99  | -99  | 100  | 73      | 59       | -100    | -100 | -100.0 | -100.0 | 108    | 00:00:58.688 | 00:00:58.484 |              |           |
| 30 | 900058.688 | 2   | 0.0  | 0.0    | -99    | -99  | -99    | -99   | -99  | -99        | -99  | -99  | -99  | -99  | -99     | -99      | -99     | -99  | -99    | -99    | -99    | 00:00:58.888 | 00:00:58.688 |              |           |
| 31 | 900058.888 | 5   | 0.0  | 14.0   | 96     | 0    | -99    | 22    | -100 | 95         | -99  | -99  | -99  | 94   | 63      | 48       | 43      | 56   | 82     | -100.0 | 95     | 00:00:59.073 | 00:00:58.888 |              |           |
| 32 | 900059.073 | 7   | 0.0  | 3.0    | 145    | -100 | -99    | 29    | -100 | 91         | -99  | -99  | -99  | -99  | -99     | -99      | 89      | 100  | 121    | -100.0 | 142    | 00:01:12.882 | 00:00:59.073 |              |           |
| 33 | 900112.882 | NA  | 61.0 | 5.8    | 130    | -100 | -130.0 | 30    | -100 | 96         | -100 | -100 | -100 | 85   | 61      | 49       | 52      | 61   | 86     | -99    | -99    | 00:01:12.884 | 00:01:12.882 |              |           |
| 34 | 900112.884 | 3   | 0.0  | 6.0    | 132    | -100 | -99    | 28    | -100 | -100       | -99  | -99  | -99  | -99  | -99     | -99      | -99     | 62   | 79     | 111    | -100.0 | -100         | 00:01:13.087 | 00:01:12.884 |           |
| 35 | 900113.087 | 6   | 60.0 | 4.0    | 131    | -100 | -99    | 76    | -100 | 100        | -99  | -99  | -99  | 83   | 46      | 37       | 51      | 58   | 77     | -100.0 | 101    | 00:01:13.291 | 00:01:13.087 |              |           |

Figure 3.3 Cleaned and Structured Vitals Format from SAFE-ICU Pipeline.

### 3.3.2. Laboratory Test Results

Laboratory test results are ingested from the hospital's Laboratory Information System (LIS), typically available in structured formats such as HL7 messages. Each record includes patient identifiers, test names (e.g., creatinine, lactate, CRP), values, units, and reference ranges, along with the date and time of sample collection. During preprocessing, test names are standardized by mapping them to codes to ensure consistency across sites and over time. The pipeline aligns lab timestamps with other patient data streams to maintain temporal coherence, applies domain-specific outlier detection rules, and imputes missing values where clinically appropriate using test-wise interpolation strategies. The cleaned data are stored in a relational database indexed by patient ID and test category, making it readily accessible for modeling and longitudinal trend analysis in the SAFE-ICU data lake.

| uhid | patient_name | p_age                  | sex    | observation                            | result_val | result_unit                 | normal_range  | lab_id | ward_name | tech_ver_time | sample_collection_date |
|------|--------------|------------------------|--------|----------------------------------------|------------|-----------------------------|---------------|--------|-----------|---------------|------------------------|
|      |              | 1 year 6 months 6 days | Male   | 64_6_TLC_627228_2016                   | 10.5       | 10 <sup>9</sup> /3/ $\mu$ L | 4.00-11.00    | 11     |           |               | 08:15 am               |
|      |              | 2 years 2 months       | Female | 177_26_SERUM CRP_620364_2016           | 55.58      | mg/L                        | 0.00-6.00     | 9      |           |               | 05:02 am               |
|      |              | 14 years 8 days        | Male   | 64_38_PROTHROMBIN TIME(PT)_626629_2016 | 15.300     | sec                         | 11.00-14.00   | 11     |           |               | 03:09 am               |
|      |              | 14 years 8 days        | Male   | 64_81_INR_626629_2016                  | 1.396      |                             |               | 11     |           |               | 03:09 am               |
|      |              | 1 year 6 months 6 days | Male   | 64_27_NUCLEATED RBC_627228_2016        | 0.0        |                             | 0.00          | 11     |           |               | 08:15 am               |
|      |              | 1 year 6 months 6 days | Male   | 64_34_Hb_627228_2016                   | 8.1        | g/dL                        | 12.00-15.00   | 11     |           |               | 08:15 am               |
|      |              | 1 year 6 months 6 days | Male   | 64_2_HCT_627228_2016                   | 27.7       | %                           | 0.00-0.00     | 11     |           |               | 08:15 am               |
|      |              | 1 year 6 months 6 days | Male   | 64_4_PLATELET COUNT_627228_2016        | 60         | 10 <sup>9</sup> /3/ $\mu$ L | 150.00-400.00 | 11     |           |               | 08:15 am               |
|      |              | 5 years 3 days         | Male   | 177_6_CPN(CK)_620371_2016              | 13900      |                             | 40.00-226.00  | 9      |           |               | 05:10 am               |
|      |              | 5 years 3 days         | Male   | 177_12_LDH-TOTAL_620371_2016           | 2313.3     |                             | 200.00-420.00 | 9      |           |               | 05:10 am               |
|      |              | 9 years                | Male   | 64_38_PROTHROMBIN TIME(PT)_626605_2016 | 10.900     | sec                         | 11.00-14.00   | 11     |           |               | 09:31 pm               |
|      |              | 9 years                | Male   | 64_81_INR_626605_2016                  | 0.991      |                             |               | 11     |           |               | 09:31 pm               |
|      |              | 9 years                | Male   | 64_40_APTT_626605_2016                 | 27.700     | sec                         | 26.00-40.00   | 11     |           |               | 09:31 pm               |
|      |              | 1 year 6 months 6 days | Male   | 64_8_NEUTRO_627228_2016                | 42.9       | %                           | 40.00-80.00   | 11     |           |               | 08:15 am               |
|      |              | 1 year 6 months 6 days | Male   | 64_9_LYMPOH_627228_2016                | 49.7       | %                           | 20.00-40.00   | 11     |           |               | 08:15 am               |
|      |              | 1 year 6 months 6 days | Male   | 64_18_MONO_627228_2016                 | 4.7        | %                           | 2.00-10.00    | 11     |           |               | 08:15 am               |
|      |              | 1 year 6 months 6 days | Male   | 64_10_EOSINO_627228_2016               | 2.6        | %                           | 1.00-6.00     | 11     |           |               | 08:15 am               |
|      |              | 1 year 6 months 6 days | Male   | 64_11_BASO_627228_2016                 | 0.1        | %                           | 1.00-2.00     | 11     |           |               | 08:15 am               |
|      |              | 9 years 1 day          | Male   | 170_14_TOTAL PROTEIN_627162_2016       | 5.1        | gm%                         | 6.60-8.70     | 1      |           |               | 08:10 am               |
|      |              | 9 years 1 day          | Male   | 170_15_ALBUMIN_627162_2016             | 2.8        | gm%                         | 3.80-4.00     | 1      |           |               | 08:10 am               |
|      |              | 9 years 1 day          | Male   | 170_5_UREA_627162_2016                 | 159        | mg%                         | 10.00-50.00   | 1      |           |               | 08:10 am               |
|      |              | 9 years 1 day          | Male   | 170_16_GLOBULIN_627162_2016            | 2.3        | gm%                         | 4.00-5.50     | 1      |           |               | 08:10 am               |
|      |              | 9 years 1 day          | Male   | 170_6_CREATININE_627162_2016           | 7.5        | mg%                         | 0.50-1.80     | 1      |           |               | 08:10 am               |
|      |              | 9 years 1 day          | Male   | 170_7_CALCUIUM_627162_2016             | 7.0        | mg%                         | 9.00-11.50    | 1      |           |               | 08:10 am               |
|      |              | 9 years 1 day          | Male   | 170_17_SGOT (AST)_627162_2016          | 30         | IU/L                        | 0.00-50.00    | 1      |           |               | 08:10 am               |
|      |              | 9 years 1 day          | Male   | 170_8_PHOSPHATE_627162_2016            | 7.5        | mg%                         | 2.50-4.50     | 1      |           |               | 08:10 am               |
|      |              | 9 years 1 day          | Male   | 170_9_URIC ACID_627162_2016            | 12.3       | mg%                         | 2.00-7.40     | 1      |           |               | 08:10 am               |
|      |              | 9 years 1 day          | Male   | 170_18_SGPT (ALT)_627162_2016          | 14         | IU/L                        | 0.00-50.00    | 1      |           |               | 08:10 am               |
|      |              | 9 years 1 day          | Male   | 170_10_SODIUM_627162_2016              | 132        | mEq/L                       | 130.00-149.00 | 1      |           |               | 08:10 am               |
|      |              | 9 years 1 day          | Male   | 170_11_POTASSIUM_627162_2016           | 3.5        | mEq/L                       | 3.50-5.00     | 1      |           |               | 08:10 am               |
|      |              | 9 years 1 day          | Male   | 170_19_ALK PHOS (ALP)_627162_2016      | 285        | IU/L                        | 0.00-0.00     | 1      |           |               | 08:10 am               |
|      |              | 9 years 1 day          | Male   | 170_36_BILIRUBIN TOTAL_627162_2016     | 0.3        | mg%                         | 0.10-1.00     | 1      |           |               | 08:10 am               |
|      |              | 9 years 1 day          | Male   | 177_12_LDH-TOTAL_627162_2016           | 2418.4     |                             | 200.00-420.00 | 9      |           |               | 08:10 am               |

Figure 3.4 Raw Format Example of Laboratory Test

### 3.3.3. Prescription and Medication Data

Prescription and medication data in the SAFE-ICU pipeline are sourced from daily nursing treatment charts, which are maintained by attending doctors in Microsoft Word documents. These documents typically include semi-structured, shorthand notations such as “Rx: Inj. PCM 170mg IV BD x 5 days,” often lacking standardized formatting.

To ensure consistent data capture, a scheduled Python-based backup script extracts and stores updated treatment charts for each patient at specific time of the day. These Word files are then parsed using the “Docx-2-txt” module to convert unstructured content into plain text, making it amenable to natural language processing. Subsequent preprocessing includes sentence segmentation, tokenization, and NER using ClinicalBERT to identify drug names, dosages, administration routes, and frequencies.

|                                                            |              |                                               |                                               |                                                |
|------------------------------------------------------------|--------------|-----------------------------------------------|-----------------------------------------------|------------------------------------------------|
| Name: <input style="width: 80%;" type="text"/>             | Age: 5 years | Sex: Female                                   | Bed: <input style="width: 80%;" type="text"/> | UHID: <input style="width: 80%;" type="text"/> |
| DIAGNOSIS: Acute Meningoencephalitis with raised ICP /ADEM |              |                                               |                                               |                                                |
| Consultants: <input style="width: 80%;" type="text"/>      |              | JR.: <input style="width: 80%;" type="text"/> |                                               |                                                |

|         |                                          |
|---------|------------------------------------------|
| Date    | <input style="width: 80%;" type="text"/> |
| Weight  | 17 kg                                    |
| TFR     | 80 %                                     |
| TPV     | 1080 ml                                  |
| IVF     | 550 ml                                   |
| IVM     | 530 ml                                   |
| Glucose | 27.5 gm                                  |
| GIR     | 1.12 mg/kg/min                           |
| Feeds   | NPO                                      |
| K       | 1.94 meq/kg/day                          |
| Na      | 5.0 meq/kg/day                           |

**Feeds:** NPO

**Inotropes:**  
 .STOP Inj Nor- Adrenaline

**IVF:** Inj IVF DHS (3 ml KG: 100 ml IVF) @ 23 ml per hour (552 ml)

**Antimicrobials(230 ml):**

1. Inj. Acyclovir 340 mg in 25 ml 5 % De over 3 hour Q 8 hrly (20mg/kg/day) (75 ml) (D7)
2. Inj Meropenem 700 mg in 15 ml NS for 1 hour & hourly (45 ml) (D7)
3. Inj Vancomycin 350 mg in 50 ml 5% De over 1 hour 6 hourly (80 ml) (D7)
4. Inj Azithromycin 170 mg in 50 ml NS IV over 60 min q 24 hrly ( D4) (50ml)

**Miscellaneous: (280 ml)**

1. Stop Inj Midazolam
2. Stop Inj Fentanyl
3. Inj Phenytoin 85 mg in 20 ml NS over 60 min q 12 hourly (6 8mg/kg/day) (40 ml)
4. Inj Valproate 170 mg in 20 ml NS over 60 min q 12 hourly (6 20mg/kg/day) (40 ml)
5. STOP Inj 3% NaCl
6. Inj PCM 170 mg IV 6 hourly
7. IVIG 5% total 200 ml at 20ml IV over 30 mins followed by 20 ml IV over 30 mins next 30 ml in 1 hour and then rest at 40ml/ hour (200 ml)

**Nursing and supportive care:**

1. Input output monitoring q2h
2. Temperature monitoring q2h
3. R&S monitoring q 6 hourly
4. Clustering of procedures, Asepsis to be maintained
5. Eye care, moisten e/d 2 q 2 hrly, lacrigel ointment apply locally q 12 hrly

Figure 3.5 Raw Format Example of Prescription and Medication Data.

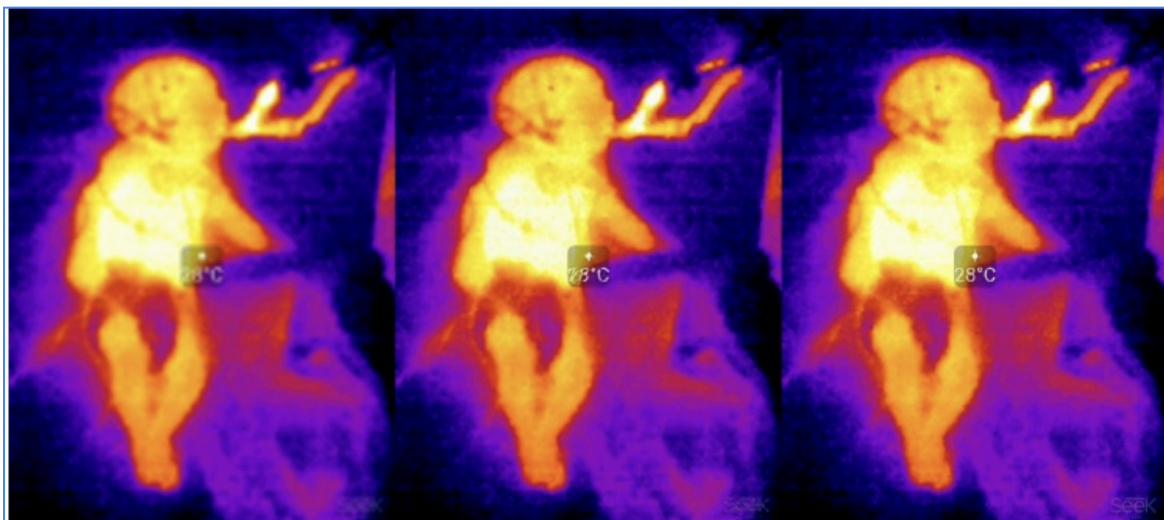
This enriched and structured prescription data are then aligned with patient timelines and made available for further summarization, drug interaction checking, and integration into downstream predictive modeling modules within the SAFE-ICU ecosystem.

| Name | Age      | Sex    | Bed | UHID | Date | Diagnosis                                           | Treatment Chart | Drugs                                           |
|------|----------|--------|-----|------|------|-----------------------------------------------------|-----------------|-------------------------------------------------|
|      | 3 years  | Female |     |      |      | Systemic onset sjIA / MAS                           | —               | Mannitol / Midazolam / Fentanyl / Vancomycin    |
|      | 8 months | Female |     |      |      | Down syndrome / LRTI / VSD / severe PAH / pneumonia | —               | Vancomycin                                      |
|      | 6 years  | Female |     |      |      | Disseminated TB / pancytopenia / shock              | —               | Mannitol / Linezolid / Omeprazole / Imipenem    |
|      | 11 years | Male   |     |      |      | Acute febrile encephalopathy / under evaluation     | —               | Vancomycin / Doxycycline / Midazolam / Fentanyl |
|      | 15 days  | Male   |     |      |      | Term / AGA / HE I / MODS / shock / sepsis           | —               | None                                            |
|      | 6 months | Male   |     |      |      | VSD / CCF / pneumonia                               | —               | Dobutamine / Dopamine / Vasopressin             |
|      |          |        |     |      |      |                                                     |                 |                                                 |

Figure 3.6 Structured Output of Prescription Data After NLP Processing.

### 3.3.4. Thermal Imaging Videos

Thermal imaging in the SAFE-ICU data pipeline serves as a non-contact modality for capturing surface temperature distributions, facilitating clinical inference, particularly in the prediction of shock risk. Videos were acquired using the Seek Thermal Compact (UW-AAA) camera, which features a thermal sensor resolution of 206×156 pixels and supports output resolutions of up to 480×640 pixels at 15 frames per second (fps). The camera operated in the “Iron Theme” mode to enhance thermal contrast. It was mounted at standardized distances depending on bed size: a median (IQR) of  $2.08 \pm 0.15$  meters for long beds and  $1.52 \pm 0.26$  meters for small beds. The ICU environment was maintained at a thermally stable 25°C, and calibration was routinely performed to ensure that the camera’s readings matched those of external probes and thermometers, thereby ensuring clinical-grade accuracy.





*Figure 3.7 Thermal Video Frame Captured for ICU Patient Monitoring.*

To ensure patient privacy and consistency, a standard operating procedure (SOP) was developed and strictly followed. The thermal camera only captures infrared radiation and cannot reveal facial or identifying features, thereby maintaining anonymity. Before video acquisition, all personnel followed ICU hygiene protocols, including donning appropriate protective gear. The camera, mounted on a wall-mounted hanger holding a smartphone interface, was positioned in the presence of nursing staff to avoid patient disturbance. Each video captured the whole body of the patient in a supine position using a consistent color scale. Metadata such as patient UHID, bed number, date, and time were logged in Excel sheets and utilized for structured storage.

Thermal videos were recorded multiple times per patient on different ICU days and under various conditions. This longitudinal, heterogeneous capture strategy ensured that variation in shock status over time could be analyzed without introducing biases related to fixed patient attributes, such as age or sex. Following capture, all video files were securely transferred to a centralized server daily and stored in patient-specific folders for downstream analysis. These videos are later processed and utilized in the ThermalShockNet module for clinical modeling and multimodal prediction in SAFE-ICU.

### **3.3.5. Microbiology Reports**

Microbiology and AST data in the SAFE-ICU system are acquired in a structured tabular format from the hospital's LIS or microbiology lab exports. These datasets include key clinical variables such as Organism Name, Sample Type (e.g., blood, urine, tracheal aspirate), Antibiotic Tested, Sensitivity Result (e.g., Sensitive, Intermediate, Resistant), and Test Date. To ensure consistency and usability across datasets, a reference dictionary of organism names and antimicrobial compounds is maintained and updated on a monthly basis. This dictionary is used to map all entries to their standardized forms, supporting robust downstream analysis and temporal comparisons.

[illegible]

Figure 3.8 Raw Format of Microbiology Reports from the ICU Laboratory System.

The complete data are stored across two files: `patients.csv`, which includes metadata such as patient ID, demographics, and hospitalization details; and `test.csv`, which holds the complete AST test records for each isolate. These two files are linked using patient identifiers and timestamped test events, enabling the construction of longitudinal antibiograms and resistance trend analyses. This modular structure allows real-time visualization and supports decision-making tools in SAFE-ICU's **AMRTrack** module, enhancing infection control and empirical therapy planning.

*Table 3.1 Summary Table of Multimodal Clinical Data*

|                                         |           |
|-----------------------------------------|-----------|
| <b>Vital Signals</b>                    |           |
| Duration                                | 2016-2024 |
| Unique Patients                         | 1,174     |
| Median Patient Hours                    | 23.99     |
| <b>Laboratory Test Results</b>          |           |
| Duration                                | 2015-2024 |
| Unique Patients                         | 212,301   |
| Total Samples                           | 1099      |
| <b>Prescription and Medication Data</b> |           |
| Duration                                | 2016-2024 |
| Unique Patients                         | 1416      |
| Total Samples                           | 13375     |
| <b>Thermal Imaging Videos</b>           |           |
| Duration                                | 2019-2020 |
| Unique Patients                         | 466       |
| Total Samples                           | 120       |
| <b>Microbiology Reports</b>             |           |
| Duration                                | 2015-2025 |
| Unique Patients                         | 1157      |
| Total Samples                           | 545       |



# Chapter- 4

## Predictive Modeling in Intensive Care

### 4.1 Early prediction of hypothermia in pediatric intensive care units using machine learning

#### 4.1.1 Introduction

Hypothermia is a major complication associated with sepsis in the ICUs and is especially in children ([Rumbus et al., 2017](#)). Hypothermia is a situation when the human body temperature is below 35°C (95°F) ([Fears, 2011](#)). It arises when the human body loses more heat compared to what it produces, thus leading to a drop in the core body temperature. There may be some warning signs such as Shivering, Confusion, and Slurred speech, whereas, in the worst scenario, this may eventually lead to organ failure and may also cause death. In resource-limited situations, hypothermia is increasingly identified as a major cause of newborn illness and mortality ([Ibrahim et al., 2021](#)). Also, children are more vulnerable to hypothermia than adults because children have a larger body surface area in proportion to their bodily weight, allowing them to lose heat more quickly ([Singer, 2021](#)).

The pathophysiology of Hypothermia involves cold diuresis, which can result in loss of fluid, culminating in the risk of hypovolemia ([Brown et al., 2012](#)). The treatment for this circulatory malfunction involves medication with vasopressors. Repeated and prolonged usage of these medications may result in worsened outcomes ([Martin et al., 2015](#); [Vincent et al., 2018](#)). A regulated rewarming is recommended in hypothermia; however, a quick rewarming can result in a shift in electrolyte balance of the body, resulting in arrhythmias and cardiac arrest ([Dietrichs et al., 2020](#)). Therefore, a decision support system for an early assessment is required for regulating the rewarming cycles.

In critically ill patients, the temperature is not only an important clinical indication of illness severity but is also an independent predictor of morbidity and mortality. Hypothermia may be a significant and modifiable factor linked to a higher risk of death in critically ill individuals. Close monitoring and regulation of body temperature to minimize extremes are especially vital in severely ill patients ([Faulds and Meekings, 2013](#)). Also, preventive measures must be taken to avoid the grueling consequences of hypothermia ([Paal et al., 2018](#)). A poor prognosis can increase the risk of Hypothermia occurrence and associated complications. Thus, to avoid delayed identification, Hypothermia monitoring needs to be advanced with the use of artificial intelligence. AI is already transforming medicine, where algorithms have been surpassing the clinical accuracy of disease prediction in the ICU ([Bohr and Memarzadeh, 2020](#)). However, none of the studies, to our knowledge, have developed Hypothermia prediction models despite the

higher association of mortality with hypothermia ([Fatteh et al., 2021](#)). Early detection using predictive modeling of hypothermia can save lives.

Artificial intelligence algorithms, such as machine learning, can automatically rebuild associations between variables and response values from big data and enhance the interpretation of conventional techniques, such as support vector machines, random forest algorithms, and regression techniques in pinpointing crucial predictors ([Jiang et al., 2017](#)).

The abundance of continuous monitor data is complemented by applicable data from the EHR or the increasingly widespread wireless wearable devices that assess physiologic signals ([Rush et al., 2019](#)). The analytic tools of machine learning can both help the workflow of reading vital signs and provide insights into data patterns and complexities beyond the perceptual capacity of the average human clinical observer ([Chen and Asch, 2017](#); [Obermeyer and Lee, 2017](#)). In data-rich situations, ML algorithms excel at analysing complicated signals ([Krumholz, 2014](#); [Tomašev et al., 2019](#)).

Disease categorization and prediction models can be improved using machine learning techniques. These methods might be beneficial in the clinic for automatically identifying individuals with extremely morbid illnesses who could benefit from intensive risk factor treatment ([Ross et al., 2016](#)). In earlier research, machine learning and statistical modeling techniques were used to address the issues related to sepsis detection and care management ([Henry et al., 2015](#)). Several researchers employed machine learning algorithms to identify those who were most likely to die from sepsis ([Gultepe et al., 2014](#); [Al-Ansari et al., 2025](#); [Taylor et al., 2016](#)).

This research developed a real-time hypothermia prediction model using physiological vitals time-series data and ML techniques. We defined the onset of hypothermia as a time point when the core body temperature is  $< 35^{\circ}\text{C}$ . We tested a battery of machine learning models and selected the best one for the prospective validation in our pediatric ICU. Timely prediction and intervention are essential to prevent severe complications and improve patient outcomes. This chapter delves into the application of machine learning models for predicting hypothermia, with a focus on pediatric ICU settings.

#### **"Clinical Use-Case:**

ThermoGnose is designed as an early warning support tool for bedside clinicians in the pediatric ICU. The model operates in a continuous monitoring context to identify patients at risk of developing hypothermia within the next 30 minutes to 4 hours, enabling timely preventive interventions such as thermal support optimization or closer temperature surveillance. The output is intended to augment clinical judgment rather than trigger automated clinical actions."

## **4.1.2 Methods**

### **4.1.2.1 Study data and pre-processing**

The ICU data stated in this investigation work were gathered from the Paediatric ICU at AIIMS. The patient's non-public information was not required amid information warehousing, additionally, there was no modification that drained the patient's care, Consequently, the Ethics committee of the medical institute allowed assent for this investigation (IEC/NP-211/08.05.2015) ([Sethi et al., 2017](#)). No data was discarded during this process as the data was not segregated. The pediatric ICU consists of eight beds, including infant beds. Phenotype deeply, capture reliably, enable decisions, and use a systematic approach are the four principles that constitute the Sepsis Advanced Forecasting Engine for the ICU's system. The SafeICU data warehoused between the months of January 2019 to January 2020 contained 193 ICU stays (884 days) from a total of 148 patients. Additionally, the ICU data collected from the month of February 2020 to November 2020 were used for prospective validation.

#### **4.1.2.1.1 Physiological vitals records**

Health Level 7 Standards-based querying of the multi-parameter CMS Mindray™ monitors was extracted using in-house software. Physiological vitals data, i.e., RR, Sys-bp, Dia-bp, HR, SpO<sub>2</sub>, and temperature records, were used. T1 and T2 temperature probes were used for the abdomen and foot, respectively. The HL7 format data obtained from the monitoring station was stored on the server. The data were parsed into analyzable tabular forms and pre-processed for analysis and model building. Notification messages generated by Pushbullet™ and RPushbullet have been dispatched to the Android phones in case of data loss, i.e., on the off chance that there's any interference in the data streaming.

#### **4.1.2.1.2 Treatment-charts**

The in-house doctors maintain a proper note of the treatments in a Word document. The backup of these files was planned at a few settled hours each day. The parsing of text files was done using the Docx-2-txt Python module, and then it was converted into a tabular form to make it easier to use for text mining.

#### **4.1.2.2 Imputation**

Pre-processed Safe-ICU stay data were imputed using Kalman Smoothing utilizing the structural time-series (StructTS) model using the imputeTS package in R ([Moritz and Bartz-Beielstein, 2017](#)).

### **4.1.2.3 Data Preprocessing and cohort-generation**

#### **4.1.2.3.1 Cohort construction**

The cohort was constructed using the SafeICU data with a case-crossover design ([Navidi, 1998](#)), whereas multiple instances were taken from each subject at 30 min to 4 hours ahead of the onset. All patients with continuous vitals monitoring for greater than 4.2 hours of their stay in the ICU were included for cohort construction. The exclusion criteria included the following. 1) Patients with <4.2 hours of ICU stay, 2) patients with non-availability of temperature monitoring within the observation window, or 3) patients with non-availability of age and gender, 4) patients already in hypothermia within the first 4 hours of admission based upon temperature recordings 5) patients with missing data in >10% of the observation period.

#### **4.1.2.3.2 Data pre-processing and scoring**

Temperature (T1), Sys-bp, SpO<sub>2</sub>, RR, and HR were utilized as the physiological predictors. Minimal pre-processing of the vitals data was done. Data imputed using the Kalman filter was used for further modeling on the patients eligible for the development cohort. Each timestamp of the vitals time series at a 15-s resolution was evaluated, and binary labels for hypothermia (yes/no) were assigned to 30-min time epochs. The onset time of hypothermia was taken as the start time of an epoch when the median binary label of T1 value was less than 35 ([Musi et al., 2021](#)) ([Figure 4.1](#)). Observational windows of length 256 min were taken 30 min to 4 h ahead of the onset time.

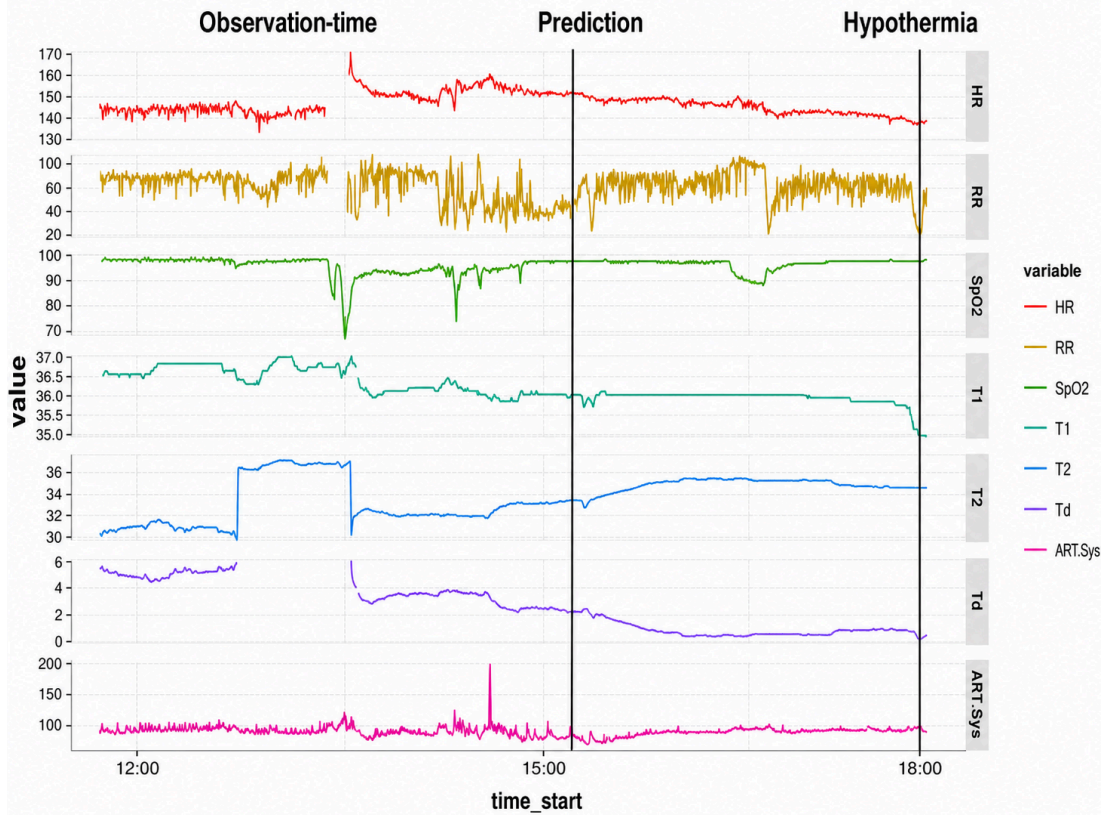


Figure 4.1 Physiological variables for prediction of Hypothermia in pediatric-ICU

**Non-linear time series feature extraction and selection.** A set of 3,939 TS features time-series features (TS features) consisting of linear and nonlinear physiological features, say Wavelet-transform coefficients, discriminative power, Fourier-transform coefficients, statistics, and other complex features, had been extracted utilizing the “tsfresh” Python package ([Christ et al., 2018](#)). Further, the Boruta feature selection algorithm was utilized to carry out the variable selection. Boruta uses a top-down approach to find important characteristics, comparing the value of original attributes to the importance achievable at random, calculating using their permuted duplicates, and gradually removing unnecessary information. TS features are explainable and also capture physiological domain understanding, as opposed to black-box features utilizing deep learning. Variable selection was performed using the Boruta algorithm with OOB, which was optimized using grid search for their n-features and n-tree combination, carried out in R.

#### 4.1.2.4 Model Development and Evaluation

The development cohort consisted of patients who developed Hypothermia within the subsequent 4 hours of the index time. The outcome variable is Hypothermia status within the subsequent 4 hours as described by Hypothermia Binary labels. Five-fold cross-validation sets were built, and hyperparameter tuning was performed to attain the best performance on the validation set utilizing grid search. We have named our pipeline as ThermoGnose and have

incorporated it in the figure and text (Figure 4.2). To overcome the problem of class imbalance, we undersampled the majority class for training the models (Bach et al., 2019). The AUROC, Precision, and Recall were assessed in the test set. A comparison of model performance was used as an indicator for different lead times. Normalization of TS features was done by the z-score computation. Random-Forest (Breiman, 2001), Adaboost (Hastie et al., 2009), Gradient boosting machine (GBM) (Chen and Guestrin, 2016), and the SVM (Noble, 2006).

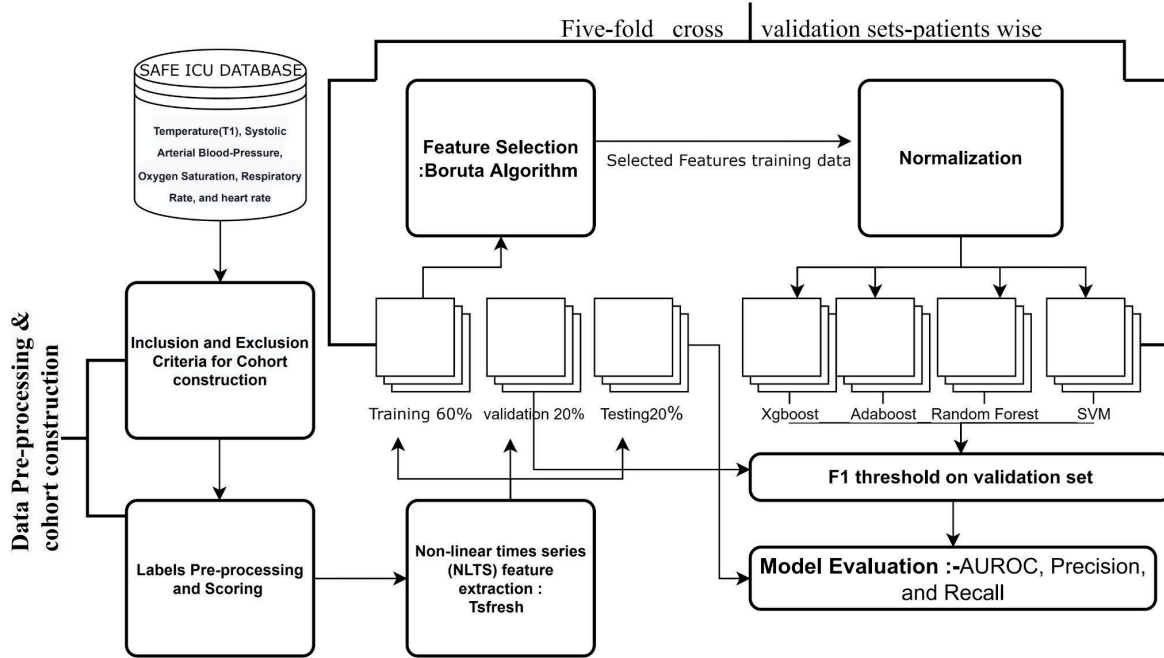


Figure 4.2 Overview of ThermoGnose pipeline

### Model Selection Rationale:

Tree-based ensemble models and support vector machines were selected for ThermoGnose due to their ability to effectively handle high-dimensional, non-linear time-series features extracted using tsfresh. These models offer robustness to multicollinearity, interpretability through feature importance analysis, and strong performance in settings with limited sample sizes. Deep learning architectures were not prioritized for this task, as the primary objective was to leverage explainable physiological features rather than black-box representations.

### Temporal Integrity of Evaluation:

All prediction windows were constructed using a lead-time-based design, with observations strictly preceding hypothermia onset. Patient-level partitioning ensured no overlap between training, validation, and test data, and random splitting across time was not used, thereby preventing look-ahead bias.



#### 4.1.2.5 Shapley additive exPlanations value analysis for model interpretability

SHAP is a method that assigns significant weight to each feature using game-theory principles ([Lundberg and Lee, 2017](#)). The SHAP values are utilized to describe how features affect the model prediction. Each feature's SHAP value allocation for all test sets was determined. On the test sets, the top 20 essential features and their relationship with SHAP values are plotted.

#### 4.1.3 Results

##### 4.1.3.1 Data Characteristics

The cohort consisted of 193 ICU stays with 2723 multiple episodes of hypothermia. Although Pneumonia (24.2%), Shock (23.4%), and Sepsis (19.7%) were the most frequent diagnoses in our cohort, there was a diverse representation of other diagnoses such as Congestive Heart Failure, Tuberculosis, Liver Failure, etc. Further, the relatively high prevalence (28%) of Hypothermia in this cohort reflects the necessity of early preventive measures in the pediatric ICU. The temperature, RR, and SpO<sub>2</sub> were found to be significantly different 1 hour before the onset of Hypothermic and non-hypothermic events ([Table 4.1](#)). It is evident that the median temperature of the hypothermic subject was 1.17 °C lower than the non-hypothermic subject 1 hour before its onset, p-value = 0.0001, while the respiratory rate was lower in the hypothermia by 1.62 bps, p-value 0.0005. Further, we note that the length of stay (LOS) is associated with the length of stay in hypothermia. A 10-minute increase in Hypothermia duration can increase the LOS by 68 minutes (0.0038) on a wide range of patients with comorbidities. This conclusion from our observational design of *ThermoGnose* will form the basis of an interventional design to directly assess the impact of the model on clinical decision-making and hard outcomes such as mortality and LOS.

*Table 4.1 Characteristics of SafeICU-cohort observational window captured (30min-4 hours) earlier than Hypothermia. Unless otherwise stated, all values are mean (SD), \*significance level at p-value  $\geq 0.01$ , W represents Wilcoxon rank-sum test (non-parametric), which is utilized once the normality assumption has been tested. The Chi-squared test of proportions is denoted by the letter C in the table.*

| Variable                           | Hypothermia<br>Mean(sd) | Non-Hypothermia<br>Mean(sd) | P-Value<br>(Significance, p-value<0.01) |
|------------------------------------|-------------------------|-----------------------------|-----------------------------------------|
| Age (months)                       | 50.63 (52.12)           | 40.67 (51.15)               | $1.75 \times 10^{-12}$ (W)              |
| Arterial-Diastolic BP (DBP), mm Hg | 86.8 (6.67)             | 87.2 (6.43)                 | 0.0592 (W)                              |
| Heart rate, per minute             | 129.69 (8.55)           | 129.84 (8.15)               | 0.9836 (W)                              |
| Respiratory rate, per minute       | 31.45 (5.35)            | 33.07 (5.5)                 | 0.0005*(W)                              |
| Oxygen Saturation                  | 93.37 (3.1)             | 92.71 (3.31)                | 0.01269 (W)                             |
| Temperature                        | 36.36 (0.51)            | 37.53 (0.8)                 | 0.0001*(W)                              |
| Gender (F%)                        | 43.6%                   | 32.2%                       | 0.0018*(C)                              |

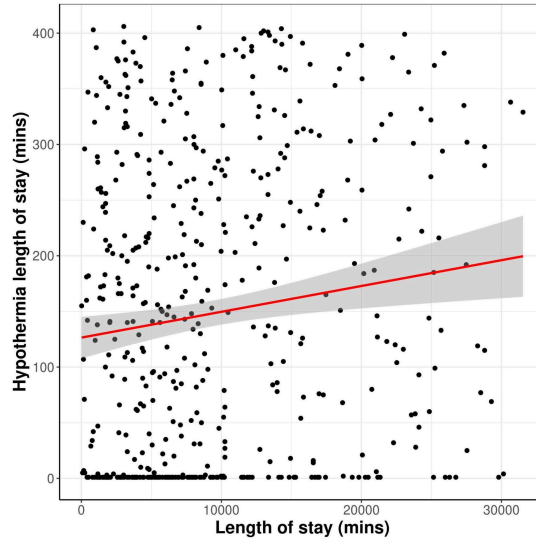


Figure 4.3 Length of stay is associated with the length of stay with hypothermia

#### 4.1.3.2 Model performance on Time-series features

A total of 726 time-series features were found to be statistically noteworthy for variable significance z-scores in the Boruta feature selection analysis for 30 min to 4 hours different lead times. Out of the Gradient boosting classifier, Adaboost, Random Forest, and support vector machine (SVM) models trained upon these TS-features, the Gradient boosting classifier performed best ([Figure 4.4 A, B](#)) with an AUROC of 85% (SD = 1.6) and a precision of 59.2% (SD = 8.8) for a 30-min lead time before the onset of hypothermia ([Figure 4.4 D, E](#)). The Average AUROC of 74.28 (SD = 2.26) and AUPRC of 47.56% (SD = 4.4) were achieved for (30 min to 4 hours) ahead of prediction, with 79.88% (SD = 9.62) of all hypothermia events identified 4 h ahead of their onset. As expected, the model showed a decline in model performance at higher lead times, such as AUROC of 77.2% (SD = 2.3) and precision of 41.34% (SD = 4.8) for 4 h ahead of Hypothermia onset. The findings of all proposed models for predicting hypothermia/non-hypothermia based on physiological factors are presented in [Table 4.1](#). The best model performance indicator ([Figure 4.4 C](#)) with an AUROC of 84.2% (SD = 2.1%) is in the age group of “0–5” years. Whereas it slightly decreased in the age group “10–15” years and “5–10” years with an AUROC of 82.3% (SD = 2.9%) and 77.2% (SD = 5.8%), respectively.



