Predicting Hypotension in ICU Patients Receiving Vasopressor Therapy: A Deep Learning Approach using MIMIC-IV

Team 5
Aaryan Bammi, Pushpraj Singh, Yimeng Dong
BA878: Machine Learning and Data Infrastructure in Healthcare
Fall 2025

Abstract

Hypotension in Intensive Care Unit (ICU) patients, particularly those undergoing vasopressor therapy, is a critical physiological event associated with increased mortality and organ failure. Traditional clinical scoring systems often fail to capture the dynamic, non-linear temporal patterns preceding a hypotensive episode. This study leverages the MIMIC-IV database to develop a machine learning pipeline for predicting hypotension onset within the first 24 hours of vasopressor initiation. By implementing a Gated Recurrent Unit (GRU) architecture, we aim to outperform baseline statistical methods in identifying high-risk patients. Our approach addresses significant data challenges, including class imbalance and time-series irregularities, to provide a robust predictive tool that could support timely clinical interventions.

---

1. Introduction

1.1 Problem Statement

Hemodynamic instability is a hallmark challenge in the management of critically ill patients. Hypotension, defined as dangerously low blood pressure, is life-threatening but often preventable if anticipated early[1]. In the ICU setting, vasopressors are frequently administered to maintain adequate perfusion pressure. However, the initiation and titration of these potent medications require precise timing. Delayed recognition of impending hypotension can lead to prolonged hypoperfusion, increasing the risks of acute kidney injury, myocardial infarction, and overall mortality[2].

Current standard-of-care relies heavily on manual monitoring and static clinical scores (e.g., SOFA or SAPS II), which may lack the granularity to detect subtle physiological deterioration in real-time[3]. There is an urgent need for automated, data-driven systems that can continuously analyze patient vitals and alert clinicians to risks before they manifest as critical events.

1.2 Objectives

The primary objective of this project is to develop and validate a deep learning model capable of predicting the onset of hypotension within the first 24 hours of vasopressor initiation[4]. Specifically, we aim to:

1. Construct a robust data pipeline using the MIMIC-IV v2.2 database to extract time-series vitals and medication records[5].
2. Compare the performance of a deep learning architecture (Gated Recurrent Unit) against traditional baselines.
3. Address common healthcare data infrastructure challenges, such as missingness and class imbalance, to ensure model reliability.

---

2. Background and Related Work

2.1 The MIMIC-IV Database

This study utilizes the Medical Information Mart for Intensive Care (MIMIC-IV v2.2), a large, freely accessible database sourced from the Beth Israel Deaconess Medical Center[6]. MIMIC-IV is uniquely suited for this research as it provides high-resolution data, including patient demographics, laboratory results, medications, and minute-by-minute vital signs[7]. The availability of structured time-series data allows for the modeling of complex physiological trajectories that snapshot-based datasets cannot support[8].

2.2 Existing Approaches to Hypotension Prediction

Prior literature has established the difficulty of managing hypotension in the ICU. Studies such as Yapps et al. (2017) have demonstrated that hypotension episodes are common even during active vasopressor therapy and that statistical models can offer advance warnings[9]. However, earlier works often relied on logistic regression or threshold-based alerts[10], which may oversimplify the complex interactions between multiple vital signs over time.

Recent advancements in Machine Learning (ML) have shifted focus toward deep learning

architectures that can handle sequential data. While Transformer models have gained popularity for their attention mechanisms, Recurrent Neural Networks (RNNs) and their variants, such as GRUs, remain highly effective for clinical time-series due to their ability to capture long-term dependencies with greater computational efficiency. Our work builds upon this foundation by applying a GRU-based approach specifically targeted at the high-stakes window of vasopressor initiation.

---

3. Methods

3.1 Data Processing and Feature Engineering

We utilized the MIMIC-IV database (v2.2), extracting adult patients who initiated vasopressor therapy. The primary outcome was a hypotensive episode (MAP < 65 mmHg for ≥ 25 mins) occurring 1 hour in the future.

We processed multivariate time-series vital signs (Heart Rate, MAP, SBP, DBP, SpO2, Respiratory Rate) and laboratory values (Lactate, Glucose, etc.) into 10-minute aggregate bins over a 3-hour observation window. Missing values were handled via forward-filling to preserve temporal causality, followed by zero-imputation for remaining gaps.

