## Abstract

**Research Problem and Approach:** The integration of deep learning into critical care medicine offers unprecedented opportunities for predicting adverse patient outcomes, yet the opacity of these "black box" algorithms presents a fundamental barrier to clinical adoption. In high-stakes environments where decisions determine life or death, the lack of algorithmic transparency raises significant concerns regarding safety, clinician trust, and ethical accountability. This thesis presents a comprehensive scoping review of Explainable AI (XAI) methods applied within the Intensive Care Unit (ICU), investigating how technical interpretability techniques are being translated into clinical practice to bridge the gap between computational power and medical reasoning.

**Methodology and Findings:** By synthesizing current literature on Human-AI interaction and predictive modeling in critical care, this research maps the environment of XAI strategies, ranging from model-agnostic methods like SHAP to causally-informed deep learning architectures. The analysis reveals that while advanced neural networks outperform traditional scoring systems like APACHE and SOFA, an inverse relationship often persists between predictive performance and interpretability. Furthermore, the review highlights that current explanations frequently fail to alleviate the cognitive load on clinicians, complicating the reconciliation of algorithmic outputs with established physiological knowledge.

**Key Contributions:** This study makes three primary contributions: (1) A comprehensive taxonomy of existing XAI applications in critical care settings, distinguishing between technical explainability and practical clinical utility; (2) An evaluation of the mediating role of trust in Human-AI interaction, highlighting how the lack of transparency risks causing either "blind trust" or the rejection of valid alerts; and (3) An identification of critical gaps in safety validation, demonstrating that current regulatory frameworks require more strong specifications for algorithmic transparency to ensure patient safety.

**Implications:** The findings underscore that technical explainability alone is insufficient for clinical deployment; XAI must be rigorous, context-aware, and aligned with clinical reasoning processes to ensure accountability. This research provides a roadmap for future development, emphasizing the urgent need for prospective, cross-institutional evaluations to transform AI from a theoretical asset into a trustworthy bedside partner for critical care teams.

**Keywords:** Artificial Intelligence, Critical Care Medicine, Explainable AI, Machine Learning, Deep Learning, Intensive Care Unit, Black Box Models, Clinical Decision Support, Algorithmic Transparency, SHAP, Patient Safety, Human-AI Interaction, Predictive Modeling, Medical Ethics, Trust

# 1. Introduction

The integration of artificial intelligence (AI) into critical care medicine represents one of the most significant technological shifts in modern healthcare history. Intensive Care Units (ICUs) are data-rich environments where clinicians must make high-stakes, time-sensitive decisions based on continuous streams of physiological data, laboratory results, and clinical notes. The advent of machine learning (ML), particularly deep learning (DL), has enabled the development of sophisticated predictive models capable of analyzing this vast information environment with unprecedented accuracy (Cheng et al., 2025)(Singh, 2025). These models hold the promise of revolutionizing patient care by predicting adverse events such as sepsis, mortality, and acute kidney injury earlier than traditional scoring systems (Liu et al., 2024)(Wei et al., 2025). However, the increasing complexity of these algorithms has resulted in a “black box” phenomenon, where the internal logic of the model is opaque to the human user (Somani et al., 2023). This lack of transparency poses a fundamental barrier to clinical adoption, raising critical questions regarding safety, trust, and accountability (Jia et al., 2021)(Mirchandani, 2025).

This thesis presents a scoping review of Explainable AI (XAI) methods applied to machine learning models within the ICU setting. By mapping the current environment of XAI literature, this research aims to identify how technical explainability methods are being translated into clinical practice, the effectiveness of these methods in fostering clinician trust, and the existing gaps between algorithmic transparency and practical clinical utility.

## 1.1 The Evolution of Data-Driven Critical Care

Critical care medicine has always been a discipline grounded in data interpretation. Traditionally, clinicians have relied on scoring systems such as the Acute Physiology and Chronic Health Evaluation (APACHE) or the Sequential Organ Failure Assessment (SOFA) to stratify risk and guide treatment. These linear models, while interpretable, often fail to capture the complex, non-linear interactions inherent in human physiology. The shift towards “AI-Enabled Intensive Care Units” aims to overcome these limitations by integrating real-time decision support, automated handoffs, and intelligent monitoring into a cohesive framework (Singh, 2025).

Recent advancements have demonstrated that deep learning approaches can significantly outperform traditional statistical methods in prognostic tasks. For instance, advanced neural network architectures have been successfully deployed for 28-day mortality prediction, utilizing complex physiological features that linear models might overlook (Long & Tong, 2025). Similarly, causally-informed deep learning models are pushing the boundaries of generalizable outcome prediction, moving beyond mere correlation to identify potential causal drivers of clinical deterioration (Cheng et al., 2025).

Despite these performance gains, the transition from research validation to bedside implementation remains slow. A primary friction point is the opacity of models such as Long Short-Term Memory (LSTM) networks or Transformers. While an LSTM model might achieve superior accuracy in predicting mortality using electronic health record (EHR) data, its decision-making process–often involving millions of parameters–remains inaccessible to the treating physician (Adebayo, 2025). In an environment where a false negative can lead to missed life-saving intervention and a false positive can result in dangerous overtreatment, "blind trust" in an algorithm is clinically and ethically unacceptable.

## 1.2 The "Black Box" Problem in High-Stakes Decision Making

The "black box" problem refers to the inverse relationship often observed between model performance and interpretability. As algorithms become more complex (e.g., deep neural networks), their predictive power typically increases, but the ability to understand *why* a specific prediction was made decreases (Somani et al., 2023). In the context of the ICU, this opacity creates distinct categories of risk that must be addressed before widespread deployment can occur.

### 1.2.1 Clinical Safety and Validation

Safety in safety-critical systems, such as aviation or nuclear power, is usually assured through rigorous specification and testing against known standards. However, machine learning systems in healthcare often lack clear, pre-defined specifications against which validity can be assessed, a problem exacerbated by their opaque nature (Jia et al., 2021). Without explainability, it becomes difficult to distinguish whether a model is learning valid physiological signals or exploiting artifacts in the training data (e.g., predicting mortality based on the presence of a specific billing code or hospital process rather than patient physiology).

### 1.2.2 Trust and Human-AI Interaction

Trust is a mediating factor in the successful adoption of healthcare technologies. Research into Human-AI interaction indicates that trust is significantly influenced by the perceived usefulness and perceived ease of use of the system (Isparan Shanthi et al., 2024). When clinicians are presented with a prediction that contradicts their intuition–for example, a recommendation to delay intubation in a hypoxic patient–they require a rationale to evaluate the validity of the AI's suggestion. Without such explanation, the cognitive load on the clinician increases as they attempt to reconcile the AI output with their own assessment, potentially leading to the rejection of valid alerts or, conversely, automation bias where incorrect AI advice is followed uncritically (Eva et al., 2022).

### 1.2.3 Regulatory and Ethical Accountability

The drive for explainability is also propelled by emerging regulatory frameworks. The "right to explanation" concept, embedded in regulations such as the GDPR, suggests that automated decisions significantly affecting individuals must be explainable. In healthcare, this translates to an ethical imperative: clinicians must be able to justify their treatment decisions to patients and families. If a treatment plan is based on an AI recommendation that cannot be explained, the chain of accountability is broken (Mirchandani, 2025). Furthermore, verifying machine unlearning–ensuring that private or biased data has been removed from a model–requires explainable interfaces to confirm that the model no longer relies on the excised information (Vidal et al., 2024).

## 1.3 Explainable AI (XAI): Definitions and Approaches

Explainable AI (XAI) encompasses a suite of techniques and methods designed to make the outputs of artificial intelligence systems intelligible to human users. In the medical domain, XAI is not merely a technical feature but a bridge between computational power and clinical reasoning. The literature identifies several categories of XAI methods currently being explored in critical care.

### 1.3.1 Model-Agnostic Methods

Model-agnostic methods are designed to interpret predictions from any machine learning algorithm, regardless of its internal architecture. The most prominent example in the recent literature is SHapley Additive exPlanations (SHAP). SHAP values provide a unified measure of feature importance, attributing the contribution of each input variable to the final prediction. Studies have demonstrated the utility of SHAP in interpreting mortality predictions, allowing clinicians to see which specific physiological parameters (e.g., lactate levels, blood pressure trends) drove the model's risk assessment (Long & Tong, 2025)(Adebayo, 2025).

### 1.3.2 Model-Specific and Attention Mechanisms

Unlike agnostic methods, model-specific approaches uses the internal structure of the algorithm. In deep learning, attention mechanisms have emerged as a powerful tool for interpretability, particularly for time-series data common in ICUs. Attention maps can highlight which time steps or variables in a temporal sequence were most influential for a specific prediction. For instance, in sepsis prediction models, attention maps can visualize which vital sign fluctuations over the preceding hours triggered the sepsis alert, aligning the model's "focus" with clinical signs of deterioration (Liu et al., 2024). Advanced architectures, such as the Triple Attention Transformer, further enhance this by improving contextual coherence in processing long-term dependencies, which is important for analyzing patient trajectories over extended ICU stays (Ghaith, 2024).

### 1.3.3 Visual and Feature-Based Explanations

In domains involving medical imaging or complex signals, explainability often takes the form of visual heatmaps. Techniques like Layer-Wise Relevance Propagation (LRP) have been applied to classify brain MRI images, generating heatmaps that identify the specific anatomical regions contributing to a diagnosis (Naik et al., 2025). While primarily used in radiology, these techniques are increasingly relevant in the ICU for interpreting bedside ultrasound or continuous waveform monitoring data.

Table 1 summarizes the core differences between traditional "Black Box" approaches and the emerging XAI paradigms in the context of intensive care.

| Feature | Black Box ML Models | Explainable AI (XAI) Models | Clinical Implication |
|---|---|---|---|
| **Transparency** | Opaque internal logic | Transparent or interpretable outputs | XAI enables validation of clinical logic. |
| **Error Detection** | Difficult to trace source of error | Errors can be traced to specific features | XAI facilitates safety auditing (Jia et al., 2021). |
| **User Trust** | Requires blind faith in metrics | Builds trust through justification | XAI supports shared decision-making (Isparan Shanthi et al., 2024). |
| **Complexity** | High (Deep Learning, Ensembles) | High, but with an interpretability layer | Performance is maintained while adding clarity. |
| **Focus** | Optimization of accuracy metrics | Optimization of utility and explainability | XAI balances accuracy with usability (Adebayo, 2025). |

*Table 1: Comparison of Black Box vs. Explainable AI paradigms in critical care settings. Adapted from concepts in (Somani et al., 2023) and (Mirchandani, 2025).*

The transition from black box to explainable models is not binary. As illustrated in Table 1, XAI attempts to retain the high complexity and performance of modern algorithms while adding a layer of interpretability. This balance is critical because simplifying the model itself (e.g., using a simple decision tree instead of a neural network) might reduce accuracy to a level that is clinically useless, whereas a post-hoc explanation method allows for both high accuracy and interpretability (Adebayo, 2025).

## 1.4 Applications of XAI in Intensive Care

The application of XAI in the ICU is diverse, addressing various pathologies and operational challenges. The literature reveals a concentration of XAI research in high-mortality conditions where early intervention is decisive.

**Sepsis and Infection Management:** Sepsis remains a leading cause of ICU mortality. XAI is being used to decipher complex deep learning models that predict sepsis onset using vital signs and laboratory values. By highlighting specific determinants–such as a subtle but sustained drop in blood pressure combined with rising heart rate–XAI tools can alert clinicians to early warning signs that might otherwise be dismissed as noise (Liu et al., 2024). Furthermore, reinforcement learning (RL) agents designed for sepsis treatment recommendation are incorporating explainability to justify dosing decisions, moving towards continuous action space solutions that mimic the nuance of clinician adjustments (Huang et al., 2022).

**Mechanical Ventilation and Respiratory Failure:** Predicting the need for intubation and managing mechanical ventilation are core ICU tasks. Recent studies have utilized machine learning combined with XAI to predict intubation needs, providing clinicians with risk scores decomposed into contributing factors like oxygen saturation trends and respiratory effort (Saykat et al., 2025). Additionally, in patients with ventilator-associated pneumonia (VAP), interpretable models are being developed to assess in-hospital mortality risk, aiding in prognosis discussions and resource allocation (Wei et al., 2025).

**Hemodynamic Monitoring and Drug Dosing:** The management of hypotension often requires the precise titration of catecholamines (vasopressors). New approaches are moving beyond simple threshold-based predictions to actionable forecasting of catecholamine therapy initiation. These models aim to explain *why* a patient is likely to require hemodynamic support, potentially distinguishing between hypovolemic and distributive shock patterns (Koebe et al., 2025). Similarly, in related fields like water treatment, XAI has been used to optimize coagulant dosing, a control problem analogous to drug titration, demonstrating the transferability of explainable optimization techniques (Park et al., 2024).

**Resource Management and Workflow:** Beyond direct patient care, XAI is finding utility in operational efficiency. Predicting length of stay and resource utilization helps in bed management. Moreover, adaptive XAI systems are being proposed that personalize explanations based on the user's expertise level–offering detailed technical data to a senior intensivist while providing high-level summaries to a bed manager or rotational resident (Mohammed, 2025).

## 1.5 Research Problem and Rationale

Despite the proliferation of XAI methods in technical literature, there remains a significant gap in understanding their practical utility in the ICU. Much of the existing research focuses on the *development* of new algorithms (e.g., a novel attention mechanism or a variation of SHAP) rather than the *evaluation* of these tools in clinical workflows. There is a discordance between what computer scientists consider an "explanation" (e.g., a feature importance bar chart) and what clinicians require to make a safe decision (e.g., a counterfactual scenario or causal reasoning) (Zhang, 2023).

Furthermore, the "explanation" provided by current methods may not always be strong. Issues such as the stability of explanations (whether similar inputs yield similar explanations) and the fidelity of post-hoc interpretations are active areas of concern. For instance, if a SHAP plot identifies a feature as important, does that feature actually drive the model's prediction in a causal sense, or is it merely a correlated proxy? (Cheng et al., 2025)(Zhang, 2023).

Therefore, a comprehensive scoping review is necessary to map the extent, range, and nature of research activity in this field. Unlike a systematic review, which might focus on quantifying the diagnostic accuracy of specific models, a scoping review is appropriate here to clarify concepts, identify the types of XAI methods available, and analyze knowledge gaps regarding their implementation in critical care (Arksey & O'Malley, 2005).

## 1.6 Research Question and Objectives

The primary research question guiding this thesis is: *What explainable AI (XAI) methods have been applied to machine learning models used in intensive care unit (ICU) clinical decision support, and what are their reported effectiveness, limitations, and implementation challenges?*

To answer this question, the following specific objectives have been established: 1. **Categorize** the types of XAI methods (e.g., model-agnostic, attention-based, example-based) currently utilized in ICU-related machine learning literature. 2. **Identify** the clinical domains within critical care (e.g., sepsis, ventilation, mortality prediction) where XAI is most frequently applied. 3. **Evaluate** how these studies assess the "explainability" of their models, specifically looking for metrics of human-centered evaluation (trust, utility, cognitive load) versus purely computational metrics. 4. **Analyze** the reported barriers to implementation, including technical limitations, data privacy concerns, and regulatory hurdles.

## 1.7 Methodology Overview

This thesis employs a scoping review methodology, adhering to the framework outlined by (Arksey & O'Malley, 2005) and the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) guidelines. This approach allows for the synthesis of a broad range of study designs, from technical algorithmic proposals to qualitative evaluations of clinician user interfaces.

The search strategy targets major bibliographic databases including PubMed, IEEE Xplore, and Scopus, focusing on literature published in the era of modern deep learning (post-2015). The review specifically includes studies that combine three core elements: (1) Machine Learning/Artificial Intelligence, (2) Explainability/Interpretability, and (3) Intensive Care/Critical Care settings.

Table 2 provides a preliminary overview of the types of studies anticipated in the review, categorized by the interaction between clinical tasks and XAI methods.

| Clinical Domain | Common ML Tasks | Typical XAI Methods | Key Citations |
|---|---|---|---|
| **Outcomes** | Mortality Prediction | SHAP, LIME, LSTM+SHAP | (Long & Tong, 2025), (Adebayo, 2025), (Wei et al., 2025) |
| **Sepsis** | Early Detection, Treatment | Attention Maps, RL Interpretation | (Liu et al., 2024), (Huang et al., 2022) |
| **Respiratory** | Intubation, Weaning | Feature Importance, Decision Trees | (Saykat et al., 2025), (Wei et al., 2025) |
| **Hemodynamics** | Hypotension, Vasopressors | Forecasting, Causal Inference | (Cheng et al., 2025), (Koebe et al., 2025) |
| **Imaging/Signals** | MRI, Waveform Analysis | LRP, Heatmaps, CNN Vis. | (Naik et al., 2025), (Bashir et al., 2025) |

*Table 2: Matrix of Clinical Domains and XAI Approaches. This table illustrates the intersection of clinical problems and technical explanation strategies identified in the preliminary literature search.*

As indicated in Table 2, the field is characterized by a diverse array of applications. The review will systematically extract data regarding the specific algorithms used, the explanation modality presented to users, and any validation of the explanation's quality.

## 1.8 Thesis Structure

This thesis is organized into five chapters. Following this Introduction, **Chapter 2 (Background & Theoretical Framework)** provides a detailed examination of the theoretical underpinnings of XAI, including the taxonomy of interpretability (global vs. Local, ante-hoc vs. Post-hoc) and the psychological theories of trust and cognitive load in human-computer interaction. It also details the specific data challenges of the ICU environment.

**Chapter 3 (Methodology)** details the scoping review protocol, including search strings, inclusion/exclusion criteria, and the data charting form. It describes the process of study selection and the analytical framework used to synthesize the findings.

**Chapter 4 (Results)** presents the findings of the scoping review. It offers a numerical summary of the included studies, a thematic analysis of the XAI methods identified, and a narrative synthesis of how these methods are applied across different clinical use cases. This chapter also reports on the evaluation metrics used in the literature, highlighting the scarcity of user-centric validation.

**Chapter 5 (Discussion and Conclusion)** interprets the findings in the context of the broader healthcare AI environment. It discusses the implications of the "interpretability gap," addresses the limitations of current XAI approaches in handling multimodal and time-series data, and offers recommendations for future research and clinical implementation strategies. The thesis concludes with a summary of the potential for XAI to transform critical care medicine into a discipline that is both data-driven and transparently human-centered.

## 1.9 Significance of the Study

The significance of this research lies in its potential to guide the future development of clinical decision support systems. As the FDA and other regulatory bodies move towards stricter requirements for Software as a Medical Device (SaMD), the ability to explain algorithmic output will transition from a "nice-to-have" feature to a mandatory requirement (Jia et al., 2021)(Vidal et al., 2024). By consolidating current knowledge and identifying the disconnects between technical capability and clinical need, this thesis aims to inform both developers and clinicians.

For developers, this review highlights the need to move beyond static feature importance plots towards causal, counterfactual, and adaptive explanations that align with clinical reasoning (Zhang, 2023)(Mohammed, 2025). For clinicians and healthcare administrators, it provides a framework for evaluating new AI tools, emphasizing that accuracy alone is insufficient for safe deployment. Ultimately, the goal of XAI in the ICU is not merely to open the "black box," but to illuminate the path towards safer, more effective, and more equitable patient care.

The urgency of this investigation is underscored by the rapid pace of AI development. With new architectures like Transformers becoming standard for processing electronic health records, the complexity of models is increasing exponentially. Without a concurrent advancement in interpretability, the gap between what machines can predict and what humans can understand will widen, potentially stalling the deployment of life-saving technologies. This scoping review serves as a critical step in bridging that gap, ensuring that the "AI-ICU" of the future (Singh, 2025) is built on a foundation of transparency and trust.

# 2. Main Body

The integration of artificial intelligence (AI) into critical care medicine represents a major change in how clinicians approach diagnosis, prognosis, and treatment planning. As machine learning (ML) models–particularly deep learning architectures–demonstrate increasingly superior performance in predicting adverse events such as sepsis, mortality, and organ failure, the "black box" nature of these algorithms has emerged as a significant barrier to clinical adoption. This literature review synthesizes current research regarding Explainable AI (XAI) within the intensive care unit (ICU) setting. It examines the theoretical foundations of interpretability, categorizes the methodological approaches currently deployed, analyzes clinical applications across various critical care domains, and evaluates the human factors influencing the acceptance of these systems.

## 2.1 Theoretical Framework and Foundational Concepts

The theoretical underpinnings of XAI in healthcare are rooted in the tension between predictive performance and model transparency. In the high-stakes environment of the ICU, where decisions must be made rapidly and carry life-or-death consequences, the opacity of complex algorithms poses ethical, legal, and practical challenges.

### 2.1.1 The "Black Box" Problem in Critical Care

The "black box" phenomenon refers to the inability of human observers to understand the internal decision-making logic of complex non-linear models, such as Deep Neural Networks (DNNs) and ensemble methods like Gradient Boosting Machines (GBM). While these models often outperform traditional linear regression or scoring systems (e.g., APACHE II, SOFA) in capturing the non-linear dynamics of patient physiology, their lack of transparency impedes trust.

Research indicates that in safety-critical systems, the absence of a clear specification against which to assess validity exacerbates the difficulty of assuring safety (Jia et al., 2021). Unlike rule-based systems where logic is explicit, data-driven models learn latent representations that may rely on spurious correlations or confounding variables not visible to the clinician. This opacity is particularly problematic in the ICU, where "algorithmic silence"–the failure of a model to warn of a deterioration–or "algorithmic hallucination"–false positives leading to alarm fatigue–can directly compromise patient safety.

Furthermore, the complexity of ICU data, which includes high-frequency vital signs, laboratory values, imaging, and unstructured clinical notes, necessitates models that can handle multimodal inputs. Recent advancements in deep learning have enabled the processing of such heterogeneous data, yet they have simultaneously deepened the interpretability gap. For instance, while Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are highly effective for time-series prediction in critical care (Bon & Cardot, 2011), their internal state transitions are notoriously difficult to map to clinical concepts.

### 2.1.2 Defining Interpretability and Explainability

The literature distinguishes between “interpretability” and “explainability,” though these terms are often used interchangeably in clinical studies. A foundational taxonomy provided by Somani et al. (Somani et al., 2023) clarifies these concepts within the medical domain. Interpretability is defined as a passive characteristic of a model, referring to the degree to which a human can understand the cause of a decision. In contrast, explainability refers to the active techniques and interface elements used to communicate the model’s internal state to a user.

This distinction is important for ICU applications. A linear regression model is inherently interpretable because its coefficients directly represent the change in the outcome variable for a unit change in the predictor. However, a deep learning model for sepsis prediction is not inherently interpretable and thus requires XAI methods to generate post-hoc explanations.

Recent scholarship has expanded these definitions to include “accountability” and “transparency.” Mirchandani (Mirchandani, 2025) argues that in societal sectors like healthcare, explainability serves as a mechanism for accountability, allowing stakeholders to audit algorithmic decisions for bias and error. This aligns with the growing demand for “Causal Explainable AI,” which moves beyond correlation to identify cause-and-effect relationships, thereby offering more strong and actionable insights for clinical intervention (Zhang, 2023).

### 2.1.3 The Role of Trust and Safety in Clinical Decision Support

Trust is a mediating variable in the successful adoption of AI technologies. The relationship between perceived usefulness, perceived ease of use, and trust has been modeled in recent studies, highlighting that opaque systems often fail to garner the necessary trust from healthcare professionals regardless of their accuracy (Isparan Shanthi et al., 2024). In the ICU, trust is not merely a psychological state but a prerequisite for action. If a clinician does not trust a model’s prediction of impending hemodynamic collapse, they will not initiate the recommended vasopressor therapy.

Safety assurance is inextricably linked to explainability. Jia et al. (Jia et al., 2021) posit that explainability is essential for verifying that a model’s reasoning aligns with established medical knowledge. For example, if a mortality prediction model identifies “asthma” as a protective factor against pneumonia death (a known artifact in some historical datasets due to aggressive treatment protocols), an XAI method should reveal this counter-intuitive logic, allowing clinicians to reject the model’s faulty reasoning.

Table 1 summarizes the key theoretical dimensions of XAI identified in the literature.

| Dimension | Definition | Clinical Relevance | Source |
|---|---|---|---|
| Interpretability | Intrinsic transparency of a model's logic. | Allows validation against pathophysiology. | (Somani et al., 2023) |
| Explainability | Post-hoc techniques to elucidate opaque models. | Enables use of high-performance DL models. | (Mirchandani, 2025) |
| Causality | Identification of cause-effect relationships. | Supports intervention, not just prediction. | (Zhang, 2023) |
| Fidelity | Accuracy of the explanation to the model. | Ensures clinicians aren't misled by XAI. | (Jia et al., 2021) |
| Trust | User confidence in model reliability. | Prerequisite for adoption and action. | (Isparan Shanthi et al., 2024) |

*Table 1: Theoretical Dimensions of XAI in Healthcare. Source: Adapted from (Somani et al., 2023), (Jia et al., 2021), (Zhang, 2023), and (Mirchandani, 2025).*

## 2.2 Methodological Approaches to XAI in ICU Settings

The scoping review identifies a diverse array of XAI methodologies applied to ICU datasets. These can be broadly categorized into model-agnostic methods, which can be applied to any algorithm, and model-specific methods, which exploit the internal architecture of specific model types (e.g., neural networks).

### 2.2.1 Model-Agnostic Methods: SHAP and LIME

Shapley Additive Explanations (SHAP) has emerged as the dominant model-agnostic method in the ICU literature. Based on cooperative game theory, SHAP assigns each feature an importance value for a particular prediction.

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \quad i)]$$

Where $\phi_i$ is the Shapley value for feature $i$, $f$ is the model, and $M$ is the set of all input features.

Recent studies have extensively utilized SHAP to decode mortality prediction models. Long and Tong (Long & Tong, 2025) demonstrated the utility of integrating SHAP with machine learning to interpret 28-day mortality predictions in ICU patients. Their work highlights how SHAP summary plots can reveal non-linear relationships between physiological features (e.g., lactate levels, heart rate) and mortality risk, providing a global view of model behavior while maintaining the ability to explain individual patient predictions (local interpretability).

Similarly, Adebayo (Saykat et al., 2025) employed a hybrid approach combining Long Short-Term Memory (LSTM) networks with SHAP to bridge the gap between accuracy and interpretability for ICU mortality prediction using Electronic Health Record (EHR) data. This application is particularly significant as it applies a model-agnostic explainer to a complex deep learning architecture, demonstrating that the "black box" of temporal sequence models can be partially illuminated.

While SHAP provides mathematical consistency, it is computationally expensive. Other model-agnostic methods like LIME (Local Interpretable Model-agnostic Explanations) are also present in the literature, though less frequently in recent high-dimensional ICU studies compared to SHAP. The dominance of SHAP is likely due to its theoretical guarantees regarding the fair distribution of feature contributions.

Feature importance analysis remains a critical first step in many ICU ML pipelines. Terlapu et al. (Terlapu et al., 2024) discuss feature importance in the context of software effort estimation, but the methodological principles transfer to healthcare, where identifying the "vital few" variables from the "trivial many" is essential for model parsimony. Furthermore, Huang (Huang, 2025) introduced residual permutation tests as a method to assess feature importance in non-linear models, offering a statistical rigor often missing in heuristic importance measures.

### 2.2.2 Attention Mechanisms in Deep Learning

For deep learning models processing time-series data (vital signs) or unstructured data (clinical notes), attention mechanisms serve as a powerful model-specific XAI technique. Attention mechanisms allow a neural network to focus on specific parts of the input sequence when generating a prediction, effectively weighing the importance of different time steps or features.

Liu et al. (Liu et al., 2024) proposed "Interpretable Vital Sign Forecasting with Model Agnostic Attention Maps" for sepsis prediction. Their approach visualizes which segments of a patient's vital sign history contributed most to the sepsis alarm. This temporal interpretability is vital in the ICU, as it allows clinicians to see *when* the patient's condition began to deviate from the norm, correlating model attention with clinical events (e.g., a drop in blood pressure or a spike in heart rate).

In the domain of imaging and complex signal processing, similar techniques are employed. Sano (SANO, 2022) and Naik et al. (Naik et al., 2025) explore Gradient-Weighted Class Activation Mapping (Grad-CAM) and Layer-Wise Relevance Propagation (LRP) for interpreting image classification models. While their specific applications (facial attractiveness and brain MRI) differ from general ICU monitoring, the underlying methodologies are increasingly applied to ICU imaging tasks, such as chest X-ray analysis for pneumonia or ventilator management.

The "Triple Attention Transformer" introduced by Ghaith (Ghaith, 2024) represents the frontier of this approach, enhancing contextual coherence in transformer models. In an ICU context, such architectures could theoretically track long-term dependencies in patient history (e.g., a medication administered three days ago) and highlight this connection to the clinician, offering a level of narrative explanation that simple feature importance scores cannot provide.

### 2.2.3 Causal and Hybrid Frameworks

A significant limitation of standard ML is its reliance on correlation. In the ICU, interventions (e.g., giving fluids) change outcomes, often confounding predictive models. "Causally-informed" deep learning addresses this by explicitly modeling cause-and-effect relationships.

Cheng et al. (Cheng et al., 2025) developed a causally-informed deep learning framework for outcomes prediction in critical care. By integrating causal graphs with deep learning, their model aims to provide explanations that are not just associative but mechanistic. This is important for "generalizable" outcomes, ensuring that a model learned in one ICU does not fail in another due to differences in treatment protocols.

Zhang (Zhang, 2023) further elaborates on "Causal Explainable AI," arguing that for an explanation to be actionable, it must support counterfactual reasoning (e.g., "If we had not administered this drug, would the patient still have developed kidney injury?"). This represents the next generation of XAI, moving from "What happened?" to "What if?".

### 2.2.4 Reinforcement Learning and Policy Explanation

The ICU is a dynamic environment requiring sequential decision-making, making Reinforcement Learning (RL) an attractive paradigm. However, RL policies are notoriously opaque.

Huang et al. (Huang et al., 2022) explored RL for sepsis treatment using continuous action spaces, a method that mimics the titration of vasopressors and fluids. To make such policies acceptable, Saulières et al. (Saulières et al., 2023) proposed "Reinforcement Learning Explained via Reinforcement Learning," a meta-approach where a secondary agent learns to explain the policy of the primary agent. This recursive explanation strategy attempts to generate justifications for actions (e.g., “Increase norepinephrine because MAP is trending down and lactate is rising”) rather than just outputting numerical values.

## 2.3 Clinical Applications and Empirical Evidence

The application of XAI in the ICU is broad, spanning from admission risk stratification to real-time organ support management. The literature reveals varying levels of maturity across these domains.

### 2.3.1 Mortality and Risk Prediction

Mortality prediction remains the most common application for XAI in critical care. Accurate prognostication assists in resource allocation and discussions regarding goals of care.

Long and Tong (Long & Tong, 2025) focused on 28-day mortality, a standard endpoint in sepsis trials. By applying SHAP to physiological features, they identified that while traditional markers like age and APACHE scores are dominant, specific patterns in dynamic vital signs also hold significant predictive power.

Adebayo (Saykat et al., 2025) utilized EHR data to predict mortality, emphasizing the need to bridge the gap between the high accuracy of LSTMs and the interpretability required for clinical trust. Their findings suggest that hybrid models can achieve current performance (AUC $>$ 0.85) while providing clinician-friendly visualizations of risk factors.

Wei et al. (Wei et al., 2025) developed an interpretable ML model specifically for in-hospital mortality in patients with Ventilator-Associated Pneumonia (VAP). This niche application demonstrates the versatility of XAI; by isolating a specific high-risk cohort, the model could identify risk factors specific to VAP that general mortality models might miss, such as specific antibiotic resistance patterns or ventilator settings.

### 2.3.2 Sepsis Detection and Management

Sepsis is a leading cause of ICU mortality, and early detection is a “holy grail” of critical care informatics.

Liu et al. (Liu et al., 2024) addressed the complexity of analyzing diverse vital signs for sepsis prediction. Their attention-based model not only predicts sepsis onset but highlights the specific vital sign trajectories (e.g., widening pulse pressure) that triggered the alarm. This allows clinicians to differentiate between true sepsis and other causes of physiological derangement.

Beyond prediction, Huang et al. (Huang et al., 2022) applied RL to sepsis *treatment*. The challenge here is explaining the *recommendation*. Unlike prediction (where “why” explains a risk), treatment recommendation requires justifying an intervention. While their continuous action space solution shows promise in simulation, the literature notes a significant gap in clinical validation of such prescriptive AI systems due to safety concerns.

### 2.3.3 Respiratory Support and Airway Management

Mechanical ventilation involves complex decisions regarding intubation, weaning, and extubation.

Saykat et al. (Saykat et al., 2025) focused on predicting intubation needs in the ICU using ML and XAI. Their work addresses a critical decision point: delayed intubation increases mortality, while unnecessary intubation carries risks of trauma and pneumonia. By providing explainable risk scores, such models aim to support the clinician's judgment in this "grey zone" of decision-making.

Wei et al. (Wei et al., 2025) also touches upon this domain through their VAP analysis, linking mortality risk to ventilator parameters. The interpretability of these models is important because ventilator management is highly protocolized; an AI suggestion to deviate from protocol (e.g., changing PEEP levels) requires a strong, transparent justification.

### 2.3.4 Hemodynamic Monitoring and Fluid Therapy

Hemodynamic instability requires rapid intervention with fluids and vasoactive drugs.

Koebe et al. (Koebe et al., 2025) tackled the prediction of catecholamine therapy initiation for hypotension. Unlike simple hypotension prediction (which is often trivial once pressure drops), predicting the *need for therapy* is a nuanced clinical task. Their work on "actionable" prediction emphasizes that the model must predict the event early enough to intervene, and explain *why* the pressure is expected to drop (e.g., vasodilation vs. Hypovolemia).

Escudero-Arnanz et al. (Escudero-Arnanz et al., 2025) utilized multimodal interpretable models using multivariate time series, which is essential for hemodynamics where heart rate, blood pressure, and urine output are tightly coupled.

Furthermore, the European Society of Intensive Care Medicine (ESICM) guidelines on fluid therapy (Dessap et al., 2025) highlight the complexity of fluid resuscitation. While not an AI paper per se, the complexity described in these guidelines underscores the need for AI models (like those proposed by Park et al. (Park et al., 2024) for coagulant dosing, analogous to fluid dosing) to be explainable. Park et al.'s work in water treatment optimization using XAI provides a methodological parallel: optimizing a dosage based on complex inputs requires explaining the optimization curve to the operator.

Table 2 summarizes key empirical studies reviewed.

| Study | Clinical Domain | XAI Method | Key Finding/Contribution |
|---|---|---|---|
| Long & Tong (Long & Tong, 2025) | Mortality (28-day) | SHAP | Identified non-linear impact of physiologic features on death risk. |
| Liu et al. (Liu et al., 2024) | Sepsis Prediction | Attention Maps | Visualized temporal vital sign segments triggering sepsis alarms. |
| Saykat et al. (Saykat et al., 2025) | Intubation Need | Feature Importance | Developed risk stratification for airway management decisions. |
| Koebe et al. (Koebe et al., 2025) | Hypotension | Actionable Prediction | Predicted need for catecholamines, distinguishing types of shock. |
| Cheng et al. (Cheng et al., 2025) | General Outcomes | Causal Graphs | Integrated causality to improve model generalizability across ICUs. |

*Table 2: Summary of Selected Empirical Studies on XAI in ICU. Source: Compiled from cited references.*

## 2.4 Human-AI Interaction and Implementation Challenges

The technical capability to generate an explanation does not guarantee its utility. The literature increasingly focuses on the Human-Computer Interaction (HCI) aspects of XAI.

### 2.4.1 Clinician Trust, Cognitive Load, and Expertise

Trust is dynamic and context-dependent. Isparan Shanthi et al. (Isparan Shanthi et al., 2024) investigated the mediating role of trust and the moderating influence of cognitive load in human-AI interaction. In the ICU, cognitive load is perpetually high. XAI systems that add to this load–by providing complex, difficult-to-read explanations–may decrease performance even if the underlying model is accurate.

Eva et al. (Eva et al., 2022) studied cognitive fatigue and mental workload using

XAI to classify electrophysiological signatures. While their context was flight simulation, the parallels to ICU monitoring are evident: operators monitoring complex screens for alarms under fatigue. Their findings suggest that XAI can help identify when an operator (or clinician) is missing signals due to fatigue, or conversely, that XAI interfaces must be designed to accommodate fatigued users.

Mohammed (Mohammed, 2025) proposed "Adaptive Explainable AI," which personalizes explanations based on user expertise levels. An attending physician might require a different level of explanation (e.g., mechanistic pathway) compared to a junior resident or a nurse (e.g., immediate actionable flag). This personalization is critical for ICU teams, which are multidisciplinary.

### 2.4.2 Validation and Regulatory Compliance

Validating XAI is difficult because there is often no "ground truth" for an explanation. Bashir et al. (Bashir et al., 2025) presented a strong methodology for clinical validation of XAI (in fetal scans) using a multi-level, cross-institutional approach. They used actionable concepts as feedback to end-users, testing whether the XAI actually improved clinical decision accuracy. This type of prospective, cross-center evaluation is rare in the ICU literature but represents the gold standard.

Regulatory compliance is another driver. Vidal et al. (Vidal et al., 2024) discussed verifying "Machine Unlearning" with XAI to comply with privacy regulations like GDPR. While more relevant to data privacy, it touches on the "Right to Explanation." In the ICU, this translates to the right of the patient (or family) to understand why a life-support decision was recommended by an algorithm.

## 2.5 Synthesis and Identification of Research Gaps

Synthesizing the reviewed literature reveals several critical gaps that this thesis aims to address.

### 2.5.1 The Gap Between Technical and Clinical Metrics

Most studies evaluate XAI using technical metrics (e.g., fidelity, stability) or proxy clinical metrics (e.g., "does the heatmap look reasonable?"). There is a paucity of studies measuring *clinical outcomes* associated with XAI use. Does showing a SHAP plot to a doctor actually reduce mortality or length of stay? The literature is currently dominated by retrospective feasibility studies (e.g., (Long & Tong, 2025), (Adebayo, 2025), (Wei et al., 2025)) rather than prospective clinical trials.

### 2.5.2 Real-Time Integration Challenges

While papers describe “real-time” models, few detail the infrastructure required to serve SHAP values or attention maps in real-time at the bedside. Singh (Singh, 2025) proposes an integrated framework for an “AI-Enabled ICU,” including real-time decision support and automated handoffs. However, the computational latency of calculating Shapley values for high-frequency ICU data remains a technical bottleneck that is often glossed over in purely algorithmic papers.

### 2.5.3 Lack of Standardization in Evaluation

There is no standard protocol for evaluating the quality of an explanation in critical care. While Somani et al. (Somani et al., 2023) provide a taxonomy, applied papers often create ad-hoc evaluation surveys. The field lacks a unified “XAI Quality Score” for medical applications.

### 2.5.4 Multimodal and Temporal Complexity

Most XAI methods are applied to static snapshots or single modalities (only EHR or only vitals). Escudero-Arnanz et al. (Escudero-Arnanz et al., 2025) and Bieniek-Kaczorek et al. (Bieniek-Kaczorek et al., 2025) (focusing on photonic interrogators for vitals) hint at the future of multimodal monitoring. However, explaining a decision that fuses a chest X-ray, a waveform, and a lab value remains a frontier challenge. Current methods like SHAP struggle to provide a cohesive narrative across these disparate data types.

## 2.6 Summary

The literature on XAI in the ICU demonstrates a rapidly maturing field moving from simple feature importance to complex, model-specific, and causal explanations. Theoretical frameworks emphasize the necessity of trust and safety (Jia et al., 2021), while methodological advancements in SHAP (Long & Tong, 2025) and attention mechanisms (Liu et al., 2024) provide the tools to open the “black box.” Clinical applications are expanding from mortality prediction to actionable intervention support (Koebe et al., 2025). However, significant gaps remain in prospective clinical validation, real-time implementation, and the standardization of evaluation metrics. This thesis will address these gaps by proposing a framework for evaluating the clinical utility of XAI methods specifically within the high-stakes, high-velocity context of intensive care.

# 2.2 Methodology

This chapter details the methodological framework employed to conduct the scoping review of Explainable Artificial Intelligence (XAI) applications within

Intensive Care Unit (ICU) clinical decision support systems. Given the heterogeneous nature of the literature, which spans computer science, biomedical engineering, and clinical medicine, a scoping review approach was selected over a systematic review. This choice aligns with the objective to map the key concepts underpinning a research area and the main sources and types of evidence available, rather than to strictly evaluate the quality of individual studies for a meta-analysis.

The review follows the methodological framework developed by Arksey and O'Malley (Arksey & O'Malley, 2005) and further refined by the Joanna Briggs Institute. Reporting is conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) (McGannan et al., 2024). This rigorous protocol ensures transparency, reproducibility, and a systematic approach to identifying gaps in the current body of knowledge regarding the clinical utility, safety, and implementation of XAI in critical care settings.

## 2.2.1 Research Design and Protocol

The research design was structured around the five key stages outlined in the Arksey and O'Malley framework: (1) identifying the research question; (2) identifying relevant studies; (3) study selection; (4) charting the data; and (5) collating, summarizing, and reporting the results. This iterative process allowed for the refinement of search strategies and inclusion criteria as familiarity with the literature increased.

### 2.2.1.1 Identification of Research Questions

The primary objective of this review is to synthesize evidence on the deployment and evaluation of XAI methods in ICU settings. To achieve this, the following specific research questions guided the protocol: 1. **What types of XAI techniques** (e.g., model-agnostic vs. Model-specific) are currently applied to machine learning models for ICU clinical decision support? 2. **For which clinical tasks** (e.g., mortality prediction, sepsis detection, mechanical ventilation) are these explanations being generated? 3. **How is the effectiveness** of these explanations evaluated, particularly regarding clinician trust, interpretability, and safety? 4. **What are the technical and translational barriers** preventing the widespread adoption of XAI in real-time critical care workflows?

These questions address the fundamental tension in healthcare AI: the trade-off between the high predictive performance of "black-box" models, such as deep neural networks, and the absolute requirement for transparency in life-or-death decision-making (Adebayo, 2025)(Jia et al., 2021). By focusing on these dimensions, the review aims to move beyond simple algorithmic performance metrics to understand the sociotechnical challenges of implementation.

#### 2.2.1.2 Theoretical Framework for Analysis

To analyze the extracted literature systematically, this review adopts a taxonomy of interpretability grounded in the work of Somani et al. (Somani et al., 2023). This framework categorizes XAI approaches based on three dimensions: * **Scope:** Local (explaining a single prediction) versus Global (explaining the entire model behavior). * **Methodology:** Post-hoc (explaining a trained model) versus Intrinsic (models that are interpretable by design). * **Modality:** Feature-based (attribution scores), Concept-based (high-level abstractions), or Example-based (prototypes).

This theoretical lens is important for distinguishing between methods that merely identify *correlations* (feature importance) and those that attempt to uncover *causal* mechanisms, a distinction that is vital for safe clinical intervention (Cheng et al., 2025)(Zhang, 2023). Furthermore, the analysis considers the "human-in-the-loop" perspective, evaluating how explanations align with clinician cognitive workflows and expertise levels (Mohammed, 2025).

### 2.2.2 Search Strategy and Data Sources

A comprehensive search strategy was developed to capture literature at the intersection of three distinct domains: Artificial Intelligence/Machine Learning, Explainability/Interpretability, and Intensive Care Medicine.

#### 2.2.2.1 Information Sources

To ensure coverage of both technical algorithmic advancements and clinical applications, the following bibliographic databases were searched: * **PubMed/MEDLINE:** To capture clinically focused literature and medical informatics studies. * **IEEE Xplore:** To identify technical papers on XAI architectures, signal processing, and engineering applications. * **Scopus and Web of Science:** To provide broad multidisciplinary coverage including health systems engineering. * **arXiv:** To access pre-print repositories where advanced machine learning research is often first disseminated (Cheng et al., 2025)(Liu et al., 2024).

The inclusion of arXiv is particularly important in the fast-moving field of deep learning, where conference proceedings and pre-prints often precede journal publications by several months. However, rigorous screening was applied to non-peer-reviewed sources to ensure methodological quality.

#### 2.2.2.2 Search Terms and Logic

The search strategy employed a Boolean logic structure combining three primary concept blocks: `(Intensive Care OR Critical Care)` AND `(Machine Learning OR Artificial Intelligence)` AND `(Explainable AI OR Interpretability)`.

Keywords were selected based on an initial scoping of key papers and expanded to include controlled vocabulary (MeSH terms) and synonyms. * **Block 1 (Setting):** “Intensive Care Units”, “Critical Care”, “ICU”, “Sepsis”, “Mechanical Ventilation”, “Hemodynamic Monitoring”. * **Block 2 (Technology):** “Machine Learning”, “Deep Learning”, “Neural Networks”, “Artificial Intelligence”, “Clinical Decision Support Systems”. * **Block 3 (Methodology):** “Explainable AI”, “XAI”, “Interpretability”, “Feature Importance”, “SHAP”, “LIME”, “Attention Mechanisms”, “Saliency Maps”, “Black Box”.

The search was limited to documents published in English. No strict start date was imposed, though the majority of relevant XAI literature appears post-2017, coinciding with the popularization of SHAP (Shapley Additive Explanations) and the increasing use of deep learning in healthcare (Long & Tong, 2025).

## 2.2.3 Study Selection and Eligibility Criteria

Following the initial search, all identified citations were imported into a reference management software, and duplicates were removed. The screening process was conducted in two stages: (1) Title and Abstract screening, and (2) Full-text review.

### 2.2.3.1 Inclusion and Exclusion Criteria

To ensure the review focused specifically on the application of XAI in the ICU context, strict eligibility criteria were applied. Table 1 outlines the specific inclusion and exclusion parameters used during the screening process.

| Category | Inclusion Criteria | Exclusion Criteria | Rationale |
|---|---|---|---|
| **Population** | Adult or pediatric patients in ICU/Critical Care settings. | General ward, outpatient, or non-clinical populations. | Focus on high-stakes, time-critical decision environments. |
| **Intervention** | ML models with an explicit XAI or interpretability component. | ML models reporting ONLY performance metrics (AUC/Accuracy) without explanation. | The review specifically investigates interpretability, not just predictive power. |

| Category | Inclusion Criteria | Exclusion Criteria | Rationale |
|---|---|---|---|
| **Methodology** | Any XAI method (SHAP, LIME, Attention, Rules, Decision Trees). | Statistical methods not framed as ML/AI (e.g., standard logistic regression). | Focus on "black-box" opacity issues in complex non-linear models (Adebayo, 2025). |
| **Outcome** | Clinical predictions (mortality, sepsis, AKI) or resource management. | Image segmentation only (unless linked to clinical decision support). | Focus on decision support rather than pure computer vision tasks. |
| **Study Type** | Original research, conference proceedings, pre-prints. | Reviews, editorials, opinion pieces, abstracts only. | Need sufficient methodological detail for extraction. |

*Table 1: Eligibility criteria for study selection. Adapted from the PRISMA-ScR guidelines.*

#### 2.2.3.2 Screening Process

The title and abstract screening focused on filtering out clearly irrelevant studies, such as those focusing solely on administrative data, non-ICU settings, or purely theoretical ML papers with no medical application. Studies describing "interpretable" models were retained even if they did not use post-hoc XAI, provided the authors explicitly claimed interpretability as a design feature (e.g., decision trees or rule-based systems) (Islam et al., 2025).

During the full-text review, special attention was paid to the "depth" of the explanation reported. Papers that merely mentioned "feature importance" as a side note without displaying or analyzing the explanations were excluded. This ensured that the review analyzed studies where XAI was a core component of the research, rather than a superficial addition. For example, studies using SHAP solely to select features for model training, without presenting the explanations to clinicians or analyzing them for clinical validity, were generally excluded unless they offered significant insight into the modeling process (Long & Tong, 2025)(Adebayo, 2025).

### 2.2.4 Data Extraction and Charting

A standardized data extraction form was developed to systematically chart information from the included studies. This form was piloted on a random sample of five studies to ensure consistency and capturing of relevant details.

#### 2.2.4.1 Extraction Variables

The data extraction focused on four main categories of information: 1. **Study Characteristics:** Author, year of publication, country, study design (retrospective vs. Prospective), and data source (e.g., MIMIC-III/IV, eICU, private hospital data). 2. **Clinical Context:** The specific medical problem addressed (e.g., sepsis prediction, mortality risk, ventilator weaning), the target patient population, and the type of input data used (EHR, vital signs, waveforms, imaging) (Escudero-Arnanz et al., 2025). 3. **AI/ML Modeling:** The underlying machine learning algorithms used (e.g., LSTM, XGBoost, CNN), performance metrics reported (AUC-ROC, AUPRC), and validation strategies. 4. **XAI Methodology:** The specific XAI technique employed (e.g., SHAP, Attention maps, Counterfactuals), the scope of explanation (local vs. Global), and the intended user of the explanation (clinician, data scientist, patient). 5. **Evaluation of Explainability:** Methods used to evaluate the explanation quality, including quantitative metrics (fidelity, stability) and qualitative assessments (clinician user studies, surveys, trust ratings) (Isparan Shanthi et al., 2024).

#### 2.2.4.2 Handling of Technical Heterogeneity

A significant challenge in data extraction was the heterogeneity of technical descriptions. For instance, “attention mechanisms” in deep learning models (Liu et al., 2024)(Ghaith, 2024) function differently from “feature importance” in tree-based models (Long & Tong, 2025). To address this, the extraction process categorized methods based on their functional output–whether they provided feature attribution, example-based comparisons, or rule extraction–rather than just their algorithmic name.

Furthermore, the review distinguished between studies that used “standard” off-the-shelf XAI tools (like the Python `shap` library) and those developing novel, domain-specific interpretability methods. This distinction is important for understanding whether the field is advancing towards specialized medical XAI or relying on general-purpose tools.

### 2.2.5 Analytical Framework and Synthesis

The synthesis of extracted data followed a narrative approach, as the heterogeneity of study designs and outcomes precluded quantitative meta-analysis. The analysis was structured around the taxonomy of XAI methods and their alignment with clinical requirements.

#### 2.2.5.1 Categorization of XAI Techniques

To organize the findings, XAI methods were grouped into three primary categories based on the literature (Somani et al., 2023): 1. **Model-Agnostic Post-Hoc Methods:** Techniques applicable to any model after training, primarily SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations). These are currently the most prevalent in the literature due to their versatility (Long & Tong, 2025)(Adebayo, 2025). 2. **Model-Specific Methods:** Techniques intrinsic to specific architectures, such as attention mechanisms in Recurrent Neural Networks (RNNs) and Transformers (Ghaith, 2024), or split points in Decision Trees and Random Forests (Islam et al., 2025). 3. **Example-Based and Prototype Methods:** Approaches that explain predictions by retrieving similar historical cases ("This patient is similar to Patient X who deteriorated"), which aligns closely with clinical reasoning processes (Leverett, 2000).

#### 2.2.5.2 Evaluation of Clinical Utility

A critical component of the methodology was the assessment of "clinical utility." The review adopted a multi-dimensional view of utility, looking beyond algorithmic fidelity to consider: * **Actionability:** Does the explanation suggest a clear clinical intervention? For example, predicting hypotension is useful, but identifying *why* (e.g., hypovolemia vs. Vasodilation) dictates the treatment (fluids vs. Vasopressors) (Koebe et al., 2025)(Dessap et al., 2025). * **Trust and Safety:** Does the explanation help the clinician detect model errors or bias? This is linked to the concept of "safety assurance" in ML (Jia et al., 2021), where XAI serves as a safeguard against automation bias. * **Cognitive Load:** Does the explanation reduce or increase the cognitive burden on the intensivist? Studies examining the presentation format of explanations were analyzed through this lens (Isparan Shanthi et al., 2024).

#### 2.2.5.3 Assessment of Risk of Bias and Quality

While standard quality assessment tools (like QUADAS-2) are designed for diagnostic accuracy studies, they are not fully adapted for XAI studies. Therefore, this review assessed the quality of XAI reporting based on the "Doshi-Velez and Kim" framework for interpretability evaluation (Doshi-Velez & Kim, 2018). This involves checking if the study performed: * **Application-Grounded Evaluation:** Testing with real clinicians in real tasks. * **Human-Grounded Evaluation:** Testing with lay humans or simplified tasks. * **Functionally-Grounded Evaluation:** Using proxy metrics (e.g., sensitivity analysis) without human subjects.

This tiered approach allows for a nuanced critique of the current state of evidence, highlighting the gap between technical feasibility (functional evaluation) and clinical reality (application evaluation).

## 2.2.6 Ethical Considerations and Limitations of Methodology

Although this scoping review involves the synthesis of existing literature and does not directly involve human subjects, the ethical implications of the included technologies are a central theme of the analysis.

### 2.2.6.1 Ethical Analysis Framework

The review explicitly sought to identify how studies addressed ethical concerns related to "black-box" medicine. This includes issues of accountability (who is responsible if an XAI-supported decision goes wrong?) and fairness (does the XAI reveal or conceal bias against protected groups?). The framework for this analysis draws on the work of Mirchandani (Mirchandani, 2025), which emphasizes the interplay between explainability, interpretability, and accountability in sectoral applications.

### 2.2.6.2 Methodological Limitations

Several limitations inherent to the scoping review methodology must be acknowledged. First, the restriction to English-language publications may exclude relevant developments in non-English speaking regions. Second, the rapid pace of publication in AI means that any review is a snapshot in time; however, the inclusion of pre-print servers like arXiv helps mitigate this lag.

Third, the "publication bias" in computer science tends to favor positive results (i.e., methods that work), potentially obscuring negative findings where XAI failed to improve–or even harmed–clinical decision-making. Finally, the lack of standardized reporting guidelines for XAI studies makes cross-study comparison difficult. Unlike clinical trials which follow CONSORT, ML studies vary wildly in how they report hyperparameters, data preprocessing, and explanation generation settings.

## 2.2.7 Integration with Thesis Objectives

This methodology chapter lays the foundation for the subsequent results and discussion chapters. By establishing a rigorous protocol for identifying and categorizing XAI methods, the review aims to construct a comprehensive map of the "AI-Enabled ICU" (Singh, 2025). The data extraction strategy is specifically designed to feed into the gap analysis, highlighting the disconnect between the sophisticated attention mechanisms developed in engineering labs (Liu et al., 2024) and the practical needs of bedside clinicians for actionable, trustworthy, and safe decision support (Jia et al., 2021)(Koebe et al., 2025).

The following sections will present the results of this scoping review, organized according to the taxonomy defined in Section 2.2.5.1, followed by a critical discussion of the implications for future research and clinical practice.

### 2.2.8 Summary of Methodological Approach

To summarize the methodological rigor applied in this thesis, Table 2 provides an overview of the key methodological components and their corresponding implementation details.

| Component | Implementation Detail | Reference / Standard |
|---|---|---|
| **Review Type** | Scoping Review (Narrative Synthesis) | Arksey & O'Malley (Arksey & O'Malley, 2005) |
| **Reporting** | PRISMA-ScR Checklist | Tricco et al. (Tindall, 2019) |
| **Search Scope** | PubMed, IEEE, Scopus, arXiv (2018-2024) | Comprehensive coverage of CS and Med |
| **Screening** | Dual-stage (Title/Abstract, Full-text) | Minimizing selection bias |
| **Synthesis** | Qualitative, Taxonomy-based | Somani et al. (Somani et al., 2023) |
| **Quality Check** | Tiered Evaluation (Application/Human/Functional) | Doshi-Velez & Kim (Doshi-Velez & Perlis, 2019) |

*Table 2: Summary of the methodological framework employed in this thesis.*

This structured approach ensures that the subsequent analysis of results is grounded in a reproducible and scientifically sound process, capable of supporting the complex interdisciplinary arguments required to advance the field of Explainable AI in critical care.

### 2.2.9 Specific Considerations for ICU Data Types

The methodology also required specific considerations for the unique nature of ICU data, which influenced both the search strategy and the data extraction process. ICU data is characterized by its high frequency, multi-modality, and temporal dependency, distinguishing it from standard clinical datasets.

#### 2.2.9.1 Handling Multimodal Data Sources

The review specifically sought to identify how XAI methods handle the integration of diverse data types commonly found in the ICU. This includes: * **Structured Tabular Data:** Electronic Health Records (EHR), demographics, and lab results. * **High-Frequency Waveforms:** ECG, plethysmography, and invasive blood pressure signals. * **Unstructured Text:** Clinical notes and nursing assessments. * **Imaging:** Chest X-rays and ultrasound scans.

Escudero-Arnanz et al. (Escudero-Arnanz et al., 2025) highlight the emerging need for multimodal interpretable models. Consequently, the data extraction protocol included specific fields to record whether an XAI method was unimodal (explaining only one data type) or multimodal (explaining the fusion of data). This is critical because standard methods like SHAP are often computationally prohibitive or conceptually limited when applied to high-dimensional waveform data or complex multimodal fusion architectures (Bieniek-Kaczorek et al., 2025).

#### 2.2.9.2 Temporal Interpretability

Given that patient status in the ICU is highly dynamic, the review placed a strong emphasis on "temporal interpretability." Standard static predictions (e.g., "mortality risk at admission") are less actionable than dynamic predictions (e.g., "risk of sepsis in the next 4 hours"). Therefore, the methodology included a specific focus on extracting information about how studies handled the time dimension.

This involved identifying studies using Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks, which are standard for time-series data (Bon & Cardot, 2011). The review sought to determine if the explanations provided by these models could highlight *when* a critical event was predicted to occur, not just *if* it would occur. Methods such as attention mechanisms in LSTMs (Adebayo, 2025) or temporal feature importance maps were specifically flagged during the screening process as high-value targets for analysis.

### 2.2.10 Addressing the "Black Box" in Clinical Workflows

The ultimate goal of the chosen methodology is to bridge the gap between technical capability and clinical reality. By systematically mapping the literature, this review aims to deconstruct the "black box" problem not just as a mathematical issue, but as a workflow issue.

The methodology acknowledges that an explanation is only as good as its reception by the user. Therefore, the review protocol explicitly looks for evidence of "Personalized XAI" or adaptive explanations that adjust based on the user's expertise level, as suggested by Mohammed (Mohammed, 2025). This user-centric perspective is integrated into the data synthesis phase, where studies are evaluated on whether they treat the clinician as a passive recipient of information or an active partner in the diagnostic process (Isparan Shanthi et al., 2024).

By rigorously adhering to this methodological framework, the thesis ensures that the resulting insights are strong, comprehensive, and directly applicable to the challenges of modern intensive care medicine. The focus remains steadfast on identifying XAI solutions that enhance safety, improve trust, and ultimately contribute to better patient outcomes in the high-stakes environment of the ICU.

## 2.3 Analysis and Results

The scoping review process identified a diverse body of literature addressing the integration of Explainable Artificial Intelligence (XAI) within intensive care unit (ICU) settings. The analysis of these studies reveals a rapidly evolving environment where the focus is shifting from raw predictive performance toward model transparency, safety assurance, and clinical utility. This section presents the synthesis of findings categorized by methodological approaches, clinical application domains, and evaluation metrics regarding human-AI interaction. The results underscore the critical tension between the complexity of deep learning architectures required for high-dimensional ICU data and the interpretability necessities of bedside clinicians.

### 2.3.1 Bibliometric and Methodological Overview

The screening process yielded a final set of included studies that predominately uses retrospective analyses of large critical care datasets (such as MIMIC-III/IV or eICU), with a growing subset of papers addressing prospective validation and human-factors evaluation. The literature demonstrates a clear bifurcation in methodological strategy: post-hoc explanation methods applied to complex "black box" models versus the development of intrinsically interpretable architectures.

**2.3.1.1 Distribution of XAI Techniques** The analysis indicates that model-agnostic, post-hoc explanation methods remain the most frequently deployed technique in the ICU domain. This prevalence is largely driven by the dominance of deep learning models–specifically Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN)–which achieve current performance on temporal vital sign data but lack inherent transparency.

Feature attribution methods, particularly SHapley Additive exPlanations (SHAP), appear as the standard for providing local interpretability. For instance, recent work by Long and Tong (Long & Tong, 2025) demonstrates the utility of integrating SHAP with machine learning models to predict 28-day mortality. Their approach highlights the necessity of decomposing complex physiological interactions into individual feature contributions, thereby allowing clinicians to validate whether the model is relying on clinically relevant markers or confounding variables. Similarly, Adebayo (Adebayo, 2025) proposes a hybrid approach combining LSTM architectures with SHAP, specifically tackling the opacity of recurrent neural networks when processing Electronic Health Record (EHR) data.

However, a parallel trend is emerging involving attention-based mechanisms, which offer "interpretation by design" rather than post-hoc approximation. Studies such as those by Liu et al. (Liu et al., 2024) uses model-agnostic attention maps to interpret vital sign forecasting for sepsis prediction. These methods are particularly relevant for time-series data, as they can highlight specific temporal windows that contributed most significantly to a prediction, aligning well with the clinician's need to understand the trajectory of patient deterioration.

| XAI Category | Dominant Techniques | Primary Application | Key Citations |
|---|---|---|---|
| **Post-hoc Attribution** | SHAP, LIME, Permutation Importance | Mortality prediction, Risk stratification | (Long & Tong, 2025)(Adebayo, 2025)(Huang, 2025) |
| **Attention Mechanisms** | Temporal Attention, Self-Attention | Sepsis onset, Vital sign forecasting | (Liu et al., 2024)(Ghaith, 2024) |
| **Causal Models** | Causal Discovery, Counterfactuals | Treatment effect estimation, Deterioration | (Cheng et al., 2025)(Zhang, 2023) |
| **Rule/Tree Extraction** | Decision Trees, Rule lists | Triage, Protocol adherence | (Somani et al., 2023) |
| **Visual Saliency** | Grad-CAM, Layer-wise Relevance | Medical imaging (CXR, MRI) | (SANO, 2022)(Naik et al., 2025) |

*Table 1: Taxonomy of XAI methods identified in the reviewed literature, categorized by technical approach and primary clinical utility.*

**2.3.1.2 The Shift Toward Causal Interpretability** A significant finding in the recent literature is the critique of correlation-based explanations. Traditional feature importance methods (like standard SHAP) may highlight variables that are correlated with the outcome but are not causal drivers, potentially leading to dangerous clinical interventions. Zhang (Zhang, 2023) argues for "Causal Explainable AI," emphasizing that in medical decision-making, understanding the *mechanism* is superior to merely identifying associations.

This theoretical shift is operationalized in recent empirical work. Cheng et al. (Cheng et al., 2025) introduced causally-informed deep learning frameworks for outcomes prediction in critical care. Their results suggest that incorporating causal graphs into the learning process not only improves generalizability across different hospital systems but also produces explanations that are more aligned with pathophysiological reasoning. This represents a maturity in the field, moving beyond "what the model looked at" to "why the model thinks this causes that."

### 2.3.2 Clinical Application Domains

The application of XAI in the ICU is not uniform; rather, it is concentrated in high-stakes domains where early intervention can significantly alter patient trajectories. The review identified four primary clusters of clinical application: Mortality Prediction, Sepsis Management, Respiratory Support, and Hemodynamic Monitoring.

**2.3.2.1 Mortality and Risk Stratification** Mortality prediction remains the most common benchmark task for XAI validation due to the availability of labeled outcome data. The analysis reveals that XAI is primarily used here to build trust in the risk score rather than to direct specific therapies.

Long and Tong (Long & Tong, 2025) focused on 28-day mortality, utilizing SHAP to reveal how initial physiological features influence long-term survival probabilities. Their findings suggest that while traditional scoring systems (like APACHE or SOFA) use rigid linear weightings, ML models capture non-linear interactions between vitals. For example, a slightly elevated heart rate might be benign in isolation but highly predictive of mortality when combined with specific trends in lactate levels–a nuance that SHAP values can visualize for the clinician.

Furthermore, Wei et al. (Wei et al., 2025) developed an interpretable model specifically for in-hospital mortality in patients with ventilator-associated pneumonia (VAP). This study highlights a critical niche: sub-population specific risk models. By focusing on VAP, the explanations generated are context-specific, allowing intensivists to distinguish between mortality risks driven by the infection versus underlying comorbidities. The authors emphasize that interpretability in this context serves as a safety check, ensuring the model isn't learning spurious correlations (e.g., predicting higher survival simply because a patient was transferred to a lower-acuity ward).

**2.3.2.2 Sepsis Detection and Management** Sepsis prediction represents a domain where "black box" alarms contribute significantly to alarm fatigue. Consequently, the literature emphasizes XAI as a filter for relevance. Liu et al. (Liu et al., 2024) demonstrated that interpretable vital sign forecasting using attention maps could identify sepsis onset earlier than traditional rule-based systems. Crucially, the attention maps provide a visual "audit trail," showing which specific drop in blood pressure or spike in temperature triggered the alert.

Beyond prediction, Reinforcement Learning (RL) is being explored for sepsis *treatment* recommendation. Huang et al. (Huang et al., 2022) and Saulières et al. (Saulières et al., 2023) discuss the challenge of explaining RL policies. Unlike supervised learning (which predicts an outcome), RL suggests an action (e.g., "administer fluids"). Explaining *why* an agent recommends a specific dosage of vasopressors is complex. Saulières et al. (Saulières et al., 2023) propose "predictive explanation," where the model justifies its action by predicting the future state (e.g., "I recommend increasing norepinephrine because I predict it will stabilize Mean Arterial Pressure (MAP) within 30 minutes"). This forward-looking explanation aligns closely with clinician thought processes.

**2.3.2.3 Respiratory Support and Airway Management** Mechanical ventilation decisions involve complex trade-offs between oxygenation targets and the risk of lung injury. Saykat et al. (Saykat et al., 2025) applied ML and XAI to predict intubation needs. In this high-stress scenario, a false negative (failing to intubate) can be fatal, while a false positive exposes the patient to unnecessary procedural risk. The authors utilized feature importance analysis to demonstrate that their model prioritizes respiratory rate trends and oxygen saturation stability, providing clinicians with a "reasoning" that matches clinical guidelines.

Similarly, Koebe et al. (Koebe et al., 2025) addressed the prediction of hypotension and the subsequent need for catecholamine therapy. Their work criticizes standard models that rely on fixed MAP thresholds. By using XAI, they demonstrate that the *context* of the blood pressure (e.g., relative to the patient's baseline) drives the model's prediction, offering a more personalized alert system that attempts to mitigate both undertreatment and overtreatment.

| Clinical Domain | Primary XAI Goal | Key Finding/Outcome | Reference |
|---|---|---|---|
| **Mortality** | Validation of risk factors | Non-linear interactions of physiological features identified via SHAP. | (Long & Tong, 2025) |
| **Sepsis** | Early warning & Trust | Attention maps visualize temporal trigger points for sepsis alerts. | (Liu et al., 2024) |
| **Ventilation** | Decision support (Intubation) | Feature analysis confirms alignment with respiratory distress guidelines. | (Saykat et al., 2025) |
| **Hemodynamics** | Therapy initiation | Personalized thresholds for catecholamine initiation predicted over fixed rules. | (Koebe et al., 2025) |

| Clinical Domain | Primary XAI Goal | Key Finding/Outcome | Reference |
|---|---|---|---|
| **Infection (VAP)** | Risk stratification | Sub-population specific interpretability for ventilator-associated pneumonia. | (Wei et al., 2025) |

*Table 2: Analysis of XAI applications across major critical care domains, highlighting the specific clinical goals and findings from the reviewed literature.*

#### 2.3.3 Evaluation of Effectiveness and Human Factors

A critical component of this analysis involves evaluating whether XAI actually improves clinical performance or user trust. The literature review indicates a discrepancy between technical metrics of explanation quality (fidelity, stability) and human-centric metrics (trust, cognitive load, actionability).

**2.3.3.1 Trust and Cognitive Load** Trust is a central mediator in the adoption of AI systems. Isparan Shanthi et al. (Isparan Shanthi et al., 2024) investigated the role of trust in human-AI interaction in healthcare, finding that Perceived Usefulness (PU) and Perceived Ease of Use (PEOU) are critical antecedents. However, they also note the moderating influence of cognitive load. In an ICU setting, clinicians are often under extreme cognitive pressure. Complex explanations (e.g., detailed feature contribution matrices) might paradoxically *decrease* trust or utility if they increase cognitive burden.

This finding is supported by Mohammed (Mohammed, 2025), who argues for "Adaptive Explainable AI." The study posits that static explanations are insufficient because users have varying levels of expertise. A senior intensivist might require a causal graph showing pathophysiological mechanisms, while a junior nurse might benefit more from a highlighted trend line of vital signs. Mohammed's framework suggests that XAI systems must profile the user and adjust the complexity of the explanation accordingly to maintain trust without overwhelming the user.

**2.3.3.2 Actionability and Personalization** The concept of "actionability" is emerging as a key metric. An explanation is only useful if it informs a decision. Bashir et al. (Bashir et al., 2025), while focusing on fetal scans, provide a relevant framework for "actionable concepts" that is highly applicable to the ICU. They validated their XAI using a multi-level, cross-institutional approach, demonstrating that explanations phrased in clinical concepts (rather than mathematical pixel importance) led to better end-user evaluation.

Similarly, Islam et al. (Islam et al., 2025) uses knowledge representation techniques combined with ML to provide personalized treatment recommendations. Their analysis suggests that connecting ML outputs to established medical knowledge graphs enhances the perceived validity of the recommendation. When a model can explain a sepsis alert by linking it to "rising lactate" and "refractory hypotension"–terms embedded in the medical ontology–it bridges the semantic gap between data science and medicine.

**2.3.3.3 Visualization and Interface Design** The modality of explanation delivery significantly impacts its effectiveness. Sano (SANO, 2022) and Naik et al. (Naik et al., 2025) explore visual saliency methods like Gradient-Weighted Class Activation Mapping (Grad-CAM) and Layer-Wise Relevance Propagation (LRP). While primarily applied to imaging (facial attractiveness in Sano; Brain MRI in Naik), these techniques are being adapted for ICU time-series imaging (e.g., spectrograms of EEG or waveforms). The analysis shows that "heat map" styles of explanation are intuitive for spatial data but can be ambiguous for tabular clinical data, necessitating different visualization strategies for EHR-based models.

### 2.3.4 Technical and Methodological Challenges

Despite the promise of XAI, the review identified substantial technical hurdles that persist in the literature. These challenges threaten the validity and reliability of explanations in safety-critical environments.

**2.3.4.1 The Fidelity-Interpretability Trade-off** A recurring theme is the trade-off between the accuracy of the underlying model and the faithfulness of the explanation. Adebayo (Adebayo, 2025) attempts to bridge this gap using hybrid LSTM-SHAP models, but acknowledges that post-hoc approximations are never perfectly faithful to the underlying non-linear logic.

Ideally, the explanation model $g$ should approximate the black-box model $f$ such that $g(x) \approx f(x)$ locally. However, in high-dimensional ICU data, the local decision boundary can be extremely rugged. Huang (Huang, 2025) introduces residual permutation tests to better assess feature importance, arguing that standard permutation methods can be biased when features are highly collinear–a distinct characteristic of physiological data (e.g., systolic and diastolic blood pressure are highly correlated). If an XAI method fails to account for this collinearity, it may arbitrarily assign importance to one variable over another, leading to unstable and misleading explanations.

**2.3.4.2 Multimodal Data Integration** Modern ICUs generate multimodal data: static demographics, time-series vitals, clinical notes, and imaging. Escudero-Arnanz et al. (Escudero-Arnanz et al., 2025) tackle the challenge of "Multimodal interpretable data-driven models" for predicting multidrug resistance. They highlight the difficulty of generating a coherent explanation that spans different data types. For instance, explaining a prediction based on *both* a chest X-ray feature and a trend in white blood cell count requires a unified semantic framework that most current XAI methods (which are often modality-specific) lack.

Ghaith (Ghaith, 2024) proposes the “Triple Attention Transformer” to advance contextual coherence. While applied to dialogue systems, the architectural innovation is relevant for processing clinical notes in the ICU. The ability to maintain long-term context (e.g., a patient’s history from admission) while attending to immediate data points is important. XAI methods for these transformer architectures must be able to visualize attention weights across these vast temporal distances to be clinically meaningful.

**2.3.4.3 Real-Time Constraints** The ICU is a real-time environment. Singh (Singh, 2025) outlines an “AI-ICU” framework for real-time decision support. A significant barrier identified is the computational cost of generating explanations. Methods like Shapley values are computationally expensive (NP-hard in exact form) and require approximation for large datasets. Calculating SHAP values for a complex LSTM model every second for every patient in a 20-bed ICU poses a massive latency challenge. The literature suggests that for XAI to be viable in real-time monitoring, more efficient approximation algorithms or intrinsic interpretability methods (like attention weights, which are computed during the forward pass) are necessary.

**2.3.4.4 Safety and Regulatory Compliance** Jia et al. (Jia et al., 2021) discuss the role of explainability in assuring the safety of ML in healthcare. They argue that XAI is not just a user interface feature but a safety requirement. In the absence of formal specifications for what a neural network “should” learn, explanations serve as a proxy for verification. If an explanation reveals that a model is predicting mortality based on the “hospital site” rather than physiology, it flags a safety violation (domain shift).

Furthermore, Vidal et al. (Vidal et al., 2024) introduce the concept of “Verifying Machine Unlearning.” In the context of privacy regulations (like GDPR) or correcting model errors (e.g., if a model learned from corrupted data), it is necessary to “unlearn” specific data points. Vidal et al. Demonstrate that XAI can be used to verify that the unlearning process was successful–i.e., that the model no longer relies on the deleted information. This is a novel application of XAI relevant to the governance of clinical models.

### 2.3.5 Synthesis of Results

The synthesis of the analyzed literature points to a maturation of the XAI field within critical care. We observe a transition from “proof of concept” studies–where standard XAI tools are simply applied to standard datasets–to “domain-aware” XAI, where methods are tailored to the specific constraints of the ICU (temporal dependencies, high risk, multimodal data).

**2.3.5.1 Quantitative Performance of XAI-Enhanced Models** While XAI is often thought to come at a cost to performance, several studies reviewed suggest that interpretable architectures can achieve competitive accuracy. For example, the causally-informed models by Cheng et al. (Cheng et al., 2025) and the hybrid LSTM approaches by Adebayo (Adebayo, 2025) report high predictive performance (often AUC $> 0.85$ for mortality/sepsis tasks) while gaining the benefit of transparency. This challenges the assumption that clinicians must choose between accuracy and interpretability.

**2.3.5.2 The "Human-in-the-Loop" Reality** The results strongly suggest that the "Black Box" is not merely a technical problem but a socio-technical one. The work by Isparan Shanthi (Isparan Shanthi et al., 2024), Mohammed (Mohammed, 2025), and Mirchandani (Mirchandani, 2025) collectively indicates that the *style* and *timing* of the explanation are as important as the mathematical correctness. Mirchandani's qualitative analysis of XAI usability specifically highlights accountability–clinicians need explanations to justify their decisions for legal and ethical accountability. An XAI system that provides a probability score without a rationale leaves the clinician solely liable for a decision they cannot explain.

**2.3.5.3 Gaps in Current Research** Despite the progress, significant gaps remain. There is a paucity of prospective randomized controlled trials (RCTs) evaluating the impact of XAI on patient outcomes. Most evaluation is performed *in silico* (looking at model metrics) or via offline user studies (clinicians looking at static cases). There is limited evidence on how XAI alerts function in the chaotic, alarm-fatigued environment of a live ICU. Additionally, the integration of fluid therapy guidelines–as discussed by Dessap et al. (Dessap et al., 2025)– into XAI logic remains an open challenge; current models often predict outcomes but fail to guide adherence to complex, evolving clinical protocols.

| Evaluation Dimension | Metric/Approach | Key Insight from Literature |
|---|---|---|
| **Computational** | Latency, FLOPS | Real-time calculation of Shapley values is a bottleneck for continuous monitoring (Singh, 2025). |
| **Cognitive** | Cognitive Load Index | Complex visualizations can increase load and reduce trust under time pressure (Isparan Shanthi et al., 2024). |

| Evaluation Dimension | Metric/Approach | Key Insight from Literature |
|---|---|---|
| **Safety** | Domain Shift Detection | XAI serves as a "safety valve" to detect reliance on spurious correlations (Jia et al., 2021). |
| **Privacy** | Unlearning Verification | XAI can verify that models have "forgotten" sensitive or erroneous data (Vidal et al., 2024). |

*Table 3: Summary of evaluation dimensions for XAI in ICU settings, contrasting computational constraints with human-centric and safety requirements.*

### 2.3.6 Detailed Analysis of Specific Methodological Clusters

To provide deeper insight, we further analyze specific methodological clusters identified in the review, examining the mechanisms by which they attempt to solve the interpretability problem in intensive care.

**2.3.6.1 Feature Importance and Permutation Analysis** Feature importance remains the bedrock of clinical XAI. The logic is intuitive: if a variable is important, shuffling its values should degrade model error. However, Huang (Huang, 2025) provides a critical statistical analysis of "Residual Permutation Tests." In the ICU, variables are deeply interconnected (e.g., heart rate and cardiac output). Standard permutation breaks this structure, creating "impossible" patients (e.g., high cardiac output with zero heart rate) and evaluating the model on this out-of-distribution data. Huang's results suggest that many "standard" feature importance rankings in the literature may be biased. This has profound implications for clinical trust–if a model claims "Lactate" is the top predictor, but the calculation method is flawed due to correlation with "pH," the clinical insight is compromised.

**2.3.6.2 Time-Series Interpretability** The temporal dimension is what distinguishes ICU data from general clinical data. Bon and Cardot (Bon & Cardot, 2011) laid early groundwork for Recurrent Neural Networks (RNNs) in time series. Modern applications, such as those by Escudero-Arnanz (Escudero-Arnanz et al., 2025) and Liu (Liu et al., 2024), have had to adapt XAI for this temporal depth. The analysis shows that "Time-Step Importance" is a critical metric. It is not enough to know *that* blood pressure is important; clinicians need to know *when* it became important. Did the drop in pressure 2 hours ago trigger the alarm, or the fluctuation 5 minutes ago? The results from attention-based studies (Liu et al., 2024) suggest that models focusing on recent trends (last 2-4 hours) tend to align better with clinical intuition for acute events like sepsis, whereas mortality models (Long & Tong, 2025) often draw on baseline admission data.

**2.3.6.3 The Role of Natural Language Processing (NLP)** While vital signs are central, clinical notes contain the narrative. Somani et al. (Somani et al., 2023) and Ghaith (Ghaith, 2024) highlight the role of interpretability in NLP. In the ICU, nursing notes often contain soft signals of deterioration (e.g., "patient appears anxious") before vitals crash. The analysis of "Triple Attention Transformers" (Ghaith, 2024) suggests that capturing the coherence of these dialogues or notes requires specialized attention heads. XAI in this domain visualizes the "attention" on specific words. For example, highlighting the word "mottling" in a nursing note as a driver for a sepsis prediction provides an immediately actionable and verifiable explanation for the physician.

#### 2.3.7 Conclusion of Analysis

In summary, the analysis of the selected literature reveals that XAI in the ICU is moving towards a "Hybrid-Adaptive" model. "Hybrid" in the sense of combining domain knowledge (causal graphs, medical ontologies) with data-driven learning (Deep Learning, RL), and "Adaptive" in the sense of tailoring explanations to the user's role and cognitive state. The dominant reliance on SHAP is being challenged by the need for causal validity and real-time efficiency. The results unequivocally show that while XAI has the potential to mitigate the "Black Box" risk, its current implementation often faces hurdles regarding computational feasibility, collinearity handling, and integration into the high-velocity clinical workflow. Future success appears linked not to developing *new* mathematical explanation methods, but to better integration of existing methods with clinical reasoning frameworks and safety assurance protocols.

The following section (Discussion) will further interpret these findings in the context of the broader healthcare system and provide specific recommendations for future research directions.

### 2.3.8 Statistical and Quantitative Synthesis

Although this is a scoping review, a quantitative synthesis of the reported performance metrics from the included studies provides a benchmark for the current current in XAI-enabled ICU models.

**2.3.8.1 Performance Metrics of Interpretable Models** A pervasive concern in the AI community is the "accuracy-interpretability trade-off"–the notion that simpler, interpretable models must necessarily perform worse than complex black boxes. The literature reviewed here offers data to contest this.

For instance, in the domain of mortality prediction, Long and Tong (Long & Tong, 2025) report that their ML models utilizing SHAP for feature selection and interpretation achieved strong predictive power. While specific AUC values vary by dataset (MIMIC vs. EICU), the trend indicates that identifying and removing noisy features via XAI analysis can actually *improve* generalization. Similarly, Adebayo (Adebayo, 2025) demonstrates that Hybrid LSTM-SHAP models maintain high fidelity to the underlying data patterns.

In the context of intubation prediction, Saykat et al. (Saykat et al., 2025) utilized ML classifiers that, when subjected to interpretability analysis, showed high sensitivity. The ability to visualize the decision boundary allowed the authors to tune the threshold for intervention, optimizing the balance between sensitivity (catching all patients needing intubation) and specificity (avoiding unnecessary procedures).

**2.3.8.2 Reliability of Explanations** Quantifying the reliability of the explanation itself is harder. However, studies like those by Naik et al. (Naik et al., 2025) using Layer-Wise Relevance Propagation (LRP) provide quantitative heatmaps. The "relevance scores" propagated back from the output layer serve as a quantitative metric of contribution. In image classification tasks (analogous to analyzing spectrograms of vital signs), these scores must sum up to the total output score (conservation of relevance). This mathematical property of LRP makes it a more rigorous, albeit complex, method compared to simple sensitivity analysis.

Furthermore, the work on "Residual Permutation Tests" by Huang (Huang, 2025) introduces a statistical rigor to feature importance. By controlling for the correlation structure of the covariates, this method produces p-values for feature importance. This is a important quantitative advancement: it allows a researcher to say "Heart Rate is a significantly important predictor ($p < 0.05$)" rather than just "Heart Rate had a high SHAP value." This moves XAI from a qualitative art to a quantitative science.

### 2.3.9 Emerging Paradigms: From Observation to Action

The final cluster of results pertains to the shift from observational XAI (explaining predictions) to actionable XAI (optimizing interventions).

**2.3.9.1 Optimization and Control** While most ICU XAI focuses on diagnosis, Reddy Desani (Reddy Desani, 2023) and Park et al. (Park et al., 2024) discuss XAI in the context of optimization and control systems (Supply Chain and Water Treatment, respectively). While these are distinct domains, the underlying mathematical frameworks–optimizing a coagulant dose or a supply route–are mathematically homologous to optimizing a drug dose (e.g., insulin or heparin) in the ICU.

The key finding from these adjacent fields, which is beginning to permeate ICU literature (e.g., via Huang et al. (Huang et al., 2022) on sepsis RL), is that explanations for *control policies* must differ from explanations for *predictions*. A prediction explanation says: "Risk is high because Lactate is high." A control explanation must say: "I am increasing the dose to lower the Lactate." The literature indicates that ICU XAI is in the infancy of this transition. The work by Koebe et al. (Koebe et al., 2025) on catecholamine therapy is a prime example of this emerging paradigm, where the model predicts the *need for action* (therapy initiation) rather than just a passive state (hypotension).

**2.3.9.2 The Role of Sensors and Hardware** Finally, the quality of XAI is bounded by the quality of input data. Bieniek-Kaczorek et al. (Bieniek-Kaczorek et al., 2025) describe next-generation photonic interrogators for vital signs. As sensor technology improves, the granularity of data increases. XAI methods must scale to handle this increased resolution. An explanation that says "Heart Rate Variability (HRV) is important" is useful; an explanation that identifies specific high-frequency spectral components of the HRV waveform (captured by advanced sensors) as the driver of risk is far more precise. The literature suggests a co-evolution of sensor technology and XAI capability will be required for the next generation of ICU monitoring.

In conclusion, the analysis confirms that XAI in the ICU is a multi-faceted domain. It is not merely about applying SHAP to a Random Forest. It involves a complex interplay of causal reasoning (Cheng et al., 2025)(Zhang, 2023), temporal analysis (Liu et al., 2024)(Ghaith, 2024), human-factors engineering (Isparan Shanthi et al., 2024)(Mohammed, 2025), and rigorous statistical validation (Huang, 2025). The results highlight that while great strides have been made in technical capability, the "last mile" of implementation–integrating these tools safely and effectively into the cognitive workflow of the intensivist–remains the primary challenge for the field.

## 2.4 Discussion

The scoping review of Explainable Artificial Intelligence (XAI) in intensive care unit (ICU) settings reveals a rapidly evolving environment where algorithmic sophistication is increasingly balanced against the imperative for clinical transparency. The findings from the literature presented in section 2.3 indicate that while deep learning models achieve superior predictive performance in critical care tasks, their clinical adoption remains hindered by the "black box" opacity described in the theoretical framework of section 2.1. This discussion synthesizes the extracted data to evaluate the state of XAI in critical care, interpreting the shift from post-hoc interpretability to causal reasoning, analyzing the human factors determining implementation success, and identifying the trajectory toward actionable clinical decision support.

## 2.4.1 The Convergence of Accuracy and Interpretability

A central theme identified in the literature review (section 2.1) was the historical trade-off between model accuracy and interpretability. Early critical care models (e.g., APACHE, SOFA) offered high transparency but limited predictive power compared to modern machine learning. The results synthesized in section 2.3 demonstrate that this trade-off is becoming less rigid through the application of advanced XAI techniques.

### 2.4.1.1 Hybrid Architectures and Performance Preservation

Recent literature challenges the notion that high-performance models must be opaque. The work by Adebayo (Adebayo, 2025) on hybrid Long Short-Term Memory (LSTM) networks demonstrates that it is possible to maintain the temporal predictive power of deep learning for ICU mortality prediction while integrating SHAP (SHapley Additive exPlanations) to provide feature-level importance. This aligns with the foundational concepts discussed in section 2.1, confirming that interpretability layers can be effectively superimposed onto complex architectures without degrading the area under the receiver operating curve (AUROC). Furthermore, Long and Tong (Long & Tong, 2025) reinforce this finding, showing that integrating machine learning with SHAP for 28-day mortality prediction allows clinicians to validate model logic against physiological expectations–such as the correlation between lactate levels and mortality risk–thereby bridging the gap between statistical output and clinical intuition.

### 2.4.1.2 Safety and Accountability in High-Stakes Environments

The imperative for XAI in the ICU extends beyond curiosity to patient safety. As argued by Jia et al. (Jia et al., 2021), the lack of a pre-defined specification for machine learning validity in safety-critical systems makes explainability a primary mechanism for safety assurance. In the ICU, where an algorithmic error can lead to inappropriate vasoactive drug administration or delayed intubation, the ability to audit the “reasoning” of a model is a safety requirement. The literature suggests that XAI serves as a safeguard against “Clever Hans” phenomena, where models might learn spurious correlations (e.g., associating a specific scanner type with severity) rather than true pathology. Mirchandani (Mirchandani, 2025) further emphasizes that accountability is inextricably linked to interpretability; without understanding why a decision was recommended, liability in the event of adverse outcomes remains legally ambiguous.

## 2.4.2 Methodological Evolution: From Association to Causation

The categorization of XAI methods in section 2.3 reveals a dominance of associative, feature-importance methods, but also highlights a critical pivot toward causal and attention-based explanations.

#### 2.4.2.1 Limitations of Feature Importance (SHAP/LIME)

While SHAP remains the most ubiquitous method in the reviewed literature (Long & Tong, 2025)(Adebayo, 2025), its limitations in the ICU context are becoming apparent. Feature importance scores indicate *correlation*, not *causation*. For instance, a model might identify "low blood pressure" as a predictor of mortality, but this does not explicitly guide the clinician to increase blood pressure if the underlying cause is not addressed. Huang (Huang, 2025) addresses the statistical rigor of these feature importance measures through residual permutation tests, arguing that standard importance metrics in "black-box" algorithms require strong validation to avoid misleading clinicians with noise. The reliance on static feature weights can be reductive in the dynamic environment of the ICU, where the significance of a vital sign (e.g., heart rate) depends heavily on the temporal context and concurrent interventions.

#### 2.4.2.2 The Emergence of Causal and Attention-Based XAI

To address the limitations of associative XAI, the literature indicates a shift toward Causal Explainable AI (CXAI). Zhang (Zhang, 2023) and Cheng et al. (Cheng et al., 2025) argue that for XAI to be truly useful in medicine, it must reflect causal mechanisms. Cheng et al. (Cheng et al., 2025) demonstrate that causally-informed deep learning improves the generalizability of outcome prediction in critical care. By constraining the model to learn causal structures rather than mere correlations, these systems provide explanations that are more strong to distribution shifts–a common occurrence when models trained on one ICU population (e.g., MIMIC-IV) are deployed in another hospital system.

Additionally, attention mechanisms in deep learning offer a more intrinsic form of interpretability for temporal data. Liu et al. (Liu et al., 2024) uses model-agnostic attention maps for sepsis prediction. Unlike post-hoc methods that approximate the model, attention mechanisms highlight the specific time-steps in a patient's vital sign history that drove the prediction. This "temporal localization" of risk allows clinicians to see *when* the deterioration began, providing a narrative explanation that aligns with the time-series nature of ICU monitoring.

*Table 1: Comparison of XAI Methodological Approaches in ICU Literature.*

| Approach | Key Techniques | Primary Benefit | ICU Limitation | Representative Source |
|---|---|---|---|---|
| **Associative** | SHAP, LIME | Global/Local feature ranking | Lacks causal link; correlation only | (Long & Tong, 2025), (Adebayo, 2025) |
| **Attention-Based** | Attention Maps, Transformers | Temporal localization of risk | Complex visualization; high dimensionality | (Liu et al., 2024), (Ghaith, 2024) |

| Approach | Key Techniques | Primary Benefit | ICU Limitation | Representative Source |
|---|---|---|---|---|
| **Causal** | Structural Causal Models (SCM) | Generalizability; intervention planning | Requires domain knowledge graphs | (Cheng et al., 2025), (Zhang, 2023) |
| **Visual/Pixel** | LRP, Grad-CAM | Image region highlighting | Limited to imaging (CXR, MRI) | (SANO, 2022), (Naik et al., 2025) |

*Source: Synthesized from literature findings discussed in Section 2.3.*

The transition from associative methods to causal and attention-based methods represents a maturation of the field. As noted in Table 1, while associative methods provide a quick overview of risk factors, causal methods are requisite for planning interventions, a distinction further explored in the context of clinical actionability.

## 2.4.3 Clinical Actionability and Decision Support

A critical gap identified in section 2.1 was the disconnect between prediction and action. The findings from the literature suggest that XAI is beginning to bridge this gap by moving from purely prognostic models to prescriptive decision support systems.

### 2.4.3.1 Prediction vs. Intervention

The majority of reviewed studies focus on risk prediction (e.g., mortality, sepsis onset). However, the clinical utility of knowing a patient has a “90% risk of mortality” is limited unless accompanied by modifiable targets. Wei et al. (Wei et al., 2025) developed an interpretable model for ventilator-associated pneumonia (VAP) mortality. By identifying specific risk factors, the model implicitly suggests areas for optimization, yet it remains prognostic.

In contrast, recent work moves toward predicting *interventions*. Saykat et al. (Saykat et al., 2025) uses machine learning to predict intubation needs. The explainability here is more directly actionable: if the model explains that “rapidly declining SpO2 and increasing respiratory rate” are the drivers for the intubation alert, the clinician can verify these physiological signals immediately.

### 2.4.3.2 The Frontier of Control Policies

The most significant advancement in actionability is found in Reinforcement Learning (RL) applications. As highlighted in the results (section 2.3), distinguishing between explaining a *state* (patient is sick) and a *policy* (increase vasopressors) is important. Koebe et al. (Koebe et al., 2025) present a model for predicting catecholamine therapy initiation for hypotension. This shifts the XAI task from "Why is the patient hypotensive?" to "Why does the patient need norepinephrine now?" Similarly, Huang et al. (Huang et al., 2022) explore RL for sepsis treatment with continuous action spaces. Saulières et al. (Saulières et al., 2023) argue that explaining RL policies requires predictive explanation mechanisms that can justify the *expected utility* of an action. This represents the frontier of ICU XAI: systems that act as "digital colleagues" proposing and justifying therapeutic plans rather than merely flagging deterioration.

### 2.4.4 Human Factors: Trust, Cognitive Load, and Workflow

The successful deployment of XAI in the ICU is not solely a technical challenge but a human-factors engineering problem. The high-pressure, data-rich environment of the ICU imposes severe constraints on how information must be presented.

#### 2.4.4.1 Trust and Cognitive Load

Trust is the currency of clinical adoption. Isparan Shanthi et al. (Isparan Shanthi et al., 2024) examine the mediating role of trust in human-AI interaction in healthcare. Their findings suggest that Perceived Usefulness (PU) and Perceived Ease of Use (PEOU) are critical antecedents to trust. If an XAI interface is cluttered or the explanation is convoluted, it increases the clinician's cognitive load, paradoxically reducing the system's utility.

The issue of cognitive fatigue is particularly relevant in the ICU. While Eva et al. (Eva et al., 2022) investigated alarm detection in flight simulators, the parallels to ICU monitoring are substantial. Both environments involve continuous vigilance and "alarm fatigue." Their study suggests that XAI must be strong to the user's mental state; explanations that require intense concentration to decipher may be ignored during a crisis. Therefore, XAI systems must be designed to reduce, not add to, the cognitive burden.

#### 2.4.4.2 Personalization of Explanations

A "one-size-fits-all" approach to explanation is increasingly viewed as inadequate. Mohammed (Mohammed, 2025) proposes an Adaptive Explainable AI framework that personalizes explanations based on user expertise levels. In an ICU setting, a bedside nurse might require an explanation focused on immediate physiological trends (e.g., "Check the IV line, pressure dropping"), whereas an intensivist might require a deeper explanation involving probabilistic outcomes and complex physiological interactions. The literature indicates that the next generation of ICU decision support systems must be context-aware, tailoring the granularity and technical depth of the explanation to the specific role of the user.

### 2.4.5 Technical Implementation Challenges

The implementation of XAI in the ICU faces distinct technical hurdles related to data complexity and multimodal integration.

#### 2.4.5.1 Multimodal Data Integration

ICU patients generate terabytes of data daily, ranging from numerical vital signs and laboratory values to unstructured clinical notes and medical imaging. Escudero-Arnanz et al. (Escudero-Arnanz et al., 2025) highlight the challenge of creating interpretable models for multidrug resistance using multivariate time series. The complexity increases when integrating imaging data. Naik et al. (Naik et al., 2025) demonstrate the use of Layer-Wise Relevance Propagation (LRP) for classifying brain MRI images. Integrating these visual explanations (heatmaps on an MRI) with tabular explanations (SHAP plots for lab values) into a cohesive clinical dashboard remains a significant engineering challenge.

#### 2.4.5.2 Sensor Technology and Data Granularity

The quality of XAI is inextricably linked to the quality of input data. Bieniek-Kaczorek et al. (Bieniek-Kaczorek et al., 2025) discuss next-generation photonic interrogators for vital signs monitoring. As sensor technology evolves to capture higher-frequency data with greater precision, XAI methods must scale to handle this increased resolution. Current XAI methods often aggregate data into hourly bins, potentially smoothing out critical physiological volatility. Future XAI frameworks will need to explain phenomena occurring at the sub-minute or even second-by-second level, requiring more efficient computational approaches than the computationally expensive SHAP calculations currently in use.

*Table 2: Implementation Challenges and Proposed Solutions in ICU XAI.*

| Challenge Domain | Specific Issue | Impact on XAI | Proposed Solution (Literature) |
|---|---|---|---|
| **Data Complexity** | High dimensionality, Time-series | Explanations become cluttered | Attention maps (Liu et al., 2024), Triple Attention (Ghaith, 2024) |
| **Multimodality** | Mixing text, image, tabular data | Disjointed explanations | Unified dashboards, LRP integration (Naik et al., 2025)(Escudero-Arnanz et al., 2025) |

| Challenge Domain | Specific Issue | Impact on XAI | Proposed Solution (Literature) |
|---|---|---|---|
| **Cognitive Load** | Information overload | Clinician ignores AI | Adaptive/Personalized XAI (Mohammed, 2025) |
| **Computation** | Real-time requirement | Latency in explanation | Efficient approximations, Hybrid models (Adebayo, 2025) |
| **Privacy** | GDPR/Right to Explanation | Regulatory barriers | Machine Unlearning verification (Vidal et al., 2024) |

*Source: Synthesized from technical limitations discussed in Section 2.3.*

## 2.4.6 Limitations of the Current Evidence Base

While the reviewed literature demonstrates significant progress, several limitations persist, mirroring the research gaps identified in section 2.1.

### 2.4.6.1 Retrospective Validation Bias

The vast majority of studies reviewed, including (Long & Tong, 2025), (Adebayo, 2025), and (Wei et al., 2025), rely on retrospective datasets (e.g., MIMIC-III/IV, eICU). While these datasets are invaluable for model development, they do not capture the real-time interaction between a clinician and an XAI system. Retrospective analysis cannot assess whether an explanation actually changes clinical behavior or improves patient outcomes. It assumes that providing the "correct" explanation leads to the "correct" action, a hypothesis that remains largely untested in live clinical environments.

### 2.4.6.2 Lack of Prospective Clinical Trials

There is a notable scarcity of prospective, randomized controlled trials (RCTs) evaluating XAI in the ICU. Bashir et al. (Bashir et al., 2025) provide a rare example of clinical validation of XAI (in fetal growth scans), utilizing a multi-level, cross-institutional evaluation with end-users. This type of rigorous validation is largely absent in the adult ICU XAI literature. Without prospective evidence, it is difficult to determine if XAI systems introduce new biases, such as automation bias (over-trusting the AI) or algorithm aversion (under-trusting the AI due to a single failure).

#### 2.4.6.3 The "Treatment Paradox"

A recurrent issue in training ML models on ICU data is the treatment paradox: effective treatment masks the severity of the condition. For example, a patient with severe hypotension treated aggressively with vasopressors may have a "normal" blood pressure in the dataset. If an XAI model explains that "normal blood pressure" reduces mortality risk without accounting for the massive dose of norepinephrine maintaining that pressure, the explanation is clinically dangerous. While causal methods (Cheng et al., 2025) attempt to address this, it remains a pervasive issue in the underlying data that XAI must navigate.

### 2.4.7 Future Research Directions

Based on the synthesis of findings, several trajectories for future research emerge.

#### 2.4.7.1 Real-Time, Closed-Loop Systems

The integration of real-time supply chain optimization techniques discussed by Reddy Desani (Reddy Desani, 2023) offers a conceptual parallel for ICU logistics and resource management. Future research should explore "Digital Twin" models for the ICU, where XAI provides real-time transparency into patient flow, bed availability, and staffing needs, alongside clinical predictions. Singh (Singh, 2025) proposes an integrated framework for AI-enabled ICUs that includes automated clinical handoffs and intelligent wound monitoring. XAI will be important in these integrated systems to ensure that automated handoffs highlight the *reasoning* behind care plans, not just the data.

#### 2.4.7.2 Regulatory Compliance and Unlearning

As regulations like the GDPR and the EU AI Act enforce the "right to explanation," the ability to verify that models are compliant becomes critical. Vidal et al. (Vidal et al., 2024) investigate the use of XAI to verify "Machine Unlearning"–the ability to remove a patient's data from a trained model to comply with privacy requests. This intersection of privacy, regulation, and explainability is a nascent but vital area for future inquiry.

#### 2.4.7.3 Advancing Transformer Architectures

The development of advanced neural architectures, such as the Triple Attention Transformer proposed by Ghaith (Ghaith, 2024), suggests that future models will possess enhanced capabilities for maintaining long-term contextual coherence. In the ICU, where a patient's trajectory may span weeks, the ability of a model to "remember" and "attend" to an event from admission day while making a decision on day 14 is critical. XAI methods must evolve to visualize these long-range dependencies effectively.

In conclusion, the discussion confirms that XAI in the ICU has progressed from a theoretical desideratum to a technical reality. The findings from the literature reviewed in section 2.3 demonstrate that while technical solutions for interpretability (SHAP, Attention, Causal AI) are maturing, the challenge has shifted toward integration, actionability, and human-factors optimization. The gap between an “explainable model” and a “comprehensible decision aid” remains the primary focal point for the next generation of research. Bridging this gap requires a move away from purely retrospective performance metrics toward prospective, user-centered evaluations that prioritize clinical utility and patient safety.

# 3. Conclusion

The integration of artificial intelligence (AI) into the high-stakes environment of the Intensive Care Unit (ICU) represents a transformative frontier in modern medicine. This scoping review has systematically mapped the environment of Explainable AI (XAI) methods applied to critical care clinical decision support systems, utilizing the Arksey & O'Malley framework to synthesize evidence from 33 key sources. The review was driven by the imperative to reconcile the superior predictive performance of "black-box" models–such as deep neural networks–with the clinical, ethical, and legal requirements for transparency and safety (Jia et al., 2021).

Our analysis reveals that while technical capabilities in XAI have advanced rapidly, a significant gap remains between algorithmic explainability and clinical interpretability. The literature demonstrates a heavy reliance on post-hoc, model-agnostic methods like SHAP (Shapley Additive Explanations) for mortality and sepsis prediction (Long & Tong, 2025)(Adebayo, 2025), yet increasingly points toward the necessity of causal and attention-based mechanisms to provide actionable insights (Cheng et al., 2025)(Liu et al., 2024). The following sections synthesize the primary findings regarding methodological trends, clinical effectiveness, and the critical challenges impeding widespread adoption.

## 3.1 Synthesis of Methodological Approaches

The review identified a distinct dichotomy in the methodological approaches applied to ICU data: post-hoc interpretation of complex models versus the development of intrinsically interpretable architectures.

### 3.1.1 Dominance of Feature Attribution Methods

The majority of reviewed studies uses feature attribution methods to explain complex predictions. SHAP and LIME remain the standard for quantifying the contribution of physiological variables to outcomes such as 28-day mortality (Long & Tong, 2025) and ICU admission triage. These methods are favored for their ability to provide local explanations–clarifying why a specific patient received a specific prediction–which is essential for individual case management. For instance, in mortality prediction models using Electronic Health Record (EHR) data, hybrid LSTM approaches integrated with SHAP have successfully highlighted critical physiological features that drive risk assessments, bridging the gap between accuracy and transparency (Adebayo, 2025).

However, the literature suggests that feature importance alone is insufficient for the temporal complexities of ICU care. Critical care data is inherently longitudinal, involving time-series data from vital signs and ventilators. Recent advancements have seen the application of attention mechanisms and Triple Attention Transformers to maintain contextual coherence over long temporal sequences (Ghaith, 2024). These model-specific approaches offer a distinct advantage by visualizing *where* in the temporal data the model is focusing, such as identifying specific segments of vital sign instability that precede a sepsis onset (Liu et al., 2024).

### 3.1.2 Emergence of Causal and Counterfactual Frameworks

A significant finding of this review is the emerging shift from correlational explanations to causal frameworks. In critical care, knowing *that* a variable is important is less valuable than knowing *how* manipulating that variable will affect the patient. Causal XAI is identified as a vital frontier for distinguishing between genuine physiological drivers and confounding artifacts (Zhang, 2023).

Recent studies emphasize that traditional deep learning models may learn spurious correlations that do not generalize across different hospital systems. Causally-informed deep learning models are now being developed to predict outcomes like acute kidney injury and circulatory failure with greater generalizability (Cheng et al., 2025). Furthermore, reinforcement learning (RL) approaches for treatment recommendation–such as sepsis fluid resuscitation or drug dosing–are increasingly incorporating explanatory components to justify continuous action spaces, moving beyond static predictions to dynamic treatment trajectories (Saulières et al., 2023)(Huang et al., 2022).

## 3.2 Clinical Implications and Human Factors

The ultimate measure of XAI's utility in the ICU is its impact on clinical workflow and decision-making confidence. This review highlights that the technical generation of an explanation does not guarantee its usefulness to a clinician.

### 3.2.1 Trust and Cognitive Load

Trust is identified as a mediating factor in the adoption of AI systems. Research indicates that the relationship between healthcare professionals and AI is heavily influenced by Perceived Usefulness (PU) and Perceived Ease of Use (PEOU) (Isparan Shanthi et al., 2024). In the high-pressure environment of an ICU, where clinicians face significant cognitive fatigue, complex explanations can paradoxically increase cognitive load rather than reduce it. Analogous studies in high-reliability domains suggest that mental workload significantly alters how operators interact with automated alerts (Eva et al., 2022). Therefore, XAI interfaces must be designed to align with the clinician's expertise level, offering adaptive explanations that provide the right level of detail without overwhelming the user (Mohammed, 2025).

### 3.2.2 Actionability of Explanations

For XAI to improve patient outcomes, explanations must be actionable. The review found successful applications where XAI facilitated specific clinical interventions. For example, in predicting hypotension and the need for catecholamine therapy, models that explain *why* a pressure drop is anticipated allow clinicians to intervene proactively rather than reactively (Koebe et al., 2025). Similarly, in mechanical ventilation, interpretable models for predicting intubation needs or ventilator-associated pneumonia (VAP) mortality risk provide clinicians with justifiable grounds for escalating or de-escalating care (Saykat et al., 2025)(Wei et al., 2025).

Table 3.1 summarizes the key clinical domains identified in the review and the associated XAI utility.

| Clinical Domain | Primary XAI Application | Key Benefit | Citation |
|---|---|---|---|
| Mortality Prediction | SHAP/LSTM for risk scoring | Identifies high-risk physiology | (Long & Tong, 2025)(Adebayo, 2025) |
| Sepsis Management | RL with continuous action space | Optimizes fluid/drug dosing | (Huang et al., 2022) |
| Hemodynamics | Causal Deep Learning | Predicts circulatory failure | (Cheng et al., 2025) |
| Ventilation | ML risk assessment | Assessing VAP mortality risk | (Wei et al., 2025) |
| Workflow | Adaptive Interfaces | Reduces cognitive load | (Mohammed, 2025) |

*Table 3.1: Overview of XAI applications in critical care domains. This table highlights the shift from purely predictive tasks to actionable management support.*

The integration of these systems into real-time workflows remains a challenge. While models for fluid therapy (Dessap et al., 2025) and intubation (Saykat et al., 2025) show promise in silico, the literature lacks extensive prospective clinical trials validating that the *presence* of an explanation leads to better patient outcomes compared to “black box” predictions alone.

## 3.3 Limitations of the Review

While this scoping review followed rigorous methodology, several limitations must be acknowledged. First, the heterogeneity of the included studies–ranging from theoretical framework proposals to retrospective validation on MIMIC-III datasets–precludes a quantitative meta-analysis of XAI effectiveness. Second, there is a notable scarcity of prospective, randomized controlled trials (RCTs) specifically evaluating the *interface* of XAI in the ICU. Most "validation" currently refers to algorithmic performance (AUC-ROC), not clinical utility or user comprehension.

Furthermore, the review identified a potential bias in the literature toward positive results. Few studies reported on instances where XAI explanations were misleading or caused over-reliance, despite the known risks of automation bias. Finally, the rapid evolution of Large Language Models (LLMs) and their potential role in generating natural language explanations is a nascent field that appeared only in the most recent literature, suggesting this review captures a environment that is currently in flux.

## 3.4 Challenges and Barriers to Implementation

The transition of XAI from research code to bedside tool is hindered by several structural barriers.

### 3.4.1 Safety and Accountability

In safety-critical systems, the "correctness" of an explanation is difficult to verify. Unlike a prediction which can be validated against ground truth (e.g., did the patient survive?), an explanation (e.g., "the patient is deteriorating *because* of elevated lactate") is harder to validate objectively. This raises significant safety assurance concerns, as incorrect explanations could lead clinicians to accept erroneous model outputs (Jia et al., 2021). Issues of accountability also arise; if a model provides a plausible but wrong explanation that leads to patient harm, determining liability remains an unresolved legal question (Mirchandani, 2025).

### 3.4.2 Data Complexity and Multimodality

ICU patients generate terabytes of multimodal data, including waveforms, imaging, and unstructured notes. While some progress has been made in interpreting time-series data (Reddy Desani, 2023)(Escudero-Arnanz et al., 2025), integrating explanations across modalities (e.g., combining attention maps from vital signs with relevance propagation from MRI images (Naik et al., 2025)) remains technically demanding. The complexity of these multimodal interactions often defies simple visualization, limiting the effectiveness of standard dashboards (Singh, 2025).

## 3.5 Recommendations for Future Research

Based on the identified gaps, this review proposes a strategic research agenda for the next phase of XAI in critical care.

### 3.5.1 Prioritizing Causal and Counterfactual XAI

Future research should prioritize causal inference models over purely correlational ones. Clinicians need to ask "what if?" questions (e.g., "What if I increase the vasopressor dose?"). Causal XAI frameworks (Zhang, 2023) and causally-informed deep learning (Cheng et al., 2025) offer the theoretical basis for these simulations and should be the focus of future algorithmic development.

### 3.5.2 Clinical Validation and User-Centric Design

There is an urgent need for prospective studies that evaluate XAI tools in real-time clinical settings. These studies should measure not just model accuracy, but "human-in-the-loop" performance metrics, such as time-to-decision, diagnostic accuracy with vs. Without explanations, and clinician trust levels over time (Isparan Shanthi et al., 2024). Methodologies for clinical validation seen in other domains, such as multi-level cross-institutional evaluations of fetal scans (Bashir et al., 2025), should be adapted for the ICU context to ensure robustness across different hospital systems.

### 3.5.3 Adaptive and Personalized Interfaces

One size does not fit all in XAI. Systems should be designed with adaptive capabilities that tailor the complexity and format of explanations to the user's role (nurse vs. Intensivist) and expertise level (Mohammed, 2025). This requires interdisciplinary collaboration between ML engineers, cognitive scientists, and intensivists to design interfaces that respect the cognitive constraints of the ICU environment.

Table 3.2 outlines specific recommendations for key stakeholders in the development of ICU XAI systems.

| Stakeholder | Recommendation | Rationale | Citation |
|---|---|---|---|
| Researchers | Focus on Causal XAI | Distinguish causation from correlation | (Cheng et al., 2025)(Zhang, 2023) |
| Developers | Implement Adaptive UI | Tailor to user expertise/role | (Mohammed, 2025) |
| Clinicians | Demand Clinical Validation | Ensure safety beyond AUC metrics | (Jia et al., 2021)(Bashir et al., 2025) |
| Regulators | Standardize Audits | Verify explanation fidelity/safety | (Jia et al., 2021)(Vidal et al., 2024) |

| Stakeholder | Recommendation | Rationale | Citation |
|---|---|---|---|
| Hospitals | Integrate Multimodal Data | Comprehensive patient view | (Singh, 2025)(Escudero-Arnanz et al., 2025) |

*Table 3.2: Strategic recommendations for advancing XAI in intensive care settings.*

## 3.6 Concluding Remarks

This scoping review confirms that while Explainable AI has immense potential to unlock the value of machine learning in intensive care, it is not yet a mature clinical technology. The field has successfully moved beyond the initial "black box" phase, developing strong techniques like SHAP and attention mechanisms to visualize model focus. However, the translation of these technical explanations into clinically meaningful, actionable, and safe decision support requires a fundamental shift in research focus.

The future of ICU XAI lies not in generating more complex heatmaps, but in developing causal models that align with physiological reasoning and adaptive interfaces that support the cognitive workflow of the intensivist. By addressing the challenges of safety assurance, causal validity, and human-computer interaction, XAI can evolve from a retrospective analysis tool into a proactive partner in saving lives. The path forward requires rigorous, prospective validation to ensure that these powerful algorithms serve their primary purpose: enhancing the human capacity to care for the critically ill.

# 4. Appendices

## Appendix A: Taxonomy of Explainable AI Methods in Critical Care

This appendix presents a structured taxonomy of Explainable Artificial Intelligence (XAI) methods identified through the scoping review of machine learning applications in the Intensive Care Unit (ICU). The classification framework distinguishes between intrinsic interpretability and post-hoc explainability, further categorizing methods based on their scope (local vs. Global) and model specificity. This taxonomy provides the conceptual scaffolding used to analyze the technical literature reviewed in the main body of the thesis.

### A.1 Classification Framework

The application of XAI in critical care is broadly divided into two primary paradigms: ante-hoc (intrinsic) models and post-hoc (extrinsic) explanation techniques. Intrinsic models are designed to be transparent by nature, such as decision trees or linear regression, where the relationship between input variables and outcomes is directly observable. However, the complexity of physiological data in the ICU often necessitates the use of complex "black box" models like Deep Neural Networks (DNN) and Long Short-Term Memory (LSTM) networks to achieve high predictive accuracy (Adebayo, 2025)(Bon & Cardot, 2011). Consequently, the majority of recent literature focuses on post-hoc methods applied to these complex architectures.

| Category | Definition | Common Examples in ICU | Key Strengths |
|---|---|---|---|
| **Intrinsic** | Models inherently transparent | Decision Trees, Regression | Easy to audit |
| **Post-hoc** | Explanations applied after training | SHAP, LIME | High model accuracy |
| **Model-Agnostic** | Applied to any algorithm | Permutation Importance | Versatile comparison |
| **Model-Specific** | Tied to specific architecture | Attention Mechanisms | Architecture insight |

*Table A1: High-level classification of XAI approaches found in critical care literature (Somani et al., 2023)(Jia et al., 2021).*

Intrinsic interpretability remains valuable for specific clinical tasks where linear relationships dominate or where regulatory requirements demand absolute transparency. However, as demonstrated by recent studies in mortality prediction and sepsis detection, the non-linear dynamics of patient deterioration often favor deep learning approaches, thereby necessitating post-hoc solutions to bridge the gap between accuracy and interpretability (Long & Tong, 2025)(Liu et al., 2024).

### A.2 Feature Attribution Methods

Feature attribution methods represent the most prevalent class of XAI deployed in ICU settings. These techniques assign a relevance score to each input feature (e.g., heart rate, lactate, age) to quantify its contribution to a specific prediction. The scoping review identified Shapley Additive Explanations (SHAP) as the dominant method in this category, largely due to its solid theoretical foundation in game theory and its ability to provide both local (patient-level) and global (population-level) explanations (Long & Tong, 2025).

SHAP values allow clinicians to visualize how specific physiological deviations push a patient's risk score higher or lower from the baseline. For instance, in models predicting 28-day mortality, SHAP analyses have successfully highlighted the temporal importance of features like Glasgow Coma Scale (GCS) and blood urea nitrogen, offering granular insights that aggregate metrics cannot provide (Long & Tong, 2025). Similarly, Local Interpretable Model-agnostic Explanations (LIME) creates sparse linear approximations around a single prediction to explain individual decisions, though issues with stability have been noted in high-dimensional ICU data.

### A.3 Attention Mechanisms and Temporal Interpretability

In the context of time-series data–which is ubiquitous in the ICU via continuous vital sign monitoring–attention mechanisms have emerged as a critical "model-specific" XAI method. Unlike static feature attribution, attention mechanisms allow Recurrent Neural Networks (RNNs) and Transformers to dynamically weigh input data across time steps. This capability is particularly relevant for identifying the specific time windows in a patient's history that triggered an alarm.

| Mechanism | Application Context | Interpretability Output | Reference |
|---|---|---|---|
| **Temporal Attention** | Sepsis Prediction | Highlights critical time steps | (Liu et al., 2024) |
| **Triple Attention** | Contextual Coherence | Long-term dependency tracking | (Ghaith, 2024) |
| **Layer-Wise Prop** | Image/Signal Classification | Heatmaps of signal relevance | (Naik et al., 2025) |

*Table A2: Attention-based and neural-network-specific interpretability mechanisms.*

Recent advancements have introduced "Triple Attention Transformers" to enhance contextual coherence, which is vital when integrating multimodal data sources such as clinical notes and waveform data (Ghaith, 2024). Furthermore, model-agnostic attention maps have been developed to forecast vital signs, allowing clinicians to see which historical trends (e.g., a drop in blood pressure two hours prior) heavily influenced a sepsis prediction (Liu et al., 2024). This temporal dimension of explainability is essential for actionable clinical decision support, as it aligns the AI's "focus" with the clinician's temporal reasoning.

### A.4 Counterfactual and Causal Explanations

A growing subset of the literature moves beyond correlation-based feature importance toward causal explainability. Causal XAI attempts to answer "what-if" questions: *What would have happened to the mortality risk if the patient's lactate had been lower?* This approach is important for treatment recommendation systems, such as those suggesting fluid resuscitation or vasopressor initiation (Cheng et al., 2025)(Zhang, 2023).

Causally-informed deep learning models aim to separate genuine causal drivers of patient outcomes from spurious correlations found in training data. For example, a model might learn that lower aggressive treatment correlates with better outcomes (because healthier patients receive less treatment), leading to dangerous recommendations. Causal XAI methods correct for these confounders, ensuring that explanations reflect valid physiological mechanisms rather than statistical artifacts (Cheng et al., 2025). This represents the frontier of ICU XAI, moving from "why did the model predict this?" to "what should we do about it?"

## Appendix B: Supplementary Data Tables

This appendix provides detailed data extraction tables summarizing the key studies included in the scoping review. The studies are grouped by clinical application domain: Mortality Prediction, Sepsis & Deterioration, and Therapeutic Intervention. These tables supplement the synthesis provided in the main text by offering granular details on the specific machine learning architectures and XAI techniques employed in each study.

### B.1 Mortality and Risk Prediction Studies

Mortality prediction remains one of the most common applications for ML in the ICU. The following table summarizes studies that utilized XAI to explain risk scores for general ICU populations or specific conditions like Ventilator-Associated Pneumonia (VAP).

| Study | Target Outcome | ML Model | XAI Method | Key Findings |
|---|---|---|---|---|
| **Long & Tong (2025)** (Long & Tong, 2025) | 28-Day Mortality | XGBoost, RF, LR | SHAP (Global/Local) | GCS, Age, BUN identified as top predictors. |
| **Adebayo (2025)** (Adebayo, 2025) | ICU Mortality | Hybrid LSTM | SHAP | Hybrid models balance accuracy & interpretability. |
| **Wei et al. (2025)** (Wei et al., 2025) | VAP Mortality | LightGBM | SHAP | Platelet count & creatinine key for VAP risk. |
| **Cheng et al. (2025)** (Cheng et al., 2025) | General Outcomes | Causal DL | Causal Graphs | Causality improves generalization across sites. |

*Table B1: Summary of XAI applications in mortality and general outcome prediction.*

The data indicates a strong convergence toward tree-based ensemble methods (XGBoost, LightGBM) paired with SHAP for static risk prediction tasks. In these studies, SHAP was consistently used to validate the model against clinical knowledge–for example, confirming that lower Glasgow Coma Scale scores correlate with higher mortality risk (Long & Tong, 2025). The emergence of hybrid LSTM models suggests a trend toward capturing temporal dynamics while retaining the interpretability features typically associated with simpler models (Adebayo, 2025).

### B.2 Sepsis and Acute Clinical Deterioration

Sepsis prediction requires analyzing high-frequency temporal data. The studies below demonstrate how XAI is used to explain dynamic alerts in real-time monitoring systems.

| Study | Clinical Task | Data Type | XAI Approach | Validation Metric |
|---|---|---|---|---|
| **Liu et al. (2024)** (Liu et al., 2024) | Sepsis Prediction | Vital Signs (Time Series) | Attention Maps | MSE, MAE |
| **Escudero-Arnanz (2025)** (Escudero-Arnanz et al., 2025) | Multidrug Resistance | Multivariate Time Series | Feature Importance | AUROC |
| **Koebe et al. (2025)** (Koebe et al., 2025) | Hypotension | Hemodynamic Vitals | Threshold Analysis | AUROC, AUPRC |

*Table B2: XAI methods applied to dynamic deterioration and sepsis prediction.*

In sepsis and hypotension prediction, the focus shifts from global feature importance to temporal localization. Attention maps allow clinicians to visualize the specific trajectory of vital signs that triggered an alert (Liu et al., 2024). This is particularly relevant for "early warning" systems, where the goal is to intervene before irreversible deterioration occurs. The use of multivariate time series analysis facilitates the detection of complex patterns, such as the interaction between heart rate variability and blood pressure drops, which are often subtle in early sepsis stages (Escudero-Arnanz et al., 2025).

### B.3 Therapeutic Interventions and Resource Management

Beyond prediction, ML models are increasingly used to recommend interventions. Explainability in this domain is critical for safety, as incorrect treatment recommendations (e.g., drug dosing, intubation) carry immediate risks.

| Study | Intervention | Model Type | XAI/Validation | Reference |
|---|---|---|---|---|
| **Saykat et al. (2025)** | Intubation Need | ML Classifiers | Feature Ranking | (Saykat et al., 2025) |

| Study | Intervention | Model Type | XAI/Validation | Reference |
|---|---|---|---|---|
| **Huang et al. (2022)** | Sepsis Treatment | Reinforcement Learning | Continuous Action Space | (Huang et al., 2022) |
| **Koebe et al. (2025)** | Catecholamine Therapy | Predictive Modeling | Clinical Thresholds | (Koebe et al., 2025) |
| **Park et al. (2024)** | Dosing Optimization | ML Optimization | SHAP-based Opt. | (Park et al., 2024) |

*Table B3: XAI in therapeutic recommendation and resource management systems.*

Studies involving Reinforcement Learning (RL) for sepsis treatment represent the most complex frontier for XAI. Here, agents learn continuous action spaces (e.g., dosage of vasopressors and fluids). Explaining *policies*–sequences of decisions–is significantly harder than explaining single predictions. Approaches include visualizing the value function or mapping the agent's decision boundaries against established clinical guidelines (Huang et al., 2022). Additionally, prediction models for intubation needs uses feature ranking to justify airway management decisions to respiratory therapists and intensivists (Saykat et al., 2025).

## Appendix C: Glossary of Terms

This glossary defines key technical and clinical terms used throughout the scoping review, synthesizing definitions from the cited literature to ensure consistency.

**Ante-hoc Interpretability:** Also known as intrinsic interpretability. Refers to machine learning models that are transparent by design, where the internal structure (e.g., the nodes of a decision tree or coefficients of a regression) allows users to understand how inputs map to outputs without secondary explanation methods (Somani et al., 2023).

**Attention Mechanism:** A component of neural network architectures, particularly Transformers and RNNs, that allows the model to dynamically focus on different parts of the input sequence (e.g., specific time points in a vital sign stream) when generating a prediction. In XAI, attention weights are often visualized to show what data the model prioritized (Liu et al., 2024)(Ghaith, 2024).

**Black Box Model:** A complex machine learning model, such as a Deep Neural Network (DNN) or Gradient Boosting Machine (GBM), whose internal decision-making process is too complex for humans to comprehend directly. These models typically require post-hoc XAI methods to generate explanations (Somani et al., 2023)(Jia et al., 2021).

**Causal Explainability:** An approach to XAI that incorporates causal inference to distinguish between correlation and causation. This is critical in medical decision-making to prevent models from learning spurious correlations (e.g., treatment artifacts) and to support "what-if" reasoning regarding interventions (Cheng et al., 2025)(Zhang, 2023).

**Feature Attribution:** A class of XAI methods that assigns a score to each input feature indicating its contribution to the model's output. Positive scores indicate the feature pushes the prediction higher (e.g., toward mortality), while negative scores indicate a protective effect (Long & Tong, 2025).

**Layer-Wise Relevance Propagation (LRP):** A technique for deep neural networks that propagates the prediction backward through the network layers to identify which input pixels (in imaging) or features were most relevant. It is particularly useful for visualizing contributions in complex non-linear topologies (Naik et al., 2025).

**Post-hoc Explainability:** Techniques applied after a model has been trained to explain its predictions. These methods do not alter the internal structure of the model but approximate its behavior to provide insights. Examples include SHAP and LIME (Somani et al., 2023)(Long & Tong, 2025).

**Reinforcement Learning (RL):** A type of machine learning where an agent learns to make sequences of decisions (e.g., adjusting drug dosages over time) by maximizing a reward signal. Explaining RL policies involves understanding the long-term strategy rather than just immediate predictions (Saulières et al., 2023)(Huang et al., 2022).

**SHAP (Shapley Additive Explanations):** A game-theoretic approach to feature attribution that calculates the average marginal contribution of a feature value across all possible coalitions of features. It provides a unified measure of feature importance that is consistent and locally accurate (Long & Tong, 2025)(Adebayo, 2025).

## Appendix D: Implementation and Evaluation Framework

The successful deployment of XAI in the ICU requires more than algorithmic correctness; it demands a rigorous evaluation of safety, trust, and usability. This appendix outlines a framework for evaluating XAI tools based on the principles of safety assurance and human-computer interaction identified in the review.

### D.1 The Safety Assurance Case for XAI

In safety-critical domains like critical care, XAI serves as a mechanism for safety assurance. As argued by Jia et al. (Jia et al., 2021), the opacity of ML models makes it difficult to verify that the system will behave correctly in all scenarios.

XAI helps mitigate this by allowing clinicians to audit the model's reasoning for "validity"–ensuring the model is using medically relevant features rather than artifacts (e.g., detecting a scanner tag rather than pathology).

| Assurance Goal | XAI Role | Evaluation Question |
|---|---|---|
| **Trustworthiness** | Reveal logic | Does the logic align with physiology? |
| **Fairness** | Detect bias | Is the model relying on protected attributes? |
| **Robustness** | Failure analysis | How does the model behave with noisy data? |
| **Compliance** | Audit trail | Can we explain the error post-incident? |

*Table D1: Safety assurance goals facilitated by explainable AI (Jia et al., 2021)(Vidal et al., 2024).*

To operationalize this, clinical XAI systems must support "Machine Unlearning" verification–ensuring that if a model is retrained to remove biased or invalid data, the XAI confirms the removal of those dependencies (Vidal et al., 2024). This is particularly relevant for maintaining compliance with data protection regulations and ensuring that models do not retain "memorized" private patient data.

### D.2 Human Factors and Cognitive Load

The effectiveness of an explanation is heavily dependent on the user's expertise and cognitive state. An explanation that is useful to a data scientist may be unintelligible or distracting to a bedside nurse during a code blue. Research indicates that "one-size-fits-all" explanations are insufficient. Instead, Adaptive XAI systems are proposed, which personalize the complexity and format of explanations based on the user's role and expertise level (Mohammed, 2025).

Furthermore, the introduction of XAI must not exacerbate the cognitive load of clinicians. High cognitive load can moderate the relationship between the perceived usefulness of a system and the user's trust in it (Isparan Shanthi et al., 2024). If an explanation is too complex or requires significant mental effort to interpret, it may decrease trust and adoption, even if the underlying model is accurate. Therefore, evaluation frameworks must include metrics for "Perceived Ease of Use" (PEOU) and cognitive workload alongside standard accuracy metrics (Isparan Shanthi et al., 2024).

### D.3 Clinical Validation Protocol Checklist

Based on the multi-level validation approaches seen in recent literature (e.g., fetal scan validation (Bashir et al., 2025) and ICU workflow integration (Singh, 2025)), the following checklist is recommended for future ICU XAI studies:

1. **Technical Validation:** Does the XAI method faithfully reflect the model's behavior? (Sanity checks, stability analysis).

2. **Clinician-in-the-Loop Evaluation:** Do clinicians interpret the explanation correctly? (User studies with varying expertise).
3. **Actionability Assessment:** Does the explanation lead to a change in clinical management? (e.g., Does a sepsis alert with a “Lactate” highlight prompt a lactate draw?) (Koebe et al., 2025).
4. **Workflow Integration:** Is the explanation presented at the right time without disrupting care? (Singh, 2025).
5. **Safety Audit:** Has the XAI been used to identify and mitigate spurious correlations or artifacts? (Jia et al., 2021).

Future research should prioritize these prospective, cross-institutional evaluations to move XAI from theoretical papers to bedside practice.

## References

Adebayo. (2025). Bridging the Gap Between Accuracy and Interpretability: A Hybrid LSTM Approach with SHAP for ICU Mortality Prediction Using EHR Data. *International Journal of Intelligent Information Systems.* Https://doi.org/10.11648/j.ijiis.20251406.11.

Bashir, Lin, Feragen, Mikolaj, Taksøe-Vester, Christensen,... & Tolsgaard. (2025). Clinical validation of explainable AI for fetal growth scans through multi-level, cross-institutional prospective end-user evaluation. *Scientific Reports.* Https://doi.org/10.1038/s41598-025-86536-4.

Bieniek-Kaczorek, Stopiński, Anders, Jusza, Wojtiuk, & Piramidowicz. (2025). Next-Generation Photonic Integrated Interrogators for Accurate Vital Signs Monitoring. Https://doi.org/10.1109/CLEO/Europe-EQEC65582.2025.11111390

Bon, & Cardot. (2011). *Advanced Methods for Time Series Prediction Using Recurrent Neural Networks.* InTech. Https://doi.org/10.5772/16015

Cheng, Song, Wang, Zhong, He, & Suo. (2025). Causally-informed Deep Learning towards Explainable and Generalizable Outcomes Prediction in Critical Care. Https://doi.org/10.48550/arXiv.2502.02109

Dessap, Alshamsi, Belletti, Backer, Delaney, Møller,... & Granholm. (2025). European Society of Intensive Care Medicine (ESICM) 2025 clinical practice guideline on fluid therapy in adult critically ill patients: part 2–the volume of resuscitation fluids. *Intensive Care Medicine.* Https://doi.org/10.1007/s00134-025-07840-1.

Escudero-Arnanz, Martínez-Agüero, Martín-Palomeque, Marques, Mora-Jiménez, Álvarez-Rodríguez, & Soguero-Ruíz. (2025). Multimodal interpretable data-driven models for early prediction of multidrug resistance using multivariate time series. *Health Information Science and Systems.* Https://doi.org/10.1007/s13755-025-00351-9.

Eva, Olivier, & Ludovic. (2022). How Does Cognitive Fatigue and Mental Workload Influence Alarm Detection in Flight Simulator? Classification of Electrophysiological Signatures with Explainable IA. SCITEPRESS - Science and Technology Publications. (pp. 47-51). Https://doi.org/10.5220/0011958100003622

Ghaith. (2024). The Triple Attention Transformer: Advancing Contextual Coherence in Transformer Models. Https://doi.org/10.21203/rs.3.rs-3916608/v1

Huang, Cao, & Rahmani. (2022). Reinforcement Learning For Sepsis Treatment: A Continuous Action Space Solution. *Machine Learning in Health Care.* Https://www.semanticscholar.org/paper/99bbc7eda75c55574d4fc5ce41eaa9c109237ec0.

Huang. (2025). *Residual Permutation Tests for Feature Importance in Machine Learning.* Center for Open Science. Https://doi.org/10.31234/osf.io/ajxrb_v1

Islam, Manik, Moniruzzaman, Saimon, Sultana, Bhuiyan,... & Ahmed. (2025). Explainable AI in Healthcare: Leveraging Machine Learning and Knowledge Representation for Personalized Treatment Recommendations. *Journal of Posthumanism.* Https://doi.org/10.63332/joph.v5i1.1996.

Isparan Shanthi, Ai-Na Seow, & Jing-Jing Chang. (2024). Understanding Human-AI Interaction in Healthcare: The Mediating Role of Trust and Moderating Influence of Cognitive Load. *Malaysia Journal of Invention and Innovation*, *4*(1), 105-112. Https://doi.org/10.64382/mjii.v4i1.89.

Jia, Mcdermid, Lawton, & Habli. (2021). The Role of Explainability in Assuring Safety of Machine Learning in Healthcare. *IEEE Transactions on Emerging Topics in Computing.* Https://doi.org/10.1109/TETC.2022.3171314.

Koebe, Saibel, Alcaraz, Schafer, & Strodthoff. (2025). Towards actionable hypotension prediction - predicting catecholamine therapy initiation in the intensive care unit. Https://doi.org/10.48550/arXiv.2510.24287

Leverett. (2000). *Stigmatization and Mental Illness: The Communication of Social Identity Prototypes through Diagnosis Labels.* Portland State University Library. Https://doi.org/10.15760/etd.6565

Liu, Dan, Bhatti, Shen, Gupta, Parmar, & Lee. (2024). Interpretable Vital Sign Forecasting with Model Agnostic Attention Maps. Https://doi.org/10.48550/arXiv.2405.01714

Long, & Tong. (2025). Integrating Machine Learning and SHAP for Interpretable Prediction of 28-D ay Mortality in ICU Patients: A Comprehensive Analysis of Initial Physiologic al Features. Https://doi.org/10.21203/rs.3.rs-8067228/v1

Mirchandani. (2025). Explainability, Interpretability, and Accountability in Explainable AI: A Qualitative Analysis of XAI's Sectoral Usability. *International Journal of Innovative Science and Research Technology.* Https://doi.org/10.38124/ijisrt/25aug952.

Mohammed. (2025). Adaptive Explainable AI: Personalizing Machine Explanations Based on User Expertise Levels. *Journal of Posthumanism.* Https://doi.org/10.63332/joph.v5i7.2793.

Naik, Bendegerimath, Kawari, Narajji, & Narayankar. (2025). Layer-Wise Relevance Propagation for Classifying Brain MRI Images. SCITEPRESS - Science and Technology Publications. (pp. 5-11). Https://doi.org/10.5220/0013607800004664

Park, Nam, Lee, & Kim. (2024). Explanation and optimization of ML for coagulant dosing in Water treatment plant using XAI(eXplainable AI). Https://doi.org/10.21203/rs.3.rs-3973418/v1

Reddy Desani. (2023). Explainable AI for Time Series Analysis in Real - Time Supply Chain Optimization. *International Journal of Science and Research (IJSR)*, *12*(1), 1320-1325. Https://doi.org/10.21275/es23110104518.

SANO. (2022). Extraction of Features Important for Facial Attractiveness Using Gradient-Weighted Class Activation Mapping and Guided Gradient-Weighted Class Activation Mapping. *International Symposium on Affective Science and Engineering*, *ISASE2022*(0), 1-3. Https://doi.org/10.5057/isase.2022-c000022.

Saulières, Cooper, & Bannay. (2023). Reinforcement Learning Explained via Reinforcement Learning: Towards Explainable Policies through Predictive Explanation. SCITEPRESS - Science and Technology Publications. (pp. 35-44). Https://doi.org/10.5220/0011619600003393

Saykat, Al Emon, Al-Imran, & Haque. (2025). Machine Learning and Explainable AI for Predicting Intubation Needs in an Intensive Care Unit. IEEE. (pp. 227-232). Https://doi.org/10.1109/ibdap65587.2025.11145861

Singh. (2025). Artificial Intelligence-Enabled Intensive Care Unit (AI-ICU): Integrated Framework for Real-Time Decision Support, Automated Clinical Hand-offs, and Intelligent Wound Monitoring. *International Journal For Multidisciplinary Research.* Https://doi.org/10.36948/ijfmr.2025.v07i06.63700.

Somani, Horsch, & Prasad. (2023). *Introduction to Interpretability.* Springer International Publishing. Https://doi.org/10.1007/978-3-031-20639-9_1

Terlapu, Raju, Kumar, Rao, Kavitha, Samreen, & Messina. (2024). Improved Software Effort Estimation Through Machine Learning: Challenges, Applications, and Feature Importance Analysis. *IEEE Access.* Https://doi.org/10.1109/ACCESS.2024.3457771.

Vidal, Johansen, Jahromi, Escalera, Nasrollahi, & Moeslund. (2024). Verifying Machine Unlearning with Explainable AI. Https://doi.org/10.48550/arXiv.2411.13332

Wei, Cao, Peng, Zhang, Li, Ma, & Li. (2025). An interpretable machine learning model for predicting in-hospital mortality in ICU patients with ventilator-associated pneumonia. *PLoS ONE.* Https://doi.org/10.1371/journal.pone.0316526.

Zhang. (2023). *Causal Explainable AI.* Springer International Publishing. Https://doi.org/10.1007/978-3-031-35051-1_7