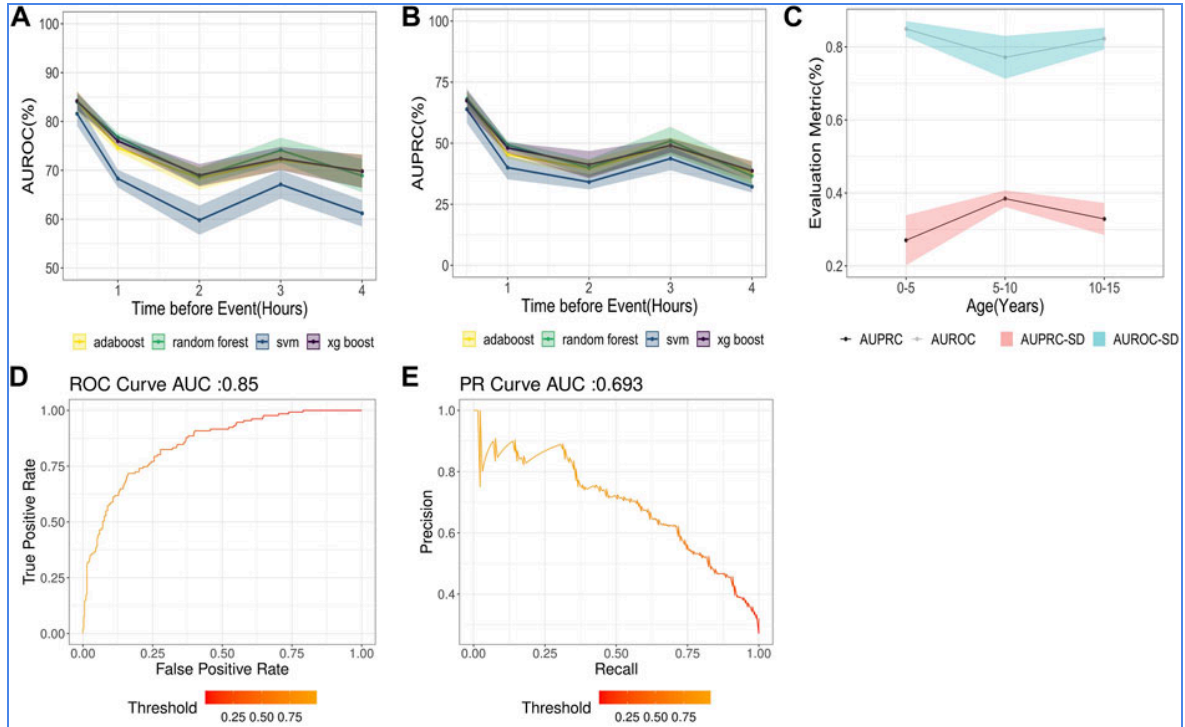


Figure 4.4 Model performance on Time-series features. (A) AUROC for different models with lead times or times before hypothermia forecast in the next 30 min to 4 h (B) AUPRC for different models with lead times or times before hypothermia forecast in the next 30 min to 4 h (C) AUROC for the Hypothermia prediction in the next 30 min (D) AUPRC for the hypothermia prediction in the next 30 min. (E) Results of the AUROC and AUPRC Models for various age groups

Table 4.2 Shows the results of the proposed models for predicting hypothermia/non-hypothermia based on physiological variables. The Time Pt. column shows the hours that passed after the observation was taken and the findings were recorded. Models used are XGboost, RF - Random Forest, Adaboost, and SVM - Support Vector Machine. Metrics used for the prediction are as follows: H/NH - Number of Hypothermia/Non-Hypothermia sequences present, AUPRC - Area Under Precision-Recall Curve, PPV - Positive predictive value, AUROC - Area Under Receiver Operating Characteristics, D - Detection, NPV - Negative predictive value, and P - Prediction

| Time Pt. | AUPRC     | AUROC     | Accuracy  | Sensitivity | Specificity | PPV       | NPV       | F1 Threshold | Model   |
|----------|-----------|-----------|-----------|-------------|-------------|-----------|-----------|--------------|---------|
|          | Mean (SE) |           |           |             |             |           |           | Mean         |         |
| 30min(p) | 68.8(6.6) | 85(1.6)   | 78.7(4.7) | 70.7(9.7)   | 81.5(9.1)   | 59.2(8.8) | 88.9(2.6) | 0.6          | XGboost |
| 1hr (P)  | 49.3(4.1) | 79.7(1.9) | 66.6(5.6) | 82.5(10.3)  | 61.3(10.2)  | 41.4(3.9) | 92(2.8)   | 0.42         | XGboost |
| 2hr (P)  | 38.8(5.9) | 75.6(4)   | 58.4(9.8) | 90.2(8.5)   | 49.7(14.3)  | 34.2(6.9) | 95.3(2.5) | 0.38         | XGboost |
| 3hr (P)  | 44(2)     | 72.1(1.5) | 65.7(3.3) | 67.1(8.5)   | 65.2(7.3)   | 40.3(2.7) | 85.3(2.1) | 0.51         | XGboost |
| 4hr (P)  | 36.9(3.6) | 73.7(2.3) | 55(4.7)   | 88.9(11.1)  | 45.4(8.8)   | 31.6(1.9) | 94.4(4.5) | 0.33         | XGboost |
| 30min(p) | 66.7(6.4) | 84.5(1.4) | 51.9(2.3) | 46.5(4.7)   | 53.8(4.6)   | 26.6(3)   | 73.7(2.9) | 0.53         | RF      |
| 1hr (P)  | 51.8(3.6) | 80.6(2)   | 69.2(5.2) | 80.6(10.4)  | 65.1(9.9)   | 43.3(2.9) | 91.8(2.9) | 0.46         | RF      |
| 2hr (P)  | 38.4(5.6) | 74.8(3.4) | 62.8(6)   | 77.9(5.3)   | 58.5(8.8)   | 34.7(5.2) | 90.6(1.2) | 0.48         | RF      |

|          |           |           |           |            |            |           |           |      |              |
|----------|-----------|-----------|-----------|------------|------------|-----------|-----------|------|--------------|
| 3hr (P)  | 45.2(2.1) | 74(1.5)   | 64.4(7)   | 71.7(12.8) | 61.7(13.7) | 40.3(3.5) | 87(3.5)   | 0.49 | RF           |
| 4hr (P)  | 37.1(2.3) | 72.7(1.9) | 60.3(4.7) | 75.7(14.8) | 55.7(9.7)  | 32.6(1.8) | 89.8(3.6) | 0.47 | RF           |
| 30min(p) | 67.4(6.7) | 84.3(1.7) | 77.7(2.8) | 74.5(4.2)  | 78.7(4.4)  | 56(5.4)   | 89.6(1.3) | 0.63 | Adaboos<br>t |
| 1hr (P)  | 48.2(2.3) | 78.8(2.1) | 70.2(1.3) | 75.9(4.1)  | 68.3(1.9)  | 43.6(2.8) | 89.8(1.5) | 0.64 | Adaboos<br>t |
| 2hr (P)  | 37.8(6.1) | 74.5(4.1) | 64.9(4.1) | 74(2.6)    | 62.3(4.8)  | 35.5(5.3) | 89.6(1.8) | 0.67 | Adaboos<br>t |
| 3hr (P)  | 43.8(1.2) | 71.7(0.8) | 68.9(1.8) | 55.4(4.7)  | 73.6(2.5)  | 42.2(2.4) | 82.6(1.7) | 0.72 | Adaboos<br>t |
| 4hr (P)  | 35.5(2.8) | 72.3(1.8) | 64.1(0.8) | 71.5(7)    | 62(2.5)    | 34.7(1.4) | 88.6(2)   | 0.7  | Adaboos<br>t |
| 30min(p) | 62.6(6.9) | 81.9(2)   | 74.6(4.6) | 73.1(3.6)  | 75.2(7)    | 52.1(8.3) | 88.6(2.2) | 0.51 | SVM          |
| 1hr (P)  | 41(4.2)   | 69.6(2.8) | 60.9(7.1) | 68.2(13.9) | 57.9(13.8) | 35(3.3)   | 85.5(1.9) | 0.48 | SVM          |
| 2hr (P)  | 31.2(3.3) | 65.9(2.4) | 54.8(2.7) | 74.3(2.2)  | 49.4(3.8)  | 29(3.2)   | 87.3(2)   | 0.45 | SVM          |
| 3hr (P)  | 41.3(1.7) | 68.6(2.1) | 61.6(4.4) | 65.7(9.2)  | 60.2(8.3)  | 36.7(3.4) | 83.6(2.8) | 0.47 | SVM          |
| 4hr (P)  | 32.7(2.3) | 66.5(1.9) | 50.4(9.2) | 78.3(12.6) | 42.4(15.4) | 28.3(2.7) | 87.9(1.7) | 0.46 | SVM          |

#### 4.1.3.3 Model Interpretability Analysis

The best model captured interpretable and clinically meaningful features such as minimum temperature, age of the patient, mean of absolute change in temperature, and fast Fourier transform coefficients of respiratory rate ([Table 4.3](#)) as the top predictors for future Hypothermia ([Figure 4.5](#)). Detailed plots of top individual predictors for future Hypothermia are available in [Figure 4.6](#). A detailed description of each of the top nonlinear feature predictors for future Hypothermia is available in Non-linear Feature Description.

*Table 4.3 Important non-linear features*

| SI.No. | Nonlinear Feature                         | Definition                                                                                                                                   |
|--------|-------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------|
| 1.     | Absolute Energy                           | Returns the time series' absolute energy                                                                                                     |
| 2.     | Continuous Wavelet Transform Coefficients | A time-scale illustration of a signal is proposed by CWT. The length of the examined signal will aid in dynamically detecting nonlinearities |
| 3.     | Fast Fourier Transformation Coefficient   | The Fourier coefficients for the one-dimensional discrete FT are calculated using the Fourier transform algorithm                            |
| 4.     | Lag                                       | At lag=0, a complete correlation will exist for every time series. The                                                                       |

|    |                                 |                                                                                                              |
|----|---------------------------------|--------------------------------------------------------------------------------------------------------------|
|    |                                 | correlation value will drop as the time series shifts                                                        |
| 5. | Mean                            | The mean of x will be returned by this feature                                                               |
| 6. | Minimum                         | The least value among the given collection of values is the minimal number                                   |
| 7. | Quantile                        | The q quantile of x is calculated. Where the quantile divides the sample into equal-sized adjacent subgroups |
| 8. | Sum                             | The sum of the time series values will be calculated                                                         |
| 9. | Sum of re-occurring data points | This feature will return the total of all time-series data points that appear more than once                 |

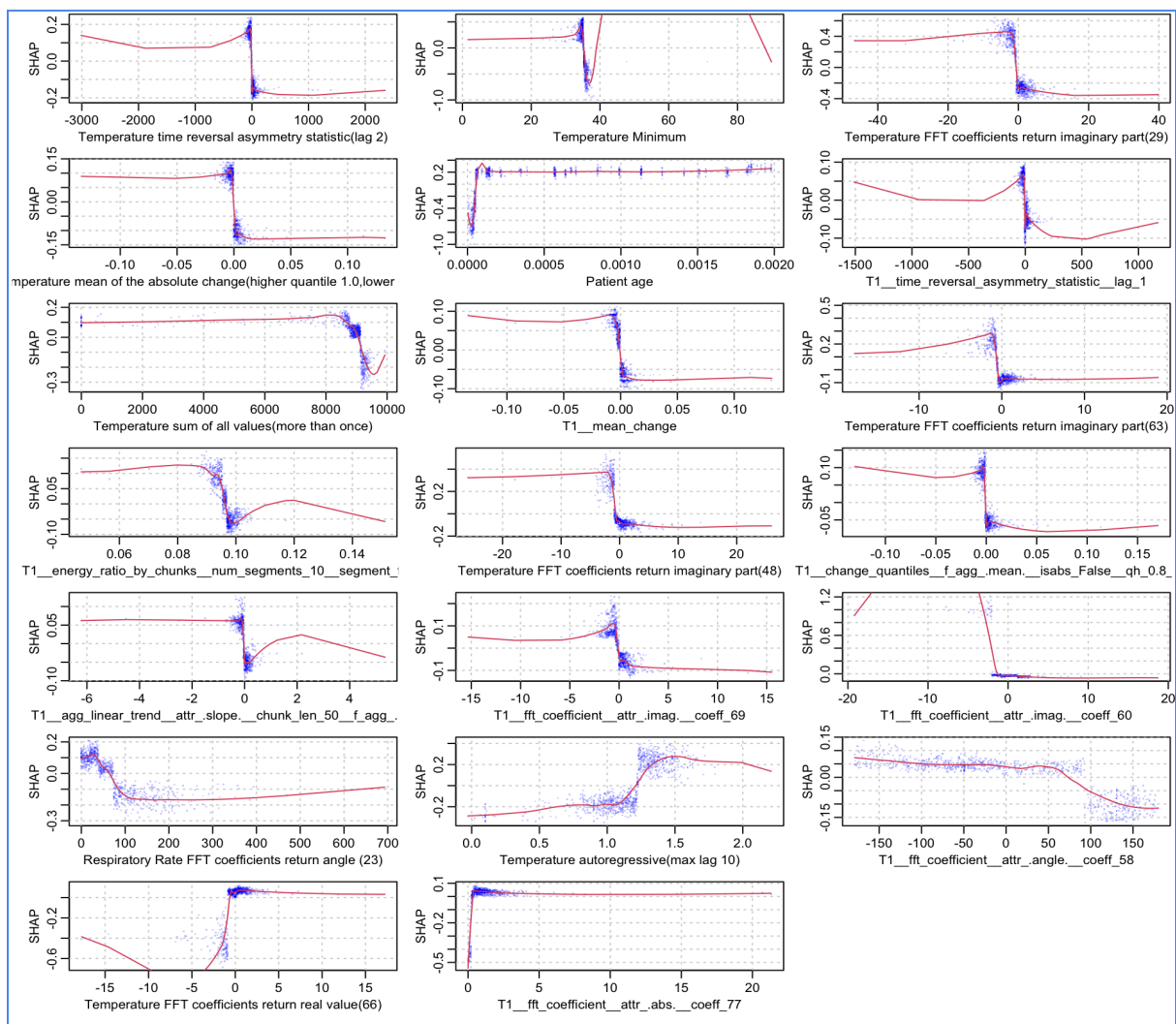


Figure 4.5 Detailed plots of individual top predictors for future Hypothermia

## Non-linear Feature Description

- **Abs\_energy (x)**

This option returns the time series' absolute energy, which is determined as the sum of squared values. The feature is calculated using only one argument, x(NumPy.ndarray), which represents the time series. The feature's floating-point value is then returned.

$$E = \sum_{i=1}^n x_i^2$$

- **cwt\_coefficients(x, param)**

The Continuous Wavelet Transform (cwt) depicts a signal's time scale. The length of the signal under investigation will aid in dynamically detecting nonlinearities. Signal non-linearities are thus exposed in a time bi-frequency domain, which also takes into account non-stationarity. The continuous wavelet transforms for the Ricker Wavelet, or the Mexican Hat Wavelet, are calculated using this characteristic. It has two input parameters: x and param. The time series that calculates the feature is x(NumPy.ndarray). And the param (list) comprises the following dictionaries: "widths": x, "coeff": y, and "w": z, according to the dictionaries. Where x is an int array, and y and z are integer types.

This function calculates three parameters, such as widths, coeff, and w. This feature collects all of the distinct widths into a single array and then calculates the single CWT for each array chosen. And finally, the values for different coefficients and widths are returned.

This is defined

$$\frac{2}{\sqrt{3a\pi^{\frac{1}{4}}}} \left(1 - \frac{x^2}{a^2}\right) e^{\frac{x^2}{2a^2}}, \text{ where } a \text{ is considered the width of the wavelet function.}$$

- **fft\_coefficient(x, param)**

The Fourier coefficients for the one-dimensional discrete Fourier transform are calculated using the Fourier transform algorithm. FT shows all the frequency components and their strength contained in any signal. That is, Fourier analysis is nothing but the extraction of a series of cosines and sines, which, when superimposed, will produce a function. It takes two parameters, x and the param. Where x(NumPy.ndarray) is the time series used to generate the feature, and param(list) is a list of dictionaries, such as "coeff": x and "attr": s. Here, x is an integer type and should be greater than or equal to zero, and the 's' is a string type and in ["real", "imag", "abs", "angle"].

The resulting coefficients are complex, and this feature can output the imaginary (attr="imag"), real (attr="real"), angle in degrees (attr="angle"), and absolute value (attr="abs").

$$A_k = \sum_{m=0}^{n-1} a_m e^{-2\pi i \frac{mk}{n}}, \quad k=0, \dots, n-1$$

- **mean(x)**

Mean is considered a mathematical average of all numbers, and sometimes it is even termed as the arithmetic mean. To calculate the mean, the sum of all the numbers will be divided by the total count of numbers in a set. This feature will return the mean of x. Where x (NumPy.ndarray) is the time series parameter to calculate the feature. It has the floating-point return type.

- **quantile(x, q)**

The q quantile of x is calculated with this feature. When the quantile divides the sample into equal-sized adjacent subgroups, the sample is divided into equal-sized adjacent subgroups. In this case, the value of x exceeds q percent of the ordered values from x. It requires the input of two parameters, x and q. Where x (numpy.ndarray) is the feature's time series parameter, and q (float) is the quantile to be calculated. Its return type is floating-point.

- **c3(x, lag)**

It demonstrates how closely the time series resembles itself. As the values are compared, a perfect correlation will be found for the all-time series at lag=0. The correlation value will drop as the time series shifts. If the time series solely contains random values, the correlation will only exist at lag=0, with no correlation elsewhere.

It requires two inputs: x and lag. Where x (numpy.ndarray) is the time series parameter used to calculate the feature, and lag (int) is the lag utilized in the feature calculation. It returns a value of the floating type.

This function will calculate the value

$$\frac{1}{n-2lag} \sum_{i=0}^{n-2lag} x_{i+2lag}^2 \cdot x_{i+lag} x_i \text{ Which will be, } L^2(X^2).L(X).X]$$

- **sum\_of\_reoccurring\_data\_points(x)**

This feature returns the total of all time-series data points that appear several times. It has only one parameter, x(numpy.ndarray), which is used to compute the feature. It returns the feature's floating-point value.

- **sum\_values(x)**

The sum of the time series values will be calculated with this function. It has the time series argument x(numpy.ndarray), which is used to calculate the feature. And the value's return type is bool.

- **min\_values(x)**

The smallest value in a set of values is called the minimum number. That is, it is either less than or equal to all of the other data values in the collection. The minimum value is a unique value by definition, but in practice, the minimum value might be repeated in the set. The last location of a minimum value of  $x$  will be returned by this feature. The length of  $x$  will be used to compute this position. To calculate the feature, it uses a single time series parameter  $x$ (numpy.ndarray). This feature's floating-point value will be returned.

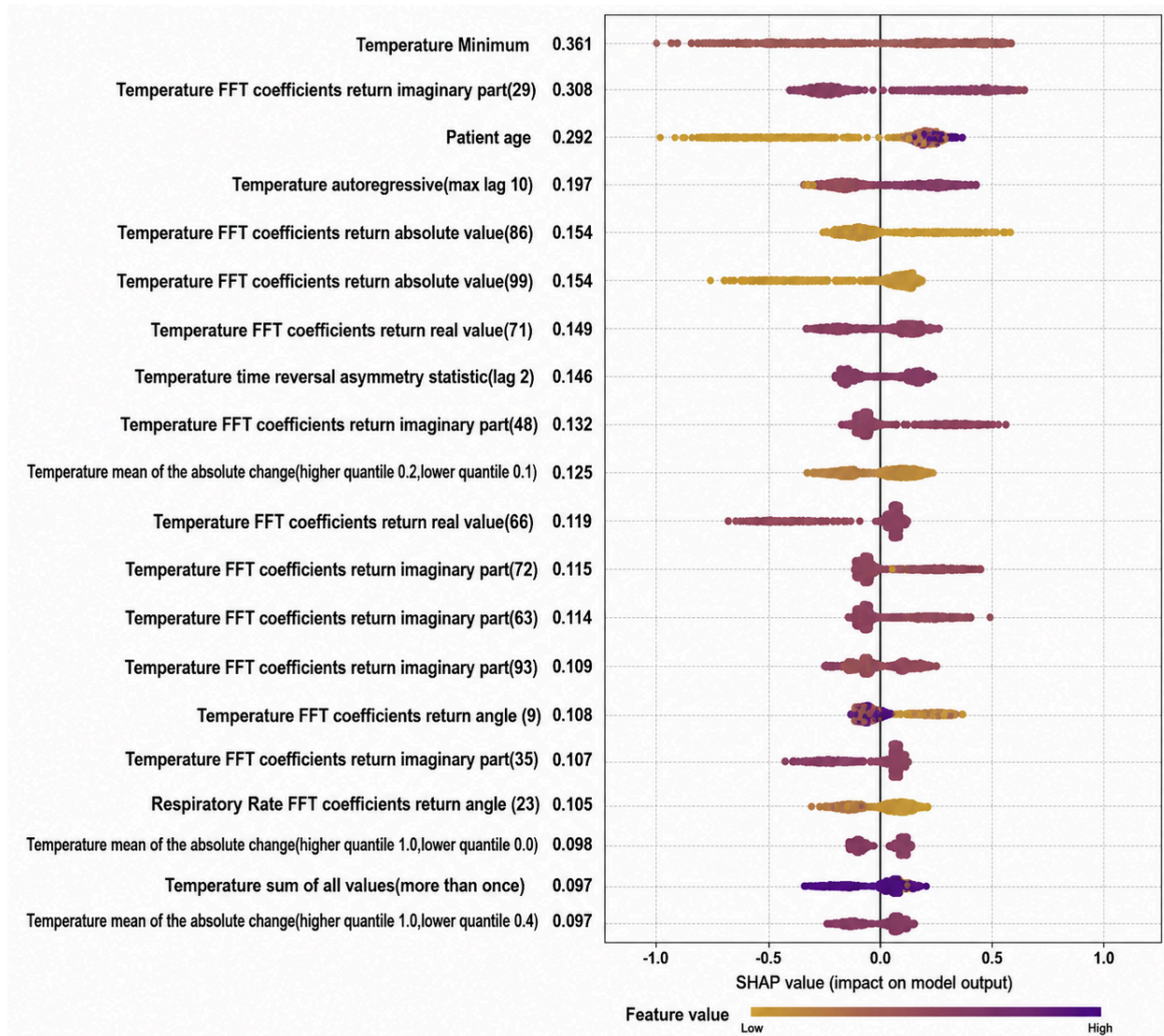


Figure 4.6 SHAP values of the top 20 Nonlinear features (descending order) generated from the five-fold test set on pediatric data

#### 4.1.3.4 Prospective Validation of Model

Gradient boosting classifier performed best with an AUROC of 79.8% and a precision of 53% for a 30-minute lead time before the onset of Hypothermia, whereas an AUROC of 69.6% and a precision of 38.52% for a (30-minute-4-hour) lead time, prospective validation of Hypothermia



(Figure 4.7A, B). The findings of prospective validation of a model for predicting hypothermia/non-hypothermia based on physiological factors are presented in Table 3.4.

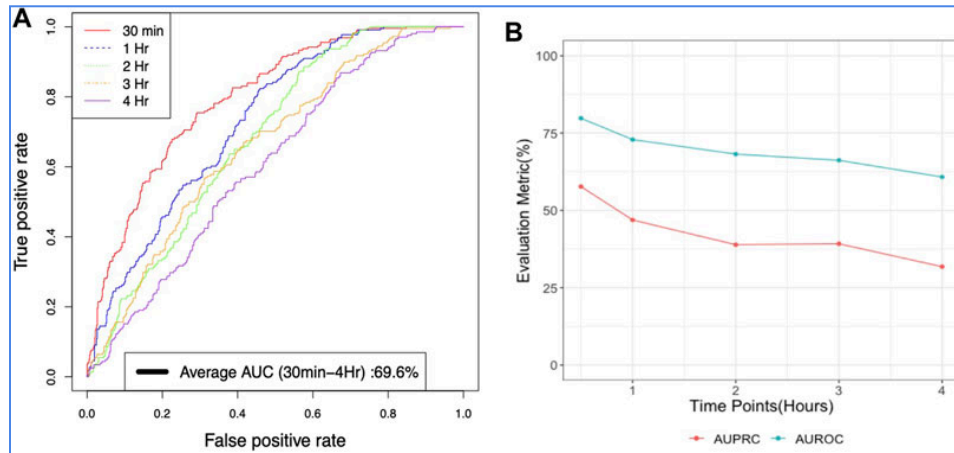


Figure 4.7 Prospective Validation of Model. A) Average AUROC performance for the xgboost model with lead times or times before for prospective validation of Hypothermia in the next 30 min to 4 h (B) AUROC and AUPRC performance for the xgboost model with lead times or times for prospective validation of Hypothermia in the next 30 min to 4 h

Table 4.4 Results of the prospective validation of a model for predicting hypothermia/non-hypothermia based on physiological variables

| Time Pt. | AUROC | AUPRC | Accuracy | Sensitivity | Specificity | PPV  | NPV  | Model   |
|----------|-------|-------|----------|-------------|-------------|------|------|---------|
|          |       |       |          | <b>Mean</b> |             |      |      |         |
| 30min(p) | 79.8  | 57.7  | 74.6     | 67          | 77.5        | 53   | 86.1 | XGboost |
| 1hr (P)  | 72.9  | 46.9  | 58       | 87.8        | 46.3        | 39.1 | 90.6 | XGboost |
| 2hr (P)  | 68.2  | 38.9  | 53.5     | 90.5        | 39.1        | 36.7 | 91.3 | XGboost |
| 3hr (P)  | 66.2  | 39.2  | 50.2     | 79.4        | 39.1        | 33.1 | 83.3 | XGboost |
| 4hr (P)  | 60.8  | 31.8  | 45       | 87.3        | 29.9        | 30.7 | 86.9 | XGboost |

#### 4.1.4 Discussion

Hypothermia in severe cases can result in mortality; studies have suggested that hypothermic subjects have a significantly higher rate of mortality when compared to non-hypothermic subjects (Kiekkas et al., 2018). Early decision-making can mitigate the risk of hypothermia and associated outcomes, yet none of the studies have attempted to predict hypothermia. This work presents a first-of-its-kind Hypothermia prediction model in pediatric patients. We leverage a SafeICU (Sethi et al., 2017) resource of the PICU patients database to build a real-time ML model for the prediction of hypothermia. Our ThermoGnose pipeline predicted the Hypothermia event with an AUC of 85% 30 min ahead of its onset and with an average AUC of 77.6% for 30 min to 4 h ahead of onset. An early therapy decision can help ICU management to initiate therapy, such as rewarming, to mitigate the risk of hypothermia (Giesbrecht, 2001). The GBM-model performed best at 30 min lead-time, and the performance declined during the higher

lead times, however, a 30-min window can provide sufficient time for early therapy such as rewarming ([Mendrala et al., 2021](#)). Our pipeline produced equal and superior results for the prospective validation, where an AUC of 69.6% was achieved for a lead-time of 30 min to 4 h. We also evaluated the interpretability of our models. Since artificial intelligence models suffer from black-box prediction, we kept our approach interpretable using SHAP analysis, which suggests the influence of each predictor on the model output ([Lundberg and Lee, 2017](#)). Our analysis suggests that the minimum temperature in a 420-minute window is the best predictor of future risk of hypothermia ([Figure 4.6](#)).

Since our main aim was to reduce the delay in early identification, we kept our choice of predictors to the readily available physiological vitals. These allowed us to make real-time predictions as we have installed in-house software to collect real-time vital data from the bedside at 15s resolution. Many studies in the intensive care unit suffer from generalizability issues due to the large number of predictor variables required to make predictions and the complexity of the data integration ([Wellner et al., 2017](#)). Studies have shown the potential of physiological vital signs for real-time prediction of sepsis. Minimal models have shown good potential for generalizability. We have, therefore, used only five physiological vitals available at the bedside in our models.

Early risk prediction of hypothermia is critical as it can act as a surrogate marker for prognostication of cold diuresis-related hypovolemia ([Brown et al., 2012](#)), and early assessment can be helpful for regulating the rewarming cycles ([Martin et al., 2015](#); [Vincent et al., 2018](#)) and reducing the LOS. We found out that the LOS is associated with the length of stay with hypothermia ([Figure 4.3](#)). Therefore, reducing hypothermia by proactive management can help reduce LOS. Thus, our prediction model can be useful from a wider perspective for clinical practice.

Prediction of hypothermia can reduce the risk of adverse outcomes and mortality in intensive care patients. Early warning systems can help in early treatment and therapy decisions for hypothermia. Our prediction models trained on the cohorts extracted from three million patient hours of data have shown the potential of prospective validation AUC of 79.8% on patients with a wide range of comorbidities; thus, our prediction models have the potential to improve patient care and save lives. The current study and the model, being observational, are not geared to evaluate the impact on clinical decision-making, which remains the main limitation of our work. However, therapeutic decisions for hypothermia are expected to be less complex, such as the use of warmers and fluids; thus, the current model lays the basis for future studies for assessing the clinical impact and outcomes in hypothermic patients based upon ThermoGnose predictions. Importantly, despite continuous monitoring of temperature, the most straightforward vital, predicting hypothermia, has remained unaddressed, and ThermoGnose provides a starting point for enabling clinical decisions to prevent and proactively treat hypothermia in the pediatric ICUs.



## 4.2 Early Prediction of Hemodynamic Shock in Pediatric Intensive Care Units With Deep Learning on Thermal Videos

### 4.2.1 Introduction

Hemodynamic shock is an acute and initially reversible condition that can act as an early signal to future end-organ failure ([Herget-Rosenthal et al., 2008](#)). It manifests due to inadequate blood perfusion than what is essentially required. Such a condition could cause tissue malfunction, leading to rapid organ deterioration, which can eventually result in death. The mortality rate due to delay in detection and treatment of such shock is as high as 34% in ICU patients admitted in developing countries ([Divatia et al., 2016](#)). Thus, proactive monitoring and therapy of hemodynamic shock can prevent the risk of impending organ failure and mortality ([Rivers et al., 2001](#); Early Goal-Directed Therapy Collaborative Group of [Zhejiang Province, 2010](#); [Vincent and De Backer, 2013](#)). However, delayed or insufficient monitoring can increase the risk of hemodynamic shock. Modern ICUs are equipped with a large number of sensors that generate humongous amounts of data, besides a large number of clinical investigations. ICU clinicians and nursing staff are exposed to this large quantity of data on a real-time basis. The overall burnout can go as high as 80.5%, which can contribute to delayed or insufficient care ([Rotenstein, 2018](#)). Moreover, the resource-limited areas have a higher risk of shock ignorance mainly due to a lack of technological advancement and a low doctor-to-patient ratio ([Bagcchi, 2015](#)). Most of the methods dealing with the shock in current times are invasive or require repeated contact, making the patient prone to hospital-acquired infections. The noninvasive ways, such as Noninvasive Blood Pressure monitoring and Ultrasonography, are not continuous but require contact with the delicate skin of infants, which is infection-prone too ([Oranges et al., 2015](#)). Techniques such as ECG and photoplethysmography (PPG) ([Liu et al., 2020](#); [Nachman et al., 2020](#)) can also provide cardiovascular parameters such as the respiratory rate and heart rate in a noninvasive and wireless way to serve as important factors for shock monitoring. However, their contact-based approach can prove to be distressing and epidermal damaging for the fragile skin of neonatal patients ([Barbosa Pereira et al., 2018](#)). Thus, the use of AI and frugal open-source technologies such as low-cost noncontact thermal monitoring can be effective for shock management.

We have created a body thermal imaging data cohort for the pediatric population under the Safe ICU initiative, which warehouses 1.5 million hours of physiological vital data, laboratory investigation, treatment charts, doctors' and nurses' investigation charts ([Sethi et al., 2017](#)). We have used these data to build an automated hemodynamic monitoring and prediction of the future onset of the hemodynamic shock pipeline. Nowadays, noninvasive vital monitoring measures have been gaining popularity because of their feasibility. Measures such as quantified capillary refill time (Q-CRT) and peripheral perfusion index (PPI) are being used for hemodynamic monitoring as noninvasive and transmitted light-based methods for assessing peripheral perfusion ([Falotico et al., 2020](#)). Vitals such as respiration and heartbeat have been known to be

measured noninvasively using a twin-core fiber (TCF)-based sensor ([Tan et al., 2019](#)). Some studies have been successful in tracking SpO<sub>2</sub> using broadband lighting and an RGB camera ([Guazzi et al., 2015](#)). [Balakrishnan et al. \(2013\)](#) take just an RGB video as input and are able to detect heart rate and beat lengths by monitoring the subtle head movements. Heart rate can also be measured through videos using DeepPhys ([Chen and McDuff, 2018](#)) and Eulerian video magnification (EVM) combined with wavelet transformation ([Froesel et al., 2020](#)).

Thermal imaging, another nonintrusive and noncontact modality, is now increasingly being used in medical literature to extract the temperature information from the body and used for studies such as breast cancer detection ([Ng, 2009](#); [Mambou et al., 2018](#)) and even diabetes mellitus (type 1) detection ([Selvarani and Suresh, 2019](#)). Thermography of the skin surfaces to study blood flow, cardiac pulse, and breath rate ([Pavlidis et al., 2007](#)) is being used as a basis for broader works such as preliminary hypertension assessment ([Thiruvengadam and Mariamichael, 2018](#)). Thermography of the human hand and foot skin surfaces has also been used to study blood flow using spectral filtering ([Sagaidachnyi et al., 2017](#)). A similar use of the thermography sensor has been proposed to make the disease activity detection in rheumatoid arthritis patients ([Pauk et al., 2019](#)). Lately, thermal imaging has also been used to detect the respiratory rate of humans ([Liu et al., 2019](#)). For instance, in cardiorespiratory signal monitoring, [Barbosa Pereira et al. \(2018\)](#) use thermal fluctuations beneath the nostrils for measuring the respiratory cycle. Since human bodies can be considered black bodies with an emissivity value of the skin close to 0.98 ([Bagavathiappan et al., 2009](#)), the infrared energy, and hence temperature information, can be suitably tapped by IR detectors such as Thermovision-550 ([Thiruvengadam and Mariamichael, 2018](#)), MWIR-Phoenix-FLIR ([FLIR Systems, 2016](#)), and SeekThermal® ([Seek Thermal, 2022](#)) easily available in the market. Even though some of the aforementioned techniques might incur a higher cost due to the nature of the thermal imaging sensors used, our study uses a simple SeekThermal smartphone camera attachment, which can feasibly be used to capture the thermal context of the body. During hemodynamic instability, redistribution of blood occurs, which causes a temperature difference in the center and periphery of the body, which acts as one of the first possible indicators of the advent of shock ([Lima and Bakker, 2005](#); [Lima et al., 2009](#); [van Genderen et al., 2012](#); [Ortiz-Dosal et al., 2014](#)). Our study leverages this temperature difference, that is, Centre-to-Peripheral Difference (CPD) ([Bourcier et al., 2016](#); [Houwink et al., 2016](#)), to study hemodynamic shock with the abdomen as the center and feet as peripheral regions ([Lima and Bakker, 2005](#); [Hasanin et al., 2017](#)). The difference observed is more in the case of a shock than of a no-shock ([Lima and Bakker, 2005](#); [Nagori et al., 2019](#)). We chose SeekThermal® Camera for our study as it has affordable, minimal hardware architecture and can be easily mounted on a smartphone. CPD has been exploited along with some vitals and the use of machine learning methods such as Histogram of Oriented Gradients features with Random Forest classifier for detection and prediction of shock with a single snapshot of an image at a time point ([Nagori et al., 2019](#)), but a single image snapshot would present only limited information as it may be an outlier caused by random fluctuation. This may reduce the robustness and reliability of the model output. In addition, using the handcrafted features for machine learning algorithms

might cost us time. To expand upon the study, we aim to reduce the manual tasks performed to calculate the features using deep learning methods and work on the continuous time-series data, improving the efficacy of detection and prediction.

In this study, we developed a future risk prediction model ([Figure 4.8](#)) using continuously streamed videos of body infrared patterns and state-of-the-art computer vision (CV) and DL techniques. Thus, the aim of our work was to build an automated, minimal-contact pipeline for proactive hemodynamic monitoring. Automated detection and prediction will help prevent the adverse outcomes related to hemodynamic instability. Hence, the impact this problem creates and the scope of extending the solution make this study worthwhile in saving the lives of many pediatric patients.

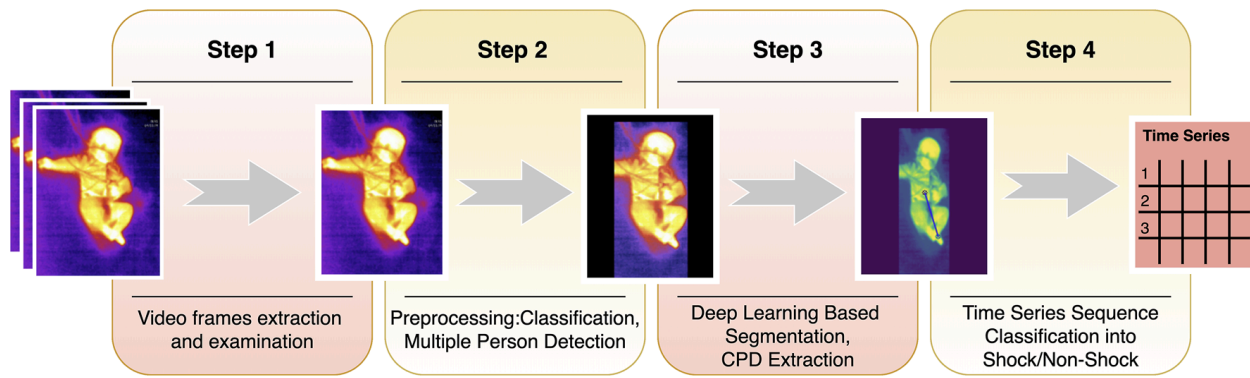


Figure 4.8 **Shock prediction pipeline.** The summary of the shock pipeline shows the steps from video frame extraction to shock prediction. Step 1 comprises sampling videos to extract frames. Step 2 classifies frames into covered or uncovered, while also finding the presence of multiple people in the frame and masking them to avoid confusion. The masked images are then input to the ResUNet-based segmentation model, and CPD is hence extracted. The series of CPDs is then passed through a time-series sequence classifier, and finally, the predictions are made for shock for the next 6 h

### "Clinical Use-Case:

The thermal video-based shock prediction model is designed for non-contact, continuous physiological monitoring in the pediatric ICU. By detecting early thermal signatures associated with circulatory compromise, the system aims to provide advance alerts to clinicians for further assessment and monitoring escalation. The model is intended as a supplementary monitoring modality and does not function as an autonomous alarm or diagnostic system."

## 4.2.2 Methods

For this study, we carried out thermal monitoring as a part of the SafeICU ([Sethi et al., 2017](#)) warehouse. We constructed an automated CPD extraction pipeline by body part segmentation on thermal videos using deep-learning models. Each time-point was labeled for shock and non-shock using an age-specific shock-index ([Acker et al., 2017](#)). The time-series data of the

extracted CPD signals, combined with the heart rate and the labels, were preprocessed for a case-control cohort design. We constructed cohorts at 0 h for detection and 1–6 h lead-time for prediction tasks. We then used multivariate LSTM ([Hochreiter and Schmidhuber, 1997](#)) models to detect and predict the future onset of the shock. We finally evaluated the performance metrics of the models using a 10-fold cross-validation method.

#### **4.2.2.1 Data Collection and Study Design**

##### **SafeICU Framework**

All the data reported in this research were collected from the Pediatric ICU of AIIMS, New Delhi, a tertiary care hospital in India. There were eight beds, including neonatal beds, in the pediatric ICU. We set up our servers to collect the real-time physiological vitals periodic data. Vitals (heart rate, respiratory rate, and blood pressure) were routinely monitored using electronic sensors; that is, the heart rate was measured by ECG sensors, blood pressure was measured using arterial catheters, and respiratory rate and oxygen saturation were measured using pulse oximeter probes. Socket programming (network protocol) based on in-house software applications was used to extract the data from the Mindray™ CMS. However, with the exception of heart rate, our study does not require these contact-based vitals for shock prediction, thus making the study minimal contact. We also warehoused laboratory investigations, daily doctor and nurse notes, treatment charts, and thermal imaging.

##### **Ethical Approval and Patient Consent**

The study was carried out with the approval of the PICU of AIIMS, New Delhi, India. Since thermal images only capture infrared radiation, these did not reveal the patient's identity, and the study did not involve any contact or change in routine patient care. Hence, a waiver of consent was sought and granted by the Institute Ethics Committee (Ref. No IEC/NP-211/08.05.2015, AA-2/09.02.2017). All experiments were performed in accordance with relevant guidelines and regulations as approved by the ethics committee.

##### **Cohort Based on Binary Shock Index**

The SAFE-ICU described earlier has warehoused over 1.5 million patient hours of monitoring data from the PICU. It is used to extract the time-stamped vital data for the patients at 0–6 h heart rate and blood pressure recordings. The SI was calculated as the ratio of the median heart rate and median arterial systolic blood pressure, considered the gold standard for BP measurement in critical care settings ([Pittman et al., 2004](#); [Lehman et al., 2013](#)). SI was computed over the median of moving sequential windows of 30 data points at a resolution of 15 s. Shock-index pediatric age-adjusted (SIPA) is used to compute shock/no-shock age-specific binarized outcomes for each patient ([Acker et al., 2017](#)).

## Thermal Imaging

Seek Thermal Compact (UW-AAA) with a resolution of 480\*640 was used to capture thermal videos at 15 fps on iron theme settings. The camera, having a thermal sensor of 206\*156 pixels, was kept at a median (IQR) distance of 2.0828 m (0.1524) for long beds and 1.524 m (0.2665) for small beds. The ICU room temperature was maintained at 25°C. Temperature calibration was performed by matching the imager temperature with other known sources. The temperature from the thermal camera was found to be accurate with temperature measurements from other sources, such as temperature probes and thermometers. Standard thermal video capturing and operating procedures were followed in order to ensure minimal disturbances by the extraneous factors, say, patient positioning, device handling, etc. (SOP). Thermal cameras only capture infrared radiation so as to make sure that the study does not reveal the patient's identity. The camera was placed properly and at a good distance from the patient so that there was neither direct contact involved nor any change in patient routine care. The thermal videos were captured in a standard color scale, guaranteeing that the full body of infants was visible. Thermal videos of every single patient were collected at different time points on different days. Thus, each patient possibly has different values for shock status, which in turn eliminates bias due to the patient's propensity characteristics, say gender, age, etc. Thermal artifacts were a lesser concern as we were calculating the relative gradient of temperature instead of the absolute temperature. Vital data with respect to the time-stamp of the videos were extracted from the data warehouse at 15-s intervals (SAFE-ICU) ([Sethi et al., 2017](#)). A comparison was made between the shock and non-shock groups using either the Wilcoxon rank-sum test or two-tailed Student's t-test after testing for normality by the D'Agostino-Pearson normality test using GraphPad Prism version 6.00, GraphPad Software, La Jolla, California, USA.

### Standard operating protocol (SOP) for capturing thermal videos

1. Before entering the ICU, make sure that one has to wear shoe covers, a scrub, and a disposable gown.
2. Need to connect the Seek Thermal Camera to the available Smartphone and then open the Seek Thermal app.
3. Ensure that the Seek Thermal app is set to a high-resolution setting and to an "Iron Theme" before capturing videos. A wall mount hanger is used to hold a smartphone. The Seek Thermal Camera is placed at a good distance from the bed and at a certain height in order to make sure that patients on the bed are not disturbed.
4. For each patient recorded in this session, the following steps were followed:
  - Making the patient comfortable is the most important part. The camera was placed properly and at a good distance from the patient so that there was no direct contact involved, nor any change in patient routine care
  - In the presence of a nurse, a smartphone is hung using a wall mount hanger



- MSEExcel sheets are used for data collection. Data such as UHID, bed number, date, and time of the patients on this session were recorded for identification purposes
- After capturing videos, the mobile hanger is removed safely, and the data is saved on a server, on a daily basis, into an identified folder with the patient's UHID, bed information, and date

### "Privacy Considerations for Thermal Video Data:

- All video data used in this study were captured using long-wave infrared thermal imaging, which records emitted heat patterns rather than visible-spectrum facial or anatomical features. As a result, the acquired thermal videos are inherently non-identifiable and do not permit recognition of individual patients. No RGB or visible-light images were collected or stored. This modality ensures patient privacy by design while enabling extraction of physiological thermal signatures relevant to model development.

#### 4.2.2.2 Classification Into Covered and Uncovered

The patients in the ICU are kept under observation for a long duration. Being a very critical area, the patients are kept covered by a blanket most of the time, which is removed only for a short period, generally when a nurse/doctor comes to provide care. A ResNet-152 ([He et al., 2016](#)) architecture was trained using the PyTorch ([Paszke et al., 2019](#)) framework in Python 3 ([Python Software Foundation, 2018](#)) to classify each frame into covered and uncovered, i.e., abdomen and feet are visible ([Figure 4.9](#)). To train the data, images were augmented and normalized by their mean and variance, especially extracted to suit the thermal data. The model was finally implemented on videos sampled at 1 fps.

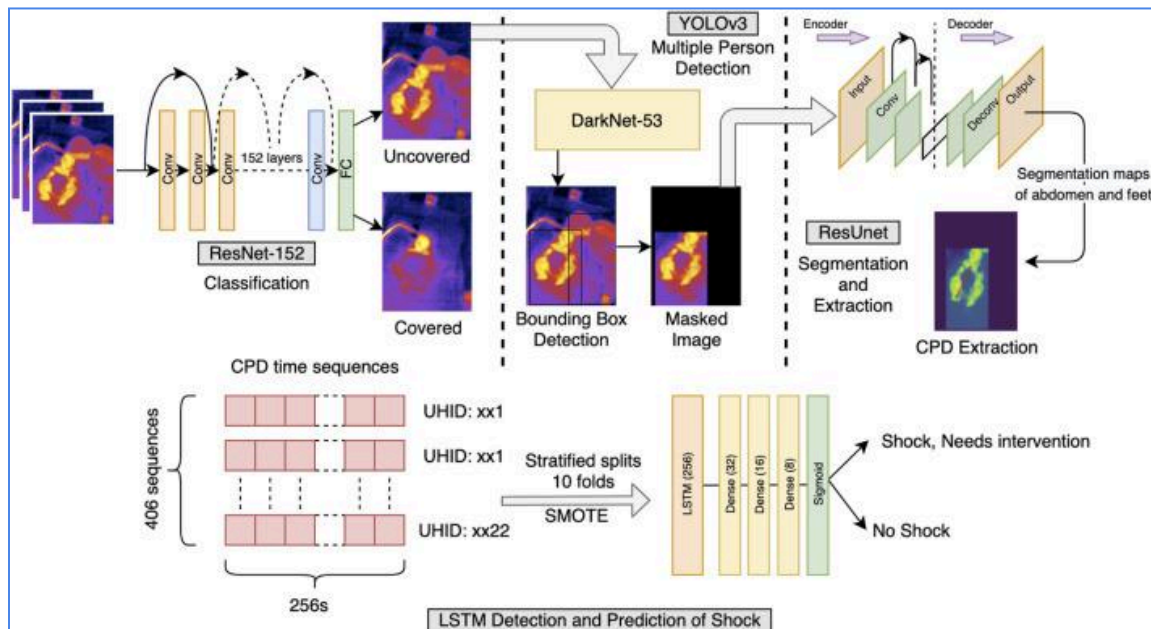


Figure 4.9 **Hemodynamic shock prediction with Deep Learning on Thermal Videos.** Illustration of the pipeline followed for the detection and prediction of shock and no-shock. Each frame of the video was examined for uncovered/covered. The uncovered frames were then filtered for the presence of people other than the patient present in the frame. The frame was finally passed from the segmentation and CPD extraction model, further collecting the sequences used for LSTM time-series classification. The time sequences had CPD and heart rate as features, and an appropriate window length of 256 s was chosen. Since the data were highly imbalanced, the SMOTE upsampling method was used in training the LSTM model. The detection and prediction of shock were carried out at 0 h and for the next 6 h, respectively.

#### 4.2.2.3 Multiple Person Detection

In the ICUs, caregivers providing care to the patient might come into the field of the camera mounted over the bed. For the CPD extraction task, bounding-box detection of multiple persons was used to filter out this additional presence so as not to confuse the algorithm. To avoid loss of crucial data where the caregiver and patient may be in the same frame, the bounding box corresponding to the larger area, that is, the patient, was kept, and the rest was masked. This region of interest was used for CPD extraction. For training purposes, a variety of images of the patient alone and along with the caretaker were manually annotated with bounding boxes using Yolo-Annotation-Tool-New ([Murugavel, 2019](#)) and augmented with linear transformations and Gaussian noise and passed through YOLOv3 ([Redmon and Farhadi, 2018](#)) based on Darknet-53 with a confidence threshold of 0.4.

The architecture had to be retrained from scratch as thermal data is very different from the COCO dataset, used in the original YOLOv3 manuscript. The YOLOv3's output was the dimensions and position of the predicted bounding boxes, detecting the people present in the frame, the correctness of which was verified through the Intersection over Union (IoU) metric by calculating the ratio of overlapping of the predicted and ground truth bounding boxes with their union area.

#### 4.2.2.4 Segmentation and CPD Extraction From Abdomen and Feet

Nagori et al. proved that the probability of shock depends directly on CPD. For this study, the abdomen was taken as the center, and the periphery was the foot. Images were annotated manually pixel-wise using the js-annotator-tool. The target maps contained three one-hot encoded layers corresponding to the abdomen, feet, and background. The input images were normalized with mean and variance, especially extracted from the distribution of the dataset in use. Appropriate image padding was carried out to ensure that the aspect ratio of images remains the same in case of any change in the input dimension. To account for a low dataset of pixel-wise segmented images for training, a ResUNet with ResNet-18, pretrained on ImageNet, was used as an encoder. UNet ([Ronneberger et al., 2015](#)) is specifically introduced to segment the less abundantly found medical data, helps to gather more local and global information even in the dearth of data, and thus efficiently segments out the images. A cutoff threshold was set on the predicted outputs to remove any weakly predicted pixels. The area detected was used as a region

of interest in the original image, and the mode of the detected probabilities was taken as the point of temperature extraction from the segmented-out abdomen and feet. The difference was divided by the abdomen value to keep CPD robust from thermal noise.

#### 4.2.2.5 LSTM Time Series Sequence Classification

The videos were sampled at 1 fps to extract the CPD data from every uncovered window possible. Windows of 256 data points corresponding to 256 s (4.26 min, padded, if necessary) were taken as input to the LSTM-based classifier. The windows greater than 256 were split in an overlapping fashion. Each CPD, along with the heart rate at its corresponding time point, was taken to finally label it with the shock index, and hence, the presence of shock/no-shock. The missing heart rate data at certain points were imputed with linear interpolation if the missing data were less than 10% of the time series length. Since the data are highly imbalanced with more non-shock sequences, the training set is augmented with SMOTE ([Chawla et al., 2002](#)). The LSTM sequence classifier was followed by a series of dense layers with a dropout of 0.2, which then passed through a sigmoid layer to output a binary shock index, and hence, shock/no-shock.

Experiments on tsfresh ([Christ et al., 2016](#)) feature classification with linear-mixed-effects models ([Bates et al., 2015](#)) and random forests ([Cutler et al., 2012](#)), and the direct end-to-end classification of images/videos is shown below.

- **Linear Mixed-Effects and Random Forest sequence classification on tsfresh features:** tsfresh features were extracted from 256-length sequences and trained on the same train and validation distributions as the previous LSTM model using the ‘Boruta’ ([Kursa and Rudnicki, 2010](#)) package from R-language. Variation Inflation Factor (VIF) was used to reduce multicollinearity in the data. If VIF exceeds 10, then the collinearity is considered problematic, and hence, the particular variable causing it should be removed. The remaining features were used to train linear mixed-effects and random forest models.
- **Direct Classification of thermal images/videos for future risk of shock:** Apart from CPD extraction, an attempt was made to classify into shock/no-shock by directly giving whole images/videos as input. In one direction, we tried to classify each video frame read at a time and conducted experiments with several modern architectures based on CNN. The concepts of TV-Chambolle denoising, data augmentation, and undersampling/oversampling were used to get the best shock detection AUROC of 0.60 using ResNet-50. Also, the information extracted from a single image frame can be very limited. So instead, we tried to use direct and continuous video samples of length 256s as an input to a conjunction of various CNN-LSTM models, trained in a time-distributed manner. Being a fundamental extension of the direct image classification problem, it suffered from similar limitations.
- **Classification into covered or uncovered patients**



Since the patients are kept under observation for long hours in intensive care conditions, it is possible for them to remain wrapped in a blanket for most of the time. Using ResNet-152, we were able to classify the video frames into covered/uncovered to get a hold of any window during which the blanket is removed. An accuracy of 95.52% was achieved, with an F1 score of 0.9589 (Precision=0.9722, Recall=0.9459), creating fewer chances of false positives.

- **Multiple person detection using YOLOv3**

The patient is barely left alone when he/she is uncovered under long hours of observation. To remove any chance of confusion between the temperatures and the body parts that the following segmentation algorithm might face, YOLOv3 was deployed for multiple person detection, which achieved an IoU of 0.71206. The other person and the background were then masked to keep the patient in focus for CPD extraction.

**Model Selection Rationale:**

LSTM-based architectures were chosen for the thermal video-based shock prediction task due to their ability to model sequential temporal dynamics in physiological signals derived from thermal imagery. Compared to static classifiers, recurrent models are better suited to capture evolving thermal patterns associated with circulatory compromise. Alternative architectures were evaluated to contextualize performance, while prioritizing models compatible with real-time, low-latency deployment in the ICU setting.

**Temporal and Patient-Level Separation:**

Video data were partitioned at the patient level, and temporal ordering was preserved such that training data preceded validation and test data. This design prevents temporal leakage and reflects realistic deployment conditions.

#### **4.2.2.6 Outcome Variable Binary Shock Index**

The SAFE-ICU initiative has enabled this research to gather the PICU data and extract vitals and corresponding timestamps at the 0th hour (time of video capturing) and the next 6 h. The shock index was taken as the median heart rate and the median arterial systolic blood pressure, for moving sequential windows of 30 data points at a resolution of 15 s. SIPA ([Acker et al., 2017](#)) was then used to compute the age-specific binary outcome for each patient.

#### **4.2.2.7 Model Evaluation**

The video data were first partitioned patient-wise, such as to keep the train, validation, and test sets unseen from each other. For the 10-fold cross-validation, the data were partitioned with a ratio of 60:20:20 into these three sets in a stratified manner, that is, keeping the distribution of low-percentage shock class comparable in all three sets.

Training data was augmented for the low-found shock class using SMOTE; the validation and test sets remain unchanged in size in each respective fold. The model analysis was mostly conducted on the AUPRC and the AUROC curve. Other standard metrics like F1-score, PPV, NPV, specificity, and sensitivity were evaluated at the Youden's Index (J) ([Ruopp et al., 2008](#)). Since there is a high significance of prevalence in the medical domain, calculating the metrics at the Youden's Index becomes important.

### 4.2.3. Results

#### 4.2.3.1 Patient Characteristics, Preprocessing, and Cohort Building

Statistical inferences of the cohort characteristics between shock and non-shock groups are depicted in Table 4.5. Shock and Non-shock conditions were derived using the age-specific shock index cutoff values. It can be observed that the most significant difference between the two categories is in the heart rate, as expected, along with the respiratory rate. We obtained a median length of stay of 6.808 days with a total of 132 shock instances and 274 non-shock instances. The study uses arterial systolic blood pressure, the gold standard for BP measurement in critical care settings ([Pittman et al., 2004](#); [Lehman et al., 2013](#)), from 22 patients monitored longitudinally as a required vital for the calculation of shock index for the training purposes which is available for only a small subset of patients. However, once the model is trained, this is no longer a limitation at inference time, as the trained model does not require blood pressure for inference (making a prediction of shock status). The pipeline uses >1,00,000 frames and the corresponding shock indices for training the prediction models.

*Table 4.5 Cohort characteristics and statistical significance of control (non-shock) vs. affected (shock) classes. The p-values were calculated using either the Wilcoxon rank sum test (W) or Student's t-test (t) after testing for normality by the D'Agostino-Pearson normality test. (n, number of sequences; IQR, interquartile range)*

| Variable                                       | Non-shock seq (n = 274) | Shock seq (n = 132) | Statistical tests |
|------------------------------------------------|-------------------------|---------------------|-------------------|
|                                                | Median (IQR)            | Median (IQR)        | p-value (W/t)     |
| Age (months)                                   | 60.12 (36.07)           | 75.89 (107.02)      | 0.6087 (W)        |
| Arterial systolic blood pressure, mm Hg        | 131.94 (19.56)          | 128.09 (25.21)      | 0.0014 (t)        |
| Systolic blood pressure, mm Hg                 | 106.00 (20)             | 102.00 (5.00)       | 0.002 (W)         |
| Heart rate, per min                            | 106.16 (19.14)          | 143.17 (63.81)      | <0.0001 (t)       |
| Respiratory rate, per min                      | 32.34 (13.14)           | 22.87 (13.20)       | <0.0001 (W)       |
| O <sub>2</sub> saturation (SpO <sub>2</sub> )% | 97.50 (2.68)            | 98.32 (3.45)        | 0.9932 (W)        |

#### 4.2.3.2 Segmentation of Abdomen and Feet Achieved a Total Dice Loss of 0.0391 Using ResUNet

Since thermal images lack texture, it is important for the model to recognize the structural shape features. ResUNet was specifically used to make the task possible on relatively less available thermal data and to

capture both the local and global information from the image. A dice loss of 0.0391 (dice coefficient = 0.9609) with a Binary Cross-Entropy loss of 0.0692 was achieved for segmenting out the abdomen, feet, and background. The mode intensities of segmented areas were then used to find out CPD and hence build the longitudinal models from continuous long-duration videos.

#### 4.2.3.3 The LSTM Model was found to be the Best Performing

We compared three models to finally arrive at the best-performing one. Linear-Mixed Effects ([Bates et al., 2015](#)), Random Forest ([Cutler et al., 2012](#)), and LSTM ([Hochreiter and Schmidhuber, 1997](#)) were tested at various time points from the observation taken. Based on various metric evaluations, LSTM was found to be the best performing on our given time series data and hence was chosen as the primary choice for our study. The F1 score comparison of the three models is shown in [Figure 4.10](#).

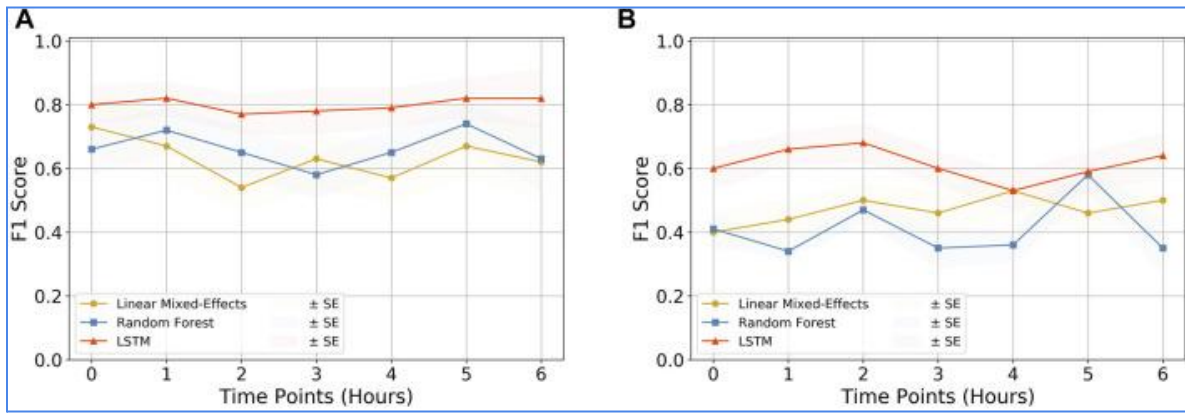


Figure 4.10 Quality assessment of models. The F1 score comparison for the three models (i.e., Linear Mixed-Effects model, Random Forest, and LSTM) tested on (A) CPD and heart rate and (B) CPD parameter only. It can be observed that the LSTM model outperforms the other two models in both cases, making it the primary choice of this research

#### 4.2.3.4 Evaluation of the LSTM Model at Different Lead Times Reveals That CPD Increases the Model Performance

The SAFE-ICU resources allowed us to match the time sequences with their corresponding states of shock/non-shock for the next 6 h since the observation was taken. We tested the LSTM model performance for 1–6 h of lead time on CPD and the heart rate, as shown in [Table 4.6](#). As shown in [Figure 4.11](#), we got the best performance at AUROC  $0.81 \pm 0.06$  and AUPRC  $0.78 \pm 0.05$  at 5 h. This gives a good time window to alert the medical practitioners to start the therapy for preventing the advent of shock.

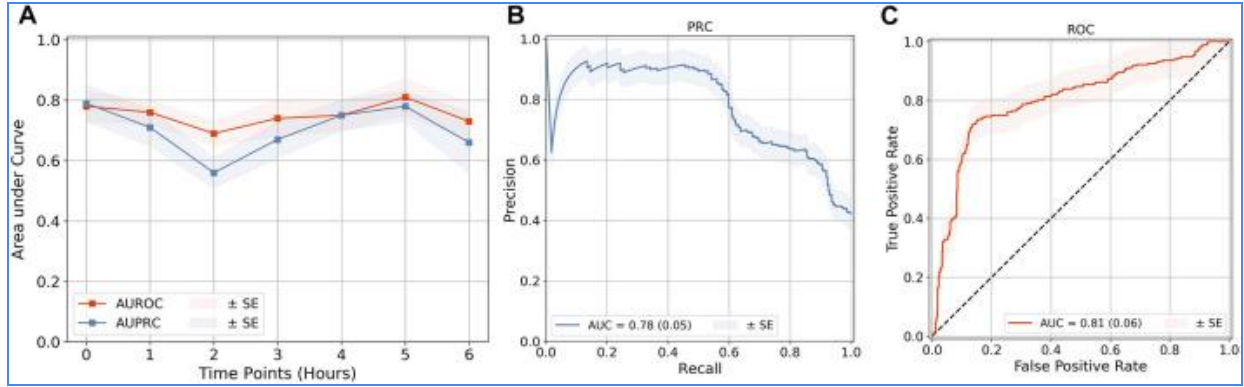


Figure 4.11 Quality evaluation on the LSTM model. (A) Quality evaluation of the LSTM time-series classification models at lead points of up to 6 h. The best performance of (B) AUPRC and (C) AUROC was obtained at a 5-h prediction

Table 4.6 Performance of the proposed model predicting the presence of shock/non-shock using automated CPD and heart rate. The time pt. column depicts the subsequent hours from the time of taking the observation, at which the results were recorded. The unequal number of shock and non-shock sequences is due to the absence of patient data with the increasing number of hours. (S/Ns, number of shock/non-shock sequences present; AUPRC, area under precision-recall curve; AUROC, area under receiver operating characteristics; PPV, positive predictive value; NPV, negative predictive value; D, detection; P, prediction)

| Time pt. | S/Ns     | AUPRC          | AUROC          | Accuracy    | Sensitivity | Specificity | PPV            | NPV            | Youden |
|----------|----------|----------------|----------------|-------------|-------------|-------------|----------------|----------------|--------|
|          |          | Mean (SE)      |                |             |             |             |                |                | Mean   |
| 0 h (D)  | 132, 274 | 0.79<br>(0.06) | 0.78<br>(0.05) | 0.85 (0.03) | 0.74 (0.06) | 0.92 (0.02) | 0.86<br>(0.05) | 0.84<br>(0.03) | 0.50   |
| 1 h (P)  | 115, 271 | 0.71<br>(0.06) | 0.76<br>(0.04) | 0.83 (0.04) | 0.83 (0.04) | 0.82 (0.06) | 0.80<br>(0.06) | 0.88<br>(0.02) | 0.42   |
| 2 h (P)  | 125, 253 | 0.56<br>(0.05) | 0.69<br>(0.04) | 0.83 (0.02) | 0.72 (0.06) | 0.90 (0.03) | 0.82<br>(0.05) | 0.84<br>(0.04) | 0.56   |
| 3 h (P)  | 133, 232 | 0.67<br>(0.06) | 0.74<br>(0.06) | 0.83 (0.04) | 0.69 (0.08) | 0.93 (0.04) | 0.88<br>(0.06) | 0.82<br>(0.04) | 0.57   |
| 4 h (P)  | 123, 242 | 0.75<br>(0.05) | 0.75<br>(0.05) | 0.81 (0.04) | 0.70 (0.06) | 0.94 (0.03) | 0.90<br>(0.05) | 0.78<br>(0.06) | 0.64   |
| 5 h (P)  | 120, 247 | 0.78<br>(0.05) | 0.81<br>(0.06) | 0.84 (0.04) | 0.76 (0.07) | 0.94 (0.02) | 0.88<br>(0.05) | 0.84<br>(0.05) | 0.62   |
| 6 h (P)  | 124, 228 | 0.66<br>(0.10) | 0.73<br>(0.06) | 0.89 (0.03) | 0.82 (0.08) | 0.92 (0.04) | 0.81<br>(0.09) | 0.95<br>(0.02) | 0.62   |

#### 4.2.4 Thermal Imaging Constraints, Failure Modes, and Scalability

ThermalShockNet relies on calibrated thermal imaging hardware operating within standardized temperature ranges. Camera calibration was performed using manufacturer-provided normalization

procedures to ensure consistency across recordings. Patient movement, such as turning or repositioning, and occlusion due to blankets or caregiver interaction may temporarily obscure the facial or torso region of interest. In such cases, the computer vision pipeline is explicitly designed to *withhold prediction* rather than generate potentially misleading outputs. These scenarios are treated as data unavailability events, not negative predictions, thereby prioritizing patient safety and minimizing false alerts during live ICU deployment.

To robustly detect such failure conditions, dedicated preprocessing models were incorporated into the pipeline. Blanket occlusion was identified using a ResNet-152-based classifier, achieving an accuracy of 95.52% and an F1-score of 0.9589, enabling reliable identification of uncovered time windows suitable for analysis. In parallel, multiple-person presence-common in pediatric ICU settings-was addressed using YOLOv3-based object detection (IoU = 0.71), ensuring that only the patient's thermal region contributed to downstream feature extraction. When these quality checks fail, the system suppresses inference rather than extrapolating unreliable thermal measurements. This layered design enables scalable deployment while explicitly accounting for real-world ICU conditions such as caregiver presence, prolonged blanket coverage, and patient movement.

#### 4.2.4 Discussion

The study presents a computer vision and deep-learning-based continuous and noncontact shock detection and prediction model that leverages CPD as one of its main parameters. Care needs to be taken to evaluate the parameters in a noninvasive way, such as not causing infections by contact. The role of noncontact, wireless health monitoring has gained much importance, especially during the times of the COVID-19 pandemic. The noncontact estimation of multiple physiological parameters can enable teleradiology and remote management of patients, at the same time decreasing the contact-based infection risk. Among the available noncontact modalities, recent advancements in thermal imaging have made thermal imaging a key technology. Affordable smartphone thermal cameras have been used to detect the breathing rate and hence the perceived mental stress ([Cho et al., 2019](#)) and even help identify sports injuries ([Quintana et al., 2015](#)). In ICUs, studies providing noncontact measurements such as ECG ([Majumder et al., 2018](#)) and other physiological measurements using noncontact cameras ([Jorge et al., 2022](#)) have been explored. Wireless sensors have shown potential for the detection of complex underlying neurological disorders, such as [Shah et al. \(2019\)](#) use wireless sensor networks (WSN) for monitoring pill-rolling tremors, which can be proven to be a precursor to Parkinson's disease. Similarly, [Liu et al. \(2018\)](#) use WSN to detect and monitor abnormal breathing patterns as an underlying condition for more gruesome diseases like hyperthyroidism and sleep apnea. Since the field of thermal image inspection has only started to be explored, it can thus be leveraged, along with some noninvasively monitored vital parameters, for the prediction of shock. The use of deep learning methods proves to be really beneficial in reducing manual preprocessing and increasing the accuracy of methods.

In this study, we have extended our previous study on hemodynamic shock prediction with longitudinal continuous monitoring of body thermal patterns, which are found to be predictive for future shock prediction ([Nagori et al., 2019](#)). Longitudinal monitoring of temperature gradient opens up a rich source of information about patient physiology. The underlying stochastic patterns can have a discriminative value for future hemodynamic shock risk. We leveraged these reasons to extract the center-to-peripheral intensity difference from the thermal images in a time-series fashion. We do so by applying our

data-specific trained screening filters for covered and uncovered patient detection models, and multiple person detection models, which are followed by segmenting the body parts into abdomen and foot using artificial intelligence-based models. The extracted CPD time series, along with vitals, were used to predict the future (1–6 h) hemodynamic shock using sequence models (LSTM). This model was also compared with Random-Forest and Linear Mixed-Effects models, known for working with longitudinal data, but LSTM outperformed the other two. We compared the results of our method with Nagori et al.'s study, too, and an observable improvement can be seen in the majority of the metrics ([Table 4.7](#)). We also trained and checked the performance of models that can classify the length of videos and images for the future risk of hemodynamic shock. CNNs, which have shown state-of-the-art performance for image classification, were used for the task. However, we found that domain features, such as CPD, performed better than direct image and video classification for the future risk of shock. Even after many experiments performed for a direct classification of images using the concepts of TV-Chambolle ([Condat, 2013](#)) denoising, data augmentation, and undersampling/oversampling, we were able to get the best AUROC of only 0.60 for shock detection using ResNet-50. This can be because of the cluttered background with diapers and tubes, and an increased region of interest for the information extraction from a limited variety of images. This also tells the importance of the domain-specific features over end-to-end CNN-based approaches. We built the models for multiple time points till 6 h, which showed the best performance at a lead time of 5 h. The metric AUPRC and F1 scores are most significant for an imbalanced dataset such as ours, as it does not get biased by the presence of true negatives and thus gives a clearer perspective of a classifier's utility. The results till 6 h show a promising window with a good prediction rate, which can prove to be helpful for the doctors to help find a buffer time prior to the shock event and therefore start the treatment.

Our study currently applies to cold shock in neonates, and young children are more commonly present with “cold shock,” a state of elevated systemic vascular resistance (SVR) and low cardiac output with cold extremities due to vasoconstriction and delayed capillary refill ([Martin and Weiss, 2015](#)). The peripheral vasoconstriction can be identified with the center-to-peripheral temperature difference using thermal imaging that we leveraged in our study to identify the cold shock. We also acknowledge the limitations and uncertainty caused by using skin temperature and core-peripheral temperature gradient in isolation ([Schev et al., 2010](#)); therefore, we introduced the heart rate as a crucial cardiovascular parameter ([Table 4.5](#)) in our CPD-based method for shock prediction. RR is also an important indicator of cardiorespiratory status ([Liu et al., 2020](#)) as indicated in [Table 4.5](#).

However, we did not encapsulate RR in our models owing to its limited availability at all-time points in our dataset and in a bid to keep the method parsimonious. Nevertheless, we did study the effects of including RR with the available measurements till 6 h, and it showed promising results as demonstrated in [Table 4.8](#), which can be taken up in our future research. Models that are predictive of future occurrences of events suffer from lower model performance indicators, and a trade-off is needed between sensitivity and specificity. While probability thresholds can be easily set to decrease the false positives (at the cost of sensitivity), our model is intended to be used as a screening tool, as a shock is a potentially lethal condition. Moreover, the FPR varies between 10 and 15% at different time points and can certainly be improved further with the incorporation of additional features. We will be taking up further performance improvement in our future research to bring it down to less than 5%.



The other limitations of our method arise during the quality data procurement. Monitoring thermal patterns could be hard at times when the patient is being covered or being operated on; caregivers can also block the recordings. However, these issues can be resolved by keeping ~4.3 minutes of uncovered slots for monitoring. Nevertheless, we have built an AI system that is efficient enough to identify the covered and uncovered images and can remove the caregiver or person other than the patient. Thus, our system is able to extract the thermal patterns efficiently. Using CPD as a co-acting parameter eliminates the use of invasive arterial monitoring and cuff-based blood pressure monitors, which can cause infections due to the repeated use on multiple patients. Our models require thermal recordings, which do not identify an individual's identity, hence are safe for patient privacy and do not take much time away from the time assigned for care. Our methods can be extended beyond the intensive care settings, that is, to help in the community-based monitoring by the developing countries' rural healthcare workers, such as the Accredited Social Health Activist (ASHA), run by women promoting accessible healthcare in rural Indian settings. This can be made possible by training them to use a specifically designed frugal mobile app for thermal shock detection and prediction, which can work with readily available Android/iOS smartphones, with only model weights preinstalled. The results can be made more robust by expanding the dataset by including more patients in the study. Also, the dataset can be made varied and generalized by including observations from multiple other clinical sites. Thus, the study performed illustrates a great noninvasive and minimal feature architecture that promises to be a lifesaver by informing clinicians about shock well in advance.

*Table 4.7 Comparison between the results of machine learning techniques for hemodynamic shock prediction by Nagori et al. and our deep learning methods, although on different sample sizes, is shown in the table below. The metrics not shown in the table below weren't available for Nagori et al, and hence the results cannot be compared for them. It can be observed that our deep learning techniques, along with being end-to-end without the need for handcrafted features, show an improvement in each metric by ~10%, which proves the efficacy of our method. Only a slight improvement in AUROC can be because of the imbalanced data in our case. (ML - Machine Learning, DL - Deep Learning, AUROC - Area Under Receiver Operating Characteristics, PPV - Positive predictive value, NPV - Negative predictive value, D - Detection, P - Prediction. All values are in Mean(SE))*

| Time Pt. | AUROC                 |                | Accuracy              |                | Sensitivity           |                | Specificity           |                | PPV                   |                | NPV                   |                |
|----------|-----------------------|----------------|-----------------------|----------------|-----------------------|----------------|-----------------------|----------------|-----------------------|----------------|-----------------------|----------------|
|          | DL                    | ML             | DL                    | ML             | DL                    | ML             | DL                    | ML             | DL                    | ML             | DL                    | ML             |
| 0hr (D)  | <b>0.78</b><br>(0.05) | 0.75<br>(0.03) | <b>0.85</b><br>(0.03) | 0.73<br>(0.03) | <b>0.74</b><br>(0.06) | 0.58<br>(0.06) | <b>0.92</b><br>(0.02) | 0.81<br>(0.06) | <b>0.86</b><br>(0.05) | 0.74<br>(0.05) | <b>0.84</b><br>(0.03) | 0.75<br>(0.03) |
| 3hr (P)  | 0.74<br>(0.06)        | 0.77<br>(0.04) | <b>0.83</b><br>(0.04) | 0.73<br>(0.03) | <b>0.69</b><br>(0.08) | 0.65<br>(0.06) | <b>0.93</b><br>(0.04) | 0.82<br>(0.03) | <b>0.88</b><br>(0.06) | 0.75<br>(0.04) | <b>0.82</b><br>(0.04) | 0.74<br>(0.05) |
| 6hr (P)  | <b>0.73</b><br>(0.06) | 0.68<br>(0.03) | <b>0.89</b><br>(0.03) | 0.69<br>(0.02) | <b>0.82</b><br>(0.08) | 0.58<br>(0.07) | <b>0.92</b><br>(0.04) | 0.74<br>(0.06) | <b>0.81</b><br>(0.09) | 0.79<br>(0.04) | <b>0.95</b><br>(0.02) | 0.64<br>(0.02) |

*Table 4.8 Performance of the proposed model predicting the presence of Shock/Non-shock using automated CPD, heart rate and respiratory rate. The Time Pt. column depicts the subsequent hours from the time of taking the observation, at which the results were recorded. The unequal number of shock and non-shock sequences is due to the absence of patient data with the increasing number of hours. (S/NS - Number of Shock/Non-Shock sequences present, AUPRC - Area Under Precision-Recall Curve, AUROC - Area Under Receiver Operating Characteristics, PPV - Positive predictive value, NPV - Negative predictive value, D - Detection, P - Prediction)*

| <b>Time Pt.</b> | <b>S/NS</b> | <b>AUPRC</b>   | <b>AUROC</b>   | <b>Accuracy</b> | <b>Sensitivity</b> | <b>Specificity</b> | <b>PPV</b>  | <b>NPV</b>     | <b>Youden</b> |
|-----------------|-------------|----------------|----------------|-----------------|--------------------|--------------------|-------------|----------------|---------------|
| <b>0hr</b>      | 120, 220    | 0.7 (0.04)     | 0.7 (0.06)     | 0.77<br>(0.05)  | 0.56 (0.08)        | 0.96 (0.02)        | 0.92 (0.03) | 0.72<br>(0.05) | 0.61          |
| <b>1hr</b>      | 116, 215    | 0.75<br>(0.05) | 0.77<br>(0.05) | 0.84<br>(0.03)  | 0.70 (0.07)        | 0.92 (0.05)        | 0.91 (0.04) | 0.83<br>(0.04) | 0.47          |
| <b>2hr</b>      | 120, 206    | 0.44<br>(0.07) | 0.61<br>(0.06) | 0.80<br>(0.04)  | 0.66 (0.05)        | 0.90 (0.07)        | 0.83 (0.08) | 0.82<br>(0.03) | 0.55          |
| <b>3hr</b>      | 121, 190    | 0.84<br>(0.05) | 0.81<br>(0.05) | 0.88<br>(0.03)  | 0.75 (0.08)        | 0.97 (0.01)        | 0.95 (0.02) | 0.86<br>(0.03) | 0.54          |
| <b>4hr</b>      | 125, 184    | 0.69<br>(0.03) | 0.68<br>(0.04) | 0.86<br>(0.03)  | 0.72 (0.05)        | 0.99 (0.01)        | 0.98 (0.02) | 0.79<br>(0.05) | 0.5           |
| <b>5hr</b>      | 113, 198    | 0.81<br>(0.03) | 0.88<br>(0.03) | 0.93<br>(0.02)  | 0.90 (0.04)        | 0.95 (0.02)        | 0.94 (0.02) | 0.94<br>(0.02) | 0.4           |
| <b>6hr</b>      | 111, 197    | 0.82<br>(0.09) | 0.87<br>(0.07) | 0.85<br>(0.07)  | 0.86 (0.06)        | 0.85 (0.09)        | 0.82 (0.10) | 0.95<br>(0.02) | 0.31          |



## 4.3 A Prospectively Validated Generalizable Model for Outcome Prognostication Using Shock Index in Intensive Care Units

### 4.3.1 Introduction

Delays in the identification of critical events in intensive care units contribute to irreversible organ dysfunction and mortality([Armen et al., 2016](#)). Shock Index, calculated as the ratio of heart rate and systolic blood pressure, is a validated indicator for bedside examination, risk stratification, and prognostication([Koch et al., 2019](#), [Cheng et al., 2020](#)). SI was originally designed to evaluate hemodynamic stability, with values  $>0.7$  indicating shock. Shock is one of the highest contributors to higher mortality rates, and SI has been used to assess Shock in ICUs([Myint et al., 2018](#)). The mortality rate in patients who develop shock in the ICU is as high as 34% in developing countries([Divatia et al., 2016](#)). While in developing countries in Europe and North America, the mortality rates are at 38% due to septic shock([Vincent et al., 2019](#)). Thus, early identification is critical for appropriate management([Armen et al., 2016](#)), improved patient outcomes, and reduction of mortality([Seymour et al., 2016](#), [Herget-Rosenthal et al., 2008](#), [Bai et al., 2014](#), [Berger et al., 2013](#)).

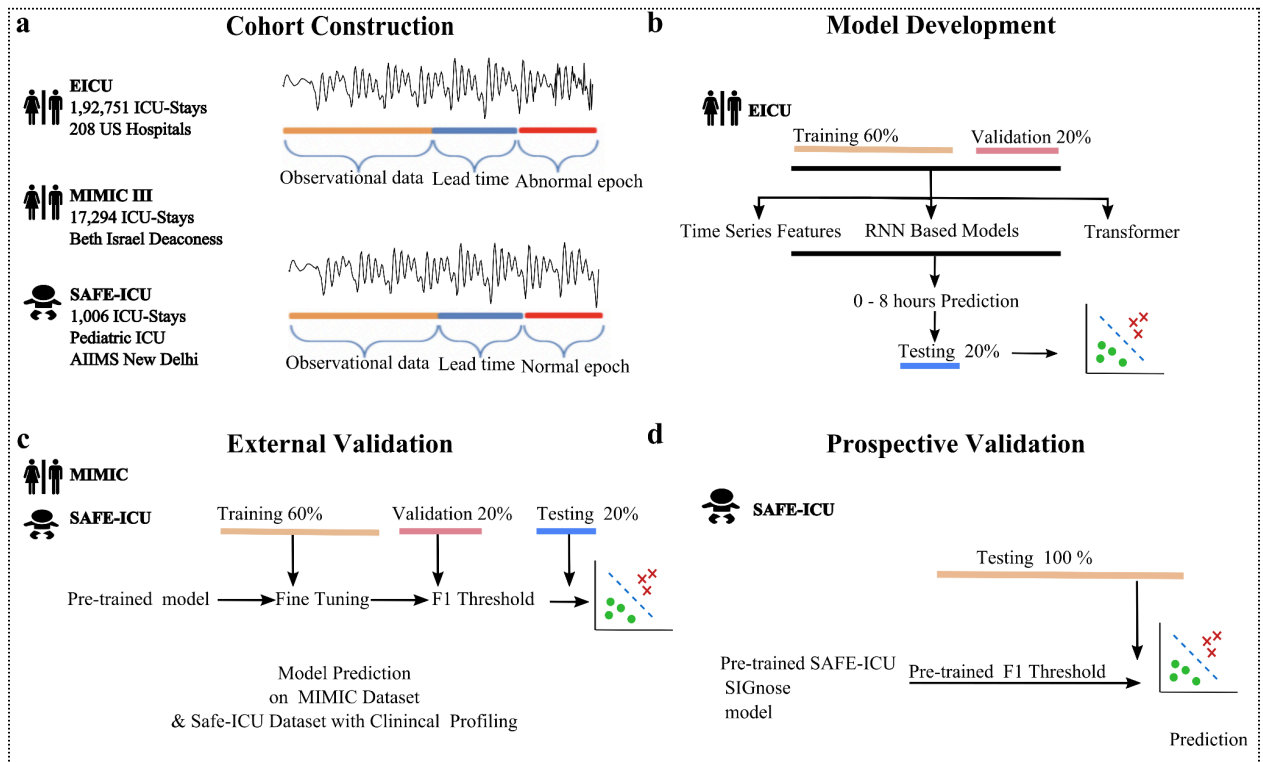
More recently, SI has been validated as an indicator of outcomes and their prognostication in conditions such as sepsis, pneumonia, acute myocardial infarction, stroke, advanced cancer, pulmonary embolism, and trauma([Koch et al., 2019](#), [Cheng et al., 2020](#), [Myint et al., 2018](#), [Berger et al., 2013](#)). These conditions are more likely to respond to interventions in their initial stages, whereas if a potential irreversible organ failure starts, mortality is high([Maheshwari et al., 2020](#)). In this regard, SI has been gaining evidence as an effective indicator for preventing mortality due to delayed identification in sepsis, shock, and Acute Kidney Injury (AKI)([Barbar et al., 2014](#)). Further, unlike complex scoring criteria, this index does not rely upon variables that require clinical judgment and is straightforward to calculate based on vitals and age alone. These factors make SI an ideal indicator for building predictive models for prognosticating outcomes. Earlier studies have used different thresholds for SI for specific outcomes. Since the goal of our study was early detection, the threshold of  $SI > 0.7$  was chosen. Other studies have used higher thresholds, such as  $SI > 0.9$  or  $SI > 1$ , which may be more relevant for downstream outcomes such as organ dysfunction but may not be sensitive enough to capture early shock.

Several machine learning and artificial intelligence-based approaches have been developed for predicting ICU outcomes. However, the generalizability of these models([Potes et al., 2017](#), [Nemati et al., 2018](#)) is difficult due to the variability in data types, temporal resolution, and the difficulty in defining the onset time. Physiological time series from vitals monitoring have shown potential for multicenter generalization and for use in ICUs([Hyland et al., 2020](#), [Mao et al., 2018](#)). Further, the current best-performing models based on accuracy on large datasets use features that may be available only at infrequent intervals, thus compromising the real-time potential of the models. On the other hand, vitals time-series are routinely monitored and

aggregated every few seconds, thus having the potential to forecast critical outcomes([Mao et al., 2018](#), [Desautels et al., 2016](#)). This study bridges the gap using the vitals time series to develop real-time generalizable models.

In parallel, the emergence of LLMs such as LLaMA has opened new possibilities for learning patient representations from structured or semi-structured data by converting them into text prompts ([Zhang & Ni, 2024](#), Future AGI, n.d.). These models, pretrained on vast textual corpora, can capture complex temporal and contextual patterns when patient vitals are framed as sequential inputs. However, the generalizability of such approaches across diverse datasets remains a critical challenge. In this study, we investigated the utility of LLaMA-based embeddings for shock prediction and benchmarked their performance against SIGnose, a specialized non-linear time-series model, to assess their comparative effectiveness and cross-cohort adaptability.

We hypothesized that the generalizability gap could be addressed using a transfer learning approach where a novel model built on eICU data (a multicenter ICU database)([Pollard et al., 2018](#)) validated on MIMIC-III adult dataset([Johnson et al., 2016](#)) from the USA and safeICU([Sethi et al., 2017](#)) pediatric data from New Delhi, India. The resulting model was validated prospectively in the New Delhi pediatric setting. [Figure 4.12](#) summarizes the study approach and expectations.



*Figure 4.12 Summary of the pipeline for generalizing the prediction of hemodynamic instability. 1a) shows Cohort Construction, where time-series of Heart rate, systolic arterial blood pressure, diastolic arterial blood*

*pressure, respiratory rate, and oxygen saturation were used. 1b) shows the pipeline for constructing models for predicting hemodynamic instability 0-8 hours in advance and evaluation. 1c) shows the generalization of our best model, SIngnose learned models on MIMIC-III data and Indian settings' SAFE-ICU data. 1d) shows the prospective validation on the prospective Safe-ICU dataset.*

### **"Clinical Use-Case:**

Signose is intended to support early identification of hemodynamic deterioration in critically ill pediatric patients. The model provides advance warning of abnormal shock index up to 8 hours prior to onset, allowing clinicians to consider closer monitoring, diagnostic evaluation, or escalation of care. The predictions are designed to assist situational awareness and risk stratification rather than replace clinical decision-making."

## **4.3.2 Methods**

### **4.3.2.1 Dataset description & pre-processing**

We leveraged two publicly available datasets from the USA and a safeICU, our in-house dataset with 0.23 million patient hours from a tertiary care setting in India. The multicenter eICU data, comprising 200,859 unit admissions of which 192,751 contained vitals time-series data, collected from 208 hospitals from 2014 to 2015 ([Pollard et al., 2018](#)), was used for cohort construction. For external validation, we used (i) safeICU pediatric data warehoused between February 2016 to January 2020, containing 1006 ICU-stays of 782 patients. The safeICU study is an in-house data warehouse at the PICU of the AIIMS, New Delhi. (ii) MIMIC-III v1.4 data collected between 2001 to 2012. The MIMIC-III numeric time-series contains adult population records of 17,294 ICU-stays of 10,269 patients. The study was approved by the Institute Ethics Committee, AIIMS, New Delhi (IEC/NP-211/08.05.2015) and involved no change in routine patient care. In addition to external validation, data collected from February 2020 to August 2020 in the safeICU data were used for prospective validation and generalization of models developed on eICU data.

### **4.3.2.2 Cohort Construction**

**Development cohort.** eICU data were used to construct the model development cohort and included data categorized as *medical (MICU)*, *surgical (SICU)*, *medical-surgical ICU*, *coronary care (CCU)*, *cardiac ICU*, *cardiac surgery(CSICU)*, *cardiothoracic (CTICU)*, and *neuro-ICU*. Our inclusion criteria for development cohort construction included all patients with continuous vitals monitoring for a minimum of 7.5 hours in their ICU stay. The exclusion criteria were the following: (i) patients with non-availability of invasive arterial blood pressure monitoring in the observation window, (ii) patients with non-availability of patient demographics (age, gender, or age greater than 89 years), (iii) patients with abnormal SI within the first 7 hours of admission to avoid biasing the model with patients already in shock, and finally (iv) patients with missing data in >10% of the observation period.

**External validation cohort.** MIMIC-III and SAFE-ICU data were used as external validation cohorts with similar inclusion and exclusion criteria as described in the development cohort. Since SAFE-ICU is a pediatric data resource, different SI thresholds were used for different age groups (SI >2.3 (ages ≤ 3 months), >1.7 (4-6 months), >1.5 (7-12 months), >1.2 (13-36 months), >1.15 (37-72 months), >0.95 (73-144 months), >0.77 (>144 months))(21).

**Prospective validation cohort.** Prospective validation cohorts were constructed using prospectively collected data from the pediatric ICU facility at AIIMS, New Delhi, using the same inclusion and exclusion criteria.

#### 4.3.2.3 Pre-processing and Scoring

Heart rate, systolic arterial blood pressure, diastolic arterial blood pressure, respiratory rate, and oxygen saturation were used as the physiological vitals to be used as model predictors. Minimal pre-processing of the vitals data was done. Vitals data from patients eligible for the development cohort were imputed using the Kalman filter. The SI score was calculated for each time stamp in the pre-processed time series. For labelling the time series for clinical deterioration of SI, we divided each time series into 30-minute time epochs. The onset time of abnormal SI ( $t_{(ASI)}$ ) is defined as the starting time of an epoch in which the median SI value was greater than 0.7 ([Berger et al., 2013](#), [Birkhahn et al., 2005](#)). The length of training windows was 7 hours, with the index time defined with a lead time of up to 8 hours from the onset time ([Figure 4.13](#)). Exploratory data analyses were carried out in R software version 3.6 for computation of data characteristics and assessing the associations of first three days of SI status with length of stay and mortality incidence.

#### Ground Truth Definition (Signose)

Ground truth labels for clinical shock were derived exclusively from **physiologically measured vital signs**, operationalized through the Shock Index (SI). SI was computed at each time point using heart rate and systolic arterial blood pressure, and abnormal shock onset was defined using a clinically established threshold (SI > 0.7). Administrative diagnostic codes (e.g., ICD codes) were **not used** to define shock labels in this study. This physiologically grounded labeling strategy avoids the known limitations of retrospective coding, such as misclassification and temporal delay, and ensures that model training reflects real-time clinical deterioration.

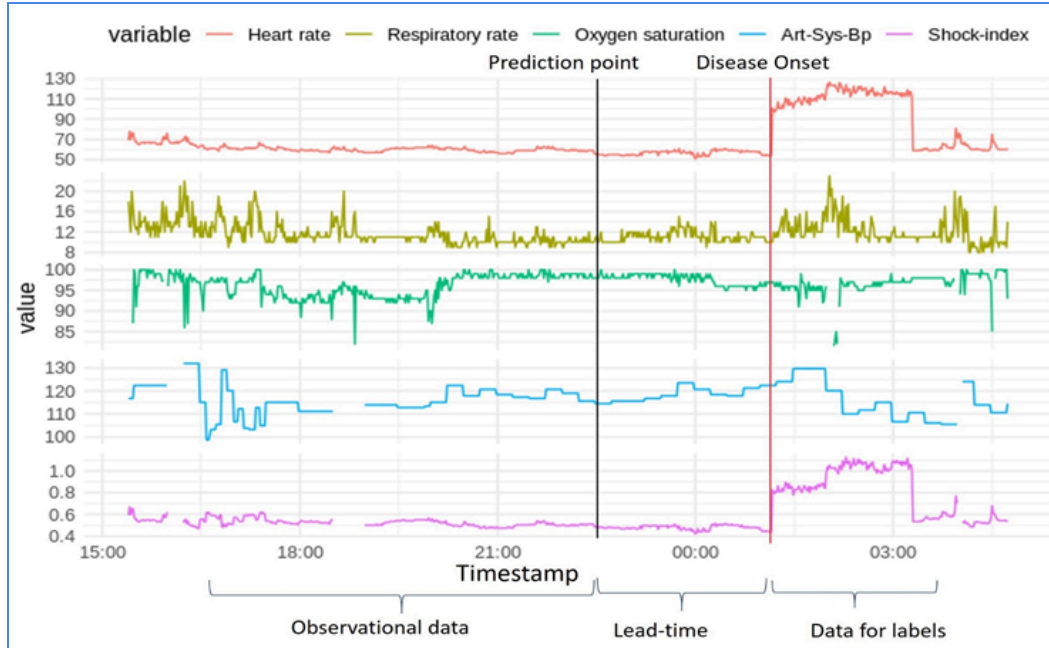


Figure 4.13: Each 30-minute time window is labeled as shock-positive or shock-negative. A 420-minute observation window is used to predict shock status up to 8 hours ahead

#### 4.3.2.4 Engineering Explainable Feature and Feature Selection

A set of 3970 TS-features, including linear and nonlinear physiological features such as statistics, periodicity, wavelet, entropy, and other complex features, were extracted using *tsfresh* ([Christ et al., 2018](#)) library. TS-features are explainable and capture physiological domain understanding, as opposed to black box features using deep learning. Variable selection was performed using a random forest based Boruta algorithm ([Kursa & Rudnicki, 2010](#)), implemented in R.