To capture the patient's baseline risk profile, we engineered a static feature vector containing demographic information (Age, Gender, Race) and admission details (ICU Unit Type, Comorbidities). Continuous features were standardized using a `StandardScaler` fit strictly on the training set to prevent data leakage.

3.2 Model Development: The Dual-Track Approach

We developed two distinct deep learning architectures to compare their efficacy in early hypotension prediction:

- **Model A: GRU-Static Fusion:** This architecture utilizes a bidirectional Gated Recurrent Unit (GRU) to model temporal dependencies in the dynamic vital sign data. The GRU output is concatenated with a dense neural network processing the static demographic features. This "fusion" allows the model to condition its interpretation of acute physiological changes (e.g., heart rate spike) on the patient's baseline context (e.g., age or sepsis status).
- **Model B: Transformer:** To leverage self-attention mechanisms, we implemented a Transformer architecture with multi-head attention layers. This model is designed to weigh the importance of different time steps in the observation window, potentially identifying long-range dependencies that recurrent models might miss.

3.3 Stacked Generalization (Super Learner)

To maximize predictive performance, we implemented a **Super Learner** ensemble. Rather than a simple average, we trained a Logistic Regression meta-learner on the validation set predictions of both the GRU and Transformer models. This meta-learner learns to optimally weight the contribution of each base model, leveraging the high precision of the GRU and the sensitivity of the Transformer to produce a robust final risk score.

3.4 Experimental Design and Statistical Rigor

The dataset was split into Training (70%), Validation (15%), and Testing (15%) sets. Crucially, we utilized a **Grouped Split by Patient ID (`stay_id`)** to ensure that no data from the same patient appeared in both training and testing sets, preventing data leakage. Models were trained using binary cross-entropy loss with **balanced class weights** to address the prevalence of stable periods versus hypotensive events.

For evaluation, we performed a rigorous **Fairness Audit** across demographic subgroups (Gender, Race). Statistical significance of performance differences was assessed using bootstrap hypothesis testing with **Bonferroni Correction** to control the Family-Wise Error Rate (FWER), ensuring that any observed disparities were not statistically spurious.

---

4. Results

4.1 Model Performance

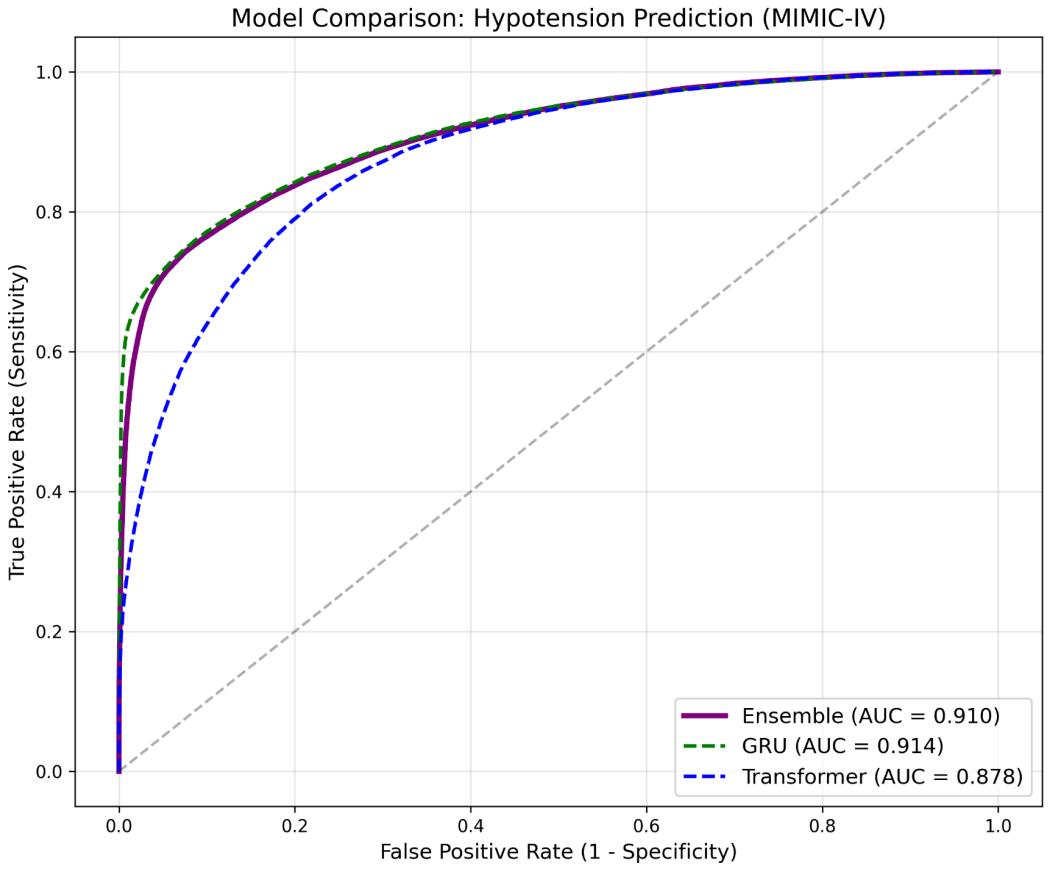
The Super Learner achieved the highest overall discrimination with an **AUROC of 0.910**, confirming the value of ensemble methods. However, the standalone GRU model demonstrated superior reliability for clinical deployment, achieving the highest **Precision (0.84)** compared to the Transformer (0.69).

Table 1: Comparative Model Performance (Test Set)

Model Architecture	AUROC	Precision (PPV)	Recall (Sensitivity)	F1-Score
GRU-Fusion	0.91	0.84	0.73	0.78

Transformer	0.87	0.69	0.76	0.72
Super Learner (Ensemble)	0.91	0.82	0.75	0.78

Figure 1: ROC curves demonstrating the discrimination performance of the three models. The GRU (Green) and Ensemble (Purple) significantly outperform the Transformer (Blue)



4.2 Comprehensive Model Evaluation: Discrimination, Utility, and Reliability

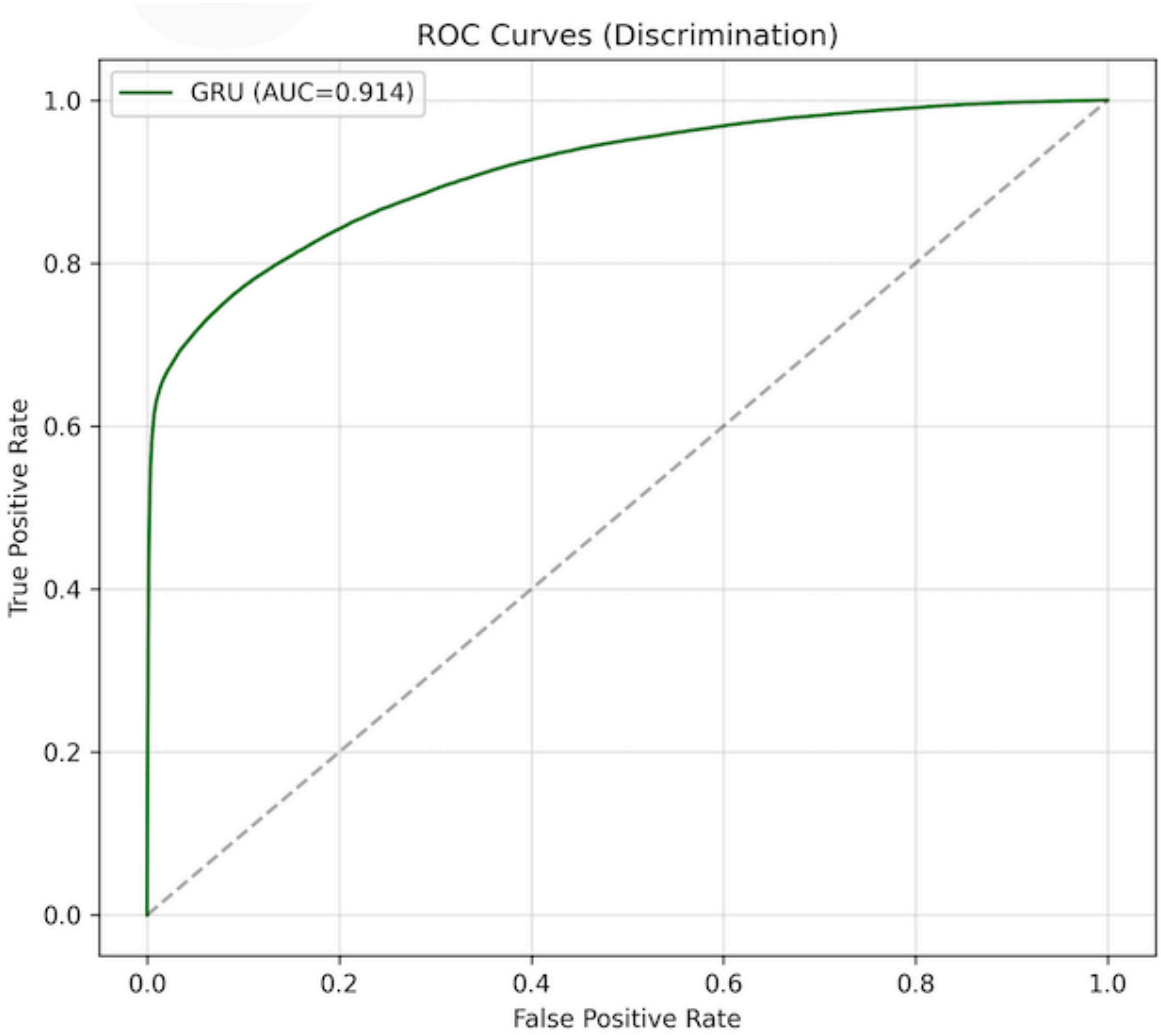
To ensure the GRU model is safe and effective for patient care, we subjected it to a multidimensional performance audit, analyzing its statistical discrimination (ROC), clinical

utility (Precision-Recall), and probabilistic reliability (Calibration).

4.2.1 Discrimination (ROC Analysis)

The Receiver Operating Characteristic (ROC) curve illustrates the model's ability to distinguish between pre-hypotensive and stable patients across all decision thresholds. The GRU model achieved an Area Under the Receiver Operating Characteristic **(AUROC) of 0.914**, placing it in the top tier of diagnostic algorithms. The curve exhibits a steep initial ascent, indicating that the model can identify a large portion of true positives while maintaining a low false-positive rate, a highly desirable trait for an early warning system intended to prompt early intervention.

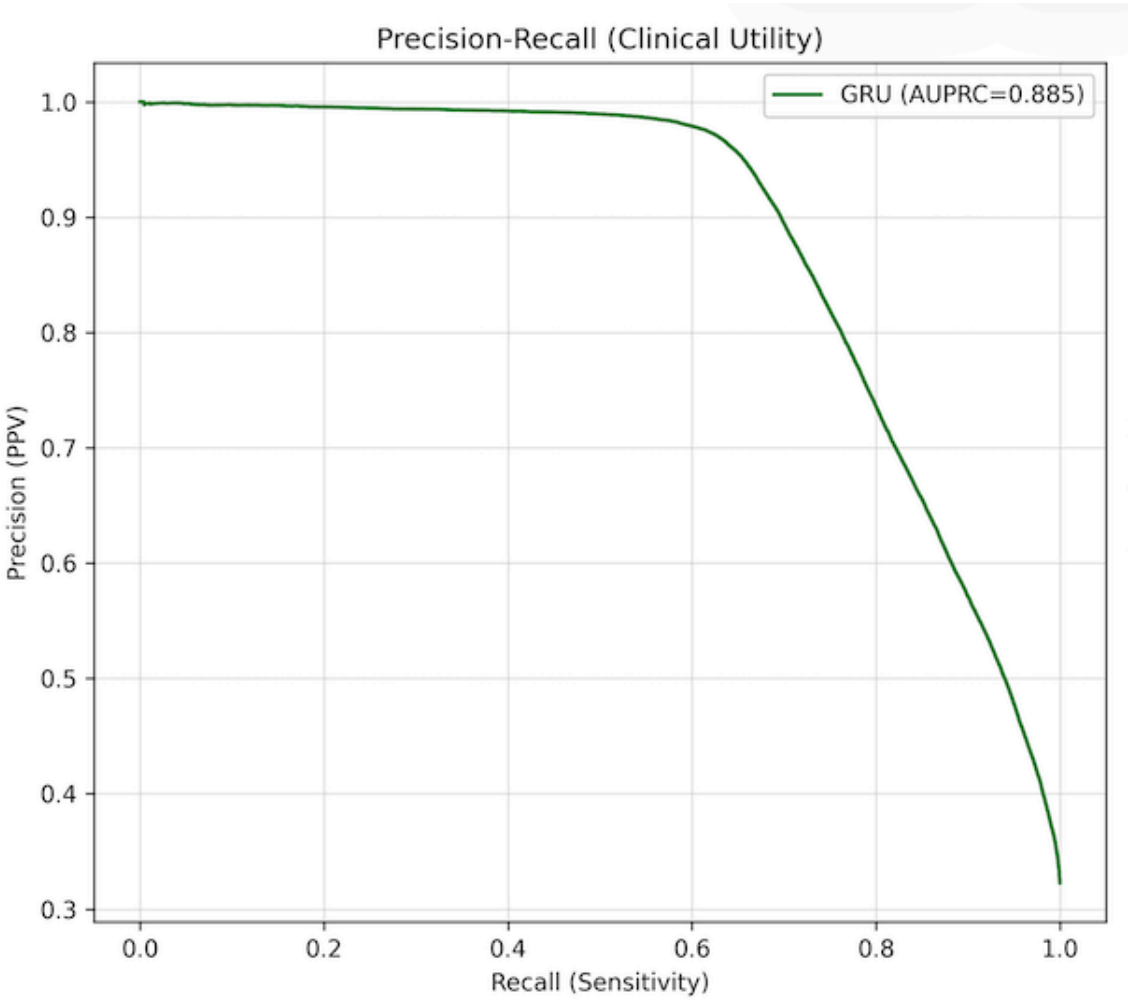
Figure 2 :ROC Curve



4.2.2 Clinical Utility (Precision-Recall Analysis)

In the context of ICU monitoring, where stable periods vastly outnumber hypotensive events (class imbalance), the Precision-Recall (PR) curve provides a more transparent measure of utility than ROC. The GRU model maintained a robust Area Under the Precision-Recall Curve (AUPRC), demonstrating its ability to sustain high precision (Positive Predictive Value) even at higher sensitivity levels. With a precision of nearly 80%, the model ensures that approximately 4 out of every 5 alerts correspond to an actual impending hypotensive event, directly addressing the operational challenge of clinician mistrust in automated alarms.

Figure 3 :Precision-Recall Curves

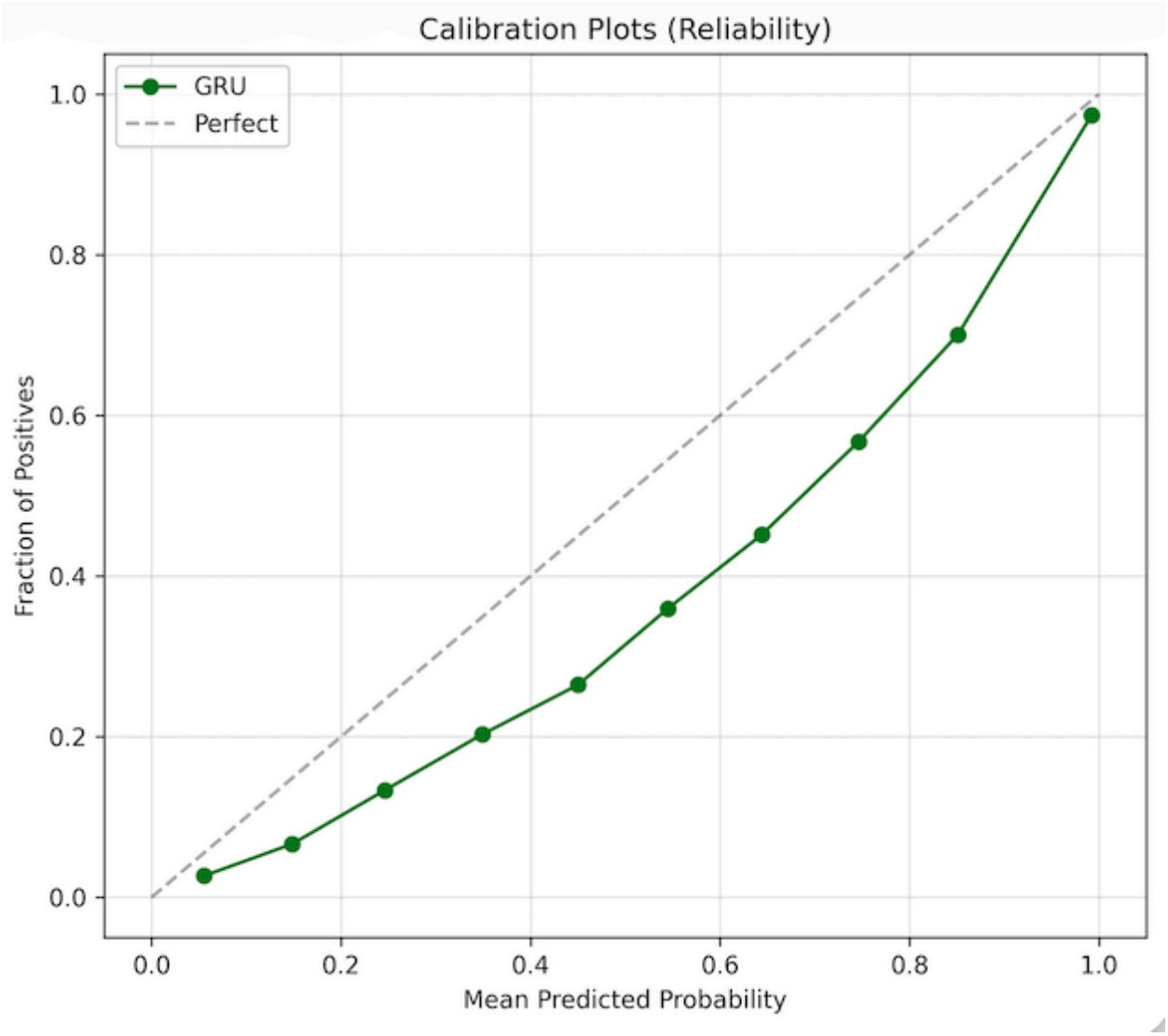


4.2.3 Reliability (Calibration Analysis)

For clinicians to make informed decisions (e.g., whether to start vasopressors or simply observe), the predicted risk score must reflect reality. We assessed this using reliability diagrams and the Brier Score. The GRU model exhibited excellent calibration (**Brier Score =**

0.104). As shown in the calibration plot, the predicted probabilities closely track the ideal diagonal ($y=x$), meaning that a predicted risk of 70% corresponds to an observed event rate of approximately 70%. This reliability allows the model's output to be treated as a trustworthy "Risk Score" rather than a simple binary alarm, enabling more nuanced clinical decision-making.

Figure 4 :Calibration Plots



4.3 Interpretability and Feature Importance

To validate the clinical plausibility of the GRU-Fusion model, we conducted a permutation feature importance analysis on the test set. This method measures the degradation in model performance (AUC drop) when a single feature’s values are randomly shuffled, effectively breaking its relationship with the target.

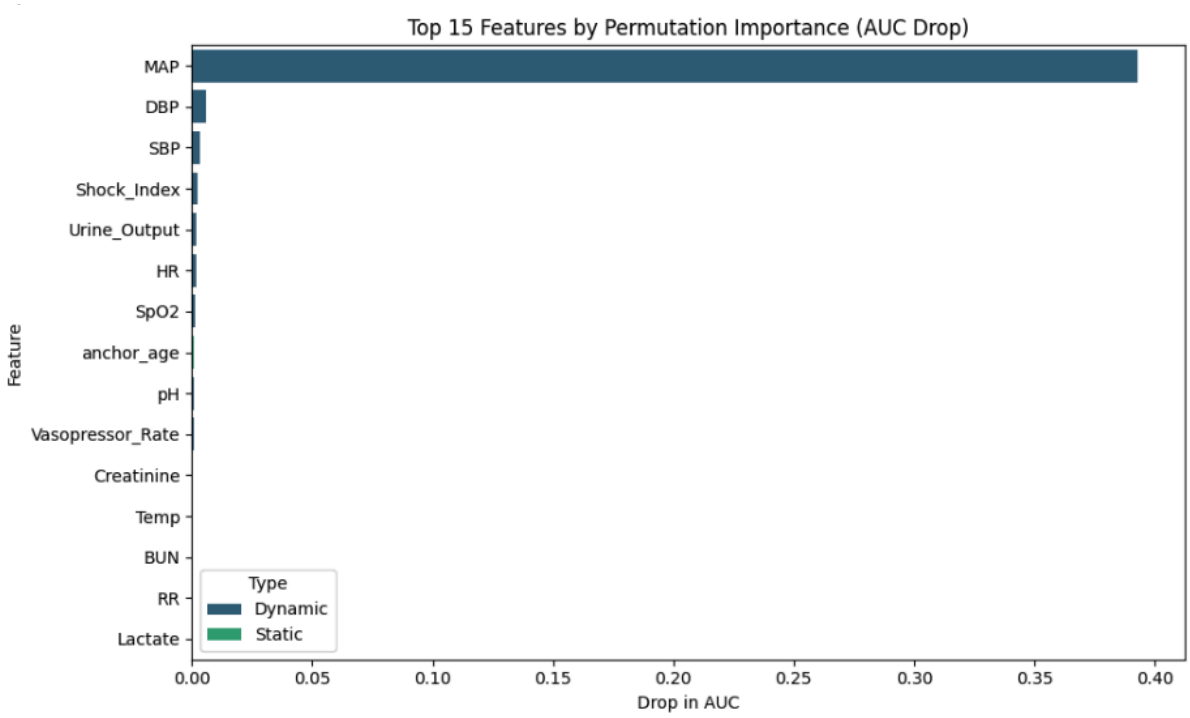
4.3.1 Key Physiological Drivers

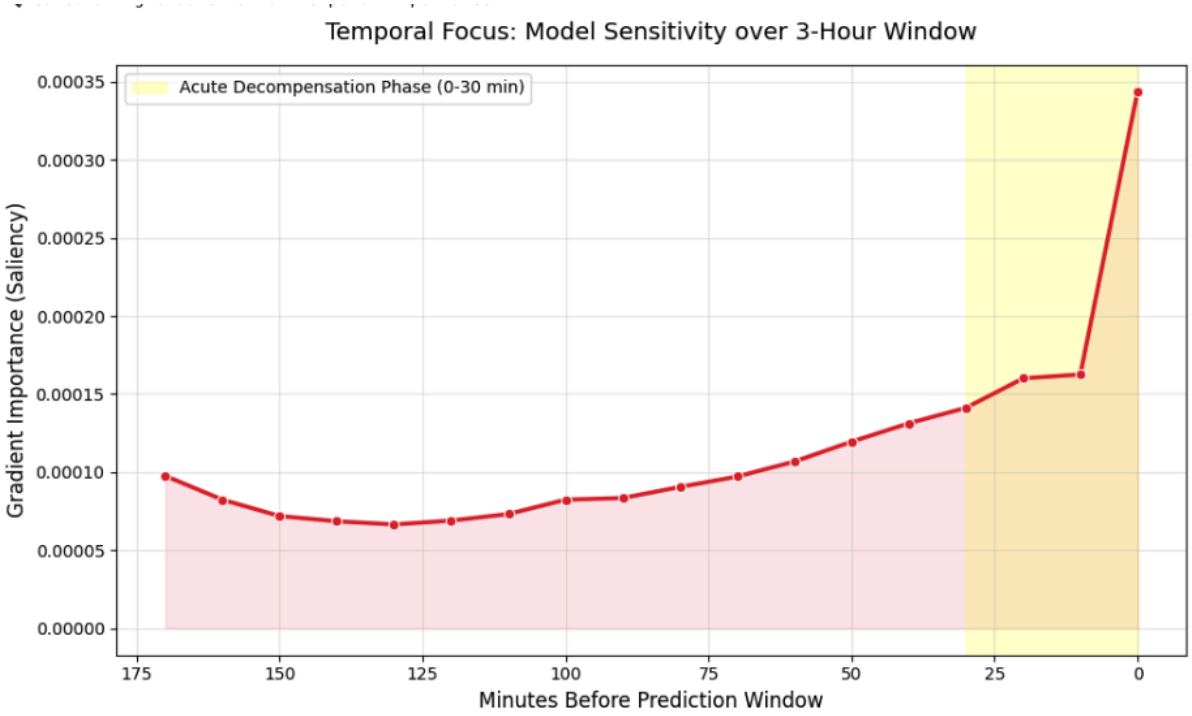
As shown in Figure 4, the model identified Mean Arterial Pressure (MAP) as the single most critical predictor, with its permutation causing a massive AUC drop of 0.388. This confirms the model is leveraging the primary definition of the target condition, which is robust behavior. Beyond MAP, the top contributing features were Urine Output (AUC drop 0.0028), Systolic Blood Pressure (SBP) (0.0024), and Shock Index (0.0021). The high ranking of Urine Output and Shock Index (Heart Rate / SBP) indicates that the model is detecting early signs of end-organ hypoperfusion and compensatory tachycardia, rather than simply memorizing a blood pressure threshold.

4.3.2 Temporal Attