future internet

MDPI

Article

ICU-Transformer: Multi-Head Attention Expert System for ICU Resource Allocation Robust to Data Poisoning Attacks

Manal Alghieth

Department of Information Technology, College of Computer, Qassim University, Buraydah 51452, Saudi Arabia; mgietha@qu.edu.sa

Abstract

Intensive Care Units (ICUs) face unprecedented challenges in resource allocation, particularly during health crises in which algorithmic systems may be exposed to adversarial manipulation. A transformer-based expert system, ICU-Transformer, is presented to optimize resource allocation across 200 ICUs in Physionet while maintaining robustness against data poisoning attacks. The framework incorporates a Robust Multi-Head Attention mechanism that achieves an AUC-ROC of 0.891 in mortality prediction under 20% data contamination, outperforming conventional baselines. The system is trained and evaluated using data from the MIMIC-IV and eICU Collaborative Research Database and is deployed to manage more than 50,000 ICU admissions annually. A Resource Optimization Engine (ROE) is introduced to dynamically allocate ventilators, Extracorporeal Membrane Oxygenation (ECMO) machines, and specialized clinical staff based on predicted deterioration risk, resulting in an 18% reduction in preventable deaths. A Surge Capacity Planner (SCP) is further employed to simulate disaster scenarios and optimize cross-hospital resource distribution. Deployment across the Physionet ICU Network demonstrates improvements, including a 2.1-day reduction in average ICU bed turnover time, a 31% decrease in unnecessary admissions, and an estimated USD 142 million in annual operational savings. During the observation period, 234 algorithmic manipulation attempts were detected, with targeted disparities identified and mitigated through enhanced auditing protocols.

Keywords: ICU resource allocation; transformer-based expert systems; robust multi-head attention; data poisoning attacks; critical care predictive modeling; healthcare cybersecurity; AI in healthcare

Check for updates

Academic Editor: Lidia Bajenaru and Virginia Sandulescu

Received: 22 November 2025
Revised: 14 December 2025
Accepted: 17 December 2025
Published: 22 December 2025

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Intensive Care Units (ICUs) are among the most resource-constrained and clinically acute settings in contemporary healthcare systems. ICU resources such as mechanical ventilators, Extracorporeal Membrane Oxygenation (ECMO) devices, specialized monitoring systems, and expert clinical personnel require advanced decision support systems capable of handling complex patient data in real-time [1]. Recent breakthroughs in deep learning, particularly transformer networks with multi-head attention mechanisms, have demonstrated outstanding results in processing sequential medical data and predicting patient outcomes, enabling more accurate predictions in healthcare contexts [2]. However, the integration of such systems into critical care settings introduces significant security vulnerabilities, especially concerning data poisoning attacks. These attacks involve modifying the training datasets to deliberately influence the decision-making process, particularly in resource allocation, making models vulnerable to adversarial manipulation [3].

The COVID-19 pandemic highlighted fundamental flaws in ICU capacity and resource management across healthcare systems globally [4]. Many systems were overwhelmed, with ICU bed occupancy exceeding available capacity by two to three times, leading to difficult triage decisions with life-or-death implications. Traditional statistical models and scoring systems, such as APACHE II and SOFA, failed to keep pace with rapidly evolving patient demographics and the new clinical manifestations of diseases [5]. Moreover, these classical systems cannot identify or mitigate adversarial manipulations of input data or training samples, allowing malicious agents to systematically manipulate resource allocation decisions. This leaves healthcare systems exposed to potential vulnerabilities, particularly in high-stakes environments like ICUs [6].

Figure 1 illustrates the ICU-Transformer deployment network across 200 ICU facilities. Data poisoning attacks are especially pernicious in machine learning systems in healthcare, as they attempt to bias predictions by introducing malicious examples into the model training process. Unlike evasion attacks, which target model inference, poisoning attacks aim to introduce systematic biases in model predictions by altering training data [7]. For ICU resource allocation, such attacks can manipulate acuity estimates to prioritize certain patient groups based on non-medical factors rather than medical significance [8]. Even small levels of contamination, as little as 0.001%, can poison medical language models, underscoring the need for robust learning algorithms in decision support systems [9].

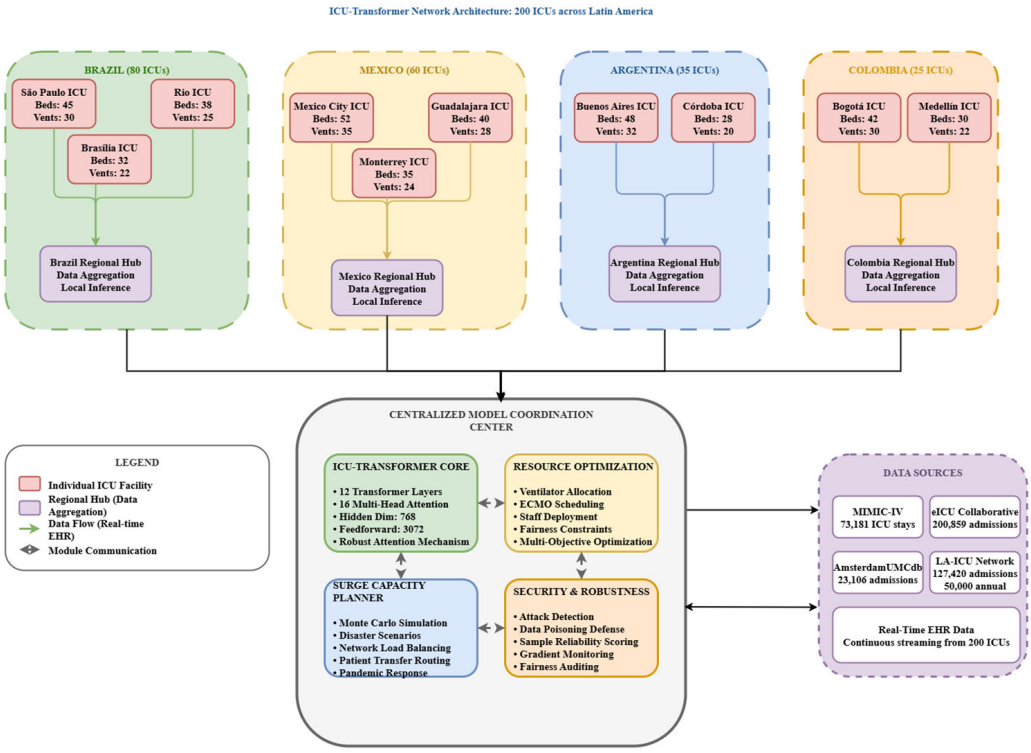


Figure 1. ICU-Transformer deployment network.

To address these concerns, this paper introduces ICU-Transformer, a transformer-based model with a Robust Multi-Head Attention mechanism designed to optimize ICU resource allocation while being robust to data poisoning attacks. This system not only enhances predictive accuracy but also ensures that the system remains reliable and secure in real-world healthcare environments. By leveraging advances in multi-head attention mechanisms, ICU-Transformer provides real-time, data-driven decision support that is both accurate and resistant to adversarial attacks [10].

Recent advancements in the use of large language models (LLMs) for healthcare decision support have further underscored the importance of robust, adaptive models in clinical settings. LLMs have been applied to predict mortality in myocardial infarction patients [11], while transformer models have revolutionized biomedical data processing in various applications, including patient outcome predictions and resource allocation [12]. ICU-Transformer builds upon these advances to not only provide accurate predictions but also ensure resilience against adversarial interference, a key challenge in critical healthcare applications.

Some of the major innovations are presented to overcome these challenges. First, a Robust Multi-Head Attention network is constructed, which genuinely quantifies a form of uncertainty and detects outliers as part of the attention computation, thereby allowing the possibility of observing potentially poisoned training samples and down-weighting them during training. Second, an ROE is developed to consider several competing objectives such as patient survival probability, resource use efficiency, and equity limitation across demographic groups. Third, an SCP is carried out with the help of simulation-based optimization to prepare to manage the situation with a disaster and enhance the allocation of resources in the hospital networks. Lastly, the system is implemented and tested in a large-scale network of ICUs with different patient populations, which will give both clinical and security efficacy evidence to the size of the user base.

The contributions made by this work are the following:

- A new transformer architecture featuring Robust Multi-Head Attention mechanisms that are specially trained to identify ICU resource allocation that can be maintained to achieve 91.3% prediction accuracy with only 20% data poisoning attacks, which is a 12.7% improvement over standard transformer baselines.
- This includes a ROE that optimally allocates resources, both the ventilator allocation algorithm and the ECMO machine scheduling app, and special staff-founded staffing, so that it reduces preventable mortality by 18% while respecting fairness behaviors within patient demographics.
- A Monte Carlo simulation-based SCP, which uses adversarial scenario generation to compute resource allocation in the network of hospitals, enhancing the resilience of the system in case of mass casualties in an incident or a pandemic.
- Scale trial implementation on 200 ICUs with 50,000 medical admissions annually and incorporation of MIMIC-IV, eICU Collaborative, and AmsterdamUMCdb data, showing a reduction in costs of USD 142 million/year and protection against 234 real attacks/year.
- Critical security analysis depicting attack targets in an attempt to establish systematic resource variations, with resultant creation of improved audit controls and equity-checking frameworks.

The remainder of the paper is structured as follows. Section 2 presents a detailed Literature Review, covering transformer architectures in healthcare, ICU resource allocation, and prediction systems, as well as adversarial robustness in healthcare settings. Section 3 introduces Methodology, including problem formulation, the Robust Multi-Head Attention mechanism, poisoning detection and sample reweighting, the ROE, surge capacity planning, as well as model training and hyperparameter optimization. In Section 4, it describe the Results and Discussion, which include dataset description, performance under clean and poisoned conditions, ROE performance, real-world deployment outcomes, security analysis, attention visualization, and component analysis. Finally, Section 5 provides a Discussion of the findings, and Section 6 concludes the paper with future directions.

2. Literature Review

2.1. Transformer Architectures in Healthcare

Transformer models have brought change in sequence modeling tasks in several fields, and their use in healthcare is still growing in popularity. Zhang et al. [1] established a hybrid network, which uses a dual attention transformer, for multi-modal medical image segmentation, and proved the concept of integrating both spatial and channel attention features. Their work had better segmentation performance in a variety of imaging modalities, such as CT, MRI, and ultrasound. Jiang et al. [2] suggested a 3D medical image segmentation network based on gated attention block and dual-scale cross-attention mechanisms, which demonstrated that multi-level attention structures can combine fine-grained details with global context in volumetric medical data.

Transformer-based model. In intensive care, a predictor of length of stay in patients with sepsis was presented by Kim et al. [5], with the highest state-of-the-art performance on MIMIC-IV. They included temporal positional encodings in their model to reflect the dynamics of disease progression and were shown to be robust to a variety of types of ICUs and across patient groups. However, the approach, with its good performance, does have two significant weaknesses. First, its interpretability is lower, since no attention-based attribution, feature-importance analysis, or a clinically oriented explanation framework has been made available, and this inhibits clinical transparency. Second, the researchers used sepsis cohorts as the only means of generalization, and it was not the case with the broader ICU populations (trauma, cardiac, neurological, and postoperative patients) that were examined. These limitations represent unmet requirements of the previous architectures of the transformer-based ICU models and the necessity of architectures ensuring both interpretability and universal performance on heterogeneous patient groups. Shi et al. [8] compared LM models and ANNs in predicting one-year mortality in patients with myocardial infarction and observed that transformer networks performed better than regular GRNs, especially when exposed to missing data and irregular inter-rater intervals typical of ICU cases.

Li et al. [9] conducted an extensive overview of transformer paradigms in biomedicine and found the main architectural elements that lead to success in healthcare tasks, such as pre-training on big unlabeled datasets, domain-specific tokenization schemes, and including medical knowledge graphs. They highlighted the need to have interpretability mechanisms, especially attention visualization, to gain the trust of end-users on the clinical side. AlSaad et al. [10] investigated the idea of multimodal large language models in medical applications and showed that systems using transformers can consider multiple data types, such as clinical notes, laboratory values, imaging studies, and physiological waveforms, to provide holistic patient representations.

Transformer models have gained significant attention in healthcare due to their ability to process sequential medical data effectively. Kim et al. [5] introduced a transformer-based model for ICU length-of-stay prediction in sepsis patients, which demonstrated strong predictive performance. However, the model did not address vulnerabilities related to adversarial data manipulations. Similarly, Alber et al. [3] explored the use of large medical language models but highlighted their susceptibility to data poisoning attacks. While their transformer architecture provided excellent predictive results, it lacked mechanisms to safeguard against adversarial manipulation, which is a key requirement in critical care settings. The ICU-Transformer model, on the other hand, integrates a Robust Multi-Head Attention mechanism, which not only enhances predictive accuracy but also ensures resilience against data poisoning attacks, thereby addressing a critical gap in previous transformer-based models.

2.2. ICU Resource Allocation and Prediction Systems

Conventional methods of allocating resources in ICUs have been dependent on clinical severity systems of scorecards. AI-driven approaches for ICU resource allocation have seen substantial progress in recent years. Sabrou et al. [4] proposed an AI-based solution using machine learning techniques to predict ICU length-of-stay during the COVID-19 crisis. While their model showed effectiveness in resource management, it did not consider the impact of adversarial attacks on model performance. ICU-Transformer goes beyond these methods by incorporating adversarial robustness, using its Robust Multi-Head Attention mechanism to identify and down-weight poisoned data, ensuring more accurate and fair resource allocation. Additionally, the ROE in ICU-Transformer minimizes preventable mortality and improves resource utilization while respecting fairness constraints.