#### 4.3.2.5 Model Development, Evaluation, and Interpretability

The development cohort consisted of patients who had abnormal SI within the next 8 hours of the index time. The outcome variable- abnormal SI within the next 8 hours was defined by SI scores. Since generalizability requires the models to perform across different sites, the training and testing sets were designed to contain non-overlapping clinical sites. Five-fold cross-validation sets were constructed, and hyperparameter tuning was performed to achieve the best performance on the validation set using grid search. F1 thresholds were used as probability cutoffs derived from the validation set. Results were reported as the median and standard deviation on fivefold test sets. The AUROC, the AUPRC, and Recall & Precision were evaluated on the test set. Since only a small subset of patients has invasive arterial blood pressure monitoring, two variants of the model (with and without invasive monitoring) were constructed and compared on the same number of samples. The model output threshold of the pre-trained model was tuned to the validation set to improve model generalization on unseen data. Model

performance for different types, i.e., medical, coronary care, surgical, cardiac surgery, and cardiothoracic ICUs, was evaluated. A comparison of model performance indicators for different lead times, age groups, and time since admission was also done. Further, we compared standard ML models with feature engineering with LSTM([Hochreiter & Schmidhuber, 1997](#)) models without feature engineering. For the first approach, vitals TS-features were computed, z-score normalized, and used for training GBM([Friedman, 2001](#)), random forest([Breiman, 2001](#)), and multilayer-perceptron models using the sklearn( [Pedregosa et al., 2012](#)) library in Python version 3.6.3. The catalogue of TS-features includes statistical, frequency domain, and nonlinear features ([Table 4.9](#)). In the second approach, Bidirectional LSTM (Bi-LSTM)([Schuster & Paliwal, 1997](#)) and Convolutional Neural Network LSTM (CNN-LSTM) models were fitted on the raw time-series from vitals in Python 3.6 using Keras (version 2.3.1), Tensorflow (version 2.1.0), and a softmax activation was used to generate probability scores. In the third approach, we used LLaMA-7B ([Grattafiori et al., n.d.](#)) with structured vital sign time-series data to evaluate the use of LLMs for shock prediction. Each patient record included demographic information (age, gender) and physiological variables recorded at 5-minute intervals: oxygen saturation, heart rate, respiratory rate, and arterial systolic and diastolic blood pressure. These data were preprocessed and converted into structured textual prompts designed to describe patient vitals and their temporal progression. The prompts followed a standardized instruction-based format and were input into LLaMA-7B to generate contextual embeddings representing each patient's physiological state.

The prompt was

Instruction: Analyze the following medical data and predict the outcome.  
*Instruction: Analyze the following medical data and predict the outcome.*  
*Input: Age: {age}, Gender: {gender}, This data includes observations at 5-minute intervals containing - Oxygen saturation (%): {os2}, Heart rate (bpm): {hrate}, Respiratory rate (brpm): {resp\_rate}, Arterial systolic blood pressure (mmHg): {ss}, Arterial diastolic blood pressure (mmHg): {sd}*

These embeddings were then used as features in multiple machine learning models SVM, XGBoost, LightGBM, and AdaBoost for the downstream classification task of clinical shock prediction. This setup allowed us to compare the embedding-based approach with traditional time-series modeling techniques.

Age and gender were included as covariates in all the approaches. The winning approach and the fine-tuned models were tested for generalization and prospective validation. EQUATOR (Enhancing the QUALity and Transparency Of Health Research) Network's TRIPOD([Collins et al., 2015](#)) guidelines were followed for developing the machine learning models and for reporting results.



Table 4.9 Non-linear time-series features Description

| Sl.No | Feature                                                         | Description                                                                                                                                      |
|-------|-----------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|
| 1     | Absolute Energy (abs_energy)                                    | Returns the absolute energy of the time series                                                                                                   |
| 2     | Continuous Wavelet Transform Coefficients (cwt_coefficients)    | CWT proposes a time scale illustration of a signal. The duration of the examined signal will assist in dynamically identifying non-linearities.  |
| 3     | Fast Fourier Transformation Coefficient (fft_coefficient)       | This feature uses the Fourier transformation algorithm to calculate the Fourier coefficients for the one-dimensional discrete Fourier transform. |
| 4     | Mean                                                            | This feature will return the mean of x.                                                                                                          |
| 5     | Quantile                                                        | Calculates the q quantile of x. The quantile divides the sample into equal-sized adjacent subgroups.                                             |
| 6     | Lag (c3)                                                        | For all the time series, perfect correlation will be at lag=0. The shift in the time series will decrease the correlation value.                 |
| 7     | Sum of re-occurring data points (sum_of_reoccurring_datapoints) | Returns the sum of all the time series data points that occur more than once.                                                                    |
| 8     | Sum (sum_values)                                                | Will calculate the sum over the time series values.                                                                                              |
| 9     | Minimum                                                         | The minimum number is the least value among the given set of values.                                                                             |
| 10    | Energy ratio by chunks (energy_ratio_by_chunks)                 | This is expressed as a ratio of the sum of all the squares of chunk i over the whole series of chunks.                                           |
| 11    | Partial Autocorrelation (Partial_autocorrelation)               | This provides the partial correlation for a time series under stationary conditions.                                                             |
| 12    | Aggregated linear trend (Agg_linear_trend)                      | This calculates the value for an aggregation function of a linear least squares regression.                                                      |
| 13    | Linear trend (linear_trend)                                     | This will calculate the linear least-squares regression over one less than the length of the time series.                                        |
| 14    | Asymmetric statistic (time_reversal_asymmetry_statistic)        | It means that the entropy will increase as time passes.                                                                                          |
| 15    | Approximate entropy                                             | It calculates the non-stationary and chaotic time series.                                                                                        |

|    |                                                            |                                                                                                                                                                                                                                               |
|----|------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|    | (approximate_entropy)                                      |                                                                                                                                                                                                                                               |
| 16 | Binned entropy (binned_entropy)                            | It will first bin the values and then measure the entropy value.                                                                                                                                                                              |
| 17 | Maximum                                                    | This feature calculates the highest among the time series values.                                                                                                                                                                             |
| 18 | Mean change (mean_change)                                  | It provides the mean value of the difference between time series values.                                                                                                                                                                      |
| 19 | Median                                                     | This feature provides the median over time series values.                                                                                                                                                                                     |
| 20 | Change quantiles (change_quantiles)                        | First, select the corridor on the y-axis, and then within this corridor, calculate the mean of the absolute change in the series.                                                                                                             |
| 21 | Ratio beyond sigma (ratio_beyond_r_sigma)                  | These are the ratios beyond the r-sigma and are away from the mean of the series.                                                                                                                                                             |
| 22 | Index mass quantile (index_mass_quantile)                  | These calculate the relative index where x% of the mass of the time series is present at the left of the relative index.                                                                                                                      |
| 23 | Sum of reoccurring values (sum_of_reoccurring_values)      | These are the sum of all the repeated values in the series.                                                                                                                                                                                   |
| 24 | Longest strike below mean (longest_strike_below_mean)      | This is the length of the longest strike (consecutive subsequence) which is below the mean of the time series.                                                                                                                                |
| 25 | Continuous Wavelet Transform peak (number_cwt_peaks)       | After the smoothing of time series with a Ricker wavelet function, this feature provides the number of peaks in the time series.                                                                                                              |
| 26 | Autocorrelation                                            | This provides the partial correlation for a time series under a stationary condition.                                                                                                                                                         |
| 27 | Largest fixed point of dynamics (max_langevin_fixed_point) | When the time series is fitted to the Langevin model (with deterministic dynamics), this parameter gives the largest fixed point of the dynamics.                                                                                             |
| 28 | percentage_of_reoccurring_values_to_all_values             | In the given time series, this parameter gives a normalized ratio of all the unique values that are there in the time series more than once.                                                                                                  |
| 29 | ratio_value_number_to_time_series_length                   | This measure is closely related to percentage_of_reoccurring_values_to_all_values, as this gives an output equal to 1 if every single value in the time series occurs only once. If the case isn't so, the output value would be less than 1. |



|    |                                                        |                                                                                                                                                                                                                                                                                                                         |
|----|--------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 30 | Standard_deviation                                     | Calculates the standard deviation of the time series. Expresses a measure that signifies how much all the values of the time series differ from the mean value of the time series.                                                                                                                                      |
| 31 | Variance                                               | Calculates the variance of the time series. It is the square of the standard deviation.                                                                                                                                                                                                                                 |
| 32 | Autoregressive coefficient (ar_coefficient)            | In a given time series, this first fits an autoregressive process's unconditional maximum likelihood to it and then continues to calculate the AR coefficients.                                                                                                                                                         |
| 33 | sample_entropy                                         | This parameter provides us with a measure of complexity. Closely related to approximate entropy, it has several advantages over it, such as data length independence and lower computational cost.                                                                                                                      |
| 34 | fft_aggregated                                         | This feature contains several of the properties of the time series, such as skew, variance, etc.                                                                                                                                                                                                                        |
| 35 | kurtosis                                               | In a frequency distribution curve of a time series, this parameter gives a measure of how sharp the peak would be.                                                                                                                                                                                                      |
| 36 | Percentage_of_reoccurring_datapoints_to_all_datapoints | In a given time series, this parameter gives a percentage of all the unique values that are there in the time series more than one time. This is different from percentage_of_reoccurring_values_to_all_values, as this percentage is normalized with respect to the unique values, rather than all of the data points. |
| 37 | skewness                                               | Skewness gives a measure of the asymmetry in the time series, or in other terms, shows on which side the time series leans towards (left or right).                                                                                                                                                                     |
| 38 | Cid_ce :                                               | This is a function calculator that estimates the complexity of time series.                                                                                                                                                                                                                                             |
| 39 | Last_location_of_minimum :                             | The location of the minimum value of x I'd calculated with respect to the length of x.                                                                                                                                                                                                                                  |
| 40 | Spkt_welch_density                                     | In this feature calculator, it first shifts the time series x from the time domain to the frequency domain. Then it estimates the cross-power spectral density.                                                                                                                                                         |
| 41 | Count_below_mean                                       | This function will return the number of values of x that may fall below the mean of x.                                                                                                                                                                                                                                  |

|    |                                 |                                                                                                                                                       |
|----|---------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
| 42 | agg_autocorrelation             | This function takes different lags into consideration and then calculates the value for an (fgg) aggregation function over the (R(l))autocorrelation. |
| 43 | Number_peaks :                  | In time series x, this function will calculate the number of peaks that are at least n.                                                               |
| 44 | Absolute_sum_of_changes:        | This function returns the sum of the absolute value of the consecutive changes in the time series x.                                                  |
| 45 | Mean_abs_change                 | This function returns the mean of the absolute difference between the subsequent time series values.                                                  |
| 46 | Longest_strike_above_mean:      | This function returns the length of the longest consecutive subsequence in x that is bigger than the mean of x.                                       |
| 47 | Friedrich_coefficients:         | It is the coefficient of the polynomial $h(X)$ , where $h(x)$ is fitted to the deterministic dynamics of the Langevin model.                          |
| 48 | Augmented_dickey_fuller:        | The test, which checks if the unit root is present in a time series sample, is called the augmented Dickey-Fuller test, which is a hypothesis test.   |
| 49 | First_location_of_maximum:      | The first location of the maximum value of x is calculated with respect to the length of x.                                                           |
| 50 | First_location_of_minimum       | The first location of the minimum value of x is calculated with respect to the length of x.                                                           |
| 51 | Mean_second_derivative_central: | This function will return the mean value for the central approximation of the second derivative.                                                      |

#### 4.3.2.6 SHAP value computation for model interpretability

SHAP is a method that assigns each feature an importance value based on game-theory principles([Lundberg et al., n.d.](#)). The SHAP values represent the influence of features on model prediction. SHAP value distribution of each feature for all the test sets was computed. All the test sets were combined to plot the model explanation regarding the top 15 important features and their relationship with SHAP values.

#### 4.3.2.7 Model generalization & prospective validation

*SIgnose* model performance was validated on publicly available retrospective MIMIC-III([Johnson et al., 2016](#)) and in-house Pediatric data- SAFE-ICU ([Sethi et al., 2017](#)) for external validation. Prospective validation was performed on in-house pediatric data. Both the data were pre-processed to offset the data-specific variation, such as time resolution and data

cleaning from the HL7 format in SAFE-ICU. The high-resolution MIMIC-III data and SAFE-ICU data were summarized to a 5-minute resolution to harmonize with the requirement of our model trained on eICU data. For each dataset, three modeling scenarios were compared for validation: (i) with fine-tuning of parameters, (ii) without (w/o) fine-tuning, and (iii) siloed model i.e., developed only on the respective site's data. The model weights and the probability threshold were tuned to the external setting. Finally, prospective validation of the model was performed on the SAFE-ICU data collected from February 2020 to August 2020.

#### **Model Selection Rationale:**

A combination of traditional machine learning models and deep learning architectures was evaluated to assess trade-offs between interpretability, temporal modeling capacity, and generalizability. Tree-based models and multilayer perceptrons were selected to leverage engineered physiological time-series features, while LSTM-based architectures were explored for their ability to model long-range temporal dependencies directly from raw signals. Large language model-based embeddings were included as an exploratory comparison to evaluate their suitability for structured physiological time-series data."

#### **Temporal Integrity of Evaluation:**

Temporal separation between training and test data was ensured through the lead-time-based construction of observation windows prior to abnormal shock index onset, combined with patient- and site-level partitioning. Random splitting across time was not performed, preventing look-ahead bias.

### **4.3.3 Results**

#### **4.3.3.1 Data characteristics**

The development cohort comprised patients from the eICU database across various ICU types, including MICU, SICU, CCU, and others, with a minimum of 7.5 hours of continuous vitals monitoring. External validation was performed on both adult (MIMIC-III) and pediatric (SafeICU) datasets, using consistent inclusion and exclusion criteria. Additionally, a prospective validation cohort was constructed using pediatric ICU data from AIIMS, New Delhi, ensuring applicability across age groups and real-world clinical settings.

The constructed cohort from the eICU database consisted of 16,246 ICU-stays, with 12,109 having at least one episode of abnormal SI ([Figure 4.14](#)). The higher prevalence of abnormal SI in this cohort reflects the development cohort criteria, which include arterial blood pressure for assigning ground-truth labels. The median LOS of patients was 3.1 days, while in the abnormal SI, the LOS was significantly higher at 4.28(6.79) days, as shown in [Table 4.10](#).

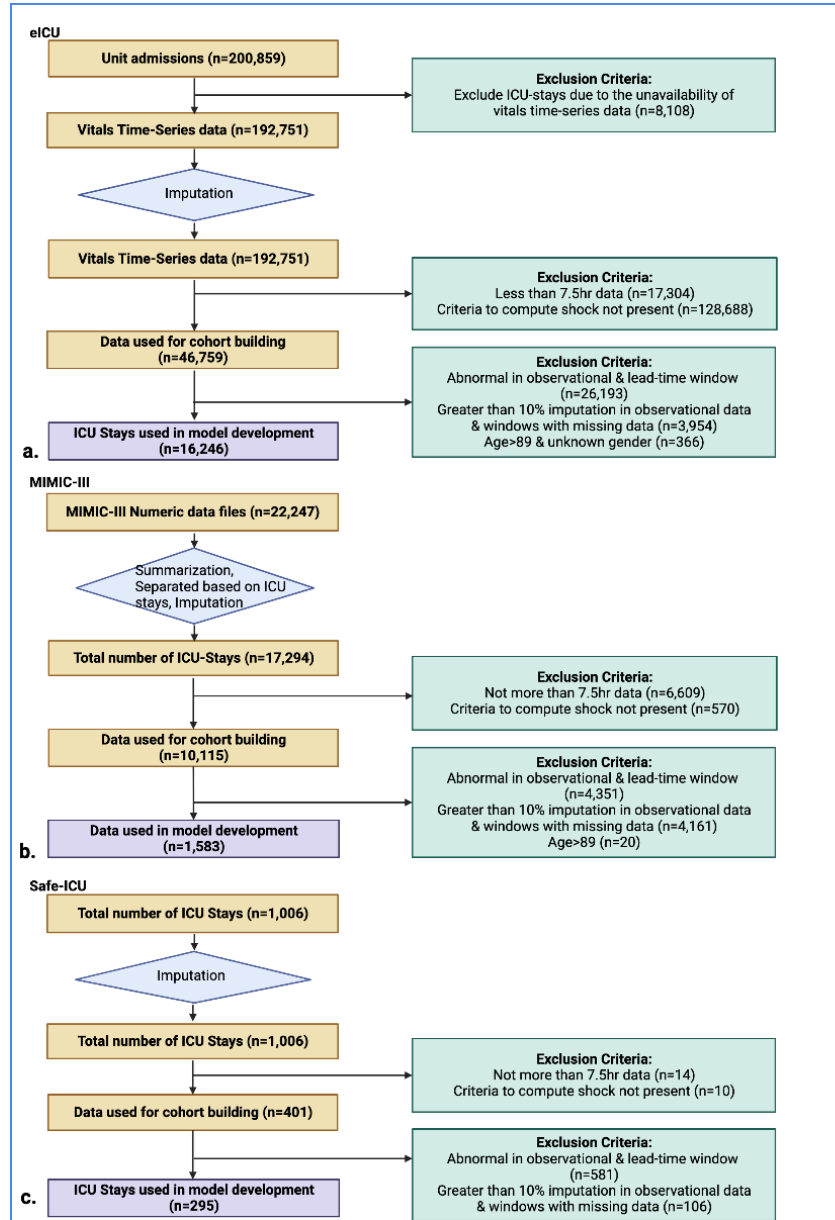


Figure 4.14 Data Preparation and Cohort Formation 4.14a.) In eICU, 200,859 admissions as a starting point, with 8108 ICU-stays removed due to the unavailability of the vitals time-series data, 4.14b.) MIMIC-III and 4.14c.) safeICU data for cohort building

Table 4.10 Characteristics eICU cohort observational window captured 4 hours earlier than the abnormal shock index

| Variables                                | Median (IQR)                 |                                 | Statistical Test       |
|------------------------------------------|------------------------------|---------------------------------|------------------------|
|                                          | Normal Shock Index<br>n=4137 | Abnormal Shock Index<br>n=12109 | Negative log10 P-value |
| Age (years)                              | 70(17)                       | 66 (19)                         | 29.53(w)*              |
| Arterial Systolic Blood Pressure, mm Hg  | 138 (26)                     | 133.5 (26)                      | 32.8(w)*               |
| Arterial Diastolic Blood Pressure, mm Hg | 58(15.5)                     | 59 (16)                         | 4(w)*                  |
| Heart rate, per minute                   | 66 (14)                      | 73 (15)                         | 216.82(w)*             |
| Respiratory rate, per minute             | 16 (4)                       | 18(6)                           | 88.31(w)*              |
| Oxygen Saturation                        | 98(4)                        | 98 (4)                          | 0.22(w)                |
| Gender (Female %)                        | 40%                          | 37%                             | 2.15(c)*               |
| Length of stay (days)                    | 1.18(1.43)                   | 4.28(6.79)                      | ~0(w)*                 |

Values are Median (IQR) unless indicated, \*significance level at negative log10 p-value  $\geq 1.3$ , Wilcoxon (W) rank sum test (non-parametric) or Student's t-test (t) (parametric) were used after testing for the normality assumption. c - Chi-squared test of proportions.

Details of included and excluded cohorts for external validation sets extracted from MIMIC-III and safelCU are shown in [Figure 4.14b, 14c](#). Data characteristics of external validation sets are shown in ([Table 4.10, 4.11](#)) SI status within the first 3 days of ICU stay was associated with both ICU mortality (all p values  $<0.05$ ) and Length of Stay (all p values  $<0.001$ ) in all three datasets ([Table 4.12, 4.13](#)). Therefore, we hypothesize that by generating the early decision-making signal for the incidence of abnormal SI, we can reduce the risk of clinical outcomes, i.e., mortality, and higher length of stay ([Table 4.12, 4.13](#)).

Table 4.11: Characteristics of the MIMIC cohort

| Variables                                   | Median (IQR)                |                                | Statistical Test          |
|---------------------------------------------|-----------------------------|--------------------------------|---------------------------|
|                                             | Normal Shock Index<br>n=482 | Abnormal Shock Index<br>n=1101 | Negative log10<br>P-value |
| Age (years)                                 | 65.76(21.06)                | 65.57(20.04)                   | 0.057(w)*                 |
| Arterial Diastolic Blood Pressure,<br>mm Hg | 60.4(6.4)                   | 59.1(5.64)                     | 0.471(w)*                 |
| Arterial Systolic Blood Pressure,<br>mm Hg  | 129(11.51)                  | 118.45(10.6)                   | 16.03(w)*                 |
| Heart rate, per minute                      | 69(4.43)                    | 86.35(4.71)                    | 84.61(w)*                 |
| Respiratory rate, per minute                | 17.2(2.64)                  | 18(2.9)                        | 4.48(w)*                  |
| Oxygen Saturation                           | 97(1.17)                    | 97(1.35)                       | 0.033(w)*                 |
| Gender (Female %)                           | 43%                         | 40%                            | 0.31(c)*                  |
| Length of stay (days)                       | 2.02(2.82)                  | 3.56(4.89)                     | ~0(w)*                    |

Table 4.12: Characteristics of the safeICU cohort

| Variables                                   | Median (IQR)               |                               | Statistical Test          |
|---------------------------------------------|----------------------------|-------------------------------|---------------------------|
|                                             | Normal Shock Index<br>n=89 | Abnormal Shock Index<br>n=206 | Negative log10<br>P-value |
| Age (months)                                | 18.5(80.09)                | 5.12(15.18)                   | 6.97(w)*                  |
| Arterial Diastolic Blood Pressure,<br>mm Hg | 50.06(3.5)                 | 48.68(0.58)                   | 1.68(w)*                  |
| Arterial Systolic Blood Pressure,<br>mm Hg  | 94.96(5.01)                | 93.06(0.96)                   | 1.01(w)*                  |
| Heart rate, per minute                      | 132.85(8.14)               | 128.5(7.12)                   | 1.53(w)*                  |
| Respiratory rate, per minute                | 31(4.53)                   | 34.98(4.38)                   | 2.33(w)*                  |
| Oxygen Saturation                           | 97(2.32)                   | 95.85(2.43)                   | 4.44(w)*                  |
| Gender (Female %)                           | 37%                        | 38%                           | 0.03(c)                   |
| Length of stay (days)                       | 4.16(5.20)                 | 9.63(11.98)                   | 196.94(w)*                |

Table 4.13: Association of mortality with the first three days of shock index status

| Outcome   | Cohort    | Odds ratio | 95%CI      | P-value  |
|-----------|-----------|------------|------------|----------|
| Mortality | eICU      | 12.01      | 8.96-16.56 | <0.0001  |
|           | MIMIC-III | 1.79       | 1.07-3.03  | 0.029756 |
|           | safeICU   | 3.26       | 1.27-10.26 | 0.021693 |

Table 4.14: Association of LOS with the first three days of shock index status

| Outcome | Cohort    | Regression Coefficient | 95%CI     | P-value               |
|---------|-----------|------------------------|-----------|-----------------------|
| LOS     | eICU      | 4.64                   | 4.41-4.88 | <2.2x10 <sup>-6</sup> |
|         | MIMIC-III | 1.20                   | 0.79-1.61 | <0.0001               |
|         | safeICU   | 4.69                   | 2.19-7.17 | 0.0002                |

#### 4.3.3.2 Model performance on eICU data

Deep learning models such as the CNN-LSTM model achieved a maximum percentage mean (SD) AUROC of 80(2), and a percentage mean (SD) AUPRC of 89(2). Bidirectional-LSTM achieved a maximum percentage mean (SD) AUROC of 83(1), and percentage mean (SD) AUPRC of 90(1). The TS-Feature set included 51 classes of 3970 linear and nonlinear time series features listed in (Table 4.9). Of these, 2,120 were statistically significant for variable importance z-scores in the Boruta feature selection analysis. From the models trained on TS-features, GBM achieved higher performance with a percentage mean (SD) AUROC of 86(1.2) and percentage AUPRC of 93(1.1) (Figure 4.17a, 4.17b, Figure 4.15a, b). It identified 92% of all the shock events with an overall precision of 81% (Figure 4.17c).

The non-invasive variant of this model (Signose non-invasive) obtained a percentage mean (SD) AUROC of 81(1), a percentage mean (SD) AUPRC of 90(1), and the overall recall was 90% with a precision of 79%. Oversampling of the minority class and under-sampling of the majority class to correct for class imbalance did not have a significant effect on the performance of this model (Figure 4.16a, b). On the basis of these characteristics, the GBM model trained upon TS-Features was taken forward as the shock early warning system (*Signose*) for validation across ICUs and validation.

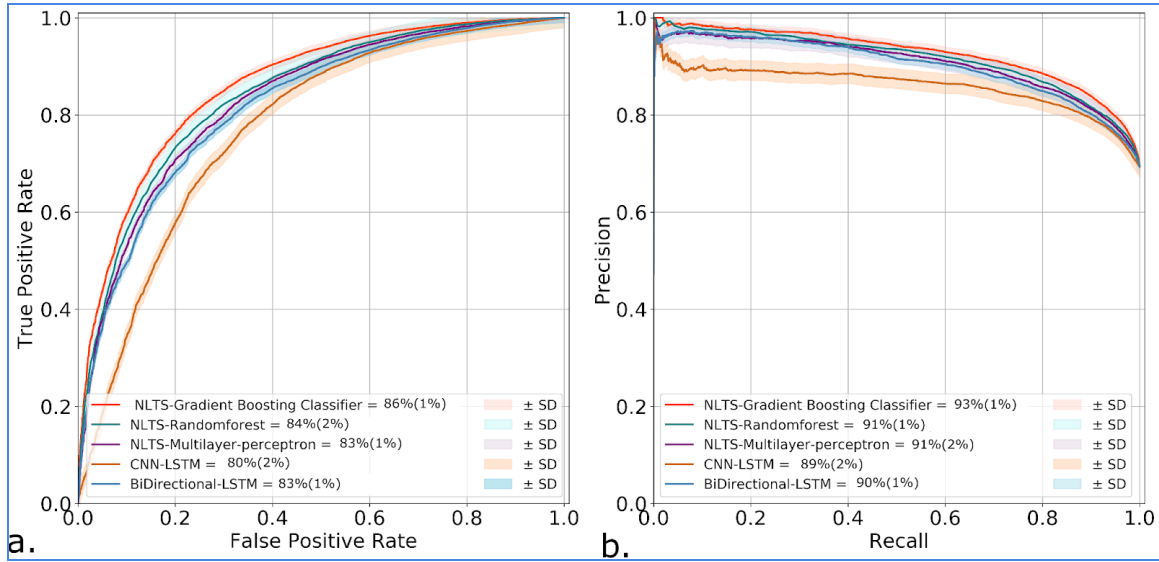


Figure 4.15 Model training on TS-features. 4.15a.) AUROC 4.15b.) AUPRC, different models trained on the eICU hospital splits. Gradient boosting classifier was found to be the best-performing model.

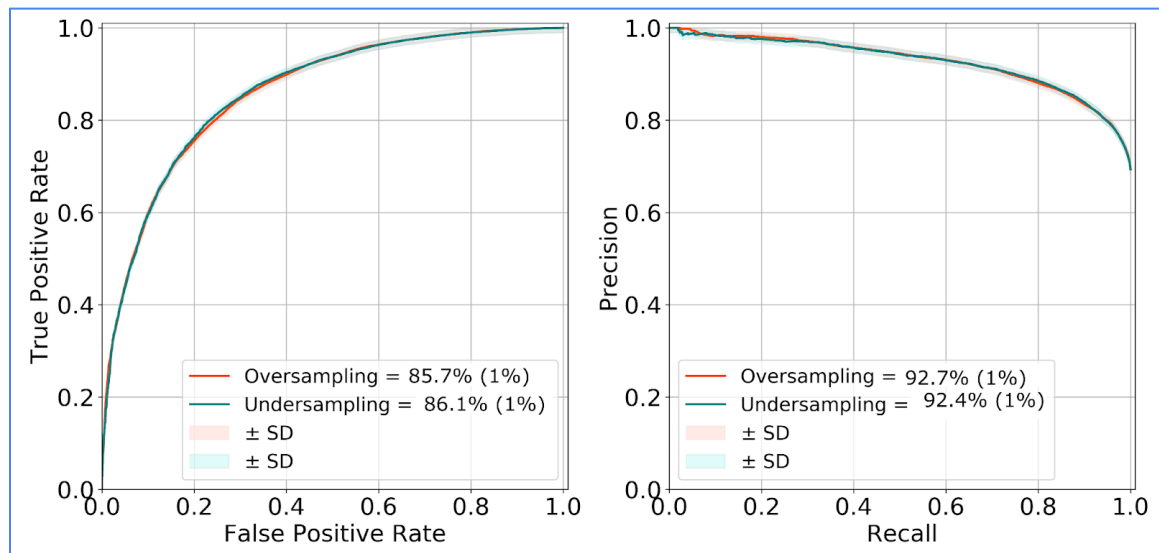


Figure 4.16 SIngnose with oversampling of the minority class, and random undersampling of the majority class 4.16a.) AUROC 4.16b.) AUPRC, SIngnose with oversampling of the minority class, and random undersampling of the majority class.

#### 4.3.3.3 Assessing LLM's predictive Performance on eICU Data

To assess the potential of LLM-based embeddings, LLaMA-7B was applied to structured 5-minute interval time-series vitals data from eICU. Text prompts encoding demographic and physiological information were input into the model to extract contextual embeddings, which were then used in machine learning classifiers. In the eICU-only setup, models trained and tested on eICU embeddings (5-fold cross-validation) showed strong performance across classifiers



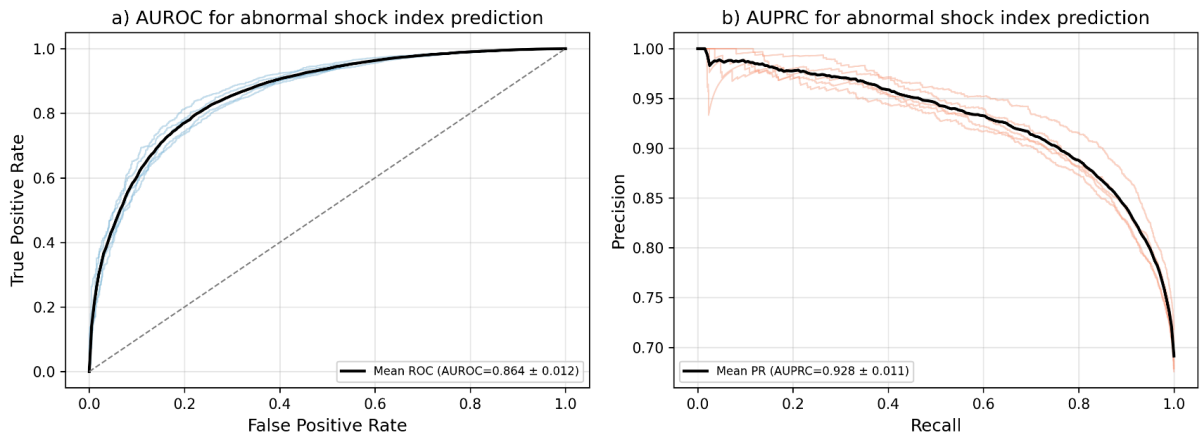
(Table 4.15). This result indicates that LLaMA embeddings capture meaningful signals when training and testing are siloed within the same dataset.

Table 4.15 LLaMA Model Performance on eICU (5-Fold Cross-Validation, Mean  $\pm$  SD)

| Model           | Accuracy                        | Precision                       | Recall                          | F1 Score                        | ROC AUC                         | AUPRC                           |
|-----------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| <b>XGBoost</b>  | 0.79 $\pm$ 0.01                 | 0.79 $\pm$ 0.01                 | 0.79 $\pm$ 0.01                 | 0.77 $\pm$ 0.01                 | 0.85 $\pm$ 0.01                 | 0.92 $\pm$ 0.01                 |
| <b>AdaBoost</b> | 0.78 $\pm$ 0.01                 | 0.77 $\pm$ 0.01                 | 0.78 $\pm$ 0.01                 | 0.77 $\pm$ 0.01                 | 0.82 $\pm$ 0.01                 | 0.90 $\pm$ 0.01                 |
| <b>LightGBM</b> | 0.78 $\pm$ 0.02                 | 0.78 $\pm$ 0.01                 | 0.75 $\pm$ 0.02                 | 0.75 $\pm$ 0.02                 | 0.83 $\pm$ 0.01                 | 0.91 $\pm$ 0.01                 |
| <b>SVM</b>      | <b>0.82<math>\pm</math>0.01</b> | <b>0.82<math>\pm</math>0.01</b> | <b>0.82<math>\pm</math>0.01</b> | <b>0.81<math>\pm</math>0.01</b> | <b>0.88<math>\pm</math>0.01</b> | <b>0.94<math>\pm</math>0.01</b> |

#### 4.3.3.4 Model performance for different lead times, age groups, unit types, and time since admission

As with all the models tested, the performance of *Signose* dropped with the increase in lead time from 0 to 8 hours. However, the model recall was greater than 89% for all lead times up to 8 hours. (Figure 4.17c). The model performance indicators AUPRC and AUROC remained consistent for different age groups, with a slight decline in higher age groups above 60 years. However, AUROC increased for age groups above 60, and the model recall remained above 90%, with precision above 79% for all age groups (Figure 4.17d). All ICU types except for cardiac surgery ICU and neuro ICU were found to have recall rates above 90%. The latter had marginally lower recall and precision, respectively (Figure 4.17e). The overall recall of the model increased with length of stay, with 90% recall after the 13th hour of stay (Figure 4.17f).



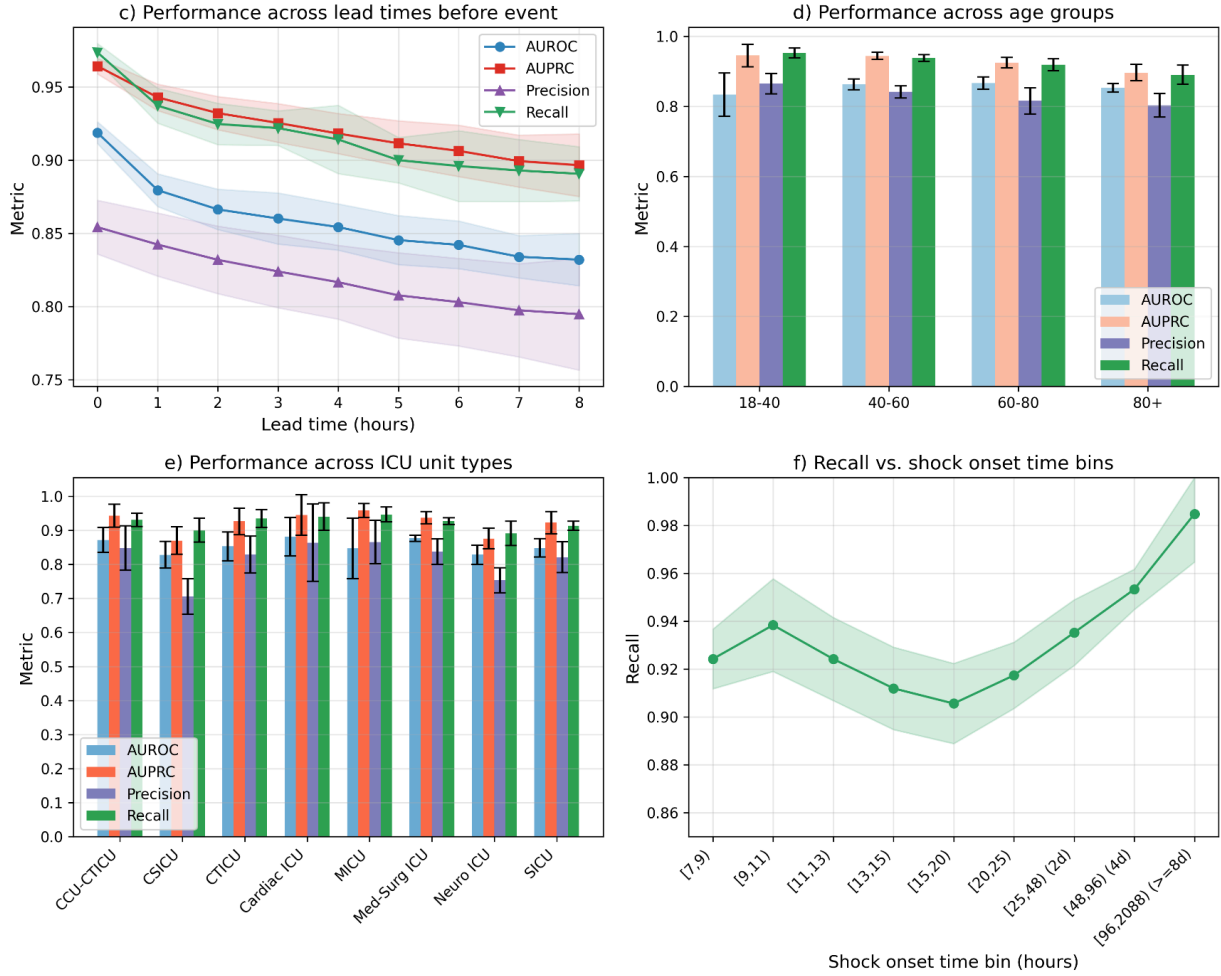


Figure 4.17 Model performance on eICU data. 4.17a) AUROC for the prediction of abnormal shock index in the next 5 min to 8 hours. 4.17b) AUPRC for the prediction of abnormal shock index in the next 5 min to 8 hours. 4.17c) Model performance for different lead times or times before the event. 4.17d) Model performance for different age groups, eICU data contains adult age groups. 4.17e) Model performance in different unit stay types (CCU – coronary care unit, CSICU- cardiac surgery ICU, CTICU- cardiothoracic ICU, MICU- medical ICU, SICU- surgical ICU). 4.17f) recall of the model for each time interval shown on the x-axis for the prediction made during different lengths of stays in the ICU.

#### 4.3.3.5 Performance of external and prospective validation

We evaluated the predictive performance of LLaMA embeddings across two datasets: MIMIC-III and eICU. When both training and testing were performed on the MIMIC dataset without fine-tuning, the SVM model achieved the highest performance (F1 score of 0.79, ROC AUC of 0.84, and AUPRC of 0.90) (Table 4.16). After fine-tuning, SVM continued to outperform others with further improvements, achieving an F1 score of 0.80 and ROC AUC of 0.85. These results highlight the benefit of fine-tuning on the target dataset, especially for SVM (Table 4.16).

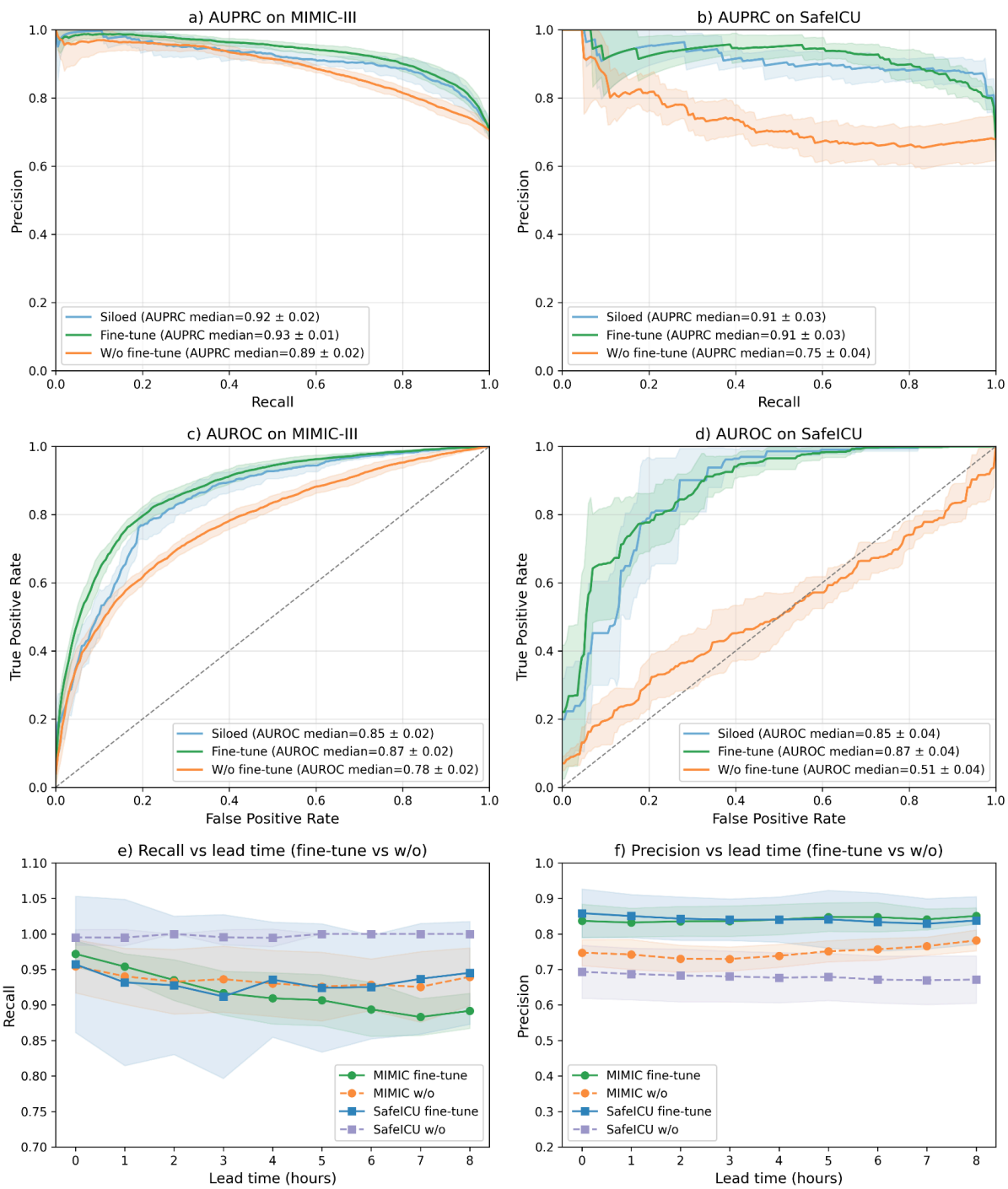
In contrast, models trained on the eICU dataset and tested on MIMIC showed a significant drop in performance, indicating limitations in cross-dataset generalizability. Without fine-tuning, accuracy ranged from 0.59 to 0.65, and ROC AUC remained below 0.60 for most models. Even with fine-tuning, performance gains were minimal, with SVM showing the weakest performance (F1 score = 0.45) and XGBoost achieving the best AUPRC of 0.75 (Table 4.16). These findings underscore the challenges of transferring LLM-based embeddings across clinical datasets and emphasize the importance of dataset-specific adaptation.

Table 4.16 Impact of Fine-Tuning on Model Generalization using LLaMA

| Llama-Train mimic and Test mimic(siloed MIMIC-5 fold(means±sd)) w/o fine tuning |                    |                    |                    |                    |                    |                    |
|---------------------------------------------------------------------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| Model                                                                           | Accuracy           | Precision          | Recall             | F1 Score           | ROC AUC            | AUPRC              |
| <b>XGBoost</b>                                                                  | 0.78±0.02          | 0.78±0.02          | 0.78±0.02          | 0.76±0.02          | 0.81±0.02          | 0.89±0.02          |
| <b>AdaBoost</b>                                                                 | 0.76±0.02          | 0.75±0.02          | 0.76±0.02          | 0.75±0.02          | 0.77±0.02          | 0.87±0.02          |
| <b>LightGBM</b>                                                                 | 0.77±0.02          | 0.77±0.02          | 0.77±0.02          | 0.73±0.03          | 0.81±0.02          | 0.89±0.02          |
| <b>SVM</b>                                                                      | <b>0.80±0.02</b>   | <b>0.80±0.02</b>   | <b>0.80±0.02</b>   | <b>0.79±0.02</b>   | <b>0.84±0.02</b>   | <b>0.90±0.02</b>   |
| Llama-Train mimic and Test mimic(siloed MIMIC-5 fold(means±sd)) fine tuning     |                    |                    |                    |                    |                    |                    |
| <b>XGBoost</b>                                                                  | 0.80 ± 0.02        | 0.79 ± 0.02        | 0.80 ± 0.02        | 0.78 ± 0.03        | 0.84 ± 0.03        | 0.91 ± 0.02        |
| <b>AdaBoost</b>                                                                 | 0.78 ± 0.02        | 0.77 ± 0.02        | 0.78 ± 0.02        | 0.77 ± 0.02        | 0.80 ± 0.02        | 0.89 ± 0.01        |
| <b>LightGBM</b>                                                                 | 0.79 ± 0.03        | 0.78 ± 0.03        | 0.79 ± 0.03        | 0.77 ± 0.04        | 0.83 ± 0.03        | 0.90 ± 0.03        |
| <b>SVM</b>                                                                      | <b>0.81 ± 0.03</b> | <b>0.81 ± 0.03</b> | <b>0.81 ± 0.03</b> | <b>0.80 ± 0.03</b> | <b>0.85 ± 0.02</b> | <b>0.91 ± 0.03</b> |
| Llama-Train eICU and Test MIMIC w/o fine tuning                                 |                    |                    |                    |                    |                    |                    |
| <b>XGBoost</b>                                                                  | 0.59               | 0.6                | 0.59               | 0.59               | 0.54               | 0.76               |
| <b>AdaBoost</b>                                                                 | 0.65               | 0.63               | 0.65               | 0.64               | 0.58               | 0.79               |
| <b>LightGBM</b>                                                                 | 0.65               | 0.63               | 0.65               | 0.64               | 0.56               | 0.76               |
| <b>SVM</b>                                                                      | 0.61               | 0.6                | 0.6                | 0.61               | 0.57               | 0.79               |
| Llama-Train eICU and Test MIMIC fine tuning                                     |                    |                    |                    |                    |                    |                    |
| <b>XGBoost</b>                                                                  | 0.62               | 0.58               | 0.62               | 0.6                | 0.48               | 0.73               |
| <b>AdaBoost</b>                                                                 | 0.56               | 0.6                | 0.56               | 0.58               | 0.54               | 0.75               |
| <b>LightGBM</b>                                                                 | 0.65               | 0.62               | 0.65               | 0.63               | 0.52               | 0.74               |

|     |      |      |      |      |      |      |
|-----|------|------|------|------|------|------|
| SVM | 0.45 | 0.55 | 0.45 | 0.48 | 0.55 | 0.79 |
|-----|------|------|------|------|------|------|

In contrast, *SIgnose*, trained solely on eICU data, generalized better to both MIMIC and SafeICU cohorts without retraining. This underscores the strength of time-series-based modeling in capturing transferable patterns of physiological deterioration and its potential for deployment in varied ICU settings. Model generalization scenario of *SIgnose* with fine-tuning was significantly higher ( $p < 0.0001$ ) than scenarios without (w/o) fine-tuning and a 2% increase over siloed models. This achieved an AUPRC of 93% on MIMIC-III ([Figure 4.18a](#)) and 91% on safeICU ([Figure 4.18b](#)) and an AUROC of 87% both on MIMIC-III and safeICU in five-fold cross-validation. Therefore, this approach demonstrated an unprecedented model performance on external validation sites, as seen in ([Figure 4.18c, 4.18d](#)). Without fine-tuning, the direct application of the *SIgnose* on MIMIC-III and safeICU achieved an AUROC of 78% and 51%, respectively, thus demonstrating the need to contextualize the pre-trained models ([Figure 4.18b, 4.18d](#)). Further, fine-tuning had a higher effect on improving the precision of the models, which improved by up to 15%, in comparison to the recall ([Figure 4.18e, 4.18f](#)). Prospective validation of the fine-tuned model on the New Delhi's safeICU site resulted in an AUROC of 89% and AUPRC of 90% ([Figure 4.18b, 4.18g, 4.18h](#)), thus validating the utility of this model for abnormal SI prediction in this setting.



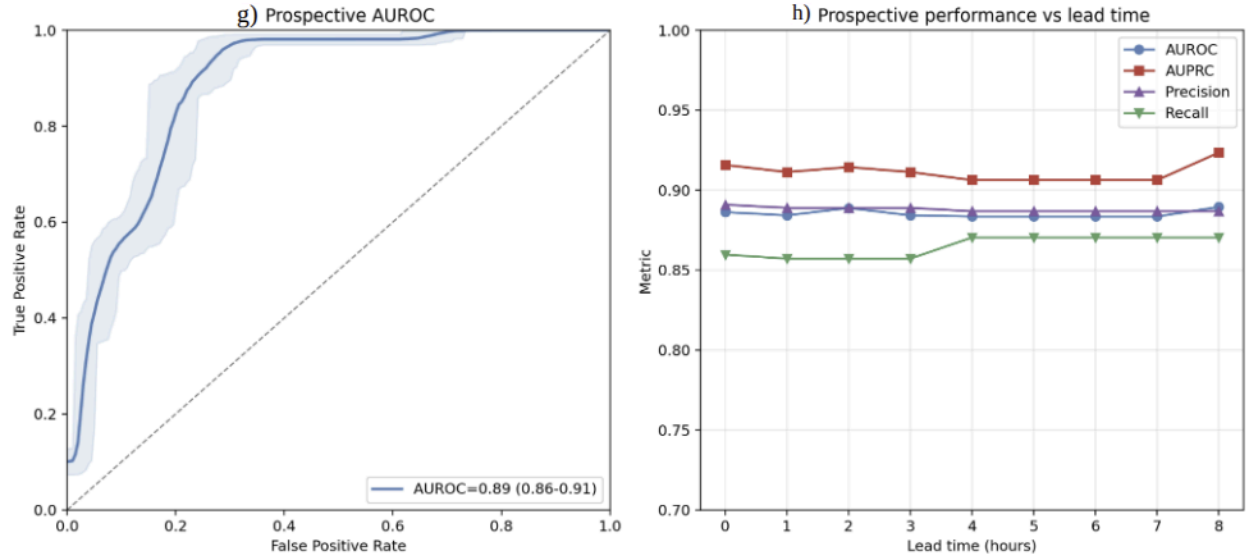


Figure 4.18: *Signose* Performance 4.18a,b) AUPRC on MIMIC-III and safeICU data, respectively, for siloed, fine-tuning, and without fine-tuning (w/o fine-tuning). 4.18c,d) AUROC on MIMIC-III and safeICU, respectively, for siloed, fine-tuning, and without fine-tuning (w/o fine-tuning). 4.18e) Recall of fine-tuned and without fine-tuned models on safeICU and MIMIC-III data. 4.18f) The precision of fine-tuned and non-fine-tuned models on safeICU and MIMIC-III data. 4.18g) AUROC for the prospective validation of safeICU data. 4.18h) Time-point-wise model performance indicators for the prospective validation of *Signose*.

#### 4.3.3.6 Model Interpretability Analysis

The *Signose* model is interpretable and clinically meaningful as it captures non-intuitive complexity features such as correlation dimension, recurrence quantification, and Benford correlations as important predictors of future development of abnormal SI, in addition to intuitive features such as minimum heart rate ([Figure 4.19a](#), [4.19b](#), [4.19c](#), [4.19d](#)). Importantly, even these intuitive features are not to perform the trivial task of phenotyping the patient, but for the non-trivial future prediction of shock, as is commonplace in other complex prediction tasks in areas such as weather forecasting. Other important features include patient age, wavelet features, and complexity features of the vital time series.

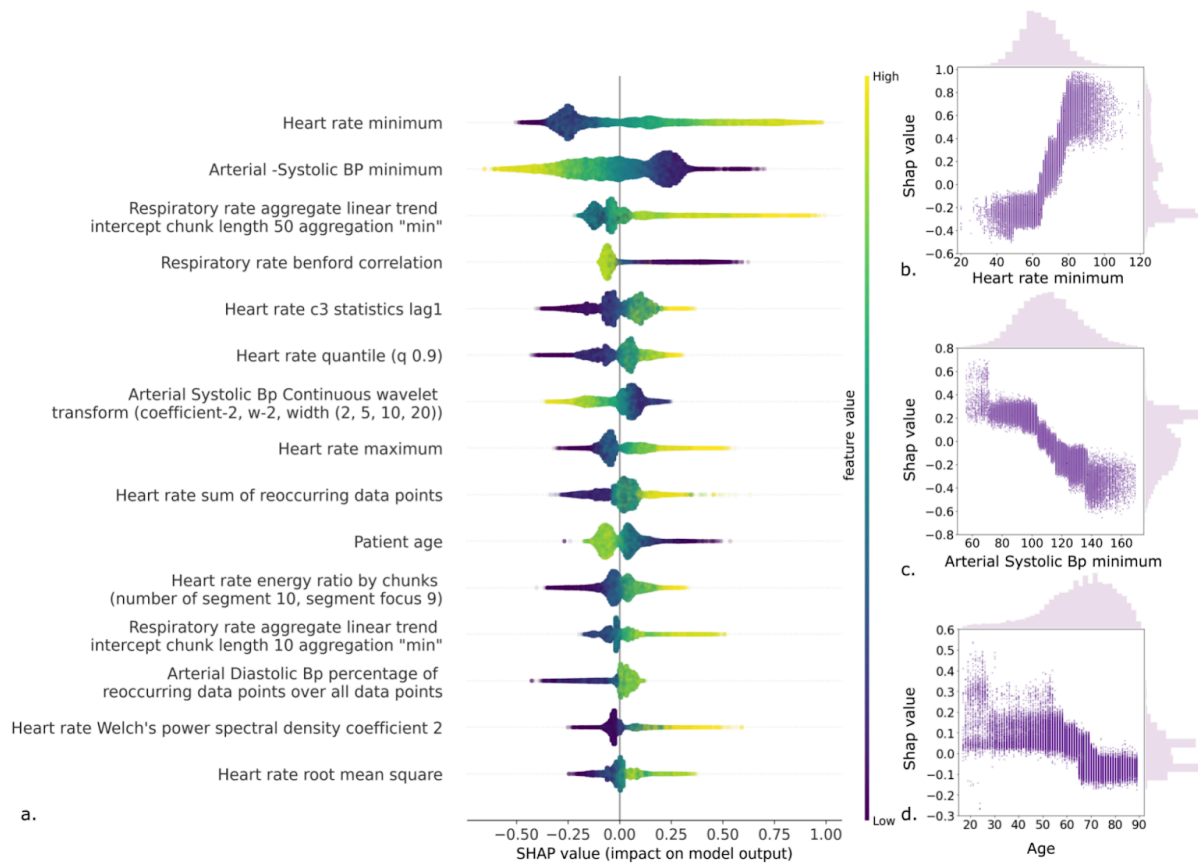


Figure 4.19: **Model Interpretability Analysis.** 4.19a) SHAP values of the top 15 TS-features (descending order) derived from the five-fold test set on eICU data. The Violin plot shows the distribution of SHAP values. The color map represents the feature value from low to high. 4.19b) A scatter plot of Heart rate minimum and corresponding SHAP values shows a positive correlation between the two. 4.19c) A scatter plot of Arterial Systolic Blood pressure minimum and corresponding SHAP values, showing a negative correlation. 4.19d) A scatter plot of patient age and corresponding SHAP values shows a negative correlation. In addition to easily interpretable features such as minimum heart rate, we find complexity features such as correlation dimension, recurrence quantification, and Benford correlations to be important predictors of the future development of abnormal shock index.

#### 4.3.3.7 Model validation across several diagnosis categories

*SIgnose* was evaluated on a range of disease-related outcomes in the MIMIC (adult setting) and safeICU (pediatric setting) datasets (Table 4.17). In the MIMIC dataset, it showed the highest sensitivity (87%) and positive predictive value (PPV) (92%) across pneumonia, atrial fibrillation, and acute kidney injury. For patients with sepsis, *SIgnose* had the highest sensitivity and positive predictive value (PPV) in the paediatric cohort (93% and 100%, respectively), whereas, for patients with hypertension, it showed higher sensitivity and PPV (88% and 94%) than MIMIC dataset hypertension (82% and 90%). This indicates that our model is broadly applicable to critical care illnesses.

Deep learning models such as the CNN-LSTM model achieved a maximum percentage mean (SD) AUROC of 80(2), and a percentage mean (SD) AUPRC of 89(2). Bidirectional-LSTM

achieved a maximum percentage mean (SD) AUROC of 83(1), and percentage mean (SD) AUPRC of 90(1). The TS-Feature set included 51 classes of 3970 linear and nonlinear time series features. Of these, 2,120 were statistically significant for variable importance z-scores in the Boruta feature selection analysis. From the models trained on TS-features, GBM achieved higher performance with a percentage mean (SD)AUROC of 86(1.2) and percentage AUPRC of 93(1.1) ([Figure 4.17a,b](#), [Figure 4.15a, 4.15b](#)). It identified 92% of all the shock events with an overall precision of 81% ([Figure 4.17c](#)).

The non-invasive variant of this model (Signose non-invasive) obtained a percentage mean (SD) AUROC of 81(1), a percentage mean (SD) AUPRC of 90(1), and the overall recall was 90% with a precision of 79%. Oversampling of the minority class and under-sampling of the majority class to correct for class imbalance did not have a significant effect on the performance of this model ([Figure 4.16a,b](#)). On the basis of these characteristics, the GBM model trained upon TS-Features was taken forward as the shock early warning system (*SIgnose*) for validation across ICUs and validation.

Table 4.17 **Model performance across several diagnosis categories**

| Precision | Recall | AUC  | AUPRC | Diagnosis                          | Sample (n patients) | Data    |
|-----------|--------|------|-------|------------------------------------|---------------------|---------|
| 0.9       | 0.82   | 0.82 | 0.88  | Hypertension                       | 993                 | MIMIC   |
| 0.94      | 0.85   | 0.83 | 0.92  | Heart failure                      | 507                 | MIMIC   |
| 0.92      | 0.84   | 0.81 | 0.9   | Atherosclerosis                    | 601                 | MIMIC   |
| 0.93      | 0.85   | 0.83 | 0.91  | Diabetes                           | 462                 | MIMIC   |
| 0.92      | 0.87   | 0.8  | 0.91  | Pneumonia                          | 448                 | MIMIC   |
| 0.91      | 0.82   | 0.84 | 0.89  | Malignant Neoplasm                 | 361                 | MIMIC   |
| 0.92      | 0.86   | 0.83 | 0.91  | Anemia                             | 442                 | MIMIC   |
| 0.91      | 0.84   | 0.8  | 0.9   | Coronary Artery                    | 553                 | MIMIC   |
| 0.93      | 0.87   | 0.83 | 0.92  | Atrial fibrillation                | 467                 | MIMIC   |
| 0.92      | 0.87   | 0.81 | 0.91  | Acute kidney failure               | 417                 | MIMIC   |
| 0.89      | 0.86   | 0.8  | 0.87  | Chronic kidney disease             | 219                 | MIMIC   |
| 0.94      | 0.86   | 0.83 | 0.93  | Respiratory failure                | 383                 | MIMIC   |
| 0.93      | 0.83   | 0.82 | 0.91  | Hemorrhage                         | 272                 | MIMIC   |
| 0.92      | 0.84   | 0.84 | 0.9   | Hypercholesterolemia               | 291                 | MIMIC   |
| 0.93      | 0.83   | 0.82 | 0.91  | Urinary tract infection            | 274                 | MIMIC   |
| 0.94      | 0.86   | 0.83 | 0.92  | Septal defect                      | 45                  | safeICU |
| 0.89      | 0.89   | 0.81 | 0.88  | Pneumonia                          | 38                  | safeICU |
| 0.96      | 0.92   | 0.79 | 0.95  | Acyanotic congenital heart disease | 27                  | safeICU |



| Precision | Recall | AUC  | AUPRC | Diagnosis                         | Sample (n patients) | Data    |
|-----------|--------|------|-------|-----------------------------------|---------------------|---------|
| 0.9       | 0.82   | 0.82 | 0.88  | Hypertension                      | 993                 | MIMIC   |
| 0.89      | 0.85   | 0.59 | 0.89  | Shock                             | 23                  | safeICU |
| 0.94      | 0.88   | 0.84 | 0.93  | Hypertension                      | 22                  | safeICU |
| 0.91      | 0.83   | 0.85 | 0.88  | Congenital heart disease          | 20                  | safeICU |
| 0.82      | 0.88   | 0.64 | 0.82  | Gastroenteritis                   | 21                  | safeICU |
| 1         | 0.93   | 1    | 1     | Sepsis                            | 14                  | safeICU |
| 1         | 0.9    | 0.95 | 0.99  | Lower respiratory tract infection | 12                  | safeICU |
| 1         | 1      | 1    | 1     | Congestive heart failure          | 9                   | safeICU |

#### 4.3.3.8 Assessing LLM's predictive Performance on eICU Data

To assess the potential of LLM-based embeddings, LLaMA-7B was applied to structured 5-minute interval time-series vitals data from eICU. Text prompts encoding demographic and physiological information were input into the model to extract contextual embeddings, which were then used in machine learning classifiers. In the eICU-only setup, models trained and tested on eICU embeddings (5-fold cross-validation) showed strong performance across classifiers ([Table 4.18](#)). This result indicates that LLaMA embeddings capture meaningful signals when training and testing are siloed within the same dataset.

Table 4.18 LLaMA Model Performance on eICU (5-Fold Cross-Validation, Mean  $\pm$  SD)

| Model           | Accuracy                        | Precision                       | Recall                          | F1 Score                        | ROC AUC                         | AUPRC                           |
|-----------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| <b>XGBoost</b>  | 0.79 $\pm$ 0.01                 | 0.79 $\pm$ 0.01                 | 0.79 $\pm$ 0.01                 | 0.77 $\pm$ 0.01                 | 0.85 $\pm$ 0.01                 | 0.92 $\pm$ 0.01                 |
| <b>AdaBoost</b> | 0.78 $\pm$ 0.01                 | 0.77 $\pm$ 0.01                 | 0.78 $\pm$ 0.01                 | 0.77 $\pm$ 0.01                 | 0.82 $\pm$ 0.01                 | 0.90 $\pm$ 0.01                 |
| <b>LightGBM</b> | 0.78 $\pm$ 0.02                 | 0.78 $\pm$ 0.01                 | 0.75 $\pm$ 0.02                 | 0.75 $\pm$ 0.02                 | 0.83 $\pm$ 0.01                 | 0.91 $\pm$ 0.01                 |
| <b>SVM</b>      | <b>0.82<math>\pm</math>0.01</b> | <b>0.82<math>\pm</math>0.01</b> | <b>0.82<math>\pm</math>0.01</b> | <b>0.81<math>\pm</math>0.01</b> | <b>0.88<math>\pm</math>0.01</b> | <b>0.94<math>\pm</math>0.01</b> |

#### 4.3.4 Discussion

The development of machine learning-based early warning systems for ICUs has been an active area of research in recent years([Nemati et al., 2018](#), [Hyland et al., 2020](#), [Mao et al., 2018](#)). However, many of these models have not been successfully implemented in clinical practice due to issues with generalizability, interpretability, and integration into clinical workflow([Kelly et al.,](#)

[2019](#)). In this study, we developed a model called *SIgnose*, which uses routinely collected vital signs data to predict the onset of abnormal SI in critical care patients. The SI is a widely accepted objective indicator of a variety of outcomes in emergency and critical care settings, and it can be easily calculated from routinely monitored vital signs data in real-time, making it potentially applicable to ICUs around the globe. Our results reiterate the value of SI in prognosticating outcomes such as mortality and length-of-stay in all three datasets.

To ensure clinical applicability, we focused on using readily available vital signs data rather than more complex and heterogeneous data, which can hinder the adoption of prediction models, particularly in conditions such as sepsis. The features used in the model were derived from time series data on five physiological vital signs and were selected due to their availability at the bedside in a variety of settings and across different age groups and geographic locations. Minimum pre-processing was carried out in order to preserve physiological signals, and the choice of Kalman filtering is based on its established ability to impute missing values in time-series data([Baddoo et al., 2021](#), [Johnson et al., 2016](#)). The winning model (gradient boosting machine, GBM) was used as our final model for the prediction of abnormal SI due to its higher performance ([Figure 4.17a,b](#), [Figure 4.18a,b](#)). Recent research on tree-based models suggested that tree-based models have different inductive biases that enable them to excel on tabular data([Grinsztajn et al., 2023](#)). The model was able to predict 92% of all abnormal SI events 8 hours ahead of the onset, with similar performance on external validation datasets. An 8-hour prediction was chosen to provide reliable predictions, with a sufficiently long actionable window, thus balancing the need for early intervention and the need for accurate prediction. The model also performed well in a pediatric population, with an average area under the receiver operating characteristic curve (AUC) of 87% and area under the precision-recall curve (AUPRC) of 92% on an external validation dataset in India, surpassing the AUC of 82% reported in a previous study that used 36 variables, including vitals and laboratory investigations, to predict the onset of hemodynamic instability in pediatric patients([Potes et al., 2017](#)). However, it is noteworthy that the study data is different, so it may not be directly comparable. However, we believe that this comparison is valuable because it allows us to assess the generalizability and accuracy of the *SIgnose* model compared to the available model for the pediatric population.

We also provided some computational guidelines for building clinically relevant predictive models, including the use of transfer learning to improve generalizability. In this study, we demonstrated the effectiveness of "feature transfer," in which the features of the model were transferred to different settings and fine-tuned to improve overall AUC and precision while maintaining a high recall rate. We performed external validation of our model on the safeICU pediatric dataset. We chose to evaluate our model on a pediatric cohort for several reasons. First, the physiological characteristics of the SI may differ between pediatric and adult patients, due to differences in body size, composition, and function. Therefore, it is clinically important to evaluate our model on a pediatric cohort to assess its generalizability and performance in this population. Secondly, the strength of the modeling approach was assessed through a feature

transfer from adult to pediatric patients, as their physiological features may share some similarities. Hence, the evaluation of models across geographies and age groups was preferred. Our model initially showed an AUROC of 51% and AUPRC of 75% on the SAFE-ICU dataset, outperforming random chance, particularly effective for imbalanced datasets, evidenced by strong sensitivity, precision (PPV), and F1 score metrics. However, when we retrained the model on the safeICU data, we observed a significant improvement in performance, with an AUROC of 88.5%. This suggests that our model can adapt to new data and generalize to different patient populations and care settings. Additionally, using SHAP analysis, we find easily interpretable features such as minimum heart rate, complexity features such as correlation dimension, recurrence quantification, and Benford correlations to be important predictors of the future development of abnormal SI.

The model was also found to have broad applicability across different critical care illnesses, with particularly high sensitivity and positive predictive value for sepsis in the pediatric population and for pneumonia, atrial fibrillation, and acute kidney injury in the MIMIC-III dataset. For all patients who may go into shock, the physiological hemodynamic instability is a final common pathway in many illnesses. We intended the model to be applicable to the early detection of this cascade that leads to multiple organ failure. We have profiled our model over a range of the most frequent diagnoses and show that the model performance is similar across the conditions ([Table 4.17](#)). A key strength of SIGnose is its performance despite the dynamic nature of physiological regulation of SI. While this confirms the presence of early warning signals before the initiation of therapeutic interventions or overt physiological decompensation, it also highlights the scope for future improvement of the model, where real-time interventions may be included for improved predictions.

In parallel with SIGnose, this work also explored the utility of Llama-based embeddings with classical machine learning classifiers for early prediction of abnormal SI across clinical datasets. While promising results were obtained within individual datasets (e.g., ROC AUC of 0.85 on MIMIC after fine-tuning), the embeddings showed poor generalizability when transferred across datasets. Performance significantly deteriorated when models trained on eICU were applied to MIMIC, even with fine-tuning, with F1 scores dropping below 0.50. This suggests that LLM-derived embeddings may encode dataset-specific biases and lack the robustness required for cross-setting clinical applications. In contrast, the SIGnose model, developed using time-series vital signs and minimal preprocessing, demonstrated strong generalization.

There are some limitations to this study. There are concerns, such as alarm fatigue, which can have an adverse impact on patient care as it can cause ICU staff to delay their response to alarms, ultimately compromising patient safety([Alsuyayfi & Alanazi, 2022](#)). The predictions are being made at 8-hour intervals; hence, our models are least likely to contribute to alarm fatigue. However, it is important to note that this strategy alone may not be sufficient to completely eliminate alarm fatigue. It is also important to consider other factors that may contribute to alarm fatigue, such as the design of the alarm system and the types of alarms that are used.

Another limitation of the *SIGnose* model is that it is not applicable to all types of shock (e.g., shock induced by symptomatic bradycardia). However, the goal of this study was the early detection of the most frequent type of shock encountered in the ICU settings. Future studies are planned to explore the use of different thresholds for defining abnormal SI and their association with a wider variety of severity of shock spectrum.

Finally, the use of a shock index, the SI, as an overall indicator of outcomes may not capture the full complexity of the critical care setting. Despite these limitations, we believe that the model presents a promising approach for predicting abnormal SI in critical care and emergency settings, and we have made the model openly available to facilitate its use and further development.

# Chapter-5

## Drug Safety and Interaction Analysis in the ICU

### 5.1 Potential Drug-Drug Interactions in the Pediatric Intensive Care Unit

#### 5.1.1 Introduction

Potential drug–drug interactions (pDDI) happen when the toxicity or efficacy of one drug is changed by another([Hassanzad et al., 2018](#)). Despite efforts to minimize the occurrence of pDDIs through the use of known pharmacological characteristics and local or national interventions, real–world data consistently demonstrate a significant prevalence of pDDIs worldwide. This suggests that pDDIs, although preventable, continue to be a persistent issue([Mousavi et al., 2017](#)). pDDIs are more common than expected in critically ill patients due to the off–label and unlicensed drug administration, thus requiring rigorous investigation([Morales et al., 2018](#), [Hassanzad et al., 2020](#)). Patients admitted to ICUs often have multi–morbid pathology and receive multiple medications, thus raising the probability of pDDIs; this places pediatric patients at an increased risk of DDIs due to the potential for pharmacokinetic and pharmacodynamic interactions. Thus, DDIs can have profound implications on patient safety and treatment outcomes, making their identification and management crucial.

Previous studies have provided insights into the prevalence and clinical implications of pDDIs in pediatric healthcare settings. A study conducted in a tertiary care children's hospital in India found that approximately 63·11% of pediatric ICU patients were exposed to at least one potential DDI, with an average of 2·27 interactions per prescription([Rao et al., 2019](#)). This high prevalence highlights the vulnerability of critically ill children to the risks associated with drug interactions. A multicentric study conducted across 43 freestanding general children's hospitals reported that approximately 75% of pediatric patients experienced at least one potential DDI, with 69% categorized as severe interactions([Dai et al., 2016](#)). Despite efforts to raise awareness about DDIs, the knowledge and understanding among healthcare professionals and patients may still be limited. Various research studies have investigated the level of knowledge among medical residents regarding DDIs. A study conducted among physicians in China revealed that a mere 15·1% of healthcare professionals demonstrated awareness of all clinically significant DDIs present in their patients' prescriptions. This finding highlights the potential gaps in knowledge and emphasizes the need for further education and awareness programs to improve understanding and mitigate the risks associated with DDIs([Yuan et al., 2021](#)).

Not all pDDIs are harmful. However, some of them could result in toxicity, ADRs([Fernandes et al., 2019](#), [Jiang et al., 2022](#)), and/ or reduced treatment efficiency. These could lead to clinical

manifestations, including unmet treatment goals, worsening of the patient's condition, or even death. Additionally, they significantly influence healthcare service consumption, particularly on hospitalization rates, and have significant economic repercussions that raise expenses for individuals and healthcare systems([Morales et al., 2018](#), [Kardas et al., 2021](#), [Smithburger et al., 2010](#)).

Although insufficient data are available for pDDIs in pediatric ICUs in India, studies in ICUs and Emergency Departments published from Mexico, Pakistan, and Ethiopia have shown the prevalence and associations of pDDIs([Hassanzad et al., 2018](#), [Morales et al., 2018](#), [Lima et al., 2020](#), [Hassanzad et al., 2020](#), [Ismail et al., 2017](#), [Getachew et al., 2016](#), [silva et al., 2013](#)) LOS was among the most significant, and a regression analysis([Lima et al., 2020](#)) showed an increase in the number of days in the PICU LOS in children with major pDDIs. Other reported associations of pDDIs in the ICUs included the number of drugs prescribed, female gender, age, and specific illness categories such as neoplasms([Krecak et al., 2023](#)). Identifying pDDIs in real-time decreases the likelihood of unexpected drug interactions and reduces the cost of drug production.

The primary goal of our study was to recognize and avoid significant pDDIs in critically ill children admitted to the PICU at AIIMS, Delhi. To achieve this, we conducted a retrospective analysis to study the frequency and outcomes of pDDIs. By understanding the prevalence and consequences of pDDIs in the vulnerable patient population, our research aims to contribute to developing tools and strategies that can help medical practitioners prevent adverse health and economic outcomes associated with pDDIs. Further, we have developed a software application named “**ContraIndicator**” to keep real-time tracking of pDDIs based on the interactions of patients admitted to ICUs. This software is designed to assist clinicians in monitoring and managing pDDIs more effectively. By providing real-time alerts about potential drug interactions, our application aims to improve the efficacy of medication regimens in critically ill children. Additionally, “ContraIndicator” allows clinicians to upload prescriptions, automatically identifying and displaying all potential drug-drug interactions among the prescribed medications, further enhancing clinical decision-making.

## 5.1.2 Methods

### 5.1.2.1 Study Setting

This retrospective cohort study was conducted at the PICU of the AIIMS, a tertiary care hospital in New Delhi, India. The study utilized routine data generated as part of standard clinical care. Data collection spanned a period of five years, from 2016 to 2021. Ethical approval for the study was obtained from the Ethics Committee under the reference number IEC/NP-211/08-05-2015. The study did not involve direct interaction with the subjects; instead, it analyzed retrospective records of the ICU data that had been routinely generated.

### 5.1.2.2 Objectives and outcome measures

The objective and outcome measures of this retrospective analysis were to identify pDDIs that are clinically relevant in pediatric patients receiving medical attention in the pediatric intensive care unit. The study aimed to determine the frequency and severity of pDDIs. It also looked into the association between the severity of pDDIs and age and the number of medicines prescribed, and identified the predictors for PICU length of stay.

### 5.1.2.3 Variables collected, Procedure, and other measures

#### SafelICU Framework

The Pediatric ICU of the tertiary care hospital AIIMS in New Delhi, India, provided all the data used in this study. In the pediatric ICU, there are eight beds in total, including beds for infants. Servers were configured to gather periodic real-time data([Vats et al., 2022](#)). The prescriptions (treatment charts) for the patients admitted to the PICU are typed on the Desktop using MS Word and saved on the same. The daily backup of this data was planned for a specific time. To enable text mining for information such as gender, age, date of prescription, drugs, and diagnosis, text files were processed using the Docx-2-txt Python package. Lab results, vital data, and thermal imaging were also warehoused.

#### Named Entity Recognition

Named Entity Recognition was performed to distill out the drug entities using MED7 ([Kormilitzin et al., 2021](#)) from the spaCy library on the treatment charts of patients admitted to the PICU. MED7 architecture has been pre-trained on the MIMIC-III([Johnson et al., 2016](#)) cohort via self-supervised learning. The model was pre-trained for predicting the next word by implementing spaCy's cloze-style word reconstruction, which was later fine-tuned on the named-entity recognition task. Metadata such as the patient's ID, gender, age (in months), diagnosis, and the length of PICU stay (in days) were combined with the drugs extracted from individual clinical notes.

#### DDInter Database

The DDInter([Xiong et al., 2022](#)), a curated database with comprehensive data, helpful pharmacological guidance, a user-friendly interface, and effective visualization for the scientific community, was used to characterize each pDDI that had been discovered. Approximately 0.24M DDI associations currently link 1833 US FDA-approved drugs in DDInter. The basic chemical and pharmacological information for each medicine and its interaction network are documented. In addition to basic research and browsing, the prescription checking feature is intended to assist professionals in determining if drug combinations can be used safely. The registered chemists examined these interactions and divided them into three categories based on severity: Minor, Moderate, and Major([Xiong et al., 2022](#)). Our study also compares DDIs for

various drug pairs, as determined by different databases, including Lexicomp, Medscape, Drugs.com, and Epocrates (**Table 5.1**).

**Table 5.1 Comparison of DDI pairs across different databases**

| <b>Pair name</b>          | <b>DDinter Database</b> | <b>Lexicomp Database</b>                                   | <b>Medscape Database</b>   | <b>Drugs.com Database</b> | <b>Epocrates Database</b>  |
|---------------------------|-------------------------|------------------------------------------------------------|----------------------------|---------------------------|----------------------------|
| Fluconazole + Fentanyl    | Major                   | Consider therapy modification                              | Serious - Use Alternative  | Major                     | Avoid/Use Alternative      |
| Fluconazole+Midazolam     | Major                   | Consider therapy modification                              | Monitor Closely            | Major                     | Avoid/Use Alternative      |
| Dexamethasone +Fentanyl   | Major                   | No interactions of Risk Level A or greater were identified | Monitor Closely            | Major                     | Caution Advised            |
| Piperacillin+Vancomycin   | Major                   | Monitor therapy                                            | Monitor Closely            | Major                     | Monitor/Modify Tx          |
| Fentanyl+Ondansetron      | Major                   | Monitor therapy                                            | No interactions            | Major                     | Avoid/Use Alternative      |
| Methotrexate+Piperacillin | Major                   | Monitor therapy                                            | No interactions            | Major                     | Avoid/Use Alternative      |
| Dobutamine+Vancomycin     | Moderate                | No interactions of Risk Level A or greater were identified | No interactions            | Moderate                  | No interactions            |
| Dopamine+Vancomycin       | Moderate                | Monitor therapy                                            | No interactions            | Moderate                  | No interactions            |
| Metoprolol+Salbutamol     | Moderate                | Monitor therapy                                            | Salbutamol Not available   | Salbutamol Not available  | Salbutamol Not available   |
| Amikacin+Vancomycin       | Moderate                | Monitor therapy                                            | Monitor Closely            | Moderate                  | Avoid/Use Alternative      |
| Amikacin+Cefoperazone     | Moderate                | Monitor therapy                                            | Cefoperazone Not available | Moderate                  | Cefoperazone Not available |
| Dexamethasone +Midazolam  | Minor                   | No interactions of Risk Level A or greater were            | Monitor Closely            | Minor                     | Caution Advised            |



|                           |       |                                                            |                           |          |                   |
|---------------------------|-------|------------------------------------------------------------|---------------------------|----------|-------------------|
|                           |       | identified                                                 |                           |          |                   |
| Ceftriaxone+Heparin       | Minor | No interactions of Risk Level A or greater were identified | Serious - Use Alternative | Minor    | Monitor/Modify Tx |
| Fluconazole+Metronidazole | Minor | No action needed                                           | No interactions           | Moderate | No interactions   |
| Dexamethasone +Isoniazid  | Minor | Monitor therapy                                            | Monitor Closely           | Minor    | Caution Advised   |
| Aminophylline+Midazolam   | Minor | Consider therapy modification                              | No interactions           | Minor    | Caution Advised   |

#### 5.1.2.4 Population, Sample, and Patient Characteristics

The study did not involve recruiting patients. The retrospective data concurs with the background rate of pediatric ICU admission for both genders. The study population includes 1069 children (prescriptions=10486) admitted to the PICU. The study sample (n=862), consisting of children between the ages of 0 and 18 years who were admitted to the PICU, stayed more than 24 hours and received at least two medications, was included in the cohort. This cohort was associated with a total of 7,355 prescriptions. The age groups are delineated as follows: newborns to infants (<1 year of age), toddlers to preschoolers ( $\geq 1$  year and <5 years), and school-age children to adolescents ( $\geq 5$  years of age). The same patient's readmission after 15 days was regarded as a new case. Of the study population, 37 children were readmitted and included in the study (n=899). Thus, the cohort consists of 8,010 prescriptions. Prescriptions of the children admitted to the PICU were analyzed in this study. The collected data included age in months, gender, date of prescriptions, diagnosis, and the medications prescribed. Topical medications were not considered in the analysis. The sample size and statistical power needed to support the primary research hypotheses were calculated using an a priori approach, guaranteeing the study's sufficient power to identify meaningful connections. The study evaluated important predictors and found that the statistical power for significant pDDI varied from 0.14 to 1.00 for factors like gender, age group, and number of medicines. Power ranged from 0.05 to 1.00 for moderate pDDI but was constantly high (0.99–1.00) for all pDDI. Depending on the predictor, power varied from 0.05 to 1.00 for length of stay. The procedure followed the EQUATOR Network's STROBE and TRIPOD criteria.

#### 5.1.2.5 Statistical Analysis

The linearity of continuous variables was checked using the Box–Tidwell test, and appropriate transformations were applied to meet the regression models' assumptions. The VIF was

calculated to assess the presence of multicollinearity among the predictors. All statistical analyses were performed using R, and statistical significance was considered if  $p < 0.05$ .

Following this, the study used the incidence of pDDI severity (major, moderate, and minor) (**Table 5.1**) to conduct a descriptive analysis of cases, accounting for variables like gender, age group, and the number of medications administered (two to four; five or more). Univariate and multivariate logistic regressions were used to determine if the severity of major pDDI, moderate pDDI, drugs prescribed, and patient age group were associated with the LOS in the PICU. Additionally, multivariate Poisson regression was employed to assess the predictors of LOS. The logistic regression analyses calculated Odds Ratios (OR) with 95% Confidence Intervals (CI), while the linear regression analysis provided Incidence Rate Ratios (IRR) with 95% CI (**Table 5.2**).

#### **5.1.2.6 ContraIndicator**

We have developed a "ContraIndicator" dashboard that effectively showcases empirical findings and raises awareness about pDDIs. This versatile tool can be accessed within an organization through a local server, enabling usage across multiple computers. Furthermore, it can be made publicly available with secure access by publishing it online, expanding its reach and impact. Additionally, "ContraIndicator" allows clinicians to upload prescriptions to the dashboard, where it automatically identifies and displays all drug interactions among the prescribed medications, enhancing its utility for real-time clinical decision-making. To comprehensively understand the "ContraIndicator" dashboard, we have included a 'proof-of-concept' using dummy data ('sample.csv') as an example. This demonstration can be accessed through our GitHub repository (<https://github.com/tavlab-iiitd/ContraIndicator>). A thorough documentation of the source code for the "ContraIndicator" application is also available in the same GitHub repository. It is important to note that our application does not retain any user-uploaded data on either the server or client sides. "ContraIndicator" solely processes the data to generate updated visualizations without storing user information.

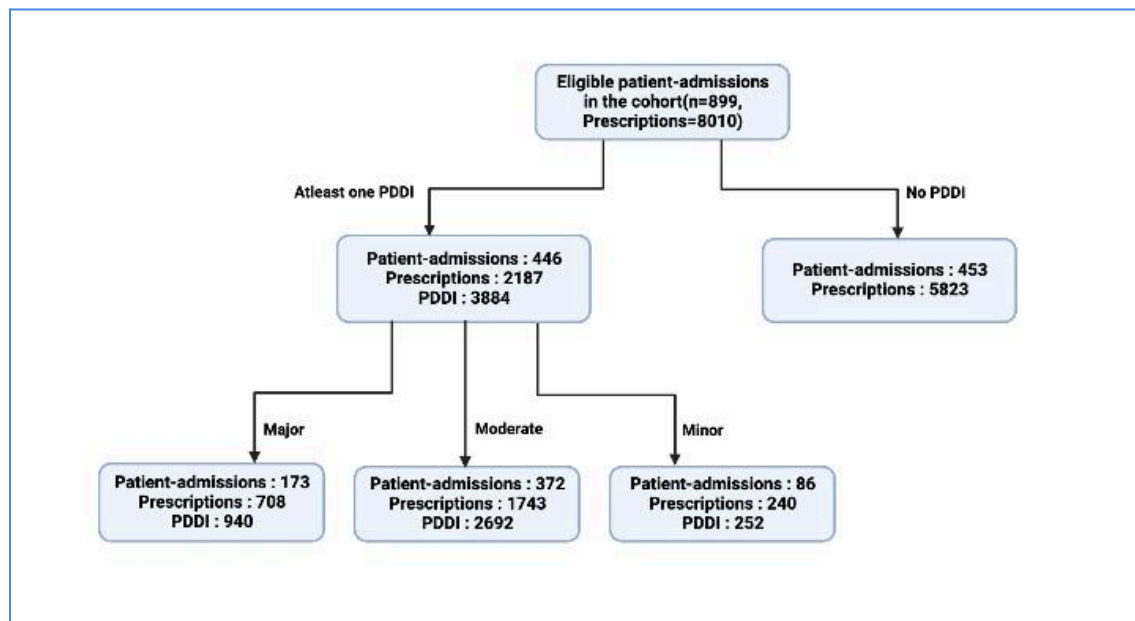
### **5.1.3 Results**

#### **5.1.3.1 Baseline characteristics**

The age of children ( $n=862$ ) admitted to the PICU ranged from one day of life to 18 years old [Median (IQR)=2.75 (0.41–8)]. A total of 8,010 drug prescriptions for 899 patient admissions were examined during their length of stay. The number of prescriptions per child ranged from one to 203 [Median (IQR)=5 (2–10)], and the number of drugs per prescription ranged from two to 13 (Median (IQR)=4 (3–5)).

Shock 10.6% (95), Pneumonia 10.6% (95), Sepsis 8.4% (75), Febrile illness 6.8% (61), and Acute kidney injury 5% (45) were the top five diagnoses among pediatric patients admitted to critical care.

The analysis of 899 patient admissions revealed that 246 unique pDDIs occurred 3,884 times. Nearly 50% (n=446) of those who were admitted to the ICU encountered at least one major, moderate, or minor interaction ([Figure 5.1](#)); at least one major pDDI was identified in 38.8% (173) of these cases ([Table 5.2](#)). The power analysis confirmed that the sample size was sufficient to detect significant associations for most key predictors in the study. For instance, the study was well-powered to detect differences in major and moderate pDDI outcomes based on age group and the number of drugs prescribed. The statistical power differed for various outcomes and predictors. In terms of significant potential pDDI, statistical power was reported as follows: 0.35 for gender differences (male vs. female), 0.84 when comparing children under 1 year old with those aged 1 to less than 5 years, and 0.14 when comparing children 5 years and older with those under 1 year old. Compared to two to four drugs, the effect of consuming five or more medicines reached a power of 1.00.



**Figure 5.1 Flow chart of potential drug–drug interactions frequency.** A summary of the frequency and classification of pDDIs among the 899 qualified patients and 8010 prescriptions. The graph shows whether pDDIs occur or not, the total amount of occurrences of pDDI (One prescription may show more than one pDDI)

Table 5.2 Patient characteristics based on exposure to pDDI

| Variables                        | Patient: n(%) | All types of interactions |      | Only major interactions |      | Only moderate interactions |      |
|----------------------------------|---------------|---------------------------|------|-------------------------|------|----------------------------|------|
|                                  |               | n                         | %    | n                       | %    | n                          | %    |
| Gender                           |               |                           |      |                         |      |                            |      |
| Male                             | 559 (62.18%)  | 278                       | 30.9 | 113                     | 12.6 | 228                        | 25.4 |
| Female                           | 340 (37.81%)  | 168                       | 18.7 | 60                      | 6.7  | 144                        | 16.0 |
| Total                            | 899 (100%)    | 446                       | 49.6 | 173                     | 19.3 | 372                        | 41.4 |
| Age group (year)                 |               |                           |      |                         |      |                            |      |
| <1                               | 349 (38.82%)  | 174                       | 19.4 | 70                      | 8.0  | 141                        | 15.7 |
| >=1 & <5                         | 228 (21.35%)  | 105                       | 11.6 | 40                      | 4.3  | 90                         | 10.0 |
| >=5                              | 322 (39.82%)  | 167                       | 18.6 | 63                      | 7.0  | 141                        | 15.7 |
| Total                            | 899 (100%)    | 446                       | 49.6 | 173                     | 19.3 | 372                        | 41.4 |
| Total number of drugs prescribed |               |                           |      |                         |      |                            |      |
| 2–4                              | 447 (49.72%)  | 120                       | 13.3 | 41                      | 4.6  | 81                         | 9.0  |
| 5 or more                        | 452 (50.27%)  | 326                       | 36.3 | 132                     | 14.7 | 291                        | 32.4 |
| Total                            | 899 (100%)    | 446                       | 49.6 | 173                     | 19.3 | 372                        | 41.4 |

When it involved Moderate pDDIs, power measurements were as follows: 0.05 for gender, 0.98 for the comparison with younger age groups, and a perfect value of 1.00 for the comparison with older age groups and number of medicines consumed. When all pDDIs were considered, power measures were either perfect or almost perfect for every predictor, with values of 1.00 for the number of medicines and both age groups and gender at 0.99.

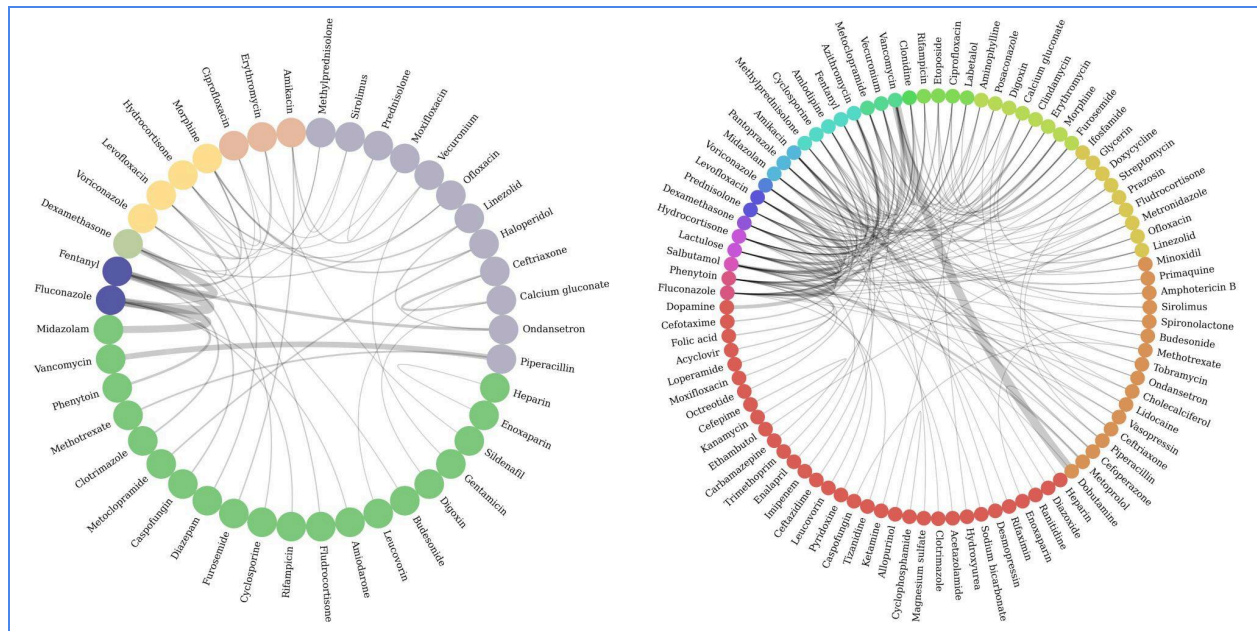
The Length of Stay power analysis showed higher values for age groups and the number of medicines (0.98 to 1.00) and lower measures for gender (0.05). High statistical power (1.00 and 0.99, respectively) was also demonstrated by the Major and Moderate pDDI outcomes about Length of Stay.

### 5.1.3.2 Drug pairs, frequency based on major, moderate, and minor pDDI

The most commonly prescribed medications during the PICU stay were Fentanyl 15.96% (5066), Midazolam 13.82% (4387), Vancomycin 9.88% (3137), Meropenem 6.51% (2066). Among these, the most common drugs involved in major pDDI were fentanyl 13.35% (519), midazolam 10.55% (410), vancomycin 7.74% (301), fluconazole 5.84% (227), and meropenem 4.55% (177). The most frequent interacting drug pairs were Fentanyl + Fluconazole, Fluconazole + Midazolam, Dexamethasone + Fentanyl, Piperacillin + Vancomycin, and Fentanyl + Ondansetron in major pDDI ([Figure 5.2](#)). [Table 5.3](#) shows some of the most frequent interacting drug pairs concerning major and moderate interactions.

Table 5.3 The frequency of drug interactions in PICU

| Severity        | Interacting drug pairs  | Frequency | N (Patients) |
|-----------------|-------------------------|-----------|--------------|
| <b>Major</b>    | Fentanyl+Fluconazole    | 175       | 49           |
|                 | Fluconazole+Midazolam   | 139       | 42           |
|                 | Dexamethasone+Fentanyl  | 131       | 54           |
|                 | Piperacillin+Vancomycin | 107       | 34           |
|                 | Fentanyl+Ondansetron    | 49        | 3            |
| <b>Moderate</b> | Dobutamine+Vancomycin   | 469       | 151          |
|                 | Dopamine+Vancomycin     | 170       | 87           |
|                 | Azithromycin+Salbutamol | 164       | 6            |
|                 | Metoprolol+Salbutamol   | 85        | 1            |
|                 | Amikacin+Vancomycin     | 78        | 35           |



**Figure 5.2 Drug Pairs, Frequency based on Major pDDI (Left) and Moderate pDDI (Right):** The linkages between various medications involved in potential interactions are shown in this Circos figure. Each drug is a data point, and the relationships between drugs are shown as ribbons linking the relevant drug pairs. The frequency of drug interactions is represented by color coding and variable line thickness. While the narrower ribbons depict the low-frequency interacting drug pairs, the broader ribbons depict the higher-frequency interacting drug pairs.

### 5.1.3.3 Predictors of pDDI

The results of the Box-Tidwell linearity test confirmed that the log-transformed continuous predictors satisfied the linearity assumption, with non-significant p-values across most models. Specifically, the interaction terms for log-transformed age had p-values of approximately 0.97 for major PDDIs, 0.38 for moderate PDDIs, and 0.62 for all PDDIs. Similarly, for log-transformed total medications, the p-values were 0.31, 0.59, and 0.95, respectively. A notable exception was found in the major PDDI model for the log-transformed length of stay interaction, with a significant p-value of 0.004, suggesting a relationship. However, this interaction remained non-significant for moderate and all PDDIs ( $p=0.66$  and  $p=0.90$ , respectively). Due to the significance of the length of stay interaction in the major PDDI model, it was excluded from the final analysis to maintain its integrity and avoid potential bias.

The VIF analysis indicated no significant multicollinearity among the predictors used in the logistic regression models, with all VIF values below the commonly accepted threshold of 5. Specifically, for major PDDIs, the GVIF values were 1.05 for age group, 1.05 for gender, and 1.05 for the number of drugs prescribed. For moderate PDDIs, the GVIF values were 1.04 for the age group, 1.00 for gender, and 1.04 for the number of prescribed drugs. For all PDDIs, the GVIF values were 1.05 for age group, 1.00 for gender, and 1.05 for the number of prescribed drugs. These findings confirm the absence of multicollinearity in the models.

A comprehensive analysis was conducted using univariate and multivariate regression analyses to examine the factors associated with pDDIs. The results, including odds ratios (OR) and p-values for each variable, are presented in [Table 5.4](#) and [Table 5.5](#). The outcomes of the univariate logistic regression analysis concerning the exposure to one or more pDDIs are displayed in [Table 5.4](#). The patient's age did not substantially affect the existence of any pDDIs ( $p=0.6$ ). Nevertheless, a significant correlation was seen between the presence of all pDDIs and the prescription of five or more medications ( $p<0.001$ ). Regarding exposure to major and moderate pDDIs, the univariate logistic regression analysis revealed the following findings. The association between major and moderate pDDIs and the patient's age was insignificant ( $p=0.3$  and  $p=0.5$ , respectively). However, a significant association was observed with the prescription of five or more drugs ( $p<0.001$ ).

**Table 5.4 Analysis of Univariate Logistic regression for at least one major, one moderate, and all pDDI**

| Characteristic                                                      | N          | All pDDI |            |         | Major pDDI |            |         | Moderate pDDI |            |         |
|---------------------------------------------------------------------|------------|----------|------------|---------|------------|------------|---------|---------------|------------|---------|
|                                                                     |            | OR/      | 95% CI/    | p-value | OR/        | 95% CI/    | p-value | OR/           | 95% CI/    | p-value |
| <b>Gender</b>                                                       | <b>899</b> |          |            |         |            |            |         |               |            |         |
| <b>M</b>                                                            | 559        | -        | -          |         | -          | -          |         | -             | -          |         |
| <b>F</b>                                                            | 340        | 0.94     | 0.68, 1.28 | 0.7     | 0.85       | 0.60, 1.19 | 0.3     | 1.07          | 0.81, 1.40 | 0.6     |
| <b>Age_years</b>                                                    | <b>899</b> |          |            |         |            |            |         |               |            |         |
| <b>&lt;1</b>                                                        | 349        | -        | -          |         | -          | -          |         | -             | -          |         |
| <b>≥1 &amp; &lt;5</b>                                               | 228        | 1.38     | 0.92, 2.07 | 0.11    | 0.94       | 0.60, 1.44 | 0.8     | 1.11          | 0.78, 1.58 | 0.6     |
| <b>≥5</b>                                                           | 322        | 1.11     | 0.78, 1.59 | 0.6     | 0.82       | 0.56, 1.19 | 0.3     | 0.91          | 0.67, 1.23 | 0.5     |
| <b>Number of drugs prescribed</b>                                   | <b>899</b> |          |            |         |            |            |         |               |            |         |
| <b>2-4</b>                                                          | 447        | -        | -          |         | -          | -          |         | -             | -          |         |
| <b>5 and more</b>                                                   | 452        | 12.7     | 8.22, 20.4 | <0.001  | 4.08       | 2.82, 6.03 | <0.001  | 8.17          | 6.03, 11.2 | <0.001  |
| / CI = Confidence Interval, OR = Odds Ratio, N = Number of patients |            |          |            |         |            |            |         |               |            |         |

Table 5.5 Analysis of Multivariate Logistic regression for at least one major, one moderate, and all pDDI

| Characteristic                                                      | N          | All pDDI |            |         | Major pDDI |            |         | Moderate pDDI |            |         |
|---------------------------------------------------------------------|------------|----------|------------|---------|------------|------------|---------|---------------|------------|---------|
|                                                                     |            | OR/      | 95% CI/    | p-value | OR/        | 95% CI/    | p-value | OR/           | 95% CI/    | p-value |
| <b>Gender</b>                                                       | <b>899</b> |          |            |         |            |            |         |               |            |         |
| Male                                                                | 559        | -        | -          |         | -          | -          |         | -             | -          |         |
| Female                                                              | 340        | 0.90     | 0.63, 1.28 | 0.6     | 0.82       | 0.57, 1.17 | 0.3     | 1.06          | 0.77, 1.44 | 0.7     |
| <b>Age group (year)</b>                                             | <b>899</b> |          |            |         |            |            |         |               |            |         |
| <1                                                                  | 349        | -        | -          |         | -          | -          |         | -             | -          |         |
| ≥1 & <5                                                             | 228        | 1.01     | 0.64, 1.58 | >0.9    | 0.74       | 0.47, 1.17 | 0.2     | 0.78          | 0.52, 1.17 | 0.2     |
| ≥5                                                                  | 322        | 0.88     | 0.59, 1.31 | 0.5     | 0.69       | 0.47, 1.03 | 0.070   | 0.69          | 0.48, 0.97 | 0.035   |
| <b>Number of drugs prescribed</b>                                   | <b>899</b> |          |            |         |            |            |         |               |            |         |
| 2-4                                                                 | 447        | -        | -          |         | -          | -          |         | -             | -          |         |
| 5 and more                                                          | 452        | 12.8     | 8.26, 20.6 | <0.001  | 4.28       | 2.94, 6.36 | <0.001  | 8.55          | 6.27, 11.8 | <0.001  |
| / CI = Confidence Interval, OR = Odds Ratio, N = Number of patients |            |          |            |         |            |            |         |               |            |         |

[Table 5.5](#) shows the findings of the multivariate logistic regression study. The patient's age did not significantly affect the existence of major pDDIs. But, it showed significance with moderate pDDI. Whereas association was significant with ≥5 prescribed drugs ( $p<0.001$ ). Major and moderate pDDIs were shown to be significantly associated with ≥5 prescription medications ( $p<0.001$ ).

#### 5.1.3.4 Predictors of length of stay

Several variables were considered in multivariate analysis to explore factors associated with the length of the stay. The odds ratio (OR) and p-value for each variable are presented in [Table 5.6](#). Prescription of five or more medicines was significantly associated with LOS and major and moderate pDDI (both  $p<0.001$ ). The median LOS for patients with any pDDI was 11 days (IQR=14) compared with four days (IQR=6) in those without pDDIs.



Table 5.6 Analysis of Univariate Linear regression for length of stay

|                                                                                |            | Length of stay |            |         |
|--------------------------------------------------------------------------------|------------|----------------|------------|---------|
| Characteristic                                                                 | N          | IRR/           | 95% CI/    | p-value |
| <b>Gender</b>                                                                  | <b>899</b> |                |            |         |
| Male                                                                           | 559        | -              | -          |         |
| Female                                                                         | 340        | 1.01           | 0.96, 1.05 | 0.8     |
| <b>Age group (year)</b>                                                        | <b>899</b> |                |            |         |
| <1                                                                             | 349        | -              | -          |         |
| ≥1 & <5                                                                        | 228        | 0.76           | 0.72, 0.81 | <0.001  |
| ≥5                                                                             | 322        | 0.58           | 0.55, 0.61 | <0.001  |
| <b>Number of drugs prescribed</b>                                              | <b>899</b> |                |            |         |
| 2-4                                                                            | 447        | -              | -          |         |
| 5 and more                                                                     | 452        | 2.06           | 1.95, 2.17 | <0.001  |
| <b>pDDI observed</b>                                                           | <b>899</b> |                |            |         |
| None or Minor                                                                  | 466        | -              | -          |         |
| Major                                                                          | 173        | 2.06           | 1.94, 2.18 | <0.001  |
| Moderate                                                                       | 260        | 1.44           | 1.36, 1.53 | <0.001  |
| / N = Number of patients, CI = Confidence Interval, IRR = Incidence Rate Ratio |            |                |            |         |

### 5.1.4 Discussion

This study provides the prevalence and significance of pDDIs in a pediatric ICU in North India. There were 246 different types of pDDIs identified on 3,884 occasions among ICU patients. This study focused on the prevalence of pDDIs among pediatric patients at a tertiary care hospital. The clinical practice adopted in this investigation to identify potential pDDI considers factors such as age, gender, and the number of drugs prescribed to a patient. While these factors may not be critical for identifying pDDIs, they are essential for assessing the clinical relevance of the interactions. Expanding the dataset and including more participants in the retrospective validation will enhance the significance of the study, as the additional data will help validate the frequency and impact of DDI interactions across the demographic region.

When administering two medications together, it has the potential to have significant clinical repercussions, such as serious adverse effects, decreased therapeutic efficacy, or other significant outcomes; this is referred to as a major drug–drug interaction. For example, co-administration with CYP450 3A4 inhibitors may raise the plasma concentrations of drugs primarily metabolized by the isoenzyme, such as fentanyl, midazolam, fluconazole, etc. Increased concentrations can worsen or prolong the adverse effects of drugs and result in respiratory depression that could be fatal([Xiong et al., 2022](#), [Labroo et al., 1997](#), [Moltke et al., 1996](#), [Peyko et al., 2017](#)). Our study revealed significant DDIs involving fentanyl with fluconazole, fluconazole with midazolam, and dobutamine with vancomycin. Specifically, the interactions between fentanyl and fluconazole and fluconazole and midazolam were observed to have major severity, while the interaction between dobutamine and vancomycin was classified as moderate in severity. These findings align with previous studies highlighting the potential risks associated with polypharmacy in critical care settings over the studied demography. Previous studies have indicated that concurrent administration of dopamine or dobutamine with vancomycin may lead to decreased vancomycin serum levels. These interactions are categorized as moderate in terms of potential drug-drug interaction (pDDI) severity ([Xiong et al., 2022](#), [Pea et al., 2000](#)). Our study has also provided a graphical user interface, “ContraIndicator,” to show the empirical findings and raise awareness about pDDIs ([ContraIndicator](#)). This tool also allows clinicians to upload prescriptions, which the system uses to automatically detect and highlight interactions between prescribed medications, facilitating better-informed clinical decisions.

In our study, the mean number of drugs prescribed in the PICU was (3·70), which is lower than in other studies([Lima et al., 2020](#), [Hassanzad et al., 2020](#)). High frequency of Fentanyl, Midazolam, Vancomycin, and Meropenem reflects common practice in other studies([Lima et al., 2020](#), [Dai et al., 2016](#)). We found that 49·6% (446) of patients between the ages of 1 day and 18 years were exposed to at least one pDDI throughout their stay in the PICU. It was comparable to Getachew et al., which reported that pediatric hospitalized patients who had been exposed to a pDDI were 45·8%(176)([Getachew et al., 2016](#)). However, these findings were higher than Hassanzad et al.([Hassanzad et al., 2020](#)) and Ismail et al.([Ismail et al., 2013](#)), in which overall interaction was 29·3% and 25·8%, respectively. In our study, moderate–pDDIs were most often found pDDI, followed by major–pDDIs. Other investigations revealed that most were moderately severe, followed by minor and major severity([Getachew et al., 2016](#)). Major interactions were discovered in 38·78% (n = 173) of the 446 patient admissions in this investigation that had at least one pDDI. These results exceeded the findings of both Getachew et al.([Getachew et al., 2016](#)) and Ismail et al([Ismail et al., 2013](#)). In Getachew et al.'s study, out of 176 pediatric patients with at least one pDDI, 19·9% (35/176) of the patients experienced a major interaction. Similarly, in Ismail et al.'s study, 10·7% of the patients encountered major interactions.

Similar to the findings previously published by Hassanzad et al.([Hassanzad et al., 2020](#)), logistic regression analysis in our study indicated that the number of medicines prescribed was

significantly associated with a greater risk of pDDIs. In other research, the primary predictors of drug interactions have been revealed to include the patient's age, the frequency of visits, the average number of prescriptions per visit, the length of the ICU stay, and the presence of chronic conditions([Hassanzad et al., 2018](#), [Ismail et al., 2017](#), [Getachew et al., 2016](#), [Langerova et al., 2013](#)). In our investigation, the LOS in the PICU ranged from one to 203 days, which was longer than the one–day to 113–day range seen in Brazilian hospitals([Lima et al., 2020](#)). Our findings indicated a link between significant pDDI severity and longer PICU stays. The median LOS for pediatric patients with pDDI was found to be 11 days, compared to 4 days for patients who had no exposure to pDDI. The team's practices may change as a result of the fact that patients with pDDI had more extended hospital stays.

There are several limitations of this study. The definition of severity levels in the DDInter database may not be appropriate for all situations worldwide. Even though Xiong G et al.([Xiong et al., 2022](#)) tried to gather data from various sources, they cannot guarantee that all of the pDDIs indicated are covered in the pDDI Inter DB. Also, the pDDI Inter DB considers interactions involving two medicines; interactions involving three or more drugs are not covered. Another limitation of this study is the retrospective design of the analysis. Retrospective analysis has been conducted using the DDInter database, a publicly available static resource. However, our robust pipeline can be fine–tuned to other dynamic databases, which are prominently updated at regular intervals. Further, this study is limited to open–sourced data related to DDI. This limitation prevents us from studying more extensive interactions in greater detail for their therapeutic importance. Fine–tuning of MED7 has been done on MIMIC–III data for the NER task. Although it achieved a very high F1–score (~0.89) for extracting the drug entities, we implemented it across treatment charts from the PICU of AIIMS, Delhi. This restricts the MED7 model from extracting drugs with an equal F1–score, as some might get missed or falsely classified.

Other limitations include generalizability, data quality, confounding variables, outcome assessment, and bias. (i) Generalizability: The characteristics of the study population at one center may not reflect the diversity seen in other settings. (ii) Data Quality: The quality of retrospective data depends on the accuracy and completeness of medical records. Incomplete or inaccurate data can compromise the reliability of study results. (iii) Confounding Variables: Controlling for all potential confounding variables in a retrospective cohort study may be challenging. Unmeasured or inadequately controlled confounders can lead to spurious associations. (iv) Outcome Assessment: Depending on the nature of the outcome, there may be limitations in diagnosing or measuring it accurately from historical data. (v) Bias: Participants may have difficulty accurately recalling past exposures, leading to recall bias.

A known limitation of conventional rule-based drug-drug interaction (DDI) checkers is alert fatigue, where large volumes of low-severity alerts reduce clinician attention to clinically critical warnings. ContraIndicator addresses this challenge through severity-based stratification of

detected interactions into major, moderate, and minor categories using curated pharmacological knowledge bases. This design enables prioritization of potentially life-threatening contraindications while contextualizing lower-severity interactions. The ContraIndicator system has been designed with deployment-oriented evaluation in mind, including planned assessment of alert burden and clinician response before and after integration into clinical workflows. However, the present work focuses on system development and feasibility, and does not claim measured reductions in alert fatigue. Prospective deployment studies are required to quantify the impact of severity-aware alerting on clinical decision-making and alert acceptance.

The drug–drug interactions identified by ContraIndicator represent pharmacologically plausible associations derived from curated knowledge bases and prescription co-occurrence patterns. These findings do not imply causal relationships between drug combinations and observed clinical outcomes. In the intensive care setting, interaction prevalence may be influenced by confounding factors such as underlying disease severity, indication bias, and polypharmacy, where critically ill patients are more likely to receive complex medication regimens. Consequently, ContraIndicator is intended as a clinical awareness and triage tool to support medication review, rather than a causal inference or prescriptive decision-making system.

This study identified 246 different pDDIs from the PICU of AIIMS, Delhi. One–third of pDDIs were classified as major. The major pDDIs were associated with increased PICU length of stay (LOS). To prevent pDDIs and ensure patient safety, pediatricians should receive education regarding clinically relevant pDDIs for highly prescribed drugs in their settings; it is desirable to have software to identify these potential interactions when prescriptions are being made. We have built a "ContraIndicator" suite to track these potential drug interactions within the PICU. The "ContraIndicator" web application will compare digital interventions to prevent pDDIs in the ICU before and after. "ContraIndicator" is a user–friendly, web–based application designed to aid healthcare providers in identifying pDDIs quickly and accurately. It allows users to input a patient's medication regimen, and the application scans a comprehensive database to identify pDDIs. By providing a reliable and user–friendly tool for identifying and managing pDDIs, this digital intervention aims to enhance patient safety and reduce the length of stay in the PICU.

# Chapter 6

## SAFE-ICU: An Integrated AI-Enabled Decision Support Platform for Intensive Care Units

### 6.1 Introduction

The need for timely, accurate, and intelligent decision support in Intensive Care Units (ICUs) is greater than ever. Clinicians face increasing complexity in managing critically ill patients, often requiring rapid interpretation of heterogeneous data streams. To address this challenge, we developed **SAFE-ICU**, an AI-enabled decision support platform that so far integrates three critical modules: (1) Predictive Modeling for Hemodynamic Shock, (2) Drug-Drug Interaction (DDI) Monitoring, and (3) Antimicrobial Resistance (AMR) Surveillance. SAFE-ICU is modular, expandable, and a scalable system that can support clinicians in real-time decision-making while also laying the foundation for more advanced AI applications.

### 6.2 Motivation

SAFE-ICU application makes the best ICU models available for real-world use and consists of three inter-related components:

- [Chapter 4](#) - **Predictive Modeling**: Hemodynamic shock prediction
- [Chapter 5](#) - **Drug Safety**: DDI monitoring module
- **AMR Surveillance**: Visualization of ICU Ast testing

The system was built with four primary goals in mind. Clinical relevance was prioritized to ensure that all tools directly address real-world ICU needs. AI readiness was incorporated into the design to facilitate seamless integration of predictive models in the future. Emphasis was also placed on usability, resulting in a simple, visual, and modular interface tailored for clinicians. Lastly, the system was developed with scalability in mind, making it easily adaptable to other ICUs or hospital settings.

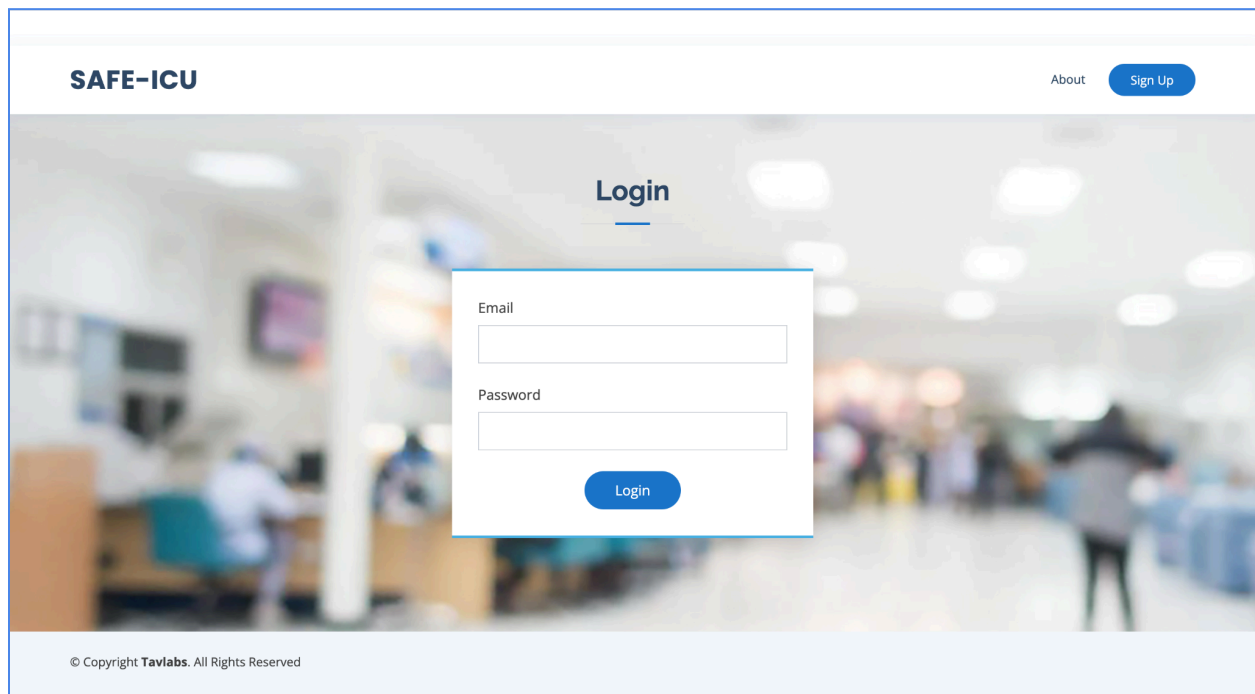
### 6.3 SAFE-ICU Functional Modules: Detailed Workflow and Implementation

The SAFE-ICU platform is built on modular microservices to interpret ICU data, enhance clinical insight, and deliver real-time, actionable alerts. Each module addresses specific information domains and is designed to operate in synergy with others through a unified dashboard. Below is a detailed explanation of core functional areas.

*Table 6.1 SAFE-ICU is a web-based application developed using the following technology stack.*

|                |                                                                               |
|----------------|-------------------------------------------------------------------------------|
| Frontend       | flask (or optionally React/Django for future expansion)                       |
| Backend        | Python-based microservices                                                    |
| Database       | PostgreSQL / SQLite                                                           |
| ML Models      | Trained offline using Scikit-learn / PyTorch and integrated as inference APIs |
| Data Ingestion | Daily ingestion of ICU data including vitals, prescriptions, and lab results  |

Each module is decoupled from the others but interacts through single database and a unified user interface.



**Figure 6.1 SAFE-ICU Login Interface.** The secure login page provides authorized access to the SAFE-ICU dashboard, ensuring that sensitive clinical data and decision-support tools are available only to verified users within the hospital network

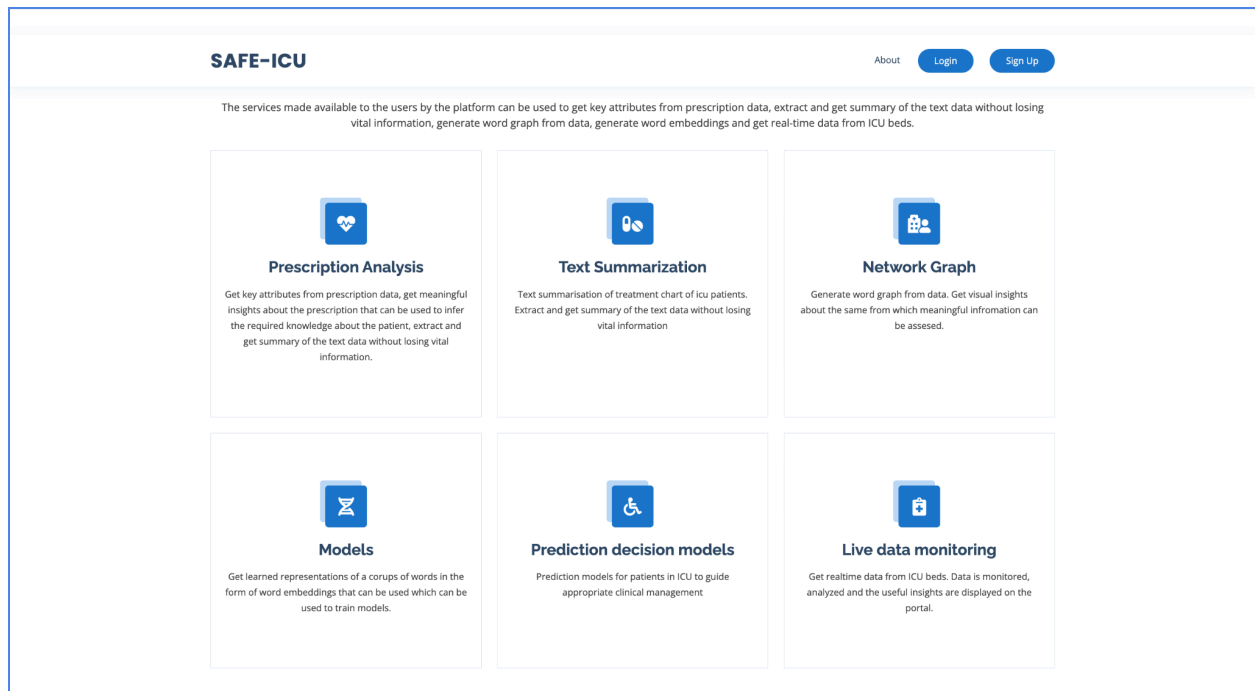


Figure 6.2 **SAFE-ICU Dashboard Interface**. The platform provides modular services for ICU data analysis, including prescription analysis, text summarization, network visualization, word embedding generation, predictive modeling, and real-time vitals monitoring. Each module is designed to enhance clinical decision-making through intuitive and actionable insights

### 6.3.1 Prescription Analysis and Drug Safety Intelligence

The **Prescription Analysis** module plays a critical role in extracting structured insights from unstructured medical orders and treatment charts. ICU prescriptions often contain overlapping drugs, non-standard nomenclature, and handwritten notes. This module uses Named Entity Recognition (NER) models to extract drug names from uploaded charts and treatment logs. These are normalized and mapped against standardized drug ontologies sourced from open repositories such as DrugBank, DDInter, and Medscape.

Key features include automatic detection of missing or duplicated dosages, inference of likely diagnoses based on prescription patterns (e.g., vancomycin and meropenem suggesting broad-spectrum therapy for suspected sepsis), and metadata extraction such as frequency, route, and timing. A prescription summary panel provides clinicians with a digest of the active medication regimen, highlighting potentially harmful overlaps or unsafe administration timelines.

This module also connects directly to **ContraIndicator**, the DDI (Drug–Drug Interaction) checker, which flags risky drug combinations. The interface allows users to interactively modify, add, or remove drugs and immediately observe how changes affect interaction severity levels

(Major, Moderate, Unknown). This feedback loop supports real-time prescribing decisions, especially in high-load or low-resource ICU environments.

### 6.3.2 Text Summarization of Clinical Notes /Treatment Charts

ICU patients accumulate dozens of pages of notes and interventions each day. The **Text Summarization** module aims to condense these verbose documents into clinically meaningful summaries. Two approaches are deployed:

- **Extractive Summarization** using Gensim’s TextRank algorithm identifies the most relevant sentences from the input corpus.

This module processes daily treatment charts, progress notes, and nursing documentation. It preserves key interventions (e.g., “intubation performed,” “noradrenaline started at 6 µg/min”) while omitting redundant or boilerplate phrases. The summaries are especially useful for shift changes, handovers, and audit trails. Clinicians can search, filter, and export these summaries as part of structured documentation or clinical decision-making support.

### 6.3.3 Network Graph Visualization for Conceptual Relationships

To enhance interpretability of clinical narratives and prescriptions, the platform offers **Network Graph Visualizations**. This component transforms extracted tokens from prescriptions and notes into conceptual maps using graph-based NLP techniques. Each node represents a term or drug, and edges denote semantic or clinical relationships determined by co-occurrence frequency, cosine similarity, or embedding proximity.

For example, antibiotics, vasopressors, and steroids form clusters that show temporal or conceptual alignment. Clinicians can explore clusters like “broad-spectrum antibiotics” or “sedation protocols” to discover patterns or outliers. Such visualizations are valuable for educational rounds, quality audits, or generating hypotheses about practice variation.

### 6.3.4 Word Embeddings and Semantic Context Generation

SAFE-ICU supports word representation learning through models like **GloVe**, **Word2Vec**, and fine-tuned domain-specific embeddings. These embeddings convert ICU vocabulary both structured and unstructured into dense vector spaces that capture semantic relationships.

For instance, embedding training on prescription data may cluster “ceftriaxone” close to “cefepime” but far from “paracetamol,” reflecting pharmacological similarity. The resulting vectors are then used in:

- Clustering similar treatments.



- Feeding into downstream models for risk prediction.
- Semantic search and auto-completion in the dashboard UI.

This module underpins other services such as the text summarizer, the DDI module, and the treatment pattern analyzer. It ensures that language-based models are attuned to ICU-specific terminology and variations.

### 6.3.5 Predictive Decision Support Models for ICU Management

The **Prediction Decision Models** module integrates AI-powered algorithms designed to forecast clinical deterioration and guide interventions. This includes modules such as:

- **SIgnose** for shock prediction,
- **ThermoGnose** for hypothermia risk

Models are trained using longitudinal data with features drawn from real-time vital signs, lab results, and treatment history. Once a threshold of risk is crossed, the dashboard alerts the clinical team and provides contextual explanations via SHAP values. The predictive horizon (e.g., “risk of shock in 0–8 hours”) is highlighted, along with contributing features like heart rate, MAP, or prior interventions.

These models are designed to be explainable, with prediction summaries, validation options, and real-time integration into existing workflows. Clinicians can validate or dispute model outputs using a binary interface, enabling a feedback loop for model retraining and continual performance tuning.

### 6.3.6 Live ICU Data Monitoring and Real-Time Alerts

The **Live Data Monitoring** service forms the operational backbone of SAFE-ICU. It continuously ingests data from bedside monitors (e.g., heart rate, respiratory rate, SpO<sub>2</sub>), EHR feeds, lab systems, and prescription records. A robust streaming architecture ensures that patient tiles on the dashboard are updated with minimal latency (usually under 1 minutes).

Each patient card displays key identifiers (age, sex, bed number), dynamic vitals, prediction outputs, and clinician notes.

Alerts are generated not just by threshold violations but by risk surges as identified by the predictive models. For example, a rising shock score over, escalating to red if the lead time prediction falls under 8 hours.

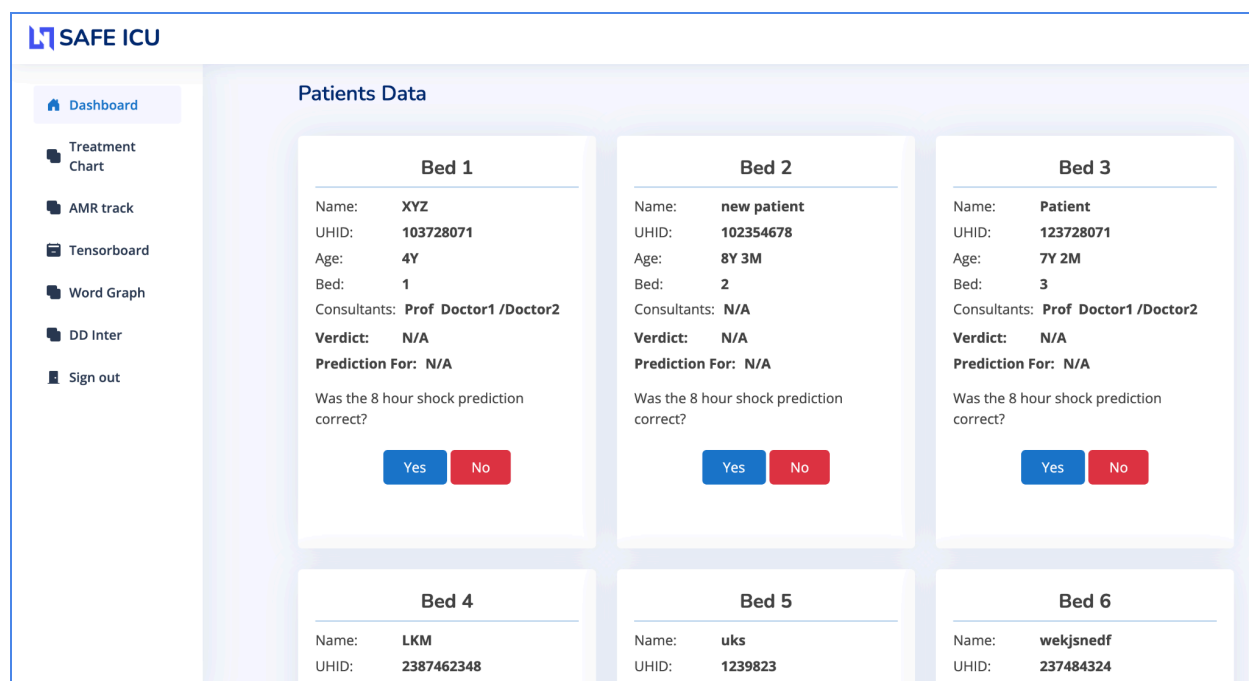


Figure 6.3 **Patient Monitoring View in SAFE-ICU Dashboard.** This interface provides an at-a-glance summary of patient details, including age, bed number, attending consultants, and prediction verdicts. Clinicians can verify the correctness of 8-hour shock predictions using quick "Yes" or "No" inputs, enabling real-time feedback for model evaluation

## 6.4 Hemodynamic Shock Prediction

This module implements a predictive model for early identification of pediatric patients at risk of hemodynamic shock. This module is an application of the work described in Chapter 3. Based on vitals time-series (e.g., heart rate, systolic pressure), the model outputs a risk score and predicted lead time.

Table 6.2 *Key Characteristics of the Hemodynamic Shock Prediction Module*

|            |                                                       |
|------------|-------------------------------------------------------|
| Model type | Trained and Finalised model (SIgnose)                 |
| Inputs     | Vitals from ICU monitors                              |
| Output     | Verdict Shock/Non-Shock, and predicted onset time     |
| UI         | Displays patient risk dashboard with alert thresholds |

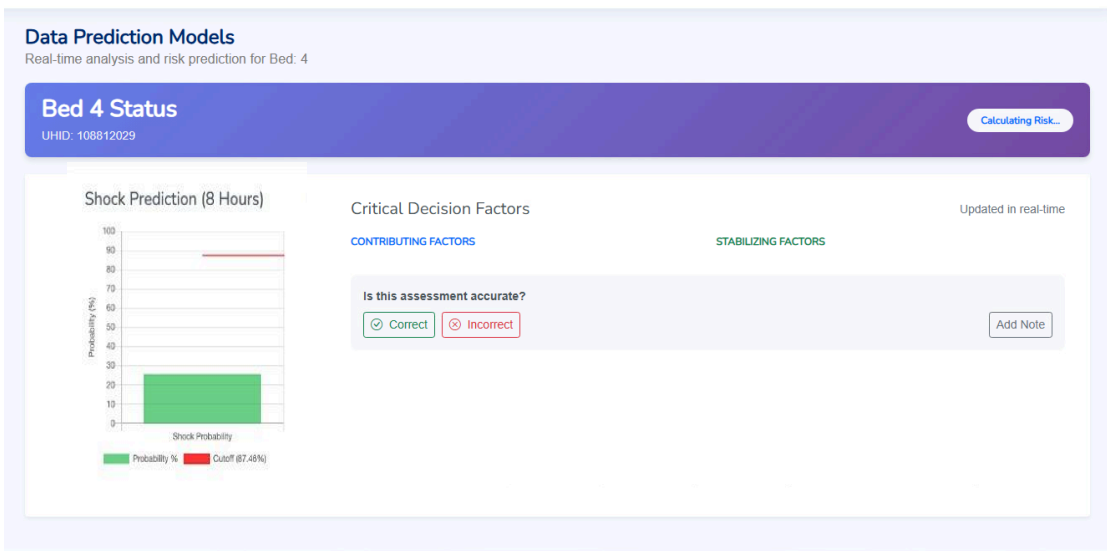


Figure 6.4 **Shock Risk Prediction Output in SAFE-ICU.** This module displays an 8-hour shock probability score computed using the *SIgnose* model. It includes the cutoff threshold, prediction time window, and contributing clinical features such as heart rate, respiratory rate, oxygen saturation, blood pressure, and demographics. Clinicians can validate the prediction accuracy and provide feedback through the integrated interface

## 6.5 Drug-Drug Interaction Monitoring

This module provides real-time alerts for potential harmful drug interactions prescribed in the ICU. This module supports safer prescription practices and is based on Chapter 4.

Table 6.3 **Key Characteristics of the Drug–Drug Interaction Monitoring Module**

|         |                                                                                                |
|---------|------------------------------------------------------------------------------------------------|
| Input   | Daily prescription data                                                                        |
| Backend | Drug interaction knowledge base from open-source databases (e.g., DDInter, Drugbank, Medscape) |
| Output  | List of flagged interactions, severity, clinical notes                                         |
| UI      | Clinician view of current patient prescriptions with interaction overlays                      |

**SAFE-ICU**

### CONTRAINDICATOR

Abaloparatide X Fluconazole X Fentanyl X Abacavir X Dexamethasone X Cetuximab X Clofedanol X Diphenoxylate X

Enter the drug name... **Add Tag** **Confirm** **Reset List**

- Abaloparatide X
- Fluconazole X
- Fentanyl X
- Abacavir X
- Dexamethasone X
- Cetuximab X
- Clofedanol X
- Diphenoxylate X

**Submit**

**Upload a Treatment-Chart**

Choose a file:

© Copyright Tavlab. All Rights Reserved

Figure 6.5 **ContraIndicator Module in SAFE-ICU**. This tool allows clinicians to input and review a list of prescribed drugs, automatically flagging potential drug–drug interactions. Users can add, remove, or reset the list of medications, and upload treatment charts for real-time safety checks, enhancing medication decision support in critical care

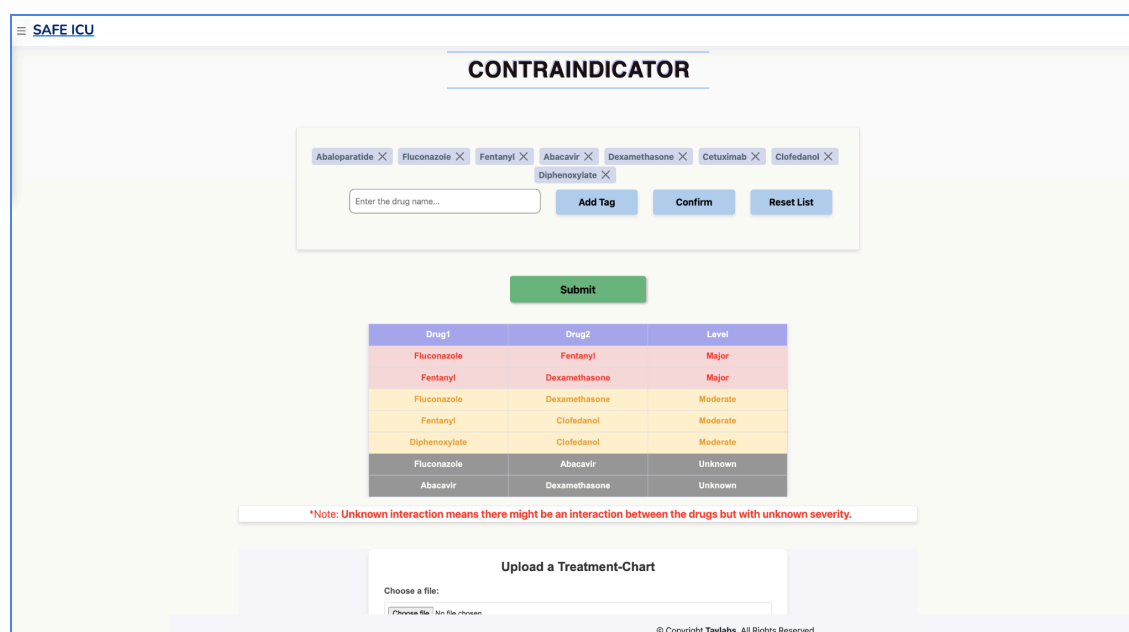


Figure 6.6 **Drug Interaction Output in ContraIndicator (SAFE-ICU)**. After drug inputs are confirmed, the system displays a tabulated summary of identified drug–drug interactions along with their severity levels (Major, Moderate, Unknown). Interactions with unknown severity are explicitly marked, prompting clinicians to exercise caution. This visualization supports safe prescribing practices by providing immediate and interpretable DDI feedback

## 6.6 AMR Surveillance

This component visualizes of historical Antimicrobial Susceptibility Testing (AST) data from the ICU. While the current system is not AI-driven, it is AI-enabled and designed to support future prediction of resistance trends.



Figure 6.7 **Login Interface of the Antimicrobial Stewardship Platform.** This platform provides secure access for registered users to AI-powered tools supporting antimicrobial resistance tracking and stewardship

Table 6.4 **Key Characteristics of the AMR Module**

|          |                                                      |
|----------|------------------------------------------------------|
| Input    | Daily ingestion of AST reports                       |
| Features | Pathogen-specific resistance trends; Heatmaps        |
| Use case | Infection control teams and empiric therapy planning |

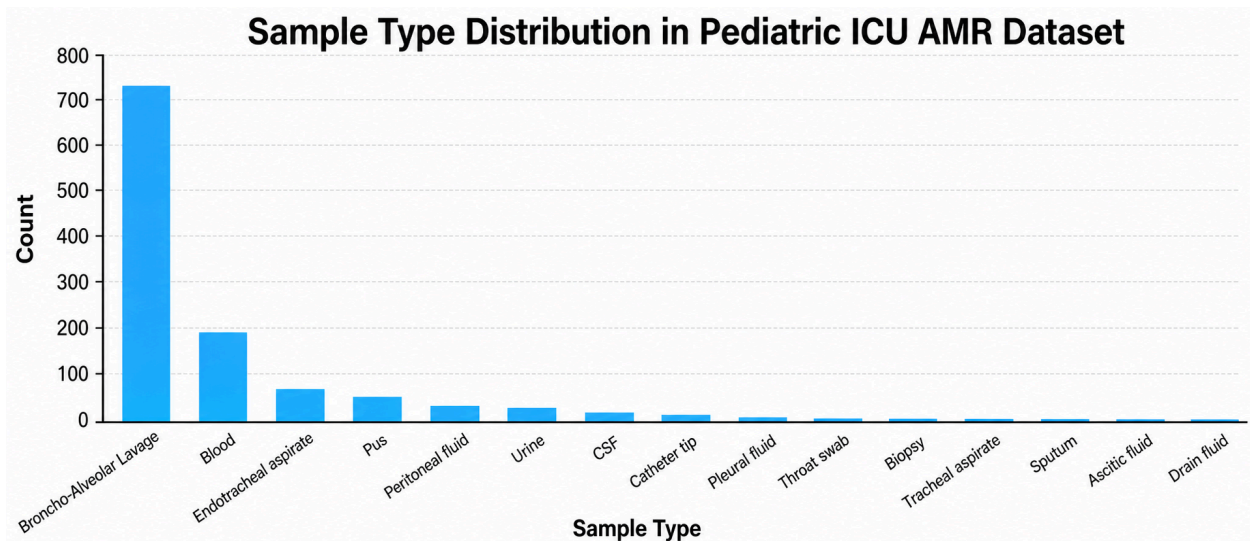
### 6.6.1 Antimicrobial Resistance (AMR) Surveillance and Clinical Visualization

The **AMR Surveillance** module in the SAFE-ICU platform serves as a vital decision-support component for empirical antibiotic selection, especially in high-acuity ICU settings. Antimicrobial resistance (AMR) has emerged as a global threat, with ICU patients being particularly vulnerable due to invasive procedures, immunocompromised states, and frequent use of broad-spectrum antibiotics. Given that initial treatments often precede microbiological confirmation, real-time awareness of local resistance trends can significantly enhance therapeutic accuracy and patient outcomes.

To demonstrate the utility of this module, a retrospective observational case study was performed using microbiological data from the **MCB-PICU** site. The dataset included **1291 Antimicrobial Susceptibility Tests (ASTs)** conducted for **550 pediatric patients** during the surveillance window, with registered clinicians linked to **clinician ID #64**. Each AST record logs the

identified organism, tested antibiotic, and the corresponding susceptibility outcome categorized as **Resistant**, **Intermediate**, or **Sensitive**.

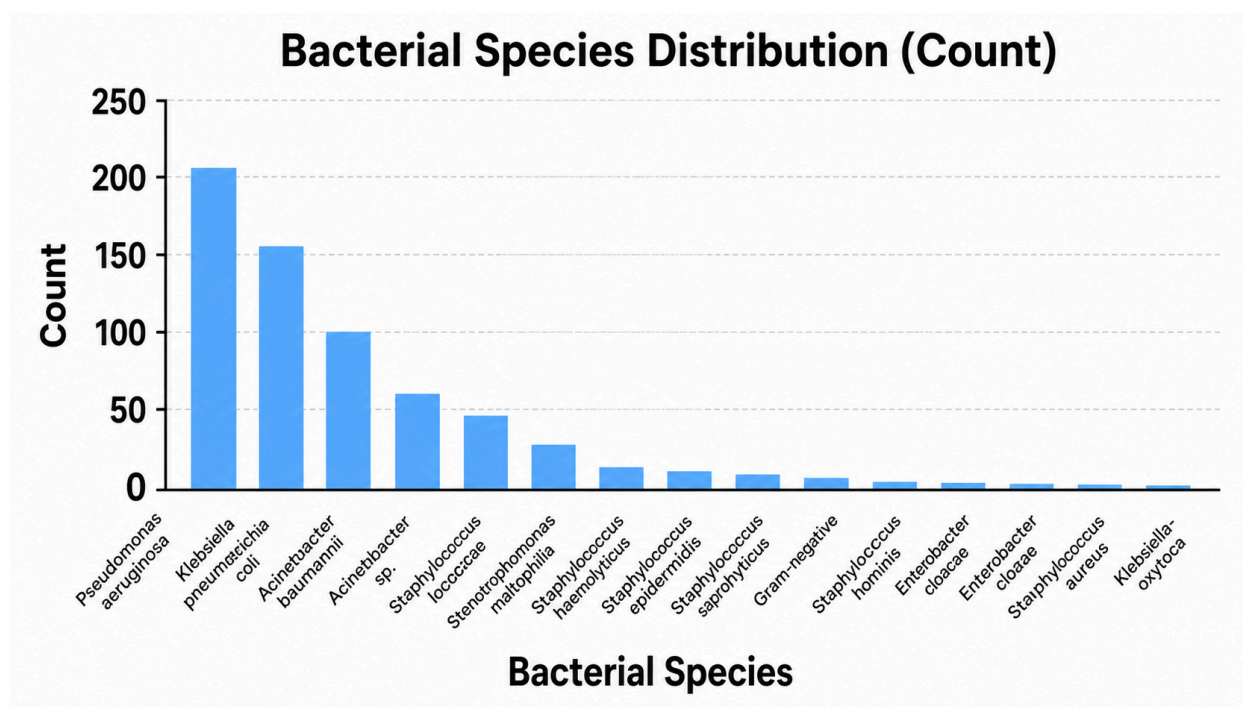
This data is processed and visualized by the SAFE-ICU AMR module to produce meaningful trends. Specifically, the system identifies the top 15 most commonly tested **sample types**, such as **blood**, **bronchoalveolar lavage**, **urine**, **Pus** , and **wound swabs**. Visualizations include bar charts representing the frequency of antimicrobial susceptibility testing across these sample types, alongside resistance breakdowns. These insights help clinicians understand the distribution of infection sources, assess site-specific resistance trends, and make more informed empirical treatment decisions based on anatomical infection loci.



**Figure 6.8:** *Sample Type Distribution in Pediatric ICU AMR Dataset.*

This visualization shows the top 15 most frequently tested isolated sample types (e.g., blood, bronchoalveolar lavage, urine, tracheal aspirate) from 1291 antimicrobial susceptibility tests conducted in the MCB-PICU. The distribution provides insight into common infection sites and helps inform sampling strategies and infection control efforts.

Simultaneously, the module visualizes the **top 15 most isolated organisms**, including **Pseudomonas aeruginosa**, **Klebsiella pneumoniae**, **E. coli**, **Acinetobacter baumannii**, and **Staphylococcus aureus**. Organism-level susceptibility profiles are rendered as resistance heatmaps, allowing clinicians to rapidly assess which pathogens exhibit multi-drug resistance (MDR) phenotypes and may require escalation of care or infection control measures.



**Figure 6.9:** Top 15 Isolated Organisms in Pediatric ICU AMR Dataset.

This chart displays the frequency of microbial organisms isolated from ICU patients, including *Klebsiella pneumoniae*, *Escherichia coli*, *Acinetobacter baumannii*, and *Pseudomonas aeruginosa*. The visualization enables clinicians to track dominant pathogens contributing to resistant infections and tailor antimicrobial strategies accordingly.

These insights are especially useful in institutional antimicrobial stewardship programs, as they support protocol refinement, outbreak identification, and long-term resistance surveillance. However, **access to these visual analytics requires registered login credentials and appropriate user permissions.**



AMRTrack

SafelCU-Dashboard

Dashboard Statistics

Antibiogram

Logout

## Hospital Antibiogram

Select Antimicrobial

- ☐ amikacin
- ☐ amoxicillin
- ☐ amoxicillin\_clavulanate
- ☐ amoxicillin\_clavulanic\_acid
- ☐ ampicillin
- ☐ ampicillin\_sulbactam
- ☐ azithromycin
- ☐ aztreonam
- ☐ carbenicillin
- ☐ cefepime
- ☐ cefixime
- ☐ cefoperazone\_sulbactam
- ☐ cefotaxime
- ☐ cefoxitin
- ☐ cefpodoxime
- ☐ ceftazidime
- ☐ ceftazidime\_clavulanic\_acid
- ☐ ceftriaxone
- ☐ ceftriaxone\_tazobactam
- ☐ cefuroxime
- ☐ cephalosporin

Select Organisms

- ☐ Achromobacter denitrificans
- ☐ Acinetobacter aeruginosa
- ☐ Acinetobacter baumannii
- ☐ Acinetobacter lwoffii
- ☐ Acinetobacter sp.
- ☐ Aeromonas sp.
- ☐ Alcaligenes faecalis
- ☐ Budding yeast
- ☐ Burkholderia cenocepacia
- ☐ Burkholderia cepacia
- ☐ Burkholderia sp.
- ☐ Chryseobacterium indologenes
- ☐ Citrobacter diversus
- ☐ Citrobacter edialis
- ☐ Citrobacter freundii
- ☐ Citrobacter koseri
- ☐ Citrobacter sp.
- ☐ Coagulase-negative staphylococci
- ☐ Commensal-flora
- ☐ Elizabethkingia meningoseptica

Select Collection Location

- ☐ 1A
- ☐ 1B
- ☐ 2A
- ☐ 2B
- ☐ 3A
- ☐ 3B
- ☐ 4A
- ☐ 4B
- ☐ AB1
- ☐ AB1-PACU
- ☐ AB2
- ☐ AB2-ICU
- ☐ AB3
- ☐ AB4
- ☐ AB4-Dialysis
- ☐ AB5
- ☐ AB5-ICU
- ☐ AB6
- ☐ AB7
- ☐ AB7-KTP
- ☐ AB7-PAR

Select Collection Type

- ☐ Ascitic fluid
- ☐ Aspirate
- ☐ Bile
- ☐ Biopsy
- ☐ Blood
- ☐ Bone
- ☐ Bone marrow
- ☐ Broncho-Alveolar Lavage
- ☐ Catheter tip
- ☐ Cervical Discharge
- ☐ Continuous ambulatory peritoneal dialysis
- ☐ CSF
- ☐ RVP

Enter Start Date

Enter End Date

Generate Sensitivity Table

Figure 6.10 **Antibiogram Query Interface in AMRTrack**. The platform enables users to generate customized antibiograms by selecting antimicrobials, organisms, sample collection locations, and sample types. Date filters allow for temporal analysis, and the generated sensitivity table supports data-driven antimicrobial stewardship decisions.

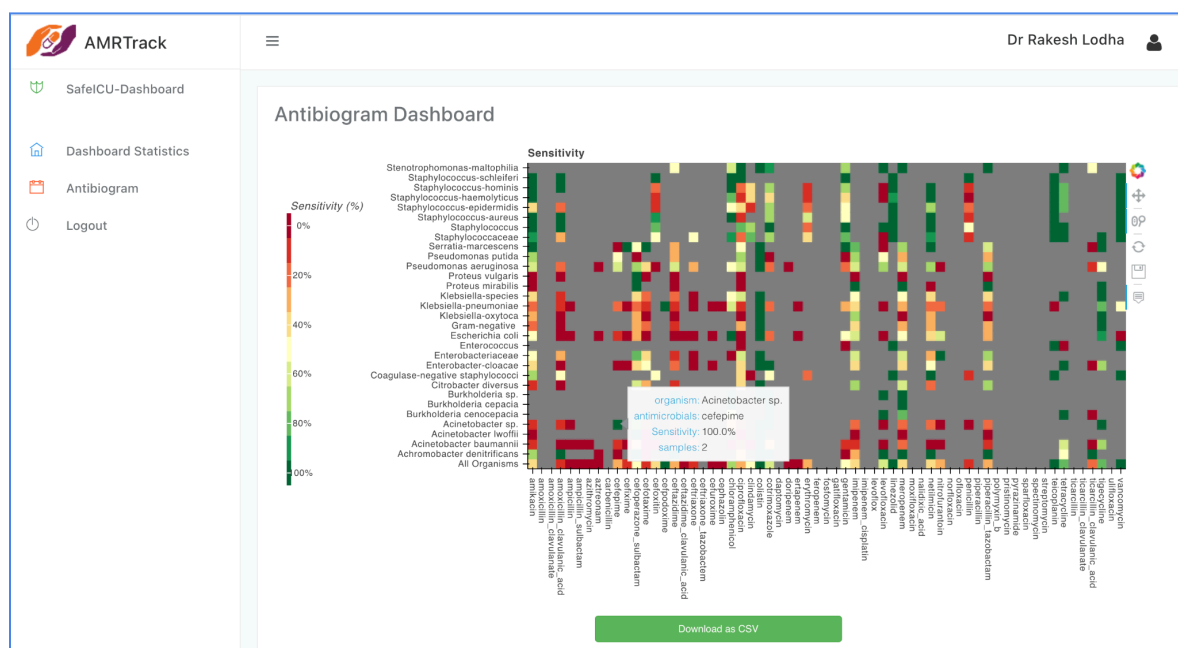


Figure 6.11 **Antibiogram Visualization in AMRTrack**. The heatmap displays antimicrobial sensitivity patterns across various organisms and antibiotics. Color gradients indicate percentage sensitivity, with green representing higher sensitivity and red indicating resistance. Grey cells denote unavailable data. This visualization aids clinicians in selecting effective antibiotics based on real-world resistance trends

## 6.7 SAFE-ICU User Interface

The SAFE-ICU dashboard presents a clean, intuitive, and clinician-friendly interface designed to support real-time decision-making in critical care environments. ICU staffs can:

- Monitor patients' shock risk: View real-time risk scores for hemodynamic shock to support timely interventions
- Review flagged drug interactions: Automatically detect and display potential drug–drug interactions from ICU prescriptions
- Explore AMR resistance visualizations: Analyze antimicrobial resistance trends using processed AST (Antibiotic Susceptibility Testing) data
- Filter by date, drugs , or organism: Easily navigate and customize views for location-specific or pathogen-specific insights
- Export reports for clinical meetings: Generate downloadable antibiogram to support case discussions and decision-making in ICU teams

## 6.8 System Architecture and Deployment Considerations

The SAFE-ICU platform is designed as a modular decision-support system integrating heterogeneous data sources, including physiological time-series, thermal imaging outputs, clinical text, and microbiology reports. Data are ingested through standardized interfaces and processed through dedicated preprocessing pipelines before being routed to task-specific predictive or analytical modules. Model outputs are presented to clinicians via a centralized dashboard to support situational awareness rather than automated decision-making.

The predictive models within SAFE-ICU are trained offline using curated historical data and are intended to be updated periodically as new data become available, rather than continuously retrained in real time. This design balances model stability with adaptability to evolving clinical practice. Latency considerations are addressed by prioritizing lightweight inference pipelines, enabling near-real-time risk estimation suitable for bedside monitoring workflows. Together, these design choices support scalability, maintainability, and safe clinical integration.

### **Integration of Additional Monitoring Devices and Waveform Data:**

Modern intensive care units increasingly employ advanced monitoring systems, including cardiac output monitors, mechanical ventilators, continuous renal replacement therapy (CRRT) machines, and processed neurological monitoring such as EEG. While these data streams are clinically valuable, their integration poses challenges related to data standardization, synchronization, proprietary device interfaces, and variable sampling frequencies. SAFE-ICU is designed as a modular framework capable of incorporating such data as they become accessible through interoperable interfaces.

Raw physiological waveform data, such as ECG, arterial pressure, or EEG signals, offer rich temporal information but require substantial storage, preprocessing, and noise management. In the current implementation, SAFE-ICU focuses on derived and clinically validated summary signals to ensure robustness and feasibility. Future extensions of the platform may incorporate waveform-level analytics and multimodal fusion approaches, contingent on data availability, infrastructure support, and clinical validation.

# Chapter-7

## Conclusion and Future Work

### 7.1 Conclusion

This thesis has presented the design, development, and evaluation of SAFE-ICU, an integrated, AI-driven decision support platform designed specifically for Pediatric Intensive Care Units (PICUs). The core objective was to bridge multiple high-impact clinical safety domains such as deterioration prediction, pharmacovigilance, and antimicrobial resistance surveillance through modular, explainable, and real-time artificial intelligence systems.

A critical foundation for these contributions was the construction of the SAFE-ICU Data Lake and warehouse infrastructure. This system enabled real-time acquisition, integration, and harmonization of multimodal clinical data including vitals, laboratory tests, thermal imaging, prescriptions, and microbiological reports into a scalable and query-efficient format. The structured data lake served not only as the input for model development but also powered the deployment of the SAFE-ICU dashboard with seamless real-time monitoring and analysis capabilities.

The ICU remains one of the most data-intensive and dynamic environments in healthcare. Patients in PICUs are particularly vulnerable, and even slight delays in detecting clinical deterioration or prescribing errors can result in significant harm. Traditional monitoring systems, although increasingly digitized, still lack the real-time analytical depth and semantic understanding necessary for nuanced decision-making. This research demonstrated how AI-driven temporal modeling, structured NLP, and LLM-based representation techniques can be synergistically integrated into a deployable decision-support system for critical care settings.

In the process, this thesis introduced and validated four core AI modules:

- **ThermoGnose**: A time-series based predictive model using tsfresh-derived features to identify pediatric hypothermia risk up to 4 hours in advance
- **ThermalShockNet**: A CNN-LSTM hybrid system that fuses thermal imaging and vital signs to predict hemodynamic shock six hours before clinical recognition
- **Signose**: A novel multi-dataset generalizable model that predicts abnormal Shock Index across Indian (AIIMS) and U.S.-based (MIMIC-III, eICU) datasets using LLM-embedded vitals and nonlinear features based shock-index prediction
- **ContraIndicator**: A named entity recognition-based pharmacovigilance pipeline that identifies and stratifies potential drug-drug interactions (pDDIs) using MED7 and curated DDI knowledge bases

## Statistical Considerations in Model Comparison:

Model performance comparisons in this thesis are based on repeated cross-validation with reporting of central tendency and variability across folds. While formal hypothesis testing between all model pairs was not uniformly conducted, observed differences were interpreted in the context of cross-validated variability and consistency across prediction horizons. The application of formal statistical significance testing for model comparison, particularly under non-independent cross-validation settings, is acknowledged as an important consideration for future work.

By integrating these modules into a clinician-friendly dashboard, SAFE-ICU enables modular AI use in PICUs, allowing for real-time, explainable alerts that can guide early interventions, optimize antimicrobial prescriptions, and flag potential medication hazards. Moreover, the platform demonstrates how AI solutions can generalize across geographical and institutional contexts, through rigorous cross-dataset validation, explainability with SHAP, and modular architecture designed for deployment. Thus, this thesis stands at the intersection of translational AI and critical care, offering practical tools with both technical rigor and clinical relevance.

## 7.2 Key Findings

### 7.2.1 Predictive Modeling of Clinical Deterioration

The platform's predictive models *ThermoGnose* and *ThermalShockNet* demonstrate the feasibility of AI to anticipate clinical deterioration events up to 24 hours in advance. *ThermoGnose*, which relies on non-linear features derived from structured ICU vitals, showed strong predictive performance for early hypothermia detection, outperforming conventional linear baselines. Meanwhile, *ThermalShockNet* extended prediction to hemodynamic instability by incorporating non-contact thermal imaging and CNN-LSTM pipelines. It effectively used Center-to-Peripheral Difference (CPD) metrics and heart rate variability, demonstrating improved early detection over vitals-only models.

The *SIgnose* model explored three approaches for modeling ICU vital signs to predict abnormal Shock Index. The first and most effective approach involved extracting non-linear time-series features from physiological signals, capturing complex temporal dynamics through domain-informed processing. These features, modeled using ensemble machine learning algorithms such as XGBoost, LightGBM, and SVM, consistently yielded high predictive performance. In contrast, the second approach using quantile-based symbolic representations, and the third approach using LLaMA-based prompt embeddings which transformed vitals into natural language input for large language model encoding did not perform as well. These results underscore the superiority of physiologically grounded, non-linear feature extraction over language-model-based or symbolic encoding strategies for clinical deterioration prediction in critical care settings.

### 7.2.2 Drug Safety Analytics via Clinical NLP

The ContraIndicator module highlighted the utility of lightweight NLP models like MED7 for real-time prescription parsing and pDDI detection. With over 8,000 ICU prescriptions analyzed, the pipeline identified drug interaction patterns stratified by age, gender, and medication count. Descriptive and regression analyses confirmed the high incidence and severity of pDDIs in pediatric care, with prescriptions involving  $\geq 5$  drugs showing significant associations with prolonged length of stay.

Although the current system used a rule-based DDI engine, future integration of transformer models such as ClinicalBERT could enable context-aware interaction analysis and event summarization.

### 7.2.3 AMR Surveillance Integration

Antimicrobial resistance surveillance in SAFE-ICU involved automated extraction and temporal visualization of AST reports. The module empowered clinicians to interpret evolving resistance patterns via organism-drug resistance heatmaps and MIC trends. While retrospective in design, this module laid the foundation for prospective resistance forecasting using ARIMA or Prophet models.

### 7.2.4 Summary of Key Contributions

[Table 7.1](#) Summary of Key Contributions. This table outlines the core modules developed as part of this thesis, highlighting their primary functions, underlying techniques, and the data sources used for model training and evaluation.

| Module                                       | Function                    | Key Techniques                                     | Data Sources                                          |
|----------------------------------------------|-----------------------------|----------------------------------------------------|-------------------------------------------------------|
| <b>SAFE-ICU Data Lake and Data Warehouse</b> | Multimodal data integration | Client-server ingest, NLP, preprocessing pipelines | ICU monitors, lab tests, thermal video, prescriptions |
| <b>ThermoGnose</b>                           | Predict Hypothermia         | tsfresh + Boruta + ML classifiers                  | SAFE-ICU vitals                                       |
| <b>ThermalShockNet</b>                       | Predict Shock               | Thermal + CNN-LSTM+ ML classifiers                 | Thermal Imaging + Vitals                              |
| <b>ContraIndicator</b>                       | pDDI Detection              | spaCy MED7 + rule-based DDI stratification         | ICU prescriptions                                     |
| <b>Signose</b>                               | Predict abnormal SI         | LLM Embedding ( LLaMA) + CNN-LSTM+ ML classifiers  | SAFE-ICU + MIMIC-III + eICU                           |

|                 |                                  |                                                |                                   |
|-----------------|----------------------------------|------------------------------------------------|-----------------------------------|
| <b>AMRTrack</b> | AMR<br>Visualization &<br>Trends | AST parsing + organism-drug<br>mapping + plots | Microbiology lab reports<br>(AST) |
|-----------------|----------------------------------|------------------------------------------------|-----------------------------------|

### 7.3 Limitations

While the thesis makes several contributions, it is important to acknowledge the following limitations:

- Model deployment is simulated and not yet integrated into live PICU settings. Prospective validation in clinical workflows remains a future direction
- Real-time scalability for certain modules like ThermalShockNet (which uses video and image processing) requires further optimization for latency
- The NLP pipeline does not currently incorporate multi-drug interaction reasoning or context-aware summarization capabilities that transformer-based LLMs could provide
- AMR surveillance in SAFE-ICU is primarily retrospective; prospective resistance forecasting was not performed
- Bias in model performance across subgroups (e.g., gender, age) was not exhaustively tested and may affect fairness.
- The predictive models developed in this thesis generate probabilistic risk estimates rather than deterministic clinical decisions. Decision thresholds were selected using validation data to optimize clinically relevant performance metrics, such as F1-score or Youden's Index, balancing sensitivity and specificity. Threshold selection reflects a trade-off between early detection and false alert burden, which is particularly important in critical care settings. Formal post-hoc probability calibration methods were not uniformly applied across all models and are acknowledged as a limitation. Prospective clinical evaluation is required to determine optimal operating thresholds for real-world deployment.
- The models were evaluated across multiple cohorts and care settings where possible, generalizability may be influenced by site-specific clinical practices, patient populations, and data acquisition protocols.
- Temporal data drift remains an important consideration, as changes in treatment strategies, monitoring devices, and case mix over time may affect model performance following deployment. Continuous performance monitoring and periodic model updates are therefore essential for safe long-term use.

## **7.4 Future Work**

The present work demonstrates feasibility and predictive utility using retrospective and observational data, prospective multicenter studies and randomized evaluations are required to determine real-world clinical impact, clinician acceptance, and outcome improvement prior to widespread adoption. Building on the foundation of this thesis, the following directions are envisioned for future work:

### **7.4.1 Clinical Deployment and Usability Testing**

- Conduct real-time, prospective validation studies at partner hospitals (e.g., AIIMS) to assess usability, false alert rates, and clinician acceptance
- Incorporate clinician-in-the-loop feedback mechanisms to improve AI interpretability and user trust

### **7.4.2 Polypharmacy Modeling with LLMs**

- Extend ContraIndicator with LLM-based polypharmacy models capable of inferring interaction context between more than two medications using transformer models like ClinicalBERT or LLaMA-2
- Explore relation extraction and adverse event summarization directly from physician notes

### **7.4.3 Real-Time Resistance Forecasting**

- Enhance the AMR dashboard with time-series forecasting algorithms (e.g., Prophet, ARIMA) to predict resistance trends and inform empirical therapy policies

### **7.4.4 Multilingual Clinical NLP**

- Adapt NLP pipelines to handle prescriptions and physician notes in regional Indian languages (e.g., Hindi, Tamil) through translation-augmented LLMs, expanding SAFE-ICU's applicability in diverse clinical settings

### **7.4.5 Fairness and Bias Audits**

- Evaluate model performance across intersectional subgroups (e.g., gender + age) to identify bias
- Integrate fairness-aware ML techniques to mitigate disparities, especially in generalizable models like SIngnose

### **7.4.6 Integration with Hospital Information Systems**



- Develop FHIR-compliant APIs to integrate SAFE-ICU with standard EHR systems
- Explore edge deployment strategies for low-resource settings using lightweight models

#### **7.4.7 AMR Surveillance**

- Future work aims to integrate predictive modeling for AMR trend forecasting, leveraging federated learning approaches to support generalizable resistance prediction across low-resource sites. As it stands, the AMR surveillance component plays a critical role in ensuring the SAFE-ICU platform delivers not only predictive risk stratification but also data-grounded antimicrobial guidance tailored to evolving microbial landscapes.

### **7.5 Final Remarks**

This thesis demonstrates that AI-driven platforms, when thoughtfully designed and validated, can enable safer, smarter, and more responsive care in pediatric ICUs. By building a unified system from predictive modeling, drug safety analytics, and resistance surveillance, SAFE-ICU provides a blueprint for next-generation AI in critical care. Its modular design allows continuous evolution, while its emphasis on explainability, real-time application, and multicenter generalizability sets a precedent for responsible, deployable clinical AI systems.

SAFE-ICU is not just a research prototype it is a step toward democratizing AI-powered safety in intensive care, laying the groundwork for data-driven transformation in pediatric critical care medicine.

# References

- [Acker N, Bredbeck B, Partrick A, Kulungowski M, Barnett C, Bensard D. \(2017\) \*Shock Index, Pediatric Age-Adjusted \(SIPA\) Is More Accurate Than Age-Adjusted Hypotension for Trauma Team Activation\*. Surgery 161 \(3\), 803–807. doi:10.1016/j.surg.2016.08.050](#)
- [Alaa M, Yoon J, Hu S, Schaar M. \(2019\) \*Personalized risk scoring for critical care prognosis using mixtures of Gaussian processes\*. IEEE Transactions on Biomedical Engineering, vol. 65, no. 1, pp. 207-218, doi: 10.1109/TBME.2017.2698602.](#)
- [Alsentzer N, Murphy R, Boag W, Weng H, Jin D, Nauman T, McDermott M. \(2019\) \*Publicly available clinical BERT embeddings\*. arXiv. <https://doi.org/10.48550/arXiv.1904.05342>](#)
- [Armen B, Freer V, Showalter W, Crook T, Whitener J, West C, et al. \(2016\) \*Improving Outcomes in Patients With Sepsis\*. Am J Med Qual. 31\(1\):56–63](#)
- [Acosta N, Falcone J, Rajpurkar P, Topol J. \(2022\) \*Multimodal biomedical AI\*. Nature Medicine, 28\(9\), 1773–1784. <https://doi.org/10.1038/s41591-022-01956-6>](#)
- [Austrian S, Jamin T, Doty R, Blecker S. \(2018\) \*Impact of an emergency department electronic sepsis surveillance system on patient mortality and length of stay\*. J Am Med Inform Assoc. 25\(5\):523-529. doi: 10.1093/jamia/ocx072. PMID: 29025165; PMCID: PMC7646853.](#)
- [Bach M, Werner A, Palt M. \(2019\) \*The proposal of undersampling method for learning from imbalanced datasets\*. Procedia Comput. Sci. 159, 125–134. doi:10.1016/j.procs.2019.09.167](#)
- [Baddoo D, Li Z, Odai N, Boni C, Nooni K, Andam-Akorful A. \(2021\) \*Comparison of Missing Data Infilling Mechanisms for Recovering a Real-World Single Station Streamflow Observation\*. Int J Environ Res Public Health. 18\(16\):8375](#)
- [Bagavathiappan S, Saravanan T, Philip J, Jayakumar T, Raj B, Karunanithi R, et al. \(2009\) \*Infrared Thermal Imaging for Detection of Peripheral Vascular Disorders\*. J. Med. Phys. 34 \(1\), 43. doi:10.4103/0971-6203.48720](#)
- [Bagchi S. \(2015\) \*India Has Low Doctor to Patient Ratio, Study Finds\*. BMJ 351, h5195. doi:10.1136/bmj.h5195](#)
- [Bai X, Yu W, Ji W, Lin Z, Tan S, Duan K, et al. \(2014\) \*Early versus delayed administration of norepinephrine in patients with septic shock\*. Crit Care.18\(5\)](#)

- Balakrishnan G, Durand F, Guttag J. (2013) *Detecting Pulse from Head Motions in Video*. IEEE Conference on Computer Vision and Pattern Recognition, 3430–3437. doi:10.1109/CVPR.2013.440
- Barbar D, Binquet C, Monchi M, Bruyère R, Quenot P. (2014) *Impact on mortality of the timing of renal replacement therapy in patients with severe acute kidney injury in septic shock: the IDEAL-ICU study*. *Trials*. 15(1):270
- Barbosa C, Czaplik M, Blazek V, Leonhardt S, Teichmann D. (2018) *Monitoring of Cardiorespiratory Signals Using Thermal Imaging: A Pilot Study on Healthy Human Subjects*. *Sensors* 18 (5), 1541. doi:10.3390/s18051541
- Barth S. *Artificial Intelligence (AI) in Healthcare & Medical Field*. ForeSee Medical. Retrieved from <https://www.foreseemed.com/artificial-intelligence-in-healthcare>
- Bates M, Mächler M, Bolker M, Walker C. (2015) *Fitting Linear Mixed-Effects Models Using {lme4}*. *J. Statist. Software* 67 (1), 1–48. doi:10.18637/jss.v067.i01
- Beger. *The crucial role of explainability in healthcare AI*. *European Journal of Radiology*. Volume 176, 111507
- Berger T, Green J, Horeczko T, Hagar Y, Garg N, Suarez A, et al. (2013) *Shock index and early recognition of sepsis in the emergency department: pilot study*. *West J Emerg Med*. 14(2):168–74
- Birkhahn H, Gaeta J, Terry D, Bove J, Tloczkowski J. (2005) *Shock index in diagnosing early acute hypovolemia*. *Am J Emerg Med*. 23(3):323–6
- Bohr A, Memarzadeh K. (2020) *The rise of artificial intelligence in healthcare applications*. *Artif. Intell. Healthc.*, 25–60. doi:10.1016/B978-0-12-818438-7.00002-2
- Bourcier S, Pichereau C, Boelle Y, Nemlaghi S, Dubée V, Lejour G, et al. (2016) *Toe-to-room Temperature Gradient Correlates with Tissue Perfusion and Predicts Outcome in Selected Critically Ill Patients with Severe Infections*. *Ann. Intensive Care* 6 (1), 63. doi:10.1186/s13613-016-0164-2
- Box P, Jenkins M, Reinsel C, Ljung M. (2015) *Time Series Analysis: Forecasting and Control* (5th ed.). Wiley
- Breiman L. (2001) *Random forests*. *Mach Learn*. 45(1):5–32
- Brown J, Brugger H, Boyd J, Paal P. (2012) *Accidental hypothermia*. *New England Journal of Medicine*, 367(20), 1930–1938. <https://doi.org/10.1056/NEJMra1114208>

- [Burdick H, Pino E, Gabel-Comeau D, McCoy A, Gu C, Roberts J, Le S, Slote J, Pellegrini E, Green-Saxena A, Hoffman J, Das R. \(2020\) \*Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals\*. BMJ Health Care Inform. 27\(1\):e100109. doi: 10.1136/bmjhci-2019-100109. PMID: 32354696; PMCID: PMC7245419.](#)
- [Chawla V, Bowyer W, Hall O, Kegelmeyer P. \(2002\) \*Smote: Synthetic Minority Over-sampling Technique\*. J. Artif. Intell. Res. 16, 321–357. doi:10.1613/jair.953](#)
- [Che Z, Purushotham S, Cho K, et al. \(2018\) \*Recurrent Neural Networks for Multivariate Time Series with Missing Values\*. Sci Rep 8, 6085. <https://doi.org/10.1038/s41598-018-24271-9>](#)
- [Cheng H, Sie D, Hsu H, Goh L, Chien Y, Chen Y, et al. \(2020\) \*Shock Index: A Simple and Effective Clinical Adjunct in Predicting 60-Day Mortality in Advanced Cancer Patients at the Emergency Department\*. Int J Environ Res Public Health. 17\(13\):4904](#)
- [Chen H, Asch M. \(2017\). \*Machine learning and prediction in medicine-beyond the peak of inflated expectations\*. N. Engl. J. Med. 376, 2507–2509. doi:10.1056/NEJMp1702071](#)
- [Chen T, Guestrin C. \(2016\) \*XGBoost: A scalable tree boosting system\*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, San Francisco, United States \(New York, NY, USA: Association for Computing Machinery\), 785–794. doi:10.1145/2939672.2939785](#)
- [Chen W, McDuff D. \(2018\) \*DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks\*. Munich, Germany: ECCV](#)
- [Choi E, Bahadori T, Schuetz A, Stewart F, Sun J. \(2016\) \*Doctor AI: Predicting Clinical Events via Recurrent Neural Networks\*. Proceedings of the 1st Machine Learning for Healthcare Conference \(pp. 301–318\). PMLR. <https://proceedings.mlr.press/v56/Choi16.html>](#)
- [Cho Y, Julier J, Bianchi-Berthouze N. \(2019\) \*Instant Stress: Detection of Perceived Mental Stress through Smartphone Photoplethysmography and Thermal Imaging\*. JMIR Ment. Health 6 \(4\), e10140. doi:10.2196/10140](#)
- [Christ M, Braun N, Neuffer J, Kempa-Liehr A. \(2018\) \*Time series FeatuRe extraction on basis of scalable hypothesis tests \(tsfresh – a Python package\)\*. Neurocomputing 307, 72–77. doi:10.1016/j.neucom.2018.03.067](#)
- [Christ M, Kempa-Liehr W, Feindt M. \(2016\) \*Distributed and Parallel Time Series Feature Extraction for Industrial Big Data Applications\*. ArXiv. e-prints: 1610.07717. doi:10.48550/arXiv.1610.07717](#)

[Collins S, Reitsma B, Altman G, Moons G. \(2015\) \*Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis \(TRIPOD\): the TRIPOD Statement\*. BMC Med. 13\(1\):1](#)

[Condat L. \(2013\) \*A Direct Algorithm for 1-D Total Variation Denoising\*. IEEE Signal Process. Lett. 20, 1054–1057. doi:10.1109/LSP.2013.2278339](#)

[Cox R. \(1972\) \*Regression models and life-tables\*. Journal of the Royal Statistical Society: Series B \(Methodological\), 34\(2\), 187–202. <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>](#)

[Cutler A, Cutler R, Stevens R. \(2012\) \*Random Forests\*. New York: Springer, 157–175. doi:10.1007/978-1-4419-9326-7\\_5](#)

[Dai D, Feinstein A, Morrison W, Zuppa F, Feudtner C. \(2016\) \*Epidemiology of polypharmacy and potential drug-drug interactions among pediatric patients in intensive care units of U.S. children's hospitals\*. Pediatr Crit Care Med.17:218–228](#)

[Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. \(2016\) \*Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach\*. JMIR Med Inform. 4\(3\):e28](#)

[Dhanalakshmi R, Sudalaimuthu T, Radhakrishnan R. \(2022\) \*Early Detection of Sepsis Using LSTM and Reinforcement Learning\*. In: Deepak, B.B.V.L., Parhi, D., Biswal, B., Jena, P.C. \(eds\) \*Applications of Computational Methods in Manufacturing and Product Design\*. Lecture Notes in Mechanical Engineering. Springer, Singapore](#)

[Dietrichs S, McGlynn K, Allan A, Connolly A, Bishop M, Burton F, et al. \(2020\) \*Moderate but not severe hypothermia causes pro-arrhythmic changes in cardiac electrophysiology\*. Cardiovasc. Res. 116, 2081–2090. doi:10.1093/cvr/cvz309](#)

[Divatia V, Amin R, Ramakrishnan N, Kapadia N, Todi S, Sahu S, et al. \(2016\) \*Intensive Care in India: The Indian Intensive Care Case Mix and Practice Patterns Study\*. Indian J. Crit. Care Med. 20, 216–225. doi:10.4103/0972-5229.180042](#)

[Xu Y, Chen H, Liu Q, Pan C, Xu P, Han B, Xie F, Huang Z, Guo M, Yang Y, Qiu B. \(2016\) \*The Effect of Early Goal-Directed Therapy on Treatment of Critical Patients with Severe Sepsis/septic Shock: a Multi-Center, Prospective, Randomized, Controlled Study\*. Anesth Analg. 2016 Aug;123\(2\):371-81. doi: 10.1213/ANE.0000000000001278. PMID: 27049857; PMCID: PMC4956677.](#)

[Falotico M, Shinozaki K, Saeki K, Becker B. \(2020\) \*Advances in the Approaches Using Peripheral Perfusion for Monitoring Hemodynamic Status\*. Front. Med. 7, 614326. doi:10.3389/fmed.2020.614326](#)

- Fan E, Brodie D, Slutsky S. (2018) *Acute respiratory distress syndrome: advances in diagnosis and treatment*. JAMA. 319(7):698–710. doi:10.1001/jama.2017.21907
- Fatteh N, Sutherland E, Santos G, Zeidan R, Gastesi P, Naranjo D. (2021) *Association of hypothermia with increased mortality rate in SARS-CoV-2 infection*. Int. J. Infect. Dis. 108, 167–170. doi:10.1016/j.ijid.2021.05.031
- Faulds M, Meekings T. (2013) *Temperature management in critically ill patients*. Continuing Educ. Anaesth. Crit. Care & Pain 13, 75–79. doi:10.1093/bjaceaccp/mks063
- Fears R. (2011) *Hypothermia*. Merck Manual Professional Version. <https://www.merckmanuals.com/professional/injuries-poisoning/hypothermia-and-hyperthermia/hypothermia>
- Fernandes M, Silva M, Sedda C, Silva P, Martins R, Oliveira G. (2019) *Assessment of the risk of QT-interval prolongation associated with potential drug-drug interactions in patients admitted to Intensive Care Units*. Saudi Pharm J. 27(2):229–234
- Finlayson G, Subbaswamy A, Singh K, Bowers J, Kuprel B, Saria S, Kohane S. (2021) *The clinician and dataset shift in artificial intelligence*. The New England Journal of Medicine, 385(3), 283–286. <https://doi.org/10.1056/NEJMc2104626>
- Finlayson G, Bowers D, Ito J, Zittrain L, Beam L, Kohane S. (2019) *Adversarial attacks on medical machine learning*. Science. 363(6433):1287-1289. doi: 10.1126/science.aaw4399. PMID: 30898923; PMCID: PMC7657648.
- FLIR Systems. (2016) FLIR Systems, 70 Castilian Dr., Goleta, California 93117. Available at: <http://www.flir.com> (Accessed December 23, 2021)
- Friedman H. (2001) *Greedy function approximation: A gradient boosting machine*. Ann Stat.
- Froesel M, Goudard Q, Hauser M, Gacoin M, Ben S. (2020) *Automated Video-Based Heart Rate Tracking for the Anesthetized and Behaving Monkey*. Sci. Rep. 10, 17940. doi:10.1038/s41598-020-74954-5
- Gerke S, Minssen T, Cohen G. (2020) *Ethical and legal challenges of artificial intelligence-driven healthcare*. Artificial Intelligence in Healthcare. 295–336. doi: 10.1016/B978-0-12-818438-7.00012-5. Epub. PMCID: PMC7332220.
- Getachew H, Assen M, Dula F, Bhagavathula S. (2016) *Potential drug–drug interactions in pediatric wards of Gondar University Hospital, Ethiopia: A cross-sectional study*. Asian Pac J Trop Biomed; 6(6):534–538

[Ghassemi M, Naumann T, Schulam P, Beam L, Chen Y, Ranganath R. \(2015\) \*Opportunities in machine learning for healthcare\*. arXiv preprint arXiv:1707.06269.](#)

[Giesbrecht G. \(2001\) \*Emergency treatment of hypothermia\*. Emerg. Med. 13, 9–16. doi:10.1046/j.1442-2026.2001.00172.x](#)

[Goldstein A, Navar M, Pencina J, Ioannidis A. \(2017\) \*Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review\*. Journal of the American Medical Informatics Association, 24\(1\), 198–208. <https://doi.org/10.1093/jamia/ocw042>](#)

[Grattafiori A, Dubey A, Jauhri A, et al. \(2024\) \*The Llama 3 herd of models\*. arXiv preprint](#)

[Greenhalgh T, Wherton J, Papoutsis C, Lynch J, Hughes G, A'Court C, Hinder S, Fahy N, Procter R, Shaw S. \(2017\) \*Beyond adoption: A new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies\*. J Med Internet Res;19\(11\):e367. doi: 10.2196/jmir.8775](#)

[Grinsztajn L, Oyallon E, Varoquaux G. \(2023\) \*Why do tree-based models still outperform deep learning on typical tabular data?\* \[https://openreview.net/forum?id=Fp7\\\_\\\_phQszn\]\(https://openreview.net/forum?id=Fp7\_\_phQszn\)](#)

[Guazzi R, Villarroel M, Jorge J, Daly J, Frise C, Robbins A, et al. \(2015\) \*Non-contact Measurement of Oxygen Saturation with an RGB Camera\*. Biomed. Opt. Express 6 \(9\), 3320–3338. doi:10.1364/BOE.6.003320](#)

[Gultepe E, Green P, Nguyen H, Adams J, Albertson T, Tagkopoulos I. \(2014\) \*From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system\*. J. Am. Med. Inform. Assoc. 21, 315–325. doi:10.1136/amiajnl-2013-001815](#)

[Guo L, Pfohl R, Fries J. et al. \(2022\) \*Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine\*. Sci Rep 12, 2726. <https://doi.org/10.1038/s41598-022-06484-1>](#)

[Hanff C, Horst A, Linden N, Vries H, Streefkerk J, Muller E, Brüggemann J. \(2022\) \*Real-time bedside decision support improves antibiotic target attainment in critically ill patients: The AutoKinetics randomized controlled trial\*. Critical Care Medicine, 50\(12\), 1736–1746.](#)

[Hanna T. \(2024\) \*What is PACS \(picture archiving and communication system\)?\* | Definition from TechTarget. TechTarget](#)

[Hao T, Huang Z, Liang L, Weng H, Tang B. \(2021\) \*Health Natural Language Processing: Methodology Development and Applications\*. JMIR Med Inform; 9\(10\):e23898. doi: 10.2196/23898. PMID: 34673533; PMCID: PMC8569540](#)



- [Harutyunyan H, Khachatrian H, Kale C. et al. \(2019\) \*Multitask learning and benchmarking with clinical time series data\*. Sci Data 6, 96](#)
- [Hardan S, Shaaban A, Abdalla J, Yaqub M. \(2024\) \*Affordable and real-time antimicrobial resistance prediction from multimodal electronic health records\*. Scientific Reports, 14, 16464. <https://doi.org/10.1038/s41598-024-66812-5>](#)
- [Hasanin A, Mukhtar A, Nassar H. \(2017\) \*Perfusion Indices Revisited\*. J. Intensive Care 5, 24. doi:10.1186/s40560-017-0220-5](#)
- [Hassanzad M, Arenas S, Baniyadi S. \(2018\) \*Potential Drug-Drug Interactions Among Critically Ill Pediatric Patients in a Tertiary Pulmonary Center\*. J Clin Pharmacol; 58\(2\):221–227](#)
- [Hassanzad M, Nejad T, Mahboobipour A, Salem F, Baniyadi S. \(2020\) \*Potential drug–drug interactions in hospitalized pediatric patients with respiratory disorders: a retrospective review of clinically important interactions\*. Drug Metab Pers Ther;35\(1\)](#)
- [Hastie T, Rosset S, Zhu J, Zou H. \(2009\) \*Multi-class AdaBoost\*. Stat. Interface 2, 349–360. doi:10.4310/SII.2009.v2.n3.a8](#)
- [He K, Zhang X, Ren S, Sun J. \(2016\) \*Deep Residual Learning for Image Recognition\*. IEEE Conference on Computer Vision and Pattern Recognition \(CVPR\), Las Vegas, United States. 770–778. doi:10.1109/cvpr.2016.90](#)
- [Henry E, Hager N, Pronovost J, Saria S. \(2015\) \*A targeted real-time early warning score \(TREWScore\) for septic shock\*. Sci. Transl. Med. 7, 299ra122. doi:10.1126/scitranslmed.aab3719](#)
- [Herget-Rosenthal S, Saner F, Chawla S. \(2008\) \*Approach to Hemodynamic Shock and Vasopressors\*. Clin. J. Am. Soc. Nephrol. 3 \(2\), 546–553. doi:10.2215/CIN.01820407](#)
- [Hochreiter S, Schmidhuber J. \(1997\) \*Long Short-Term Memory\*. Neural Comput. 9 \(8\), 1735–1780. doi:10.1162/neco.1997.9.8.1735](#)
- [Houwink I, Rijkenberg S, Bosman J, Voort J. \(2016\) \*The Association between Lactate, Mean Arterial Pressure, Central Venous Oxygen Saturation and Peripheral Temperature and Mortality in Severe Sepsis: a Retrospective Cohort Analysis\*. Crit. Care 20, 56. doi:10.1186/s13054-016-1243-3](#)
- [Hyland L, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. \(2020\) \*Early prediction of circulatory failure in the intensive care unit using machine learning\*. Nat Med; 26\(3\):364–73](#)
- [Ibrahim M, Farah M, Osman O, Hashi A. \(2021\) \*The effect of admission hypothermia for neonatal death among neonates admitted to neonatal intensive care unit at sheik hassan yabare\*](#)



[jigjiga university referral hospital in jigjiga city, Somali region, eastern Ethiopia. Res. Rep. Neonatol. 11, 43–55. doi:10.2147/RRN.S297470](#)

[Alsuyayfi S, Alanazi A. \(2022\) Impact of clinical alarms on patient safety from nurses' perspective. Journal of Taibah University Medical Sciences, 17\(5\), 788–794. https://www.sciencedirect.com/science/article/pii/S2352914822001873](#)

[Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Ng Y. \(2021\) CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence, 33\(01\), 590–597. https://doi.org/10.1609/aaai.v33i01.3301590](#)

[Ismail M, Aziz S, Noor S, Haider I, Shams F, Haq I, Khadim F, Khan Q, Khan F, Asif M. \(2017\) Potential drug–drug interactions in pediatric patients admitted to intensive care unit of Khyber Teaching Hospital, Peshawar, Pakistan: A cross–sectional study. J Crit Care. 40:243–250](#)

[Ismail M, Iqbal Z, Khan I, Javaid A, Arsalan H, Farhadullah, Khan F, Khan Z, Nasir F, Khan A. \(2013\) Frequency, Levels and Predictors of Potential Drug-Drug Interactions in a Pediatrics Ward of a Teaching Hospital in Pakistan. Trop J Pharm Res;12:401–406](#)

[Ryu J, Kim H, Lee S. Deep learning improves prediction of drug–drug and drug–food interactions. Proc. Natl. Acad. Sci. U.S.A. 115 \(18\) E4304-E4311](#)

[Ji Y, Zhou Z, Liu H, Davuluri V. \(2021\) DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics, 37\(15\), 2112–2120. https://doi.org/10.1093/bioinformatics/btab](#)

[Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. \(2017\) Artificial intelligence in healthcare: Past, present and future. Stroke Vasc. Neurol. 2, 230–243](#)

[Jiang H, Lin Y, Ren W, Fang Z, Liu Y, Tan X, Lv X, Zhang N. \(2022\) Adverse drug reactions and correlations with drug–drug interactions: A retrospective study of reports from 2011 to 2020. Front. Pharmacol. 13:923939. doi: 10.3389/fphar.2022.923939](#)

[Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. \(2014\) A survey on concept drift adaptation. ACM Comput. Surv. 46, 4, Article 44, 37 pages. https://doi.org/10.1145/2523813](#)

[Johnson W, Pollard J, Shen L, Lehman H, Feng M, Ghassemi M, Mark G. \(2016\) MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035. https://doi.org/10.1038/sdata.2016.35](#)

[Johnson W, Ghassemi M, Nemati S, Niehaus E, Clifton A, Clifford D. \*Machine Learning and Decision Support in Critical Care\*. Proc IEEE Inst Electr Electron Eng.;104\(2\):444-466. doi: 10.1109/JPROC.2015.2501978. Epub. PMID: 27765959; PMCID: PMC5066876.](#)

[Josephson B, Wong S, White A. \(2023\) \*Algorithmic fairness in healthcare: Quality, equity and justice in the era of clinical AI\*. Nature Biomedical Engineering. <https://doi.org/10.1038/s41551-023-01056-8>](#)

[Jorge J, Villarroel M, Tomlinson H, Gibson O, Darbyshire L, Ede J, et al. \(2022\) \*Non-contact Physiological Monitoring of Post-operative Patients in the Intensive Care Unit\*. npj Digit. Med. 5, 4](#)

[Kalimoultou A, Kennedy N, Feng J, et al. \(2025\) \*Optimal Vasopressin Initiation in Septic Shock: The OVISS Reinforcement Learning Study\*. JAMA. ;333\(19\):1688–1698. doi:10.1001/jama.2025.3046](#)

[Kardas P, Urbanski F, Lichwierowiez A, Chudzynska E, Czech M, Makowska K, Kardas G. \(2021\) \*The prevalence of selected potential drug–drug interactions of analgesic drugs and possible methods of preventing them: Lessons learned from the analysis of the real–world national database of 38 Million citizens of Poland\*. Front Pharmacol; 11:607852](#)

[Karimi D, Dou H, Warfield K, Gholipour A. \(2020\) \*Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis\*. Medical Image Analysis, 65, 101759. <https://doi.org/10.1016/j.media.2020.101759>](#)

[Elazar Y, Kassner N, Ravfogel S, Ravichander A, Hovy E, Schütze H, Goldberg Y. \(2021\) \*Measuring and improving consistency in pretrained language models\*. Transactions of the Association for Computational Linguistics, 9, 686–701. \[https://doi.org/10.1162/tacl\\\_a\\\_00410\]\(https://doi.org/10.1162/tacl\_a\_00410\)](#)

[Kelly J, Karthikesalingam A, Suleyman M, Corrado G, King D. \(2019\) \*Key challenges for delivering clinical impact with artificial intelligence\*. BMC Med. Oct 29;17\(1\):195.](#)

[Kiekkas P, Fligou F, Igoumenidis M, Stefanopoulos N, Konstantinou E, Karamouzou ., et al. \(2018\) \*Inadvertent hypothermia and mortality in critically ill adults: Systematic review and meta-analysis\*. Aust. Crit. Care 31, 12–22. doi:10.1016/j.aucc.2017.01.008](#)

[Khor S. et al. \(2023\) \*Fairness in predicting cancer recurrence across racial and ethnic subgroups: A colorectal cancer recurrence risk prediction model evaluation\*. JAMA Network Open, 6\(6\), e2318495. <https://doi.org/10.1001/jamanetworkopen.2023.18495>](#)

[Koch E, Lovett S, Nghiem T, Riggs A, Rech A. \(2019\) \*Shock index in the emergency department: utility and limitations\*. Open Access Emerg Med OAEM. 11:179–99](#)

- [Komorowski M, Celi A, Badawi O, Gordon C, Faisal A. \(2018\) \*The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care\*. Nature Medicine, 24\(11\), 1716–1720](#)
- [Kormilitzin A, Vaci N, Liu Q, Nevado-Holgado A. \(2021\) \*Med7: A transferable clinical natural language processing model for electronic health records\*. Artificial Intelligence in Medicine, 118, 102086](#)
- [Krečak I, Pivac L, Lucijanić M, Skelin M. \*Polypharmacy, potentially inappropriate medications, and drug-to-drug interactions in patients with chronic myeloproliferative neoplasms\*. Biomedicines. 2023;11\(5\):1301](#)
- [Krumholz M. \(2014\) \*Big data And new knowledge in medicine: The thinking, training, and tools needed for a learning health system\*. Health Aff. \(Millwood\) 33, 1163–1170. doi:10.1377/hlthaff.2014.0053](#)
- [Kursa B, Rudnicki R. \(2010\) \*Feature selection with the boruta package\*. J Stat Softw.; 36\(11\):1–13.](#)
- [Labroo B, Paine F, Thummel E, Kharasch D. \(1997\) \*Fentanyl metabolism by human hepatic and intestinal cytochrome P450 3A4: implications for interindividual variability in disposition, efficacy, and drug interactions\*. Drug Metab Dispos; 25\(9\):1072–1080. PMID: 9311623](#)
- [Langerova P, Prokes M, Konvalinka M, Furstova J, Urbanek K. \(2013\) \*Incidence of potential drug interactions in medication prescriptions for children and adolescents in the University Hospital Olomouc, Czech Republic\*. Eur J Pediatr. 2013;172\(5\):631–638](#)
- [Lehman H, Saeed M, Talmor D, Mark R, Malhotra A. \(2013\) \*Methods of Blood Pressure Measurement in the ICU\*. Crit. care Med. 41 \(1\), 34–40. doi:10.1097/CCM.0b013e318265ea46](#)
- [Lekadir K, Frangi F, Porras R, Glocker B, Cintas C, Langlotz P, et al. \(2024\) \*FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare\*. BMJ. 2025 Feb 5;388:e081554. doi: 10.1136/bmj-2024-081554. Erratum in: BMJ. 2025 Feb 17;388:r340. doi: 10.1136/bmj.r340. PMID: 39909534; PMCID: PMC11795397.](#)
- [Li M, Jiang Y, Zhang Y, Zhu H. \(2023\) \*Medical image analysis using deep learning algorithms\*. Front. Public Health. 11:1273253. doi: 10.3389/fpubh.2023.1273253](#)
- [Lima A, Bakker J. \(2005\) \*Noninvasive Monitoring of Peripheral Perfusion\*. Intensive Care Med. 31 \(10\), 1316–1326. doi:10.1007/s00134-005-2790-2](#)
- [Lima A, Jansen C, Bommel J, Ince C, Bakker J. \(2009\) \*The Prognostic Value of the Subjective Assessment of Peripheral Perfusion in Critically Ill Patients\*. Crit. Care Med. 37 \(3\), 934–938. doi:10.1097/CCM.0b013e31819869db](#)

- [Lima C, Camarinha D, Ferreira C, Panisset G, Belmino R, Silva T, Lopes C. \(2020\) \*Severe Potential Drug–Drug Interactions and the Increased Length of Stay of Children in Intensive Care Unit\*. Front Pharmacol;11:555407](#)
- [Liu R, Greenstein J, Fackler J, Bergmann J, Bembea M, Winslow R. \(2021\) \*Offline reinforcement learning with uncertainty for treatment strategies in sepsis\*. doi.org/10.48550/arXiv.2107.04491](#)
- [Wen Q, Zhou T, Zhang C, Chen W, Ma Z, Yan J, Sun L. \(2023\) \*Transformers in time series: A survey\*. Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence \(IJCAI-23\). <https://doi.org/10.24963/ijcai.2023/759arXiv+3>](#)
- [Lim C, Miliya T, Chansamouth V, Aung T, Karkey A, Teparrukkul P, et al. \(2020\) \*Automating the Generation of Antimicrobial Resistance Surveillance Reports: Proof-of-Concept Study Involving Seven Hospitals in Seven Countries\*. J Med Internet Res; 22\(10\):e19762. doi: 10.2196/19762. PMID: 33006570; PMCID: PMC7568216.](#)
- [Liu H, Allen J, Zheng D, Chen F. \(2019\) \*Recent Development of Respiratory Rate Measurement Technologies\*. Physiol. Meas. 40, 07TR01. doi:10.1088/1361-6579/ab299e](#)
- [Liu H, Chen F, Hartmann V, Khalid G, Hughes S, Zheng D. \(2020\) \*Comparison of Different Modulations of Photoplethysmography in Extracting Respiratory Rate: from a Physiological Perspective\*. Physiol. Meas. 41, 094001. doi:10.1088/1361-6579/abaaf0](#)
- [Liu L, Shah S, Zhao G, Yang X. \(2018\) \*Respiration Symptoms Monitoring in Body Area Networks\*. Appl. Sci. 8 \(4\), 568. doi:10.3390/app8040568](#)
- [Lundberg, S. M., & Lee, S.-I. \(2017\). A unified approach to interpreting model predictions. In \*Advances in Neural Information Processing Systems\* \(Vol. 30, pp. 4765–4774\).](#)
- [Lu J, Liu A, Dong F, Gama J, Zhang G. \(2018\) \*Learning under concept drift: A review\*. IEEE Transactions on Knowledge and Data Engineering, 31\(12\), 2346–2363. <https://doi.org/10.1109/TKDE.2018.2876857>](#)
- [Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, Liu Y. \(2022\) \*BioGPT: Generative pre-trained transformer for biomedical text generation and mining\*. Briefings in Bioinformatics, 23\(6\), bbac409.](#)
- [Maheshwari K, Nathanson H, Munson H, Hwang S, Yapici O, Stevens M, et al. \(2020\) \*Abnormal shock index exposure and clinical outcomes among critically ill patients: A retrospective cohort analysis\*. J Crit Care. Jun;57:5–12.](#)
- [Maithani L, Prasad S. \(2025\) \*Government of India. India Budget\*. Retrieved from <https://www.indiabudget.gov.in/economicsurvey/doc/echapter.pdf>](#)

[Majumder S, Chen L, Marinov O, Chen H, Mondal T, Deen J. \(2018\) \*Noncontact Wearable Wireless ECG Systems for Long-Term Monitoring\*. IEEE Rev. Biomed. Eng. 11, 306–321. doi:10.1109/RBME.2018.2840336](#)

[Mambou S, Maresova P, Krejcar O, Selamat A, Kuca K. \(2018\) \*Breast Cancer Detection Using Infrared Thermal Imaging and a Deep Learning Model\*. Sensors 18 \(9\), 2799. doi:10.3390/s18092799](#)

[Mao Q, Jay M, Hoffman L, Calvert J, Barton C, Shimabukuro D, et al. \(2018\) \*Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU\*. BMJ Open. 2018;8:17833.](#)

[Ribeiro M, Singh S, Guestrin C. \(2016\) \*Why Should I Trust You?: Explaining the Predictions of Any Classifier\*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining \(KDD '16\). Association for Computing Machinery, New York, NY, USA, 1135–1144. <https://doi.org/10.1145/2939672.2939778>](#)

[Martin C, Medam S, Antonini F, Alingrin J, Haddam M, Hammad E, et al. \(2015\) \*Norepinephrine: Not too much, too long\*. Shock 44, 305–309. doi:10.1097/SHK.0000000000000426](#)

[Martin K, Weiss L. \(2015\) \*Initial Resuscitation and Management of Pediatric Septic Shock\*. Minerva Pediatr. 67 \(2\), 141–158.](#)

[Al-Ansari A, Nejad A, Al-Nasr J, Prithula J, Rahman T, Hasan A, Chowdhury H, Alam F. \(2025\) \*Predicting ICU Mortality Among Septic Patients Using Machine Learning Technique\*. Journal of Clinical Medicine, 14\(10\), 3495. <https://doi.org/10.3390/jcm14103495>](#)

[Medscape Drugs & Diseases - Comprehensive peer-reviewed medical condition, surgery, and clinical procedure articles with symptoms, diagnosis, staging, treatment, drugs and medications, prognosis, follow-up, and pictures. Retrieved from <https://reference.medscape.com/>](#)

[Mendrala K, Kosiński S, Podsiadło P, Pasquier M, Paal P, Mazur P, et al. \(2021\) \*The efficacy of renal replacement therapy for rewarming of patients in severe accidental hypothermia systematic review of the literature\*. Int. J. Environ. Res. Public Health 18, 9638. doi:10.3390/ijerph18189638](#)

[Moor M, Banerjee O, Abid A, Miura Y, Wang Y, Re C, Zitnik M. \(2023\) \*Foundation models for generalist medical artificial intelligence\*. Nature, 617\(7960\), 759–765. <https://doi.org/10.1038/s41586-023-06042-2>](#)

[Moor M, Bennett N, Plečko D, Horn M, Rieck B, Meinshausen N, Bühlmann P, Borgwardt K. \(2023\) \*Predicting sepsis using deep learning across international sites: A retrospective development and validation study\*. eClinicalMedicine, 62, 102124. <https://doi.org/10.1016/j.eclinm.2023.102124>](#)

[Morales O, Jasso L, Reyes A, Garduno J, Munoz O. \(2018\) \*Potential drug–drug interactions and their risk factors in pediatric patients admitted to the emergency department of a tertiary care hospital in Mexico\*. PLoS One;13\(1\):e0190882. <https://doi.org/10.1371/journal.pone.0190882>](#)

[Moritz S, Bartz-Beielstein T. \(2017\) \*imputeTS: Time series missing value imputation in R\*. R. J. 9, 207. doi:10.32614/RJ-2017-009](#)

[Mousavi S, Ghanbari G. \(2017\) \*Potential drug–drug interactions among hospitalized patients in a developing country\*. Casp J Intern Med.; 8:282–288. <https://doi.org/10.22088/cjim.8.4.282>](#)

[Murugavel M. \(2019\) \*Yolo-Annotation-Tool-New\*. Available at: <https://github.com/ManivannanMurugavel/Yolo-Annotation-Tool-New>.](#)

[Musi E, Sheets A, Zafren K, Brugger H, Paal P, Hölzl N, et al. \(2021\) \*Clinical staging of accidental hypothermia: The revised Swiss system: Recommendation of the international commission for mountain emergency medicine \(ICAR MedCom\)\*. Resuscitation 162, 182–187. doi:10.1016/j.resuscitation.2021.02.038](#)

[Myint K, Sheng S, Xian Y, Matsouaka A, Reeves J, Saver L, et al. \(2018\) \*Shock Index Predicts Patient-Related Clinical Outcomes in Stroke\*. J Am Heart Assoc; 7\(18\):e007581.](#)

[Nachman D, Constantini K, Poris G, Wagnert-Avraham L, Gertz D, Littman R, et al. \(2020\) \*Wireless, Non-invasive, Wearable Device for Continuous Remote Monitoring of Hemodynamic Parameters in a Swine Model of Controlled Hemorrhagic Shock\*. Sci. Rep. 10, 17684. doi:10.1038/s41598-020-74686-6](#)

[Nagori A, Patel D, Soni M, Sharma P, Bhamare D, Ghosh S. \(2019\) \*Predicting hemodynamic shock from thermal images using machine learning\*. Scientific Reports, 9, 14000. <https://doi.org/10.1038/s41598-018-36586-8>](#)

[CMS. \(2024\) \*National Health Expenditures data: Historical\*. Retrieved from, <https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/historical>](#)

[Navidi W. \(1998\) \*Bidirectional case-crossover designs for exposures with time trends\*. Biometrics, 54\(2\), 596–605](#)

[Nemati S, Holder A, Razmi F, Stanley D, Clifford D, Buchman G. \*An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU\*. Crit Care Med. 2018 Apr;46\(4\):547–53.](#)

[Neupane S, Mitra S, Mittal S, Rahimi S. \(2025\) \*Towards a HIPAA compliant agentic AI system in healthcare\*. arXiv. <https://doi.org/10.48550/arXiv.2504.17669>](#)



- Ng K. (2009) *A Review of Thermography as Promising Non-invasive Detection Modality for Breast Tumor*. *Int. J. Therm. Sci.* 48 (5), 849–859. doi:10.1016/j.ijthermalsci.2008.06.015
- Noble S. (2006) *What is a support vector machine?* *Nat. Biotechnol.* 24, 1565–1567. doi:10.1038/nbt1206-1565 Available at: <https://www.nature.com/articles/nbt1206-1565>.
- Obermeyer Z, Emanuel J. (2016) *Predicting the Future Big Data, Machine Learning, and Clinical Medicine*. *The New England Journal of Medicine*, 375(13), 1216–1219. <https://doi.org/10.1056/NEJMp1606181>
- Obermeyer Z, LeenH. (2017) *Lost in thought the limits of the human mind and the future of medicine*. *The New England Journal of Medicine*, 377, 1209–1211.
- Obermeyer Z, Powers B, Vogeli C, Mullainathan S. (2019) *Dissecting racial bias in an algorithm used to manage the health of populations*. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax234>
- Oranges T, Dini V, Romanelli M. (2015) *Skin Physiology of the Neonate and Infant: Clinical Implications*. *Adv. Wound Care* 4 (10), 587–595. doi:10.1089/wound.2015.0642
- Ortiz-Dosal A, Kolosovas-Machuca S, Rivera-Vega R, Simón J, González J. (2014) *Use of Infrared Thermography in Children with Shock: A Case Series*. *SAGE Open Med. Case Rep.* 2, 2050313X1456177. doi:10.1177/2050313X14561779
- Paal P, Brugger H, Strapazzon G. (2018) *Chapter 33 - accidental hypothermia*. *Handbook of clinical neurology, thermoregulation: From basic neuroscience to clinical neurology, Part II*. Editor A. A. Romanovsky (Elsevier), 547–563. doi:10.1016/B978-0-444-64074-1.00033-1
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. (2019) *Pytorch: An Imperative Style, High-Performance Deep Learning Library*. *Advances in Neural Information Processing Systems* (Vancouver, Canada: Curran Associates, Inc.), 32, 8024–8035.
- Pauk J, Wasilewska A, Ihnatouski M. (2019) *Infrared Thermography Sensor for Disease Activity Detection in Rheumatoid Arthritis Patients*. *Sensors* 19 (16), 3444. doi:10.3390/s19163444
- Pavlidis I, Dowdall J, Sun N, Puri C, Fei J, Garbey M. (2007) *Interacting with Human Physiology. Comput. Vis. Image Underst.* 108 (1–2), 150–170. doi:10.1016/j.cviu.2006.11.018
- Pea F, Porreca L, Baraldo M, Furlanut M. (2000) *High vancomycin dosage regimens required by intensive care unit patients co-treated with drugs to improve haemodynamics following cardiac surgical procedures*. *J Antimicrob Chemother.* 45(3):329–335. doi:10.1093/jac/45.3.329. PMID: 10702552
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. (2012) *Scikit-learn: Machine Learning in Python*. *J Mach Learn Res.* 2012;12:2825–30.

[Pei L, Shashikumar P, Malhotra A, Nemati S, Wardi G. \(2023\) \*Making the improbable possible: Generalizing models designed for a syndrome-based, heterogeneous patient landscape\*. Critical Care Clinics, 39\(4\), 751–768. <https://doi.org/10.1016/j.ccc.2023.02.003>](#)

[Peyko V, Smalley S, Cohen H. \(2017\) \*Prospective comparison of acute kidney injury during treatment with the combination of piperacillin-tazobactam and vancomycin versus the combination of cefepime or meropenem and vancomycin\*. J Pharm Pract. 2017 Apr;30\(2\):209–213. doi:10.1177/0897190016628960. Epub 2016 Jul 9. PMID: 26912532](#)

[Pittman L, Ping S, Mark B. \(2004\) \*Arterial and Central Venous Pressure Monitoring\*. Int. Anesthesiol. Clin. 42 \(1\), 13–30. doi:10.1016/j.atc.2006.08.00810.1097/00004311-200404210-00004](#)

[Pollard J, Johnson W, Raffa D, Celi A, Mark G, Badawi O, Deliberato O. \(2018\) \*The eICU Collaborative Research Database, a freely available multi-center database for critical care research\*. Scientific Data, 5, 180178. <https://doi.org/10.1038/sdata.2018.178>](#)

[Potes C, Conroy B, Xu-Wilson M, Newth C, Inwald D, Frassica J. \(2017\) \*A clinical prediction model to identify patients at high risk of hemodynamic instability in the pediatric intensive care unit\*. Crit Care. 21\(1\). Available from: <https://pubmed.ncbi.nlm.nih.gov/29151364/>](#)

[Python Software Foundation \(2018\) \*Python Programming Language-Official Website\*. Available at: <http://www.python.org/> \(Accessed July 14, 2020\).](#)

[Qiu X, Bradshaw J, Li M, et al. \(2024\) \*VisionFM: a multi-modal medical foundation model for generalist ophthalmic AI\*. Biomedical Engineering Letters.](#)

[Quintana S, Fernández-Jaén T, Fernández-Cuevas I, Gómez-Carmona M, Arnaiz-Lastras J, Pérez D, et al. \(2015\) \*Infrared Thermography as a Support Tool for Screening and Early Diagnosis in Emergencies\*. J. Med. Imaging Hlth Inf. 5 \(6\), 1223–1228. doi:10.1166/jmihi.2015.1511](#)

[Rajkomar A, Oren E, Chen K, Dai M, Hajaj N, Hardt M, Liu X, Marcus J, Sun M, Sundberg P, Yee H, Zhang K, Corrado S, Shah H, Butte J, Dean J. \(2018\) \*Scalable and accurate deep learning with electronic health records\*. NPJ Digital Medicine, 1, 18. <https://doi.org/10.1038/s41746-018-0029-1> \[arxiv.org+8\]\(https://arxiv.org/abs/1808.07446\)](#)

[Rao C, Shenoy V, Udaykumar P. \(2019\). \*Potential Drug–Drug Interactions in the Pediatric Intensive Care Unit of a Tertiary Care Hospital\*. J Pharmacol Pharmacother;10\(2\):63–68.](#)

[Rasmy L, Xie Z, Hu C, Zhi D. \(2021\) \*Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction\*. NPJ Digital Medicine, 4\(1\), 86.](#)



[Rawson, Timothy M et al. \(2024\) \*Using digital health technologies to optimise antimicrobial use globally\*. The Lancet Digital Health, Volume 6, Issue 12, e914 - e925](#)

[Redmon J, Farhadi A. \(2018\) \*YOLOv3: An Incremental Improvement\*. CoRR.](#)

[Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, et al. \(2001\) \*Early Goal-Directed Therapy in the Treatment of Severe Sepsis and Septic Shock\*. N. Engl. J. Med. 345, 1368–1377. doi:10.1056/NEJMoa010307](#)

[Rockenschaub P, Hilbert A, Kossen T, Dincklage F, Madai I, Frey D. \(2023\) \*From single-hospital to multi-center applications: Enhancing generalisability of deep learning models for adverse event prediction in the ICU\* \[Preprint\]. arXiv. <https://doi.org/10.48550/arXiv.2303.15354>](#)

[Ronneberger O, Fischer P, Brox T. \(2015\) \*U-net: Convolutional Networks for Biomedical Image Segmentation\*. Munich, Germany: CoRR.](#)

[Ross M, Wills J, French A, Tivey R. \(2016\) \*Machine learning models improve risk stratification and identify high-risk patients with morbid illnesses\*. Journal of Biomedical Informatics, 61, 62–71.](#)

[Rotenstein S, Torre M, Ramos A, Rosales C, Guille C, Sen S, et al. \(2018\) \*Prevalence of Burnout Among Physicians: A Systematic Review\*. JAMA 320, 111131–111150. doi:10.1001/jama.2018.12777](#)

[Ruamviboonsuk, Paisan et al. \(2022\) \*Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study\*. The Lancet Digital Health, Volume 4, Issue 4, e235 - e244](#)

[Rudin C. \(2019\) \*Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead\*. Nat Mach Intell 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>](#)

[Rumbus Z, Matics R, Hegyi P, Zsiborás C, Szabo I, Illes A, et al. \(2017\) \*Fever is associated with reduced, hypothermia with increased mortality in septic patients: A meta-analysis of clinical trials\*. PLOS ONE 12, e0170152. doi:10.1371/journal.pone.0170152](#)

[Ruopp D, Perkins J, Whitcomb W, Schisterman F. \(2008\) \*Youden Index and Optimal Cut-point Estimated from Observations Affected by a Lower Limit of Detection\*. Biom. J. 50, 419–430. doi:10.1002/bimj.200710415](#)

[Rush B, Celi A, Stone J. \(2019\) \*Applying machine learning to continuously monitored physiological data\*. J. Clin. Monit. Comput. 33, 887–893. doi:10.1007/s10877-018-0219-z](#)

[Sagaidachnyi A, Fomin V, Usanov A, Skripal V. \(2017\) \*Thermography-based Blood Flow Imaging in Human Skin of the Hands and Feet: a Spectral Filtering Approach\*. \*Physiol. Meas.\* 38 \(2\), 272–288. doi:10.1088/1361-6579/aa4eaf](#)

[Schey M, Williams Y, Bucknall T. \(2010\) \*Skin Temperature and Core-Peripheral Temperature Gradient as Markers of Hemodynamic Status in Critically Ill Patients: a Review\*. \*Heart & Lung\* 39 \(1\), 27–40. doi:10.1016/j.hrtlng.2009.04.002](#)

[Schuster M, Paliwal K. \(1997\) \*Bidirectional recurrent neural networks\*. \*IEEE Trans Signal Process.\* 1997;45\(11\):2673–81.](#)

[Seek Thermal. \(2022\) \*Seek Thermal Compact, Seek Thermal Inc.\* Available at: <https://www.thermal.com/compact-series.html> \(Accessed January 22, 2022\).](#)

[Selvarani A, Suresh R. \(2019\) \*Infrared Thermal Imaging for Diabetes Detection and Measurement\*. \*J. Med. Syst.\* 43 \(2\), 23. doi:10.1007/s10916-018-1140-1](#)

[Sendak P, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. \(2020\) \*Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study\*. \*JMIR Med Inform.\* 8\(7\):e15182. doi: 10.2196/15182. PMID: 32673244; PMCID: PMC7391165](#)

[Sethi T, Nagori A, Bhatnagar A, Gupta P, Fletcher R, Lodha R. \(2017\) \*Validating the tele-diagnostic potential of affordable thermography in a big-data data-enabled ICU\*. \*Proceedings of the Special Collection on EGovernment Innovations in India, ICEGOV '17, March 7–March 9, 2017, 10th International Conference on Theory and Practice of Electronic Governance New Delhi AA India\* \(New York, NY, USA: Association for Computing Machinery: \), 64–69. 10.1145/3055219.3055234](#)

[Seymour W, Liu X, Iwashyna J, Brunkhorst M, Rea D, Scherag A, et al. \(2016\) \*Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock \(Sepsis-3\)\*. \*JAMA.\* 315\(8\):762–74.](#)

[Shah A, Yang X, Abbasi H. \(2019\) \*Cognitive Health Care System and its Application in Pill-rolling Assessment\*. \*Int. J. Numer. Model\* 32, 2632. doi:10.1002/jnm.2632](#)

[Shashikumar P, Wardi G, Paul P, Carlile M, Brenner N, Hibbert A, North M, Mukerji S, Robbins K, Shao P, Westover B, Nemati S, Malhotra A. \(2021\) \*Development and prospective validation of a deep learning algorithm for predicting need for mechanical ventilation\*. \*Chest.\* 159\(6\), 2264–2273.](#)

[Shickel B, Loftus J, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. \(2018\) \*DeepSOFA: A continuous acuity score for critically ill patients using clinically interpretable deep learning\*. \*Sci Rep\* 9, 1879 \(2019\). <https://doi.org/10.1038/s41598-019-38491-0>](#)

- Shortliffe H. (1976) *Computer-based medical consultations: MYCIN* (Doctoral dissertation, Stanford University). Retrieved from <https://people.dbmi.columbia.edu/~ehs7001/Buchanan-Shortliffe-1984/Chapter-15.pdf>
- Silva B, Araujo R, Arduini G, Alonso R, Shibata O, Troster J. (2013) *Adverse drug events in a pediatric intensive care unit: a prospective cohort*. *BMJ Open.*; 3(2):e001868. <https://doi.org/10.1136/bmjopen-2012-001868>
- Singer D. (2021) *Pediatric hypothermia: An ambiguous issue*. *Int. J. Environ. Res. Public Health* 18, 11484. doi:10.3390/ijerph182111484
- Singer M, Deutschman S, Seymour W, et al. (2016) *The third international consensus definitions for sepsis and septic shock (Sepsis-3)*. *JAMA*; 315(8):801–810. doi:10.1001/jama.2016.0287
- Singhal K, Azizi S, Tu T, et al. (2023) *Large language models encode clinical knowledge*. *Nature*, 616, 82–88. <https://doi.org/10.1038/s41586-023-05785-3>
- Smithburger L, Seybert L, Armahizer J, Kane L. (2010) *QT prolongation in the intensive care unit: commonly used medications and the impact of drug–drug interactions*. *Expert Opin Drug Saf.* 2010;9(5):699–712. <https://doi.org/10.1517/14740331003739188>
- Soenksen R, Ma Y, Zeng C. et al. (2022) *Integrated multimodal artificial intelligence framework for healthcare applications*. *npj Digit. Med.* 5, 149. <https://doi.org/10.1038/s41746-022-00689-4>
- Strickland E. (2019) *How IBM Watson Overpromised and Underdelivered on AI Health Care*. *IEEE Spectrum*.
- Subbaswamy A, Saria S. (2020) *From development to deployment: dataset shift, causality, and shift-stable models in health AI*. *Biostatistics*, 21(2), 345–352. <https://doi.org/10.1093/biostatistics/kxz041>
- Tan F, Chen S, Lyu W, Liu Z, Yu C, Lu C, et al. (2019) *Non-invasive Human Vital Signs Monitoring Based on Twin-Core Optical Fiber Sensors*. *Biomed. Opt. Express* 10 (11), 5940–5951. doi:10.1364/BOE.10.005940
- Tan W, Wang W, Zhou X. et al. (2024) *OntoMedRec: Logically-pretrained model-agnostic ontology encoders for medication recommendation*. *World Wide Web* 27, 28. <https://doi.org/10.1007/s11280-024-01268-1>
- Taylor A, Pare R, Venkatesh K, Mowafi H, Melnick R, Fleischman W, Hall K. (2016) *Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data–driven, machine learning approach*. *Acad. Emerg. Med.* 23, 269–278. doi:10.1111/acem.12876

[Thiruvengadam J, Mariamichael A. \(2018\) \*A Preliminary Study for the Assessment of Hypertension Using Static and Dynamic IR Thermograms\*. Biomed. Tech. 63 \(2\), 197–206. doi:10.1515/bmt-2016-0237](#)

[Tomašev N, Glorot X, Rae J, Zielinski M, Askham H, Saraiva A, Hassabis D. \(2019\) \*A clinically applicable approach to continuous prediction of future acute kidney injury\*. Nature, 572\(7767\), 116–119. <https://doi.org/10.1038/s41586-019-1390-1>](#)

[Seinen T, Fridgeirsson E, Ioannou S, Jeannetot D, John L, Kors J, et al. \(2022\) \*Use of unstructured text in prognostic clinical prediction models: a systematic review\*. Journal of the American Medical Informatics Association, Volume 29, Issue 7, Pages 1292–1302](#)

[Tonekaboni S, Joshi S, McCradden D, Goldenberg A. \(2019\) \*What clinicians want: Contextualizing explainable machine learning for clinical end use\*. Proceedings of Machine Learning Research, 1–21.](#)

[Topol J. \(2019\) \*High-performance medicine: the convergence of human and artificial intelligence\*. Nature Medicine, 25\(1\), 44–56. <https://doi.org/10.1038/s41591-018-0300-7>](#)

[Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, Monnet L, et al. \(2018\) \*WHO Pathogens Priority List Working Group. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis\*. The Lancet Infectious Diseases, 18\(3\), 318–327. \[https://doi.org/10.1016/S1473-3099\\(17\\)30753-3\]\(https://doi.org/10.1016/S1473-3099\(17\)30753-3\)](#)

[Touvron H, Lavril T, Izacard G, Martinet X, Lachaux A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample, G. \(2023\) \*LLaMA: Open and Efficient Foundation Language Models\*. arXiv preprint arXiv:2302.13971.](#)

[Truong R, Hidayat L, Bolaris A, Nguyen N L, Yamaki J, Pogue M. \(2021\) \*The antibiogram: key considerations for its development and utilization\*. JAC-Antimicrobial Resistance, 3\(2\), dlab060. <https://doi.org/10.1093/jacamr/dlab060>](#)

[Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang C, Natarajan V. \(2023\) \*Towards generalist biomedical AI\*. arXiv. <https://doi.org/10.48550/arXiv.2307.14334>](#)

[UpToDate Lexidrug \(formerly Lexicomp\). Retrieved from <https://store.wolterskluwer CDI.com/CDI>](#)

[Genderen E, Bommel J, Lima A. \(2012\) \*Monitoring Peripheral Perfusion in Critically Ill Patients at the Bedside\*. Curr. Opin. Crit. Care 18 \(3\), 273–279. doi:10.1097/MCC.0b013e3283533924](#)

- Vats V, Nagori A, Singh P, Dutt R, Bandhey H, Wason M, Lodha R, Sethi T. (2022) *Early Prediction of Hemodynamic Shock in Pediatric Intensive Care Units With Deep Learning on Thermal Videos*. *Front Physiol*;13:862411. <https://doi.org/10.3389/fphys.2022.862411>
- Calster B, McLernon J, Smeden M, Wynants L, Steyerberg W. (2019) *Calibration: the Achilles heel of predictive analytics*. *BMC Medicine*, 17, 230. <https://doi.org/10.1186/s12916-019-1466-7>
- Schaik A, Liu X, Atallah L, Badawi O. (2022) *Algorithmic fairness audits in intensive care medicine: Artificial intelligence for all?* *Critical Care*, 26, 104. <https://doi.org/10.1186/s13054-022-04197-5>
- Vayena E, Blasimme A, Cohen G. (2018) *Machine learning in medicine: Addressing ethical challenges*. *PLoS Medicine*, 15(11), e1002689. <https://doi.org/10.1371/journal.pmed.1002689>
- Ventola L. (2015) *The antibiotic resistance crisis: Part 1: Causes and threats*. *Pharmacy and Therapeutics*, 40(4), 277–283.
- Vincent L, Backer D. (2013) *Circulatory Shock*. *N. Engl. J. Med.* 369, 1726–1734. [doi:10.1056/NEJMr1208943](https://doi.org/10.1056/NEJMr1208943)
- Vincent L, Backer D, Wiedemann P. (2018) *Clinical review: Circulatory shock pathophysiology and management*. *Critical Care*, 22(1), 1-10.
- Vincent L, Nielsen D, Shapiro I, Gerbasi E, Grossman A, Doroff R, et al. (2018) *Mean arterial pressure and mortality in patients with distributive shock: A retrospective analysis of the MIMIC-III database*. *Ann. Intensive Care* 8, 107. [doi:10.1186/s13613-018-0448-9](https://doi.org/10.1186/s13613-018-0448-9)
- Vincent L, Jones G, David S, Olariu E, Cadwell K. *Frequency and mortality of septic shock in Europe and North America: a systematic review and meta-analysis*. *Crit Care*. 2019 May 31;23(1):196.
- Vo T, Nguyen N, Kha Q, Khanh L. (2022) *On the road to explainable AI in drug-drug interactions prediction: A systematic review*. *Computational and Structural Biotechnology Journal*. 20. [10.1016/j.csbj.2022.04.021](https://doi.org/10.1016/j.csbj.2022.04.021).
- Moltke L, Greenblatt J, Schmider J, Duan X, Wright E, Harmatz S, Shader I. (1996) *Midazolam hydroxylation by human liver microsomes in vitro: inhibition by fluoxetine, norfluoxetine, and by azole antifungal agents*. *J Clin Pharmacol*. 1996 Sep;36(9):783–791. [doi:10.1002/j.1552-4604.1996.tb04251.x](https://doi.org/10.1002/j.1552-4604.1996.tb04251.x). PMID: 8889898
- Wang W, Bao H, Li S, Dong L, Zhang Q, Yang F, Zhou M. (2022) *Image as a foreign language: BEiT-3 pretrain and finetune all at once*. *arXiv preprint arXiv:2208.10442*.

[Welch G, Bishop G. \(2006\) \*An Introduction to the Kalman Filter\*. University of North Carolina at Chapel Hill.](#)

[Wellner B, Grand J, Canzone E, Coarr M, Brady W, Simmons J, et al. \(2017\) \*Predicting unplanned transfers to the intensive care unit: A machine learning approach leveraging diverse clinical elements\*. JMIR Med. Inf. 5, e45. doi:10.2196/medinform.8680](#)

[Wei Y, Deng Y, Sun C, Lin M, Jiang H, Peng Y. \(2024\) \*Deep learning with noisy labels in medical prediction problems: A scoping review\*. Journal of the American Medical Informatics Association, 31\(7\), 1596–1609.](#)

[Widmer G, Kubat M. \(1996\) \*Learning in the presence of concept drift and hidden contexts\*. Mach Learn 23, 69–101 \(1996\). <https://doi.org/10.1007/BF00116900>](#)

[World Health Organization. \(2022\) \*Antimicrobial resistance\*. <https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance>](#)

[Wright G. \(205\) \*What is a laboratory information system \(LIS\)?\* | Definition from TechTarget. TechTarget. Retrieved from <https://www.techtarget.com/searchhealthit/definition/laboratory-information-system>](#)

[Zhang X, Zhao Z, Tsiligkaridis T, Zitnik M. \(/2022\) \*Self-supervised contrastive pre-training for time series via time-frequency consistency\*. In Proceedings of the 36th International Conference on Neural Information Processing Systems \(NIPS '22\). Curran Associates Inc., Red Hook, NY, USA, Article 288, 3988–4003.](#)

[Xiong G, Yang Z, Yi J, Wang N, Wang L, Zhu H, Wu C, Lu A, Chen X, Liu S, Hou T, Cao D. \(2022\) \*DDInter: an online drug-drug interaction database towards improving clinical decision-making and patient safety\*. Nucleic Acids Res.; 50\(D1\):D1200-D1207. doi: 10.1093/nar/gkab880. PMID: 34634800; PMCID: PMC8728114.](#)

[Xu L, Hao S, Ziyu N, & Li H, Zhang S. \(2024\) \*MedViLaM: A multimodal large language model with advanced generalizability and explainability for medical data understanding and generation\*. 10.48550/arXiv.2409.19684](#)

[Xu Y, Khan M, Song Y, Meijering E. \(2025\) \*Edge deep learning in computer vision and medical diagnostics: A comprehensive survey\*. Artificial Intelligence Review, 58, 93. <https://doi.org/10.1007/s10462-024-11033-5>](#)

[Yoon H, Pinsky R, Clermont G. \(2022\) \*Artificial Intelligence in Critical Care Medicine\*. Critical Care, 26\(1\), 75. <https://doi.org/10.1186/s13054-022-03915-3>](#)



[Yuan J, Shen C, Wang C, Shen G, Han B. \*Assessment of Physician's Knowledge of Potential Drug-Drug Interactions: An Online Survey in China\*. Front Med. 2021;8:650369.   
https://doi.org/10.3389/fmed.2021.650369](#)

[Zabir Al Nazi. \(2024\) \*Large Language Models in Healthcare and Medical Domain: A Review\*.   
arXiv:2401.06775v2 \[cs.CL\] 08 Jul 2024](#)

[Zhang T, Qu Y, Wang D, Zhong M, Cheng Y, Zhang M. \(2024\) \*Optimizing sepsis treatment strategies via a reinforcement learning model\*. Biomedical Engineering Letters, 14\(2\), 279–289.   
https://doi.org/10.1007/s13534-023-00343-2](#)

[Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. \(2021\) \*BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining\*. arXiv preprint   
arXiv:2110.08040. https://arxiv.org/abs/2110.08040](#)

[Zhang Z, Ni H. \(2024\) \*Critical care studies using large language models based on electronic healthcare records: A technical note\*. J Intensive Med. 5\(2\):137–150.   
doi:10.1016/j.jointm.2024.09.002. PMID: 40241837;   
PMCID: PMC11997556.   
https://futureagi.com/blogs/how-llms-are-transforming-time-series-data-analysis-in-ai-applications](#)

[Zhuang J, Huang H, Jiang S, Wang H, Wang L, Liu W, Zhou T. \(2023\) \*A generalizable and interpretable model for mortality risk stratification of sepsis patients in intensive care unit\*. BMC Medical Informatics and Decision Making, 23, 247.   
https://doi.org/10.1186/s12911-023-02279-0](#)