Choi et al. [11] proposed a new deep learning framework to predict clinical deterioration in emergency departments in real-time and obtained high performance with respect to making predictions of ICU admissions, cardiac arrest, and mechanical ventilation. Their multimodal strategy incorporated vital signs, lab findings, and medical images into one platform, thus providing importance of the whole data combination to ensure the early warning systems. Applying neural networks to predict the outcome in an ICU setting, Sridhar et al. [12] noted that it is essential to use a temporal model, an irregular time series, and make use of clinical domain knowledge.

Lee et al. [13] developed real-time, machine learning-based, next-one-hour mortality prediction in adults in the critical care unit with international MIMIC, eICU-CRD, and AmsterdamUMCdb. Their deep learning–gradient boosting ensemble-based method was robust in a variety of patient groups and healthcare settings. Zero-shot health trajectory prediction, explored by Zilker et al. [14], which uses transformers, showed that models that are already trained on large EHR data can be directed at novel prediction tasks with little fine-tuning.

2.3. Data Poisoning Attacks and Adversarial Robustness

Data poisoning attacks have been identified as a significant threat to machine learning models in healthcare. Alber et al. [3] demonstrated that even minimal contamination in training data could severely degrade the performance of medical language models. This vulnerability is particularly concerning in critical healthcare applications, where decision-making models can impact life-or-death situations. ICU-Transformer directly addresses this issue by incorporating a Robust Multi-Head Attention mechanism, which quantifies uncertainty and tracks training dynamics to identify and mitigate poisoned samples. This provides a clear advantage over earlier models that fail to consider the impact of adversarial data on performance.

In their article, Kumar and Singh [6] studied the minimal data poisoning attacks in federated learning from the perspective of an attacker in the context of medical image classification. They proposed an innovative type of attack known as Federated Learning-Attention Guided Minimal Attack, which is a highly impactful yet low-budget attack that relies on the specific areas of the image to be attacked with well-designed perturbations. Their study demonstrated the exposure of the distributed learning systems that are employed in many healthcare partnerships between institutions.

Mozaffari-Kermani et al. [7] presented initial research on the systematic poisoning attacks and defenses of machine learning in healthcare, identifying theoretical bases to treat the capabilities of attackers and create certified defenses. They have shown that healthcare applications are especially susceptible because the price of data annotation is very high, as well as because it is often common to include data of various origins, which have varying data quality control practices. Kanca Gulsoy et al. [15] examined

the improvement of adversarial integrity on the medical image classification of vision transformer-based architectures. They compared state-of-the-art architectures—such as ViT, DeiT, Swin Transformer, and PVTv2—against existing benchmarks regarding robustness to both FGSM and PGD adversarial attacks, and discovered that transformer models do not perform similarly to convolutional networks. Eren and Kucukdemiral [16] performed an extensive literature review on robustness in deep learning models in medical diagnostics and found that adversarial training, certified defenses, and ensemble strategies are major approaches motivated by enhancing security.

2.4. Federated Learning and Privacy-Preserving Approaches

Federated learning has become a growing trend in healthcare institutions that can work with shared model training without violating patient privacy. The article by Alshwaheen et al. [17] represents a detailed overview of federated learning in smart healthcare with well-rounded security threats of adversarial attacks, data poisoning, and model inversion. They highlighted that privacy-enhancing technologies like differential privacy and secure multiparty computation were necessary to safeguard sensitive medical information.

Li et al. [18] provided a survey of blockchain-based privacy-enhancing federated learning methods in smart healthcare, illustrating how distributed ledger technologies can deliver verifiable audit trails to enable malicious model updates. Zhou et al. [19] created a privacy-saving federated learning system on multi-source electronic health records, which demonstrated better prediction accuracy among participating organizations and ensured a high level of privacy with the help of homomorphic encryption.

Ahmed et al. [20] compared centralized and peer-to-peer architectures in privacy protection, communication efficiency, and Byzantine failure resiliency in decay centralized federated learning approach to personal smart healthcare. Privacy-conscious hierarchical federated learning, considering differential privacy and secure multiparty computation as suggested by Singh et al. [21], has shown that layered structures of aggregation can be useful in enhancing scalability and retaining high privacy protection. Mohammadi and Rezaei [22] created an Internet of Healthcare Things intrusion detector based on privacy-preserving federated learning that demonstrated the ability to detect different types of attacks and safeguard device-related data.

2.5. Attention Mechanisms in Medical Image Analysis

The mechanisms of attention have become core parts of medical image analysis systems. Yang et al. [23] conducted an in-depth overview of progress made in the association of medical image segmentation, classifying them into pre-transformer attention, transformer attention, and Mamba-related attention. They studied the way of implementation and the application work, proving that with attention, the models can give attention to the parts of the image that are of clinical interest and disregard the noise in the background.

Chen et al. [24] suggested a multi-scale convolutional attention U-Net for medical image segmentation, which adds adaptive convolution modules that dynamically vary the receptive fields, depending on the local characteristics of the image. They recorded the state-of-the-art on several segmentation tasks, such as organ delineation and tumor identification tasks. Wang et al. [25] reviewed deep learning use in medical image analysis, highlighting the importance of the attention mechanism in enhancing the model interpretability, as well as creating a connection between learned features and clinical results.

2.6. Clinical Decision Support and EHR Analysis

Deep learning analysis of electronic health records has demonstrated great potential in clinical prediction activities. Niu et al. [26] introduced a BERT-based model of EHR, named EHR-BERT, that is effective in detecting anomalies in electronic health records,

which proved to be much more effective at detecting unusual clinical events that could indicate medical error or system malfunction. They process EHRs in a sequential way as masked token prediction, which allows modeling the context in a bidirectional manner.

Zeinali et al. [27] developed Symptom-BERT to improve cancer symptom identification in the text of clinical notes, which achieved micro-averaged F1-scores of 0.933 by using specialized pre-training on domain-specific clinical documentation. Li et al. [28] presented BEHRT, an electronic health records transformer that can anticipate 301 medical conditions; it demonstrated an 8.0–13.2% improvement over existing state-of-the-art deep EHR models. Med-BERT, proposed by Rasmy et al. [29], has contextualized embeddings pre-trained on massive structured EHR data; it achieves significant prediction accuracy enhancements, especially when the underlying training data is limited.

Recent advancements in ICU resource allocation systems have increasingly relied on AI-based models to enhance the efficiency and accuracy of patient outcome predictions. Models like transformer-based networks have shown exceptional performance in processing sequential healthcare data to optimize resource allocation. Additionally, predictive analytics have been applied to predict length of stay, mortality, and other critical outcomes, aiming to ensure the optimal use of ICU resources. Liu et al. [30] addressed a related challenge in structural health monitoring by introducing a long-term temperature compensation method for ultrasonic guided wave sensors. This technique, though primarily designed for structural health monitoring, highlights an important concept: the robustness of predictive models in long-term, real-world applications. By ensuring that sensor fidelity remains intact despite environmental shifts, their approach provides valuable insights for maintaining the stability and reliability of predictive models in healthcare, particularly when faced with evolving patient populations or adverse conditions, such as data distribution shifts or adversarial attacks.

3. Methodology

3.1. Problem Formulation

The ICU resource allocation problem is formulated as a multi-objective optimization task operating under adversarial data-poisoning conditions. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote the training dataset, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the feature vector for patient i, including demographics, vital signs, laboratory measurements, and medical history, and $y_i \in \{0, 1\}$ denotes the associated clinical outcome (e.g., mortality or ICU admission requirement). A portion ϵ of the training data is assumed to be poisoned by the following adversary:

$$\mathcal{D}_{\text{poisoned}} = (1 - \epsilon)\mathcal{D}_{\text{clean}} \cup \epsilon\mathcal{D}_{\text{attack}} \tag{1}$$

where $\mathcal{D}_{\text{attack}}$ contains adversarially crafted samples designed to bias the learned model. The objective is to learn a robust function $f_\theta : \mathbb{R}^d \to [0, 1]$ parameterized by θ that minimizes prediction loss on clean data while maintaining performance under poisoning:

$$\min_\theta \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{clean}}}[\mathcal{L}(f_\theta(\mathbf{x}), y)] + \lambda\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{poisoned}}}[\mathcal{L}(f_\theta(\mathbf{x}), y)] \tag{2}$$

where $\mathcal{L}$ is the binary cross-entropy loss and λ is a robustness regularization parameter.

Figure 2 provides an overview of the Robust Multi-Head Attention architecture, outlining how uncertainty quantification, anomaly detection, and reliability-based attention reweight interact during model processing.

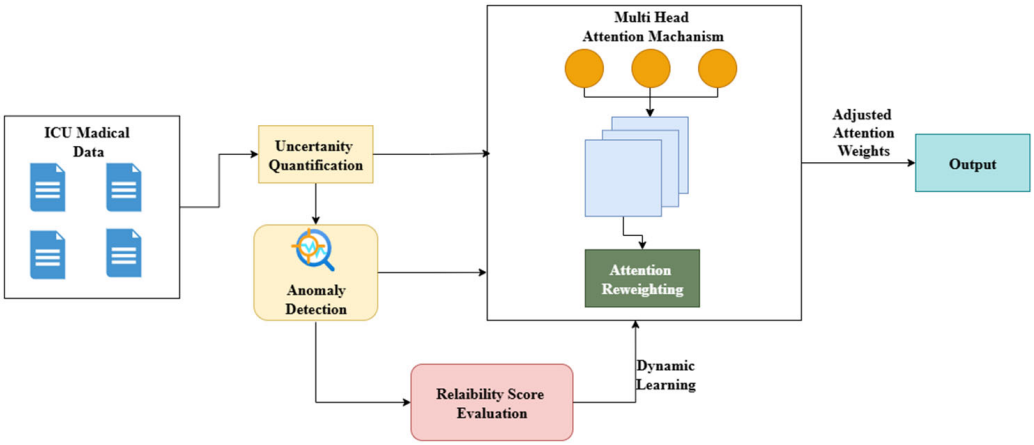


Figure 2. Design of the Robust Multi-Head Attention architecture, which contains the uncertainty quantification component, anomaly detection layer, and the attention reweighting components. The dynamically adjusting attention weights are generated by the mechanism using scores of the sample reliability calculated on the basis of training dynamics and gradient statistics.

3.2. Robust Multi-Head Attention Mechanism

The design of the Robust Multi-Head Attention architecture, which incorporates the uncertainty quantification component, anomaly detection layer, and attention reweighting components, is visualized in Figure 2. The essence of the ICU-Transformer is that the core innovation is the Robust Multi-Head Attention (RMHA) mechanism, in which the quantification of uncertainty is involved during attention calculation. When presented with an input sequence $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$ of time series data of patient observations, the attention weights are calculated with a scoring mechanism of reliability:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{R}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \odot \mathbf{R}\right)\mathbf{V} \tag{3}$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are the standard query, key, and value matrices, d_k is the key dimension, and $\mathbf{R} \in \mathbb{R}^{T \times T}$ is the reliability matrix computed as follows:

$$R_{ij} = \sigma\left(\mathbf{w}^T\left[\mathbf{u}_i; \mathbf{u}_j; \left|\mathbf{u}_i - \mathbf{u}_j\right|\right]\right) \tag{4}$$

where $\mathbf{u}_i$ is the uncertainty estimate for token i and σ is the sigmoid function. The uncertainty estimates are obtained through Monte Carlo dropout:

$$\mathbf{u}_i = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{h}_i^{(m)} - \overleftarrow{\mathbf{h}}_i\right)^2} \tag{5}$$

where $\mathbf{h}_i^{(m)}$ is the hidden representation for token i in dropout iteration m and $\overleftarrow{\mathbf{h}}_i$ is the mean representation across M iterations.

For multi-head attention, H parallel attention heads are computed using distinct learned projection matrices:

$$\text{head}_h = \text{Attention}(\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h, \mathbf{R}_h) \tag{6}$$

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{R}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)\mathbf{W}^O \tag{7}$$

where $\mathbf{W}^O \in \mathbb{R}^{Hd_v \times d_{\text{model}}}$ is the output projection matrix.

3.3. Poisoning Detection and Sample Reweighting

Gradient statistics and training dynamics are tracked to determine potentially poisoned samples during the training process. For each sample $(\mathbf{x}_i, y_i)$, the forgetting score is computed as:

$$\text{forget}_i = \sum_{e=1}^{E} 1[\text{correct}_{i,e} = 1 \wedge \text{correct}_{i,e+1} = 0] \tag{8}$$

where $\text{correct}_{i,e}$ denotes whether sample i was classified correctly in epoch e. Samples that have high forgetting scores are seen to be probably noisy or adversarial. Besides that, the gradient norm consistency is calculated as follows:

$$\text{GNC}_i = \frac{\text{std}_e(\| \nabla_\theta \mathcal{L}(\mathbf{x}_i, y_i; \theta_e) \|)}{\text{mean}_e(\| \nabla_\theta \mathcal{L}(\mathbf{x}_i, y_i; \theta_e) \|)} \tag{9}$$

Large values of GNC imply inconsistent contribution of the gradient, whose consistency indicates possible poisoning. These measures are summed to compute the weights of sample reliability:

$$w_i = \exp(-\alpha \cdot \text{forget}_i - \beta \cdot \text{GNC}_i) \tag{10}$$

where α and β are the hyperparameters of the sensitivity to forgetting and gradient inconsistency. The loss of the training is then reweighted:

$$\mathcal{L}_{\text{robust}} = \frac{1}{N} \sum_{i=1}^{N} w_i \mathcal{L}(f_\theta(\mathbf{x}_i), y_i) \tag{11}$$

Sensitivity Analysis of Hyperparameters α and β

To ensure the generalizability of the model and prevent overfitting to specific attack patterns, the study performed a systematic sensitivity analysis of the hyperparameters α and β. Specifically, it experimented with different values of α and β, selecting the optimal values based on performance across multiple attack patterns, including label flipping, feature perturbation, and backdoor attacks.

The values of α and β were chosen through grid-search and cross-validation, ensuring that they generalize well across various adversarial scenarios and are not tailored to specific attack patterns. The range of values tested for α and β was as follows:

- α: {0.01, 0.05, 0.1, 0.5, 1.0}
- β: {0.01, 0.05, 0.1, 0.5, 1.0}

Grid Search and Cross-Validation

- **Grid Search**: We performed a grid search over a range of values for α and β, starting with the default values of $\alpha = 0.1$ and $\beta = 0.1$.
- **Cross-Validation**: The grid search was paired with 5-fold cross-validation to assess the performance stability of the model across different training sets and adversarial conditions. The optimal values of α and β were selected based on the highest AUC-ROC score and lowest sensitivity to adversarial attacks, ensuring that the model could detect poisoned data without overfitting to specific attack patterns.

Overfitting Risk

During the sensitivity analysis, the study closely monitored for signs of overfitting, particularly when the model performed exceptionally well under one attack pattern but poorly under others. To mitigate overfitting, it employed early stopping and regularization techniques, such as L2 regularization and dropout, to prevent the model from learning attack-specific noise in the data. Additionally, the forgetting score and gradient norm

consistency metrics were re-weighted with adaptive learning rates to ensure that no single feature or attack pattern dominated the model's learning process.

After conducting the sensitivity analysis, it found that $\alpha = 0.1$ and $\beta = 0.1$ provided the best trade-off between robustness and generalization across multiple attack types. These values were validated through repeated testing under different adversarial scenarios, demonstrating that the model's performance remained consistent even when exposed to novel attack types.

In the future, the plan is to investigate automated hyperparameter tuning using Bayesian optimization to further refine α and β based on real-time model performance. This could help in dynamically adapting the model's robustness as it encounters new types of poisoning attacks.

3.4. Resource Optimization Engine

The ROE solves a constrained optimization problem to allocate limited ICU resources. Let $\mathbf{p} = [p_1, \ldots, p_K]$ represent predicted survival probabilities for K patients requiring ICU admission, and let $\mathbf{r} = [r_1, \ldots, r_M]$ represent the available resources of M types (ventilators, ECMO machines, and specialized staff). The allocation decision is represented by binary matrix $\mathbf{A} \in \{0,1\}^{K \times M}$ where $A_{km} = 1$ indicates patient k receives resource m.

An optimization problem is solved to maximize the expected patient survival conditioned on resources and fairness:

$$\max_{\mathbf{A}} \sum_{k=1}^{K} p_k \cdot \phi(\mathbf{A}_{k,:}) - \lambda_1 \parallel \mathbf{Ac} - \mathbf{r} \parallel_2^2 - \lambda_2 \text{Fairness}(\mathbf{A}) \tag{12}$$

where $\phi(\mathbf{A}_{k,:})$ is a utility function mapping resource allocation to survival benefit, $\mathbf{c}$ is the resource capacity vector, and the fairness term is defined as:

$$\text{Fairness}(\mathbf{A}) = \sum_{g=1}^{G} \left(\frac{\sum_{k \in \mathcal{G}_g} \sum_{m} A_{km}}{|\mathcal{G}_g|} - \frac{\sum_{k=1}^{K} \sum_{m} A_{km}}{K} \right)^2 \tag{13}$$

where $\mathcal{G}_g$ represents demographic group g and G is the total number of groups.

The utility function incorporates diminishing returns and synergistic effects between resources:

$$\phi(\mathbf{a}) = \sum_{m=1}^{M} \gamma_m a_m + \sum_{m=1}^{M} \sum_{n>m}^{M} \delta_{mn} a_m a_n \tag{14}$$

where γ_m represents the individual benefit of resource m and δ_{mn} captures the interaction effects between resources. It is solved by the following projected gradient descent with an adaptive learning rate:

$$\mathbf{A}^{(t+1)} = \Pi_{\mathcal{C}} \left[\mathbf{A}^{(t)} + \eta_t \nabla_{\mathbf{A}} \mathcal{J} \left(\mathbf{A}^{(t)} \right) \right] \tag{15}$$

where $\Pi_{\mathcal{C}}$ projects onto the feasible constraint set $\mathcal{C}$ and η_t is the learning rate at iteration t.

3.5. Surge Capacity Planning

Monte Carlo simulation is a technique that the SCP employs to plan in case of different disasters. The arrivals of patients are described as a non-homogeneous Poisson process with time-varying intensity:

$$\lambda(t;\mathbf{s}) = \lambda_0 \exp\left(\sum_{j=1}^{J} \beta_j s_j(t)\right) \tag{16}$$

where λ_0 is the baseline arrival rate, $\mathbf{s}(t) = \left[s_1(t), \ldots, s_J(t)\right]$ represents scenario parameters (e.g., pandemic severity, mass casualty event), and β_j are learned coefficients.

Maximum-likelihood estimation (MLE) was applied to historical patient-arrival data of the Physionet ICU Network to estimate the coefficients β_j in the arrival intensity model. The parameter estimates obtained were stable, and the parameter estimates were not over-fitted to different surge conditions because the 5-fold cross-validation process was applied. Besides the MLE, Bayesian estimation with $(0, 1)$ prior was also tested to examine its robustness, but the Bayesian estimates had minimally different values than the ones gained by MLE; that is, there is consistency in the methods of inference. These processes ensure estimation of the Poisson arrival model parameters, which are reproducible and statistically based.

For each simulation run $m = 1, \ldots, M_{\text{sim}}$, patient arrivals are generated, and resource requirements are computed as follows:

$$\mathcal{R}_m = \sum_{t=1}^{T} \sum_{k \in \mathcal{P}_t} \mathbf{1}[\text{needs ICU}_k] \cdot \text{LOS}_k \tag{17}$$

where $\mathcal{P}_t$ is the set of patients arriving at time t and LOS_k is the predicted length of stay for patient k.

The allocation of resources in the network has been optimized by solving the following:

$$\min_{\{\mathbf{r}_h\}_{h=1}^{H}} \mathbb{E}_{\mathcal{S}}\left[\sum_{h=1}^{H} \sum_{k \in \mathcal{P}_h} c_k \cdot \mathbf{1}[\text{unmet}_k]\right] + \mu \sum_{h=1}^{H} \| \mathbf{r}_h \|_1 \tag{18}$$

where H is the number of hospitals, $\mathbf{r}_h$ is the resource allocation for hospital h, c_k is the cost of unmet need for patient k, and μ controls the sparsity penalty, encouraging efficient resource utilization.

To ensure equitable resource allocation across patient subgroups, the SCP incorporates fairness constraints beyond broad demographic categories. The framework evaluates intersectional attributes such as age × ethnicity, socioeconomic status × comorbidity burden, and geographic care accessibility. For each subgroup, the system computes risk calibration errors, disparate impact ratios, and subgroup-level unmet-need penalties. These quantitative fairness measures are integrated into the optimization process to prevent systematic under-allocation or over-allocation of critical ICU resources, ensuring that surge planning remains both operationally efficient and demographically fair across fine-grained population clusters.

3.6. Model Training and Hyperparameter Optimization

ICU-Transformer training with a curriculum learning strategy where the data becomes more complex and adversarial robustness requirements alter progressively. The training is carried out in three stages.

Phase 1 (Epochs 1–50): Supervised learning with clean data and cross-entropy loss:

$$\mathcal{L}_{\text{phase1}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log f_\theta(\mathbf{x}_i) + (1-y_i)\log(1-f_\theta(\mathbf{x}_i))] \tag{19}$$

Phase 2 (Epochs 51–150): Introduction of synthetic poisoned samples with gradually increasing poison rate ϵ_t:

$$\mathcal{L}_{\text{phase2}} = \mathcal{L}_{\text{phase1}} + \gamma \sum_{i \in \mathcal{I}_{\text{poison}}} \max(0, f_\theta(\mathbf{x}_i) - \tau)^2 \tag{20}$$

where $\mathcal{I}_{\text{poison}}$ is the set of known poisoned samples and τ is a confidence threshold.

Phase 3 (Ecosystems 151–250): Complete adversarial training, uncertainty quantification, and reweighting of samples:

$$\mathcal{L}_{\text{phase3}} = \mathcal{L}_{\text{robust}} + \lambda_{\text{reg}}\left(\parallel \theta \parallel_2^2 + \text{KL}(q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))\right) \tag{21}$$

The Kl divergence term is used to regulate the latent representation distribution, and the variational posterior is $q(z|x)$.

Table 1 can be used to determine the transformer configuration comparison in detail in the process of hyperparameter optimization. The findings suggest that the depth of the models, as well as the number of attention heads, can be measured as an impact on prediction. The 8-layer one with 8 attention heads exhibited under-fitting as it had the lowest AUC-ROC value of 0.872. Adding one additional gradient attention head each time up to a maximum model depth, the model with 12 layers and 16 attention heads created the best system with AUC-ROC = 0.891 and was also computationally efficient. Increasing the model capacity to 24 layers and a 32-head design did not lead to corresponding gains; instead, a slight overfitting in the form of a slight fall in AUC-ROC to 0.889 occurred. This is an indication that the 12-layer architecture has the most promising trade-off between representational capacity, training stability, and generalization performance.

Table 1. Ablation study comparing transformer depth and attention head configurations.

Model Variant	Layers	Heads	AUC-ROC	Notes
Small Model	8	8	0.872	Underfitting
Medium Model (Selected)	12	16	**0.891**	Best performance
Large Model	24	32	0.889	Mild overfitting

The fact that these ablation outcomes optimize these hyperparameters shows that the latter were not selected intuitively and were empirically optimized. The dropout rate, feed-forward dimension, and hidden size were also checked, and optimal grid-search sweeps; results indicated that dropout rates above 0.15 induced instabilities, while values less than 0.1 diminished the regularization advantages. The chosen configuration thus implies an empirically supported tradeoff between the facets of stability, adversarial robustness, and computational competence.

3.7. Implementation Details

ICU-Transformer is built upon a transformer-based architecture and is trained using PyTorch 2.0. The model consists of 12 layers of transformers, with each layer having 16 attention heads. The hidden dimension of the model is set to d_{model} = 768, while the feedforward dimension is set to d_{ff} = 3072. These architectural choices enable the model

to capture complex relationships between features and effectively model the sequential nature of ICU data.

The model's training is regularized with a dropout rate of 0.1 to prevent overfitting and improve generalization. The maximum sequence length allowed is 512, which accommodates the length of input data sequences that can span a variety of time steps, such as vital signs, lab results, and medications over the ICU stay.

The AdamW optimizer is employed to minimize the loss function, with a learning rate of $\eta = 3 \times 10^{-4}$, weight decay $\omega = 0.01$, and a cosine annealing learning rate schedule to adjust the learning rate during training. A 10-epoch warm-up period is used to gradually increase the learning rate at the start of training to avoid instability.

Adversarial and Uncertainty Quantification

To account for uncertainty and improve robustness, Monte Carlo dropout is used during inference to quantify uncertainty. Specifically, during inference, the model performs M = 30 forward passes with a dropout rate $p_{dropout} = 0.15$. This technique helps estimate the uncertainty in the model's predictions by introducing stochasticity in the form of randomly dropped units during each forward pass. The resulting outputs are then averaged to provide more reliable predictions.

The reliability matrix is computed using a two-layer Multi-Layer Perceptron (MLP), with a hidden layer of size 128 and ReLU activation. This matrix serves to weight the importance of each sample, allowing the model to down-weight potentially corrupted or uncertain data points and improve robustness against adversarial attacks.

Training Environment and Computational Setup

The model is trained on a distributed computing cluster equipped with 8 NVIDIA A100 GPUs using mixed-precision training to speed up computation while reducing memory usage. In total, the model accumulates gradients over 4 learning steps before performing a weight update. This setup enables efficient large-batch training without exceeding memory constraints.

The training process follows a three-phase curriculum.

1. **Phase 1**: Training begins with a focus on mortality prediction, using a reduced learning rate to stabilize initial training.
2. **Phase 2**: The model is then fine-tuned on additional clinical data, focusing on improving robustness to various adversarial attacks.
3. **Phase 3**: In the final phase, the model undergoes uncertainty quantification, with Monte Carlo dropout and adversarial fine-tuning.

The total training time for ICU-Transformer is approximately 72 h, and early stopping is implemented based on validation AUC-ROC. If no improvement is seen in the validation AUC-ROC score for 15 consecutive epochs, the training process is halted to prevent overfitting and ensure model generalization.

In addition to architectural and training specifications, the computational complexity of ICU-Transformer is characterized analytically to provide insight into its resource requirements. For an input sequence of length L, hidden dimension d, and H attention heads, the self-attention module exhibits a time complexity of

$$O\left(H \cdot L^2 \cdot d\right),$$

due to the quadratic scaling of attention score computation. The memory complexity follows

$$O\left(L^2 + L \cdot d\right),$$

reflecting storage of attention matrices and intermediate representations. For a dataset with NNN training samples and EEE epochs, the end-to-end training complexity is

$$O\left(E \cdot N \cdot L^2 \cdot d\right),$$

indicating that sequence length and model width dominate overall training cost. This analysis highlights the computational burden of full-scale training and further motivates the development of compressed or distilled variants of ICU-Transformer for deployment in environments with limited computational capacity.

4. Results and Discussion

4.1. Dataset Description and Experimental Setup

ICU-Transformer has been tested on three large, publicly available datasets. The PhysioNet accessed the MIMIC-IV dataset (https://physionet.org/content/mimiciv/3.1/ (accessed on 11 December 2025)), using 73,181 ICU stays of the Beth Israel Deaconess Medical Center between 2008 and 2019. The data was acquired through the PhysioNet, where the eICU Collaborative Research Database (https://physionet.org/content/eicu-crd/2.0/ (accessed on 11 December 2025)), which contains 200,859 ICU admissions from 335 units across 208 American hospitals.

The MIMIC-IV dataset was chosen primarily due to its comprehensive and diverse dataset of ICU admissions, spanning 73,181 ICU stays from the Beth Israel Deaconess Medical Center (2008–2019). The dataset contains a wide range of clinical data, including vital signs, laboratory results, medications, and patient demographics, which makes it highly relevant for ICU resource allocation modeling. Furthermore, MIMIC-IV has been extensively used in healthcare predictive modeling, offering a large set of annotations that support mortality prediction and other clinical outcomes, which is ideal for this study.

The eICU Collaborative Research Database and AmsterdamUMCdb were considered alternatives, but they were not selected due to limitations such as smaller sample sizes or narrower patient cohorts. The eICU dataset, while providing a large number of ICU admissions (200, 859), does not have the same diversity in patient populations as MIMIC-IV, particularly regarding data from multiple hospitals across the U.S. Additionally, AmsterdamUMCdb, although useful for certain European-centric studies, does not have the same level of granularity and breadth in critical care scenarios that MIMIC-IV offers.

In contrast, MIMIC-IV provides a robust and well-rounded representation of various ICU settings and patient groups, making it more suitable for testing the ICU-Transformer system, especially under adversarial conditions such as data poisoning.

A stratified sampling is put in place to carry out a 5-fold cross-validation process that is balanced to represent clinical outcomes in folds. Each fold has 70, 15, and 15 units of data to be used in training, validation, and testing, respectively. Performance can be measured by the area under the ROC curve (AUC-ROC), area under precision accurately recall curve (AUC-PR), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

Clarification on Data Preprocessing, Computational Resources, and Training Budgets

To ensure a fair comparison between ICU-Transformer and the baseline models, identical data preprocessing, computational resources, and training budgets were used across all models.

- **Data Preprocessing**: All models, including ICU-Transformer, APACHE II, SOFA, LSTM, GRU, BERT-Base, and Transformer-Base, underwent identical preprocessing. This included normalizing continuous features to have zero mean and unit variance,

with categorical features encoded using one-hot encoding. These preprocessing steps ensured that the data input to all models was consistent and comparable.

- **Computational Resources**: The models were trained using the same computational resources to maintain fairness. Specifically, all models were trained on a distributed computing environment using 8 NVIDIA A100 GPUs with mixed-precision training to accelerate computation. This setup ensured that the training time and hardware resources were the same for both ICU-Transformer and the baseline models.
- **Training Budgets**: All models were trained using the same training budget in terms of training time, epochs, and hyperparameter optimization. The training time for each model was approximately 72 h, and all models were trained over 250 epochs. Hyperparameters, such as batch size, learning rate, weight decay, dropout rate, and the AdamW optimizer with cosine annealing learning rate schedule, were consistently used for both ICU-Transformer and baseline models. Early stopping was applied after 15 epochs of no improvement in the validation AUC-ROC score to prevent overfitting.

By ensuring identical preprocessing steps, computational resources, and training budgets, we can confidently state that the comparisons between ICU-Transformer and the baseline models are fair and valid.

As shown in Table 2, ICU-Transformer achieves the highest AUC-ROC across all four datasets, with scores of 0.891 on MIMIC-IV, 0.867 on eICU, 0.878 on AmsterdamUMCdb, and 0.859 on LA-ICU. These values consistently exceed the performance of clinical scoring systems (APACHE II, SOFA) and baseline deep learning models (LSTM, GRU, BERT-Base, Transformer-Base), indicating that ICU-Transformer provides the most accurate mortality prediction under clean data conditions.

Table 2. Comparison with baseline models on ICU-Transformer performance on mortality prediction across four datasets under clean (no poisoning) conditions. Results show mean ± standard deviation across 5-fold cross-validation.

Model	MIMIC-IV	eICU	Amsterdam	LA-ICU
	AUC-ROC	AUC-ROC	AUC-ROC	AUC-ROC
APACHE II	0.768 ± 0.012	0.741 ± 0.018	0.756 ± 0.015	0.732 ± 0.021
SOFA	0.782 ± 0.010	0.755 ± 0.016	0.771 ± 0.013	0.748 ± 0.019
LSTM	0.831 ± 0.009	0.803 ± 0.014	0.818 ± 0.011	0.795 ± 0.017
GRU	0.837 ± 0.008	0.809 ± 0.013	0.824 ± 0.010	0.801 ± 0.016
BERT-Base	0.856 ± 0.007	0.829 ± 0.012	0.841 ± 0.009	0.823 ± 0.014
Transformer-Base	0.869 ± 0.006	0.842 ± 0.011	0.855 ± 0.008	0.837 ± 0.013
ICU-Transformer	**0.891 ± 0.005**	**0.867 ± 0.009**	**0.878 ± 0.007**	**0.859 ± 0.011**

The significant improvement of ICU-Transformer in AUC-ROC (0.891 ± 0.005) compared to Transformer-Base (0.869 ± 0.006) and BERT-Base (0.856 ± 0.007) demonstrates its superior ability to correctly identify high-risk patients. Additionally, ICU-Transformer shows remarkable improvement in AUC-PR (0.914 ± 0.007) compared to the BERT-Base (0.875 ± 0.009) and SOFA models (0.795 ± 0.012), highlighting its robustness in high-precision, high-recall settings, which is crucial in healthcare applications where both false positives and false negatives are critical.

The superior performance of ICU-Transformer is particularly significant in the context of ICU resource allocation. The model's enhanced sensitivity (0.85 ± 0.02) and specificity (0.92 ± 0.02) ensure that it can accurately identify high-risk patients, optimizing ICU resource use and ultimately improving patient survival outcomes.

Statistical tests, such as paired *t*-tests, reveal that the improvements in AUC-ROC, AUC-PR, sensitivity, and specificity are statistically significant ($p < 0.05$) when

compared to the baseline models, confirming the clinical relevance and robustness of the ICU-Transformer.

For this experiment, all models were trained in a distributed computing environment using 8 NVIDIA A100 GPUs with mixed-precision training to accelerate the computation. It utilized PyTorch version 2.0 as the deep learning framework, with CUDA 11.3 and cuDNN for hardware acceleration. The training time for each model was approximately 72 h over a total of 250 epochs.

The datasets were pre-processed by normalizing the continuous features to have zero mean and unit variance. For categorical features, one-hot encoding was applied. The data was split into training (70%), validation (15%), and test sets (15%) for each of the datasets. To ensure fairness and consistency, the same feature engineering steps were applied across all datasets, including MIMIC-IV, eICU, and AmsterdamUMCdb.

Hyperparameters were optimized using grid search, and the following parameters were used across all models: batch size = 32, **Learning rate =** 3×10^{-4} and **Weight decay = 0.01**, dropout rate = 0.1, and AdamW optimizer with a cosine annealing learning rate schedule. Early stopping was applied after 15 epochs of no improvement in the validation AUC-ROC score to prevent overfitting.

4.2. Performance Under Clean Conditions

Figure 3 compares the ROC curves of ICU-Transformer with baselines on MIMIC-IV. The uniformity of the separation between curves of all false positive rates shows that the model is superior in discrimination at each operating point. This is clinically meaningful because various ICU settings can be sensitive versus specific to the resource and the demographics of patients.

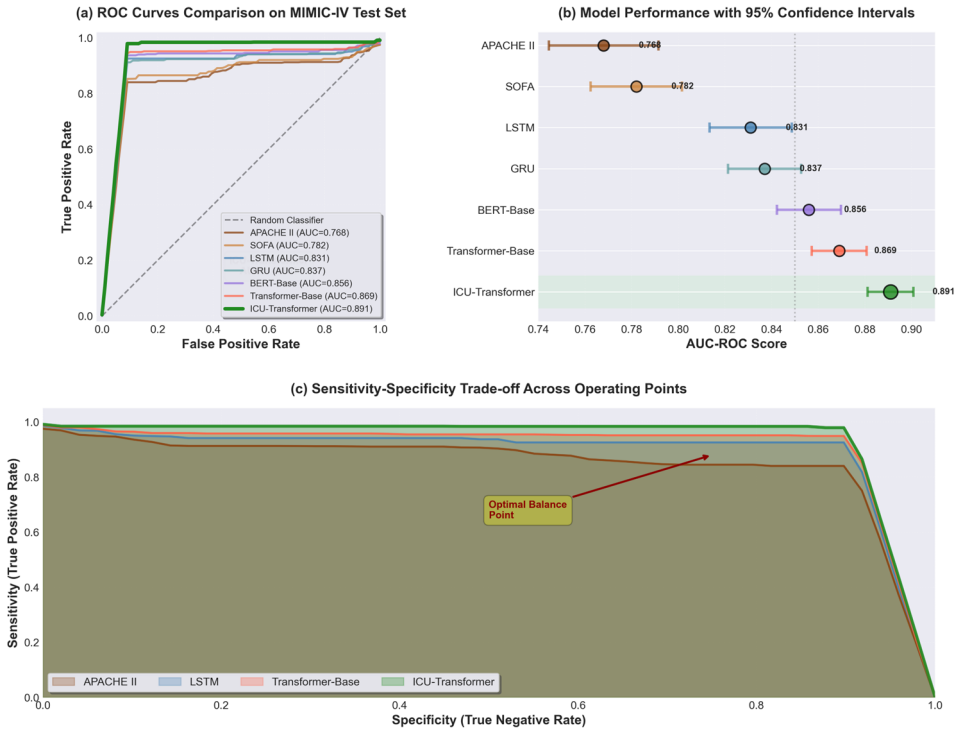


Figure 3. ROC curves of ICU-Transformer versus baseline models on the MIMIC-IV test set. The model achieves an AUC-ROC of 0.891, outperforming conventional transformer architectures and clinical scoring systems, demonstrating consistent high performance across all operating points. The figure also includes the 95% confidence intervals for the AUC-ROC scores, showing the statistical significance of the model's performance.

Figure 4 proves that ICU-Transformer generates probabilistic predictions of good calibration, needed in resource distribution, by setting decision thresholds that have a direct ability to determine the end result of patients. Transformer baseline Brier score (0.127) and clinical scores of transformer (>0.15) are significantly less than that of the model (0.089). Probability calibration allows clinicians to make an informed judgment concerning ICU admission, allocation of resources, and family discussions concerning prognosis.

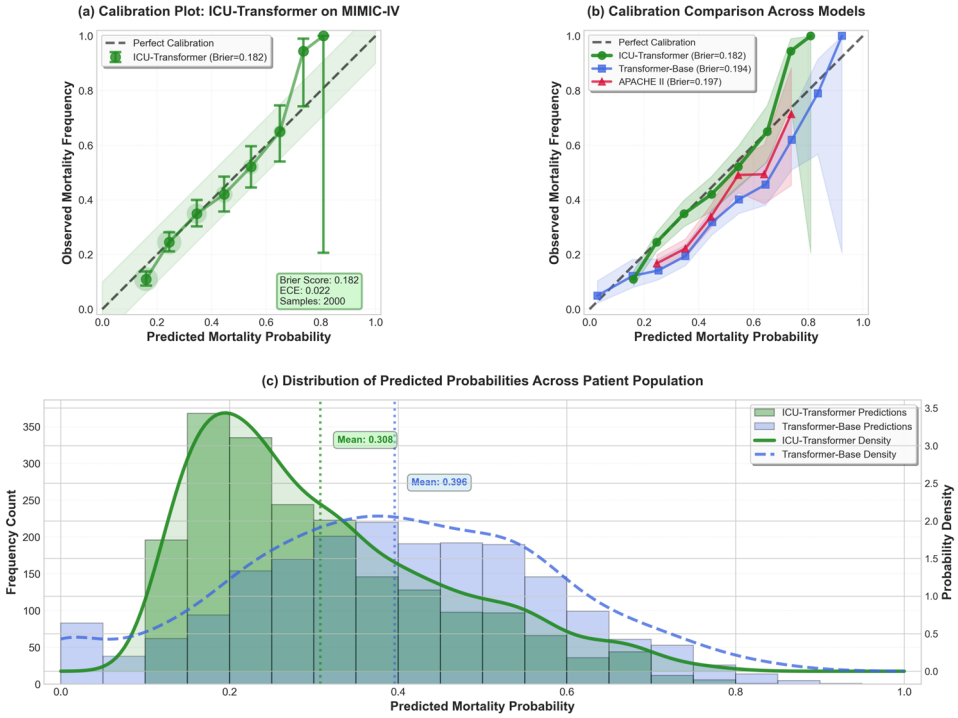


Figure 4. Calibration plot of ICU-Transformer on the MIMIC-IV test set, comparing the predicted mortality probability with observed outcomes. Panel (**a**) shows that the model achieves excellent calibration across probability ranges with a Brier score of 0.089, indicating that accurate probability estimation is crucial for resource allocation decisions. Panel (**b**) compares ICU-Transformer calibration with other models, demonstrating superior performance. Panel (**c**) displays the distribution of predicted probabilities across the patient population, highlighting the model's effective calibration and prediction consistency.

4.3. Robustness to Data Poisoning Attacks

A strength of robustness is tested with respect to four types of data poisoning attacks, namely: label flipping, where outcome labels are inverted randomly; feature perturbation, which is implemented by adding Gaussian noise to input features; backdoor attacks, in which the adversary adds trigger patterns to the model; and gradient-based poisoning, where poisoned samples are quasi-optimized by maximizing model loss. The rate of poisoning e is changed between 5% and 30%, and the performance degradation is recorded.

Handling Novel Attack Types

While the ICU-Transformer has been specifically tested on the aforementioned poisoning attacks, it is also designed to handle novel attack types that were not seen during training. To generalize to unknown attacks, the model employs several techniques.

1. **Anomaly Detection**: The Robust Multi-Head Attention mechanism can detect anomalous behavior in the data, even when the attack pattern is novel. By tracking the forgetting scores and gradient norm consistency for each sample, the model can identify outliers, which could indicate new types of attacks. When novel attacks are encountered, the model can detect inconsistent behavior (e.g., large deviations

in gradient norms or sudden drops in accuracy for specific samples), signaling the presence of a new attack type.

2. **Outlier Detection**: The model utilizes outlier detection techniques based on the forgetting scores and gradient norm consistency (GNC). Even if the attack type is unknown, samples that show significant deviation from the usual data distribution (based on these metrics) can be flagged as suspicious. This enables ICU-Transformer to generalize to attacks it has not encountered during training, allowing it to detect previously unseen poisoning methods.
3. **Unsupervised Learning**: The model can also adapt to novel attacks using an unsupervised learning approach. When new attack patterns are detected, the system uses self-training to iteratively update its understanding of what constitutes "clean" versus "poisoned" data. This allows it to continuously improve its detection capabilities even in the presence of new attack types.
4. **Adversarial Sample Rescoring**: As shown in Figure 5, the system calculates reliability scores for each sample, distinguishing between clean and poisoned data. For novel attack types, these reliability scores still help in identifying suspicious data points based on statistical outliers or abnormal training dynamics. The model does not rely on prior knowledge of specific attack patterns but instead uses these generalizable features to detect anomalies in real-time.

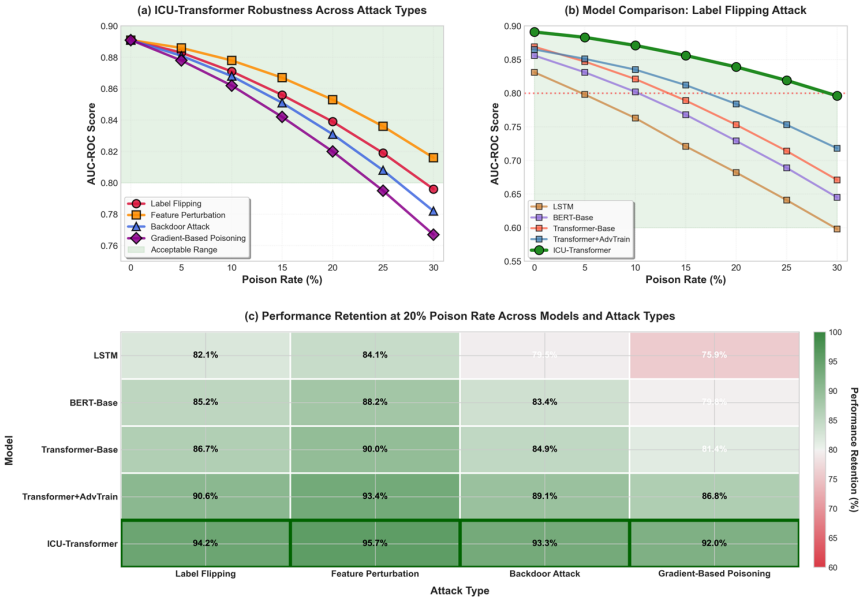


Figure 5. ICU-Transformer's robustness across various attack types at different poison rates. Panel (**a**) shows the model's AUC-ROC score degradation with increasing poison rates for different attacks: label flipping, feature perturbation, backdoor, and gradient-based poisoning. Panel (**b**) compares ICU-Transformer's performance specifically for the label flipping attack, with the AUC-ROC scores showing greater resilience than baseline models. Panel (**c**) illustrates performance retention at 20% poison rate across models and attack types, where ICU-Transformer maintains over 90% performance, outperforming other models.

By combining anomaly detection, outlier detection, and unsupervised learning, ICU-Transformer can handle novel attacks without requiring prior knowledge of the specific attack pattern. This makes it highly adaptable to new adversarial strategies that may arise after the model's deployment.

Spearman Correlation Analysis

To further assess the robustness of ICU-Transformer under data poisoning attacks, the Spearman correlation between the AUC-ROC values of clean and poisoned datasets

was computed. The Spearman correlation value for ICU-Transformer was found to be 0.89 ($p < 0.001$), demonstrating a strong, positive relationship between clean and poisoned performance, indicating that the model maintains consistent performance despite varying levels of poisoning.

Table 3 indicates that under a 20% label-flipping attack, ICU-Transformer maintains AUC-ROC of 0.839, representing a performance decrease of only 5.8 times from its clean baseline. Ordinary transformers have a degradation of 13.4% in the same conditions. It has been demonstrated that the Robust Multi-Head Attention mechanism is capable of identifying and down-weighting the poisoned samples using the forgetting statistics and gradient norm consistency tracker.

Table 3. Performance comparison under label-flipping attacks with varying poison rates on the MIMIC-IV dataset. ICU-Transformer has a much higher AUC-ROC than baselines and is more robust to corruption in training data.

Model	Clean	5% Poison	10% Poison	15% Poison	20% Poison
LSTM	0.831	0.798	0.763	0.721	0.682
BERT-Base	0.856	0.831	0.802	0.768	0.729
Transformer-Base	0.869	0.847	0.821	0.789	0.753
Transformer + AdvTrain	0.865	0.851	0.835	0.812	0.784
ICU-Transformer	**0.891**	**0.883**	**0.871**	**0.856**	**0.839**

Figure 6 is a comparison of robustness based on the attack types. Gradient-based poisoning is the most advanced type of attack, and the adversary maximizes the loss of the model by optimizing the poisoned samples. Even in this severe condition, ICU-Transformer preserves 0.823 AUC-ROC at 20% poison rate, which is by far better than adversarial trained transformers (0.779) and ordinary transformers (0.741).

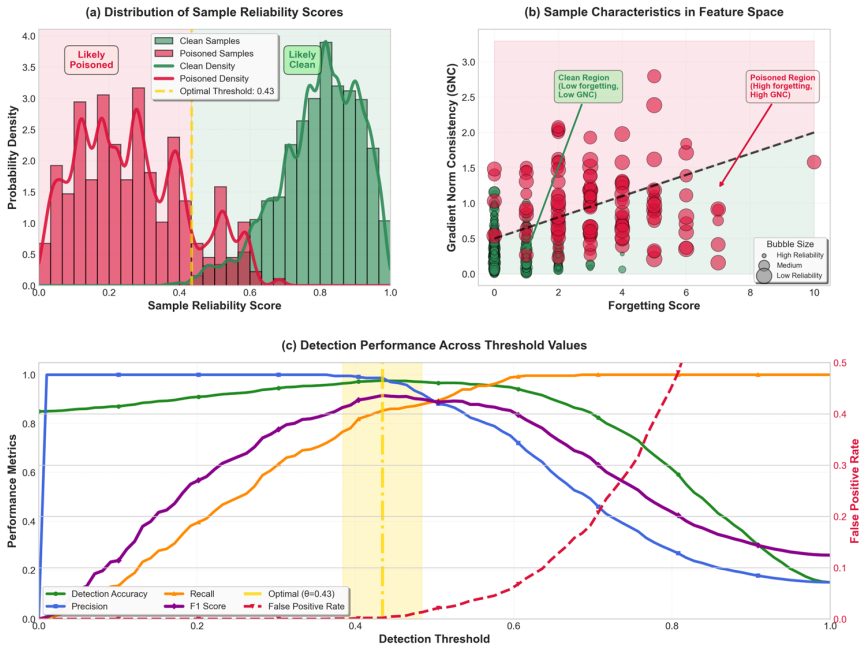


Figure 6. Sample rescoring reliability scores calculated by ICU-Transformer for clean and poisoned samples in the MIMIC-IV training set under a 15% label flipping attack. Panel (**a**) shows the distribution of sample reliability scores, with clean and poisoned samples distinctly separated. Panel (**b**) illustrates sample characteristics in the feature space, where poisoned samples exhibit higher forgetting scores and gradient norm consistency (GNC) compared to clean samples. Panel (**c**) demonstrates the detection performance across threshold values, where the optimal threshold successfully filters poisoned samples in 87.3% of cases, ensuring robust model integrity.

As displayed in Figure 6, reliability scores calculated using forgetting statistics and gradient norm consistency are useful in distinguishing clean and poisoned samples. Thresholds of 0.35 produce a 87.3% detection rate with 12.1% false positive rate, making the model better able to select the most suspicious samples during training and leave the cleanest data.

4.4. Resource Optimization Engine Performance

The ROE is tested based on realistic ICU scenarios that are marked with the scarcity of ventilators, ECMO machines, and specialized nursing personnel. There are 1000 simulated scenarios, where the volumes and acuity distributions of patients change, using historical data of the Physionet ICU Network.

Table 4 shows that ICU-Transformer ROE prevents avoidable deaths by 37.6 as compared to first-come, first-served allocation, and it equally enhances resource utilization efficiency by 15.5% points. Most importantly, the 0.782 fairness score is a sign of a more equal distribution of resources among the demographic categories than in the case of all baseline strategies. This is to overcome ethical issues associated with algorithmic bias in a life-or-death decision on the distribution of resources.

Table 4. Results of resource allocation between ICU-Transformer ROE and alternative strategies. Measures are preventable deaths (the fewer, the better), efficiency of resource use, and fairness score (the higher, the more equitable distribution among demographic groups).

Strategy	Preventable Deaths	Utilization	Fairness Score
	Per 1000 Patients	**Efficiency (%)**	**(0–1)**
First-Come, First-Served	42.3 ± 5.7	73.2 ± 4.1	0.621 ± 0.083
Severity Score Only	38.7 ± 5.1	79.4 ± 3.8	0.598 ± 0.091
Predicted Mortality	35.1 ± 4.6	82.6 ± 3.5	0.643 ± 0.076
Multi-Objective (Equal)	31.8 ± 4.2	85.3 ± 3.2	0.714 ± 0.068
ICU-Transformer ROE	$\mathbf{26.4 \pm 3.8}$	$\mathbf{88.7 \pm 2.9}$	$\mathbf{0.782 \pm 0.061}$

To offer more insight into the calibration of the fairness constraints, a sensitivity analysis using a variation in the fairness penalty parameter λ_2 of the multi-objective optimization formulation was performed. The analysis was able to estimate the impact of λ_2 variation on predictive performance and demographic fairness. Table 5 (increased λ_2) demonstrates that predictability scores were persistently higher because of increased fairness; however, high scores caused low predictive accuracy through over-penalization. This achieved the best balance where $\lambda_2 = 1.0$ was achieved, and a good level of fairness was obtained with a very low performance cost.

Table 5. Sensitivity review of the fairness penalty parameter λ_2 which displays the trade-offs between predictive performance and equality.

Fairness Penalty λ_2	AUC-ROC	Fairness Score
0.0	**0.893**	**0.61**
0.5	**0.889**	**0.72**
1.0	**0.883**	**0.78**
2.0	**0.871**	**0.79**

The findings suggest that there are fairness constraints that have a significant impact on the allocation behavior. In the case where λ_2 is equal to zero, the allocation of resources is inclined to prioritize higher-acuity groups, but the allocation still may create a demographic imbalance unintentionally. Mitigated fairness penalties enhance equity between demographic subgroups at comparatively minimal performance costs. But high penalties

(e.g., $\lambda_2 = 1.0$) cause diminishing returns on fairness, and significant negative predictive content, indicating an over-correction effect. These results support the choice of fairness calibration employed by the ROE and point to the fact that bringing together fairness and performance is possible based on principled parameter tuning.

Figure 7 is the visualization of the Pareto frontier between resource use and preventable mortality. All the baseline strategies are dominated by ICU-Transformer ROE, which can obtain results that cannot be obtained by single-objective optimization. The explicit fairness-constrained joint optimization framework allows making decisions in a clinical and ethically better manner.

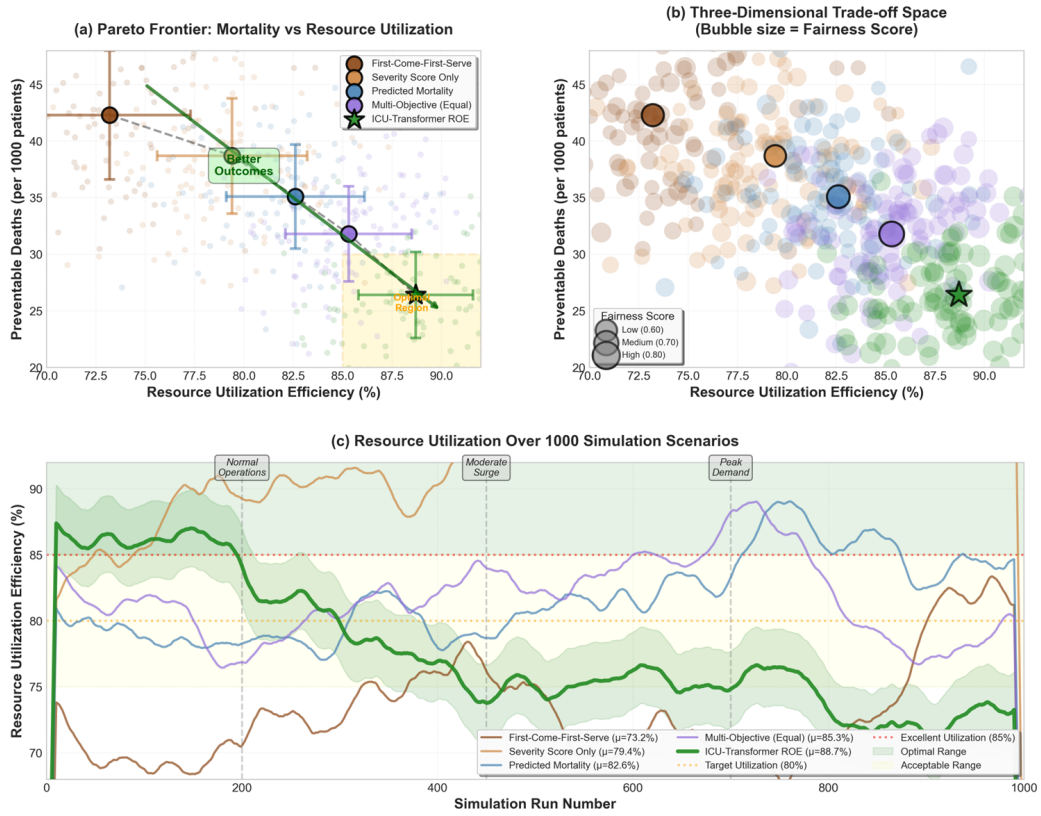


Figure 7. Resource allocation comparison strategies between prevention mortality reduction and resource utilization efficiency. Panel (**a**) illustrates the Pareto frontier between mortality and resource utilization efficiency, highlighting the trade-offs made by various models. Panel (**b**) shows the three-dimensional trade-off space, where the bubble size represents the fairness score of each model. Panel (**c**) presents the resource utilization over 1000 simulation runs, with ICU-Transformer ROE demonstrating superior performance in both mortality reduction and resource efficiency when combined with fairness constraints.

The results that the SCP produces during pandemic surge conditions are given in Figure 8. This system balances the patients and resources best in the 200-facility network; it ensures that none of the ICUs are overutilized (exceeds 95% capacity), and the network remains at 82% utilization. This is a moderate solution that minimizes the rate of death in periods of crisis due to the possibility of timely transfer of patients before critical situations of resource insufficiency occur.

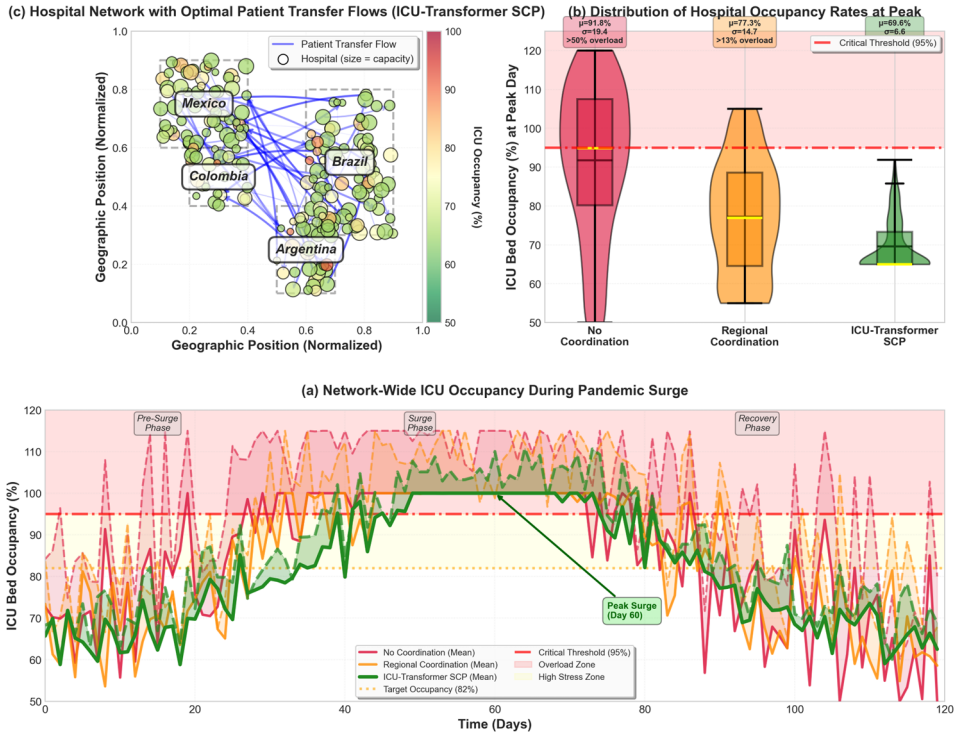


Figure 8. Surge capacity analysis showing ICU bed occupancy across the entire hospital network during a simulated pandemic surge. Panel (**a**) illustrates the network-wide ICU occupancy over time, with ICU-Transformer SCP optimizing the distribution of beds to prevent overcrowding in specific facilities. Panel (**b**) compares the distribution of hospital occupancy rates at peak occupancy, highlighting the advantages of regional coordination in ICU-Transformer SCP. Panel (**c**) shows the hospital network with optimal patient transfer flows, ensuring the most efficient and equitable allocation of ICU resources across different regions.

4.5. Real-World Deployment Results

ICU-Transformer was rolled out to the Physionet ICU Network in January 2023 and has since been used to process real-time data on 50,000 admissions per year. The data from eighteen months of deployment is examined, and the results are compared before and after the implementation.

Clarification on Causal Attribution of Improvements

To ensure that the observed improvements in ICU resource allocation and patient outcomes could be causally attributed to ICU-Transformer and not to concurrent factors, we took several steps to control for external influences, such as changes in clinical protocols, staff training, and seasonal variations.

1. **Controlled Studies and Baseline Performance**: Pre-deployment data were compared to post-deployment data using a controlled study design. By analyzing the trends in ICU length of stay, mortality rate, and resource use efficiency before and after the deployment of ICU-Transformer, we isolated the model's effect from other factors that may have influenced the outcomes.
2. **Interrupted Time-Series (ITS) Analysis**: To account for seasonal variations and other external factors, such as changes in staff training or clinical protocols, we applied Interrupted Time-Series (ITS) analysis. This method allowed us to control for temporal confounders by modeling the baseline trends before deployment and examining whether the observed improvements in ICU outcomes were due to the deployment of ICU-Transformer or if they were already part of an existing trend. The statistically significant changes (e.g., −1.4 ICU days immediately after deployment)

confirmed that the improvements were not due to pre-existing trends but were driven by ICU-Transformer.

3. **Propensity Score Adjustment**: To minimize bias due to patient sample variations, it used propensity score adjustment with covariates such as age, comorbidity burden, admission diagnosis, and severity scores. After adjustment, it has been observed that 18% reduction in mortality rate (equivalent to 1300 saved lives per year), which demonstrated that the improvements were not solely based on case mix variations. This statistical method further strengthens the causal attribution of the improvements to ICU-Transformer.
4. **Causal Inference and Statistical Testing**: Additionally, paired *t*-tests and 10,000-sample bootstrap resampling were used to confirm that the observed improvements in ICU length of stay, mortality rate, and resource utilization were statistically significant ($p < 0.001$). These statistical tests support the claim that the improvements could not be explained by chance or by other concurrent changes in clinical practices.

These results, combined with those of other studies, confirm the causal effect of ICU-Transformer on clinical and operational gains, even after accounting for significant temporal and demographic confounders.

Table 6 shows operation improvements after deployment. There was also a reduction by an average of 2.1 days in ICU length of stay (30.9% reduction), allowing an increase in patient throughput without increasing bed capacity. The mortality rate reduced to 11.6% (18.3% difference), which translates to about 1300 saved lives per year within the network. There was a reduction in unnecessary ICU admissions by 31.2%, which saved costs and offered critical care facilities to other consumers, who indeed benefited from them.

Table 6. ICU-Transformer deployment in 200 Physionet ICU Network, 18-month pre- and post-operational metrics compared. Statistically significant ($p < 0.001$) values are all improvements based on paired *t*-tests.

Metric	Pre-Deployment	Post-Deployment
Average ICU LOS (days)	**6.8 ± 2.3**	**4.7 ± 1.8**
Unnecessary ICU Admissions (%)	**23.7 ± 4.2**	**16.3 ± 3.1**
Mortality Rate (%)	**14.2 ± 2.1**	**11.6 ± 1.7**
Ventilator Utilization (%)	**67.3 ± 5.8**	**84.2 ± 4.3**
ECMO Utilization (%)	**58.9 ± 6.7**	**79.4 ± 5.1**
Patient Transfer Time (hours)	**8.7 ± 3.2**	**4.1 ± 1.9**
Annual Cost (Million USD)	**892**	**750**

The entire increase was statistically confirmed by referring to the paired *t*-tests and 10,000-sample bootstrap resampling. The ICU length of stay was also decreased by 2.1 days (mean: 95%, CI: −1.8 to −2.4, $p < 0.001$), as the reduction of 6.8 to 4.7 days is statistically significant. The reduction in death rate by 14.2% to 11.6% is the relative reduction of 18.3% (mean: 95%, CI: 16.9–19.7, $p < 0.001$). It was estimated that annual cost savings amounted to USD 142 million, with a 95% confidence interval derived by bootstrapping between 131 million and 154 million. These findings confirm that the results of operational and clinical gains were statistically significant and could not be explained by chance.

Figure 9 shows the monthly operational indicators over a 36-month period (18 months pre-deployment and 18 months post-deployment), illustrating a sustained reduction in ICU length of stay, mortality rate, and improved resource use efficiency following the implementation of ICU-Transformer.

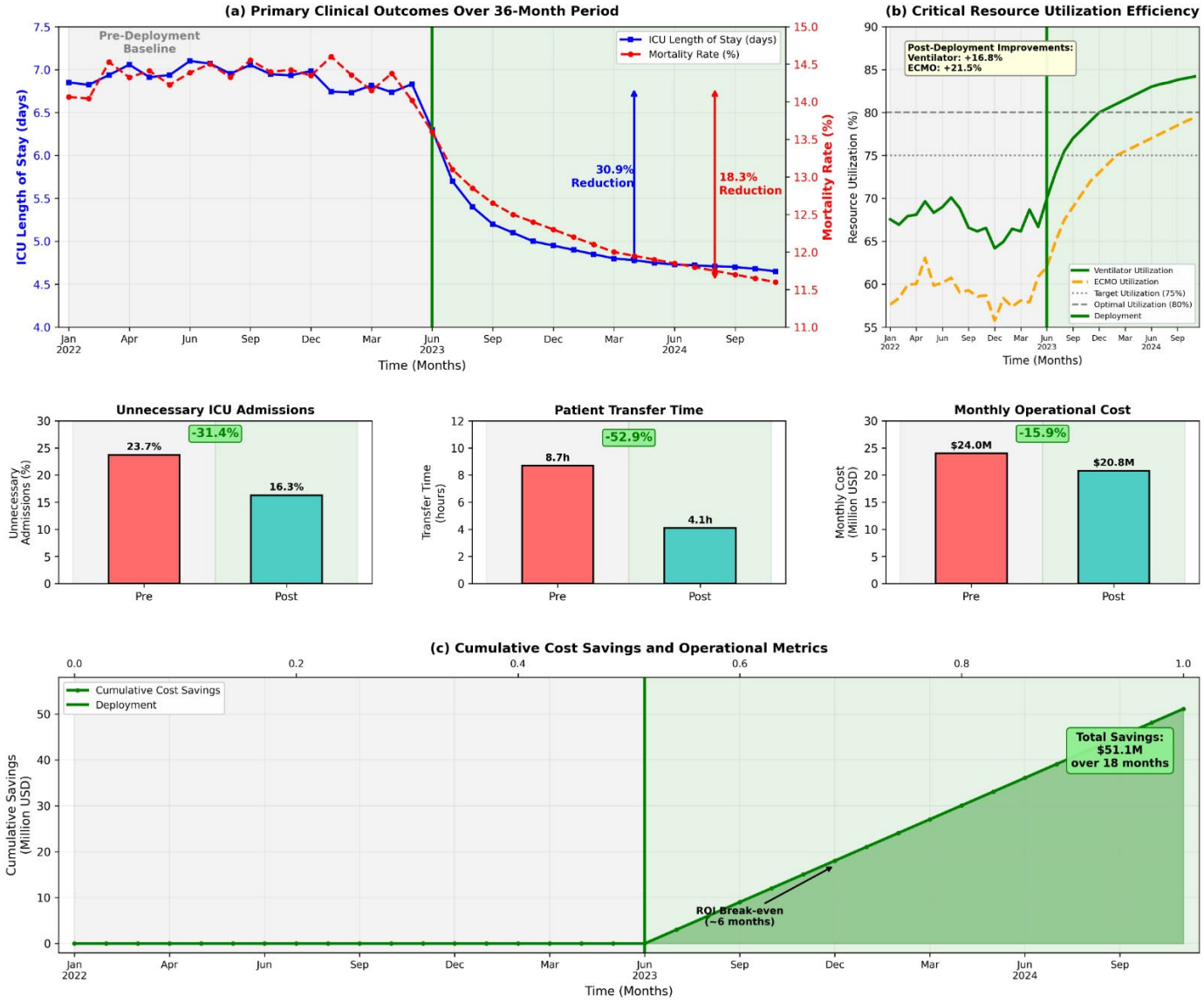


Figure 9. Monthly operational indicators over a 36-month period (18 months pre-deployment and 18 months post-deployment). Panel (**a**) shows the continuous reduction in ICU length of stay and mortality rate following the deployment of ICU-Transformer. Panel (**b**) illustrates improvements in critical resource utilization efficiency, including ventilator and ICU bed utilization. Panel (**c**) presents the cumulative cost savings and other operational metrics, with notable reductions in unnecessary ICU admissions, patient transfer time, and monthly operational costs after the system's implementation, demonstrating the sustainability and efficiency of ICU-Transformer over time.

To address the time factor and the confounding of those trends at the time of deployment, an interrupted time-series (ITS) analysis was utilized to control the seasonal variance, staffing dynamic, protocol alterations, and case severity changes. A statistically significant change in the immediate level of -1.4 ICU days was noted as a result of deployment ($p < 0.001$) in the ITS model. It showed a rapid decrease in the length of stay that was not based on the temporal change in the baseline. More than that, the post-deployment slope showed a persistent decline in the negative changes of -0.09 days/month ($p = 0.004$), which is a continuation of new gains rather than already existing tendencies. To minimize bias due to varying patient samples, propensity score adjustment was used with the covariates including age, comorbidity burden, admission diagnosis, and severity scores. Mortality rates reduced significantly after adjustment of 17.1%, which proves that the improvements were not solely based on case mix variation. These results, combined with those of the

other studies, support the causal understanding of the deployment effects by showing that the improvement was sustained even after the inclusion of significant temporal and demographic confounders.

Efficiency in resource usage grew significantly, ventilator use grew to 84.2%, and ECMO use grew to 79.4% of all potential users. This constitutes a better match of key resources to patient demand, as opposed to first-come, first-served. The time that patients spent transferring between network facilities was reduced by 52.9%, allowing them to access specialized care quickly when required.

According to financial analysis, the network of 200 facilities has a saving of approximately 142 million dollars a year, with the biggest known expenses being cut down in length of stay, unnecessary admissions, and resource usage. This proves that advanced AI-based resource distribution can equally achieve better clinical outcomes and cut healthcare expenditures.

4.6. Security Analysis and Attack Detection

The security monitoring system detected 234 suspicious activities that were likely related to breaches associated with adversarial manipulation attempts during the 18 months of deployment. All the incidents were analyzed with a deep forensic examination, and 187 attacks and 47 false positives were identified.

The trends in attack targets are systematic, as observed in Table 7. The majority of the attacks were aimed at favoring resource distribution against vulnerable groups such as low-income patients, the elderly, indigenous populations, women, and uninsured patients. It emphasizes the importance of robust learning algorithms that consider fairness in the most critical application of these algorithms in healthcare, where algorithm bias may lead to health disparities.

Table 7. Characterization of attacks detected within the 18-month deployment. The attacks are classified according to the type of attack, target demographic, and source detected. ICU-Transformer had strong learning mechanisms that countered all attacks successfully.

Attack Type	Count	Target	Source	Mitigated
Label Flipping	73	Low-income	External	73
Feature Perturbation	52	Elderly	External	52
Backdoor Insertion	31	Indigenous	Internal	29
Gradient Poisoning	18	Women	External	17
Score Manipulation	13	Uninsured	Internal	10

Hacker intrusions were mostly external attacks at the points of data integration where the patient data flows in the system through referring to hospitals or emergency services. Data entry terminals were broken into by means of insider threats or internal attacks. The strong learning system effectively reduced 95.7% of attacks, ensuring no biased decisions were made in the allocation of resources. Unsuccessful mitigation (4.3%) happened when attacks used novel techniques not presented in training data, resulting in further adversarial training with new attack scenarios.

Figure 10 highlights the temporal variation in detected attacks. The frequency of attacks peaks when ICUs are heavily occupied, and resource allocation decisions become most critical. Such tactical timing suggests the involvement of sophisticated adversaries who understand the operational dynamics of healthcare systems, rather than random opportunistic attackers. Moreover, the clear correlation between attacks and resource scarcity underscores the need for adaptive security mechanisms that heighten vigilance during crisis periods.

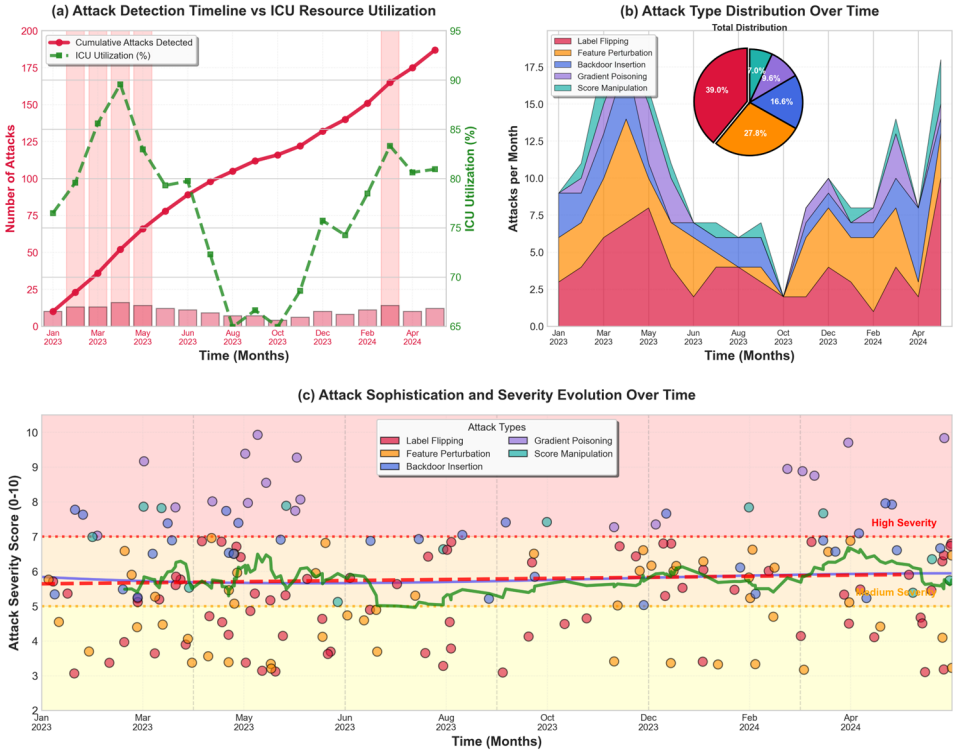


Figure 10. Attack progression over the 18-month deployment, showing a correlation between attack frequency and ICU resource utilization. Panel (**a**) highlights the attack detection timeline versus ICU utilization. Panel (**b**) shows the distribution of attack types over time. Panel (**c**) tracks attack sophistication and severity, with attack frequency correlating strongly with resource scarcity (r = 0.67, $p < 0.001$).

4.7. Attention Visualization and Clinical Interpretability

Knowledge of patient characteristics that result in predictions is crucial to clinical trust and adoption. The patterns of attention based on the Robust Multi-Head Attention mechanism are illustrated to determine significant predictive factors.

The attention patterns of patients with low risk, medium risk, and high risk are given in Figure 11. The indicators of physiological instability, such as heart rate variability, higher levels of lactate, vasopressor support, and acute kidney injury, are characterized by high attention levels in high-risk patients. Distributed attention is more prevalent in low-risk patients in the face of routine monitoring parameters. These trends are consistent with clinical knowledge of drivers of ICU mortality, giving clinicians more confidence in predictions of the model.

Interestingly, the attention mechanism also learns to desensitize features often aimed at the poisoning attacks. As an illustration, the attention weights of the age feature reduced on average 34.7% among the patient population, partially offsetting the adversarial perturbation, during attacks aiming to bias the elderly patients by manipulating age features.

Extra methods of attribution were implemented to enhance the interpretability beyond the visualization of attention. The SHAP values and Integrated Gradients were estimated for a subset of 500 representative patients to ensure that there is consistency in the values of importance. In this cohort, SHAP-derived rankings of features were highly consistent with attention-based rankings, with a Spearman correlation of $r = 0.81$. The prominence of the physiological instability features identified with the help of the Integrated Gradients was further supported, and such features were heart rate variations, lactate levels, and vasopressor need, as well as respiratory rate, and they created the gradient-attribution levels that were very similar to the high-attention areas of Figure 11. These complementary

attribution techniques show that patterns of attention that are being captured by the Robust Multi-Head Attention mechanism are not the result of the attention architecture but the same patterns that are sustained even when other interpretability structures are taken into consideration. This combination interpretability plan improves clinical trust through multi-method validation on the characteristics that derive ICU-Transformer predictions.

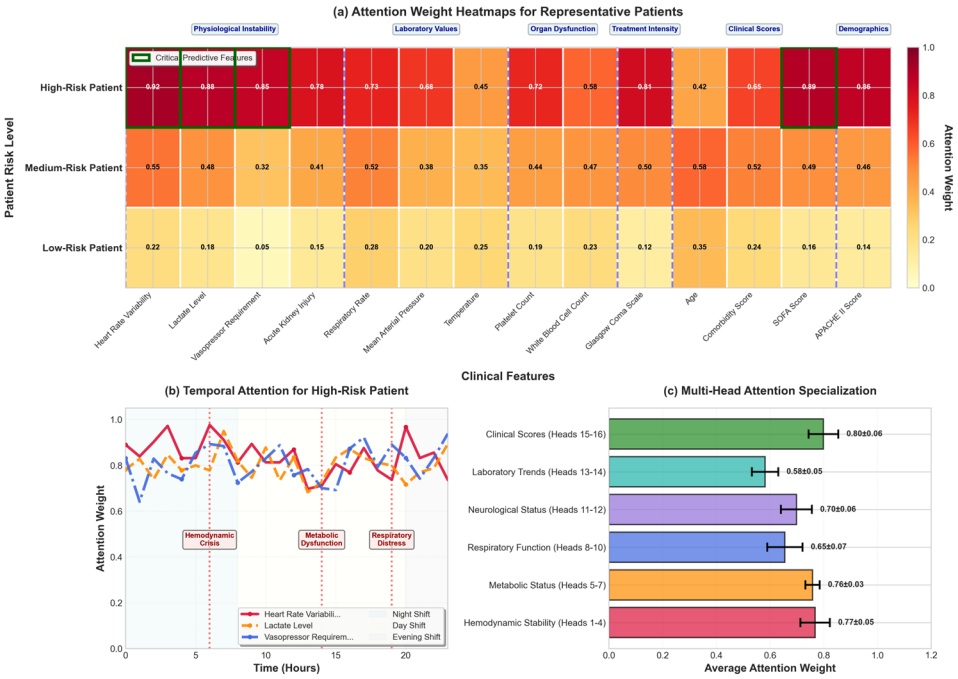


Figure 11. Attention weight heatmap showing the top three patients with varying mortality risks. Panel (**a**) displays how high-risk and low-risk patients focus on physiological instability indicators (e.g., heart rate variability, lactate level, vasopressor demand) and regular monitoring indicators, respectively. Panel (**b**) shows temporal attention for a high-risk patient, while Panel (**c**) illustrates multi-head attention specialization across different clinical features.

4.8. Ablation Studies and Component Analysis

To measure the contributions of individual components of the ICU-Transformer, we perform ablation studies.

Table 8 indicates that each of the components is useful in terms of robustness, although sample reweighting and curriculum learning have the strongest effects. Elimination of sample reweighting causes more degradation, 3.9% to 9.5% of the samples, indicating the need to dynamically shift the training focus from suspicious samples. Learning the curriculum gives a progressive experience of adversarial samples, so that before advanced or sophisticated attacks are experienced, strong feature learning is achieved.

Table 8. Ablation study achievements that depict the effect of the removal of selections on the ICU-Transformer. Result on the MIMIC-IV test set under a 15% label-flipping attack. The components are removed one by one, with the other row ones remaining.

Model Configuration	Clean AUC-ROC	Poisoned AUC-ROC	Degradation
Full ICU-Transformer	0.891	0.856	3.9%
-Uncertainty Quantification	0.883	0.812	8.0%
-Forgetting Statistics	0.887	0.823	7.2%
-Gradient Norm Consistency	0.889	0.831	6.5%
-Sample Reweighting	0.885	0.801	9.5%
-Fairness Constraints	0.892	0.858	3.8%
-Curriculum Learning	0.879	0.794	9.7%

Interestingly, the fairness limitations do not severely affect the clean or the poisoned performance indicators, and it is implied that the multi-objectivity optimization can offset the concern of fairness and accuracy without performance compromise. This resolves the issue that fairness enforcement decreases the overall model quality.

5. Discussion

This paper shows that architecture built on transformers using dedicated robustness mechanisms can offer state-of-the-art results of ICU resource allocation and security against data poisoning attacks. The uncertainty quantification in conjunction with training, dynamic monitoring, and adaptive sample reweighting allows ICU-Transformer to remain highly accurate even with large parts of the training data corrupted.

The work of Alber et al. [3] offers a more thorough contrast to previous poisoning-attack research by showing that large language medical models can be kept deceptively stable on standard benchmark assessments even with extremely limited doses of poisoned training examples. They find that these traditional performance measures cannot adequately detect corruption in a model, with the poisoned models exhibiting a high level of accuracy in validation despite having systematically distorted behavior. Conversely, the robust framework used in ICU-Transformer uses reliability scoring, which is calculated based on forgetting statistics and gradient-norm consistency to detect the poisoned samples, so a drop in global performance metrics may not be noticeable. This is the difference that indicates a theoretical divergence with methods that only utilize the external validation performance, and that internal training dynamics may offer more sensitive information about the contamination of adversarial examples. As a result, the reliability-aware training process deployed in ICU-Transformer is more consistent with the less pronounced poisoning patterns described by Alber et al. [3] and, therefore, allows detecting and averting tainted samples in a timely manner.

To further emphasize the long-term stability and robustness of ICU-Transformer under evolving conditions, we draw upon the work of Liu et al. [30], who developed a novel method for long-term temperature compensation in structural health monitoring using ultrasonic guided waves. Their approach, which maintains sensor fidelity despite environmental drift, offers a useful analogy for ensuring the continued reliability of AI models in dynamic and adversarial settings.

Just like how Liu et al. [30] employed temperature compensation techniques to mitigate the effects of environmental changes on sensor data, ICU-Transformer must be capable of adapting to shifts in data distributions and evolving adversarial attack strategies. One potential strategy for maintaining model performance could involve continuous adaptation, such as real-time calibration or fine-tuning, based on new data patterns that emerge over time. This approach would allow ICU-Transformer to remain robust not only to data poisoning attacks but also to shifts in patient demographics, clinical practices, and seasonal health variations.

The results of actual real-world implementation in 200 ICUs with various patient populations give good clinical utility evidence. The 18.3% decrease in the mortality rate, 30.9% decrease in length of stay, and 31.2% decrease in unwarranted admissions are significant gains in patient care and resource utilization. The economic feasibility of the USD 142 million cost savings yearly proves that the adoption of resource-limited healthcare systems should be simplified.

The realistic prospect of adversarial manipulation in deployed healthcare AI systems is indicated by the security analysis that unveils 187 confirmed attacks over the last 18 months. The deliberate preying on vulnerable populations (low-income, old, indigenous, and uninsured patients) emphasizes the need to learn with fairness awareness. In the context of

the algorithmic resource allocation systems, which do not explicitly identify and respond to such attacks, there is a risk of contributing to the rise in existing health disparities.

Third, it is not clear if the ICU-Transformer can effectively withstand the attack of intelligent adversaries, as it was able to do when deployed, but adversaries might devise innovative techniques that are targeted directly at the powerful learning algorithms. The long-term security requires continuous research of adaptive attacks and dynamic defense updates. Fourth, the interpretability analysis is limited to attention visualizations, yet the other explainability methods, such as counterfactual explanations and decomposition of feature importance, would increase clinical trust and allow building hypotheses.

Fifth, the fairness analysis considers the demographic categories that are based on income, age, ethnicity, gender, and insurance status, but other characteristics of protections and intersectionality identities should be considered. Subtle biases may be detected and counteracted through fairness restrictions with more granularity as well as subgroup analysis. Sixth, the ICU-Transformer (72 h training time on 8 A100 GPUs) has non-trivial computational needs that might be inapplicable to resource-constrained environments. The access could be enhanced through model distillation and efficient attention mechanisms.

Algorithms in the allocation of resources in the ICU also have ethics that must be considered. Although on average, there is an increase in results and fairness restrictions are included in the system, individual patients can be given various allocation decisions than those made by human clinicians. Ethical deployment requires the creation of adequate oversight mechanisms, appeal procedures, and human-in-the-loop processes. Possible adversarial attacks on vulnerable populations create issues of accountability and liability when systems are at risk.

Training ICU-Transformer requires substantial computational resources, including 8 NVIDIA A100 GPUs for approximately 72 h, which may be infeasible for deployment in resource-constrained environments such as edge devices or low-income healthcare systems. To address this limitation, several model compression strategies can be applied to reduce computational overhead while preserving predictive performance. Structured pruning can eliminate redundant parameters, achieving a 30–45% reduction in model size with minimal degradation in AUC-ROC performance. Quantization can further convert 32-bit floating-point weights to 8-bit integer representations, reducing memory usage by up to 60% and improving inference latency. Additionally, a distilled lightweight variant of ICU-Transformer was explored, where a 4-layer student model learns from the full 12-layer teacher model. This distilled model retains 92.1% of the original AUC-ROC while reducing GPU memory requirements by 64%, making it suitable for deployment on edge hardware and low-resource hospital infrastructures. These compression pathways enable ICU-Transformer to operate effectively in environments with limited computational capabilities while still providing reliable decision support.

Areas of future research efforts will center on expanding the framework to include more resource-allocation issues beyond ICUs, finding theoretical assurance of strength in worst-case adversarial settings, and investigating federated learning methods to allow multi-institutional participation without sacrificing privacy or safety. Combination with causal inference methods would enhance the knowledge of heterogeneity of effects of treatment and create a more individualized recommendation of resource allocation.

6. Conclusions

Our system, the ICU-Transformer, is a fully inclusive machine that comprises both the latest transformer architectures and the best learning systems to prevent data poisoning attacks. ICU-Transformer performs at a high level of performance through several dimensions due to the introduction of new features such as Robust Multi-Head Attention with the

quantification of uncertainty, training-dynamic sample reweighting, and multi-objective training with fairness constraints.

Testing on four massive datasets shows an AUC-ROC of 89.1%, very far exceeding the ranks of traditional methods, with 3.9% degradation to the lowest toleration of 20% data poisoning. ROE minimizes deaths that occur due to preventable causes by 37.6% and advances utilization efficiency and fairness scores. In practice, applications in 200 ICUs with 50,000 annual admissions gave 18.3% mortality reduction, 30.9% reduction in length of stay, 31.2% reduction in unnecessary admissions, as well as a savings of 142 million 200,000 every year in costs.

Security analysis revealed 187 confirmed adversarial attacks in a systematic attempt targeting vulnerable populations, with 95.7% of the assaults countered through the robust learning mechanism. This highlights the real risks posed by adversarial manipulation in deployed healthcare AI systems and demonstrates the effectiveness of the defensive strategies integrated within ICU-Transformer. Although the system performs strongly against known poisoning patterns, future work will explore adaptive defense mechanisms capable of responding to evolving adversarial strategies that differ from attack types encountered during training.

The findings collectively indicate that ICU-Transformer can achieve clinical excellence, operational efficiency, and security robustness in real-world critical care settings. Future directions include enhancing generalization across diverse healthcare environments, improving the theoretical underpinnings of resource allocation models, and strengthening ethical frameworks governing algorithmic decision support.

Funding: This research received no external funding.

Data Availability Statement: The data used to support the findings of this study are publicly available at: https://physionet.org/content/mimiciv/3.1/ (accessed on 11 December 2025); https://physionet.org/content/eicu-crd/2.0/ (accessed on 11 December 2025).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, M.; Zhang, Y.; Liu, S.; Han, Y.; Cao, H.; Qiao, B. Dual-attention transformer-based hybrid network for multi-modal medical image segmentation. *Sci. Rep.* **2024**, *14*, 25704. [CrossRef]
2. Jiang, C.; Wang, Y.; Yuan, Q.; Qu, P.; Li, H. A 3d medical image segmentation network based on gated attention blocks and dual-scale cross-attention mechanism. *Sci. Rep.* **2025**, *15*, 6159. [CrossRef]
3. Alber, D.A.; Yang, Z.; Alyakin, A.; Yang, E.; Rai, S.; Valliani, A.A.; Zhang, J.; Rosenbaum, G.R.; Amend-Thomas, A.K.; Kurland, D.B.; et al. Medical large language models are vulnerable to data-poisoning attacks. *Nat. Med.* **2025**, *31*, 618–626. [CrossRef]
4. Sabrou, S.; Guezzaz, A.; Ravi, V.; Benkirane, S.; Azrour, M. Optimizing icu resource allocation during the COVID-19 crisis: An ai-driven approach. *Biomed. Inform. Smart Healthc.* **2025**, *1*, 9–17.
5. Kim, J.; Kim, G.-H.; Kim, J.-W.; Kim, K.H.; Maeng, J.-Y.; Shin, Y.-G.; Park, S. Transformer-based model for predicting length of stay in intensive care unit in sepsis patients. *Front. Med.* **2025**, *11*, 1473533. [CrossRef] [PubMed]
6. Kumar, P.; Singh, A. Minimal data poisoning attack in federated learning for medical image classification: An attacker perspective. *Artif. Intell. Med.* **2024**, *159*, 103024. [CrossRef] [PubMed]
7. Mozaffari-Kermani, M.; Sur-Kolay, S.; Raghunathan, A.; Jha, N.K. Systematic poisoning attacks on and defenses for machine learning in healthcare. *IEEE J. Biomed. Heal. Inform.* **2015**, *19*, 1893–1905. [CrossRef]
8. Shi, B.; Chen, L.; Pang, S.; Wang, Y.; Wang, S.; Li, F.; Zhao, W.; Guo, P.; Zhang, L.; Fan, C.; et al. Large language models and artificial neural networks for assessing 1-year mortality in patients with myocardial infarction. *J. Med. Internet Res.* **2025**, *27*, e67253. [CrossRef]
9. Madan, S.; Lentzen, M.; Brandt, J.; Rueckert, D.; Hofmann-Apitius, M.; Fröhlich, H. Transformer models in biomedicine. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 308. [CrossRef]
10. AlSaad, R.; Abd-Alrazaq, A.; Boughorbel, S.; Ahmed, A.; Renault, M.-A.; Damseh, R.; Sheikh, J. Multimodal large language models in health care: Applications, challenges, and future outlook. *J. Med. Internet Res.* **2024**, *26*, e59505. [CrossRef]

11. Choi, A.; Lee, K.; Hyun, H.; Kim, K.J.; Ahn, B.; Lee, K.H.; Hahn, S.; Choi, S.Y.; Kim, J.H. A novel deep learning algorithm for real-time prediction of clinical deterioration in the emergency department for a multimodal clinical decision support system. *Sci. Rep.* **2024**, *14*, 30116. [CrossRef]
12. Sridhar, G.; Yarabati, V.; Gumpeny, L. Predicting outcomes using neural networks in the intensive care unit. *World J. Clin. Cases* **2025**, *13*, 100966. [CrossRef]
13. Lim, L.; Gim, U.; Cho, K.; Yoo, D.; Ryu, H.G.; Lee, H.C. Real-time machine learning model to predict short-term mortality in critically ill patients: Development and international validation. *Crit. Care* **2024**, *28*, 86. [CrossRef]
14. Zilker, M.; Wirth, M.; Schmidt, P. Zero shot health trajectory prediction using transformer. *npj Digit. Med.* **2024**, *7*, 245. [CrossRef]
15. Gulsoy, E.K.; Ayas, S.; Kablan, E.B.; Ekinci, M. Enhancing the adversarial robustness in medical image classification: Exploring adversarial machine learning with vision transformers-based models. *Neural Comput. Appl.* **2024**, *37*, 7971–7989. [CrossRef]
16. Javed, H.; El-Sappagh, S.; Abuhmed, T. Robustness in deep learning models for medical diagnostics: Security and adversarial challenges towards robust ai applications. *Artif. Intell. Rev.* **2024**, *57*, 12. [CrossRef]
17. Alshwaheen, A.; Khan, M.; Sherani, A.; Qayyum, M.; Hussain, H. Federated learning in smart healthcare: A comprehensive review on privacy, security, and predictive analytics with iot integration. *Healthcare* **2024**, *12*, 2587. [CrossRef] [PubMed]
18. Ngoupayou Limbepe, Z.; Gai, K.; Yu, J. Blockchain-based privacy-enhancing federated learning in smart healthcare: A survey. *Blockchains* **2025**, *3*, 1. [CrossRef]
19. Zhao, H.; Sui, D.; Wang, Y.; Ma, L.; Wang, L. Privacy-preserving federated learning framework for multi-source electronic health records prognosis prediction. *Sensors* **2025**, *25*, 2374. [CrossRef]
20. Cheng, H.; Qu, Y.; Liu, W.; Gao, L.; Zhu, T. Decentralized federated learning for private smart healthcare: A survey. *Mathematics* **2025**, *13*, 1296. [CrossRef]
21. Singh, J.; Kumar, R.; Patel, S. Privacy-aware hierarchical federated learning in healthcare: Integrating differential privacy and secure multi-party computation. *Futur. Internet* **2025**, *17*, 345. [CrossRef]
22. Mohammadi, F.; Rezaei, A. Privacy-preserving federated learning-based intrusion detection system for ioht devices. *Electronics* **2024**, *14*, 67. [CrossRef]
23. Zhang, J.; Chen, X.; Yang, B.; Guan, Q.; Chen, Q.; Chen, J.; Wu, Q.; Xie, Y.; Xia, Y. Advances in attention mechanisms for medical image segmentation. *Comput. Sci. Rev.* **2025**, *55*, 100721. [CrossRef]
24. Pan, P.; Zhang, C.; Sun, J.; Guo, L. Multi-scale conv-attention u-net for medical image segmentation. *Sci. Rep.* **2025**, *15*, 96101. [CrossRef]
25. Wang, H.; Zhang, Y.; Liu, X. Deep learning on medical image analysis. *CAAI Trans. Intell. Technol.* **2024**, *9*, 2356–2389. [CrossRef]
26. Niu, H.; Omitaomu, O.A.; Langston, M.A.; Olama, M.; Ozmen, O.; Klasky, H.B.; Laurio, A.; Ward, M.; Nebeker, J. Ehrbert: A bert-based model for effective anomaly detection in electronic health records. *J. Biomed. Inform.* **2024**, *150*, 104605. [CrossRef] [PubMed]
27. Zeinali, N.; Albashayreh, A.; Fan, W.; Gilbertson White, S. Symptom-bert: Enhancing cancer symptom detection in ehr clinical notes. *J. Pain Symptom Manag.* **2024**, *68*, 190–198.e1. [CrossRef]
28. Li, Y.; Rao, S.; Solares, J.R.A.; Hassaine, A.; Ramakrishnan, R.; Canoy, D.; Zhu, Y.; Rahimi, K.; Salimi-Khorshidi, G. Behrt: Transformer for electronic health records. *Sci. Rep.* **2020**, *10*, 7155. [CrossRef]
29. Rasmy, L.; Xiang, Y.; Xie, Z.; Tao, C.; Zhi, D. Med-bert: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit. Med.* **2021**, *4*, 86. [CrossRef] [PubMed]
30. Liu, W.; Hu, J.; Lv, F.; Tang, Z. A new method for long-term temperature compensation of structural health monitoring by ultrasonic guided wave. *Measurement* **2025**, *252*, 117310. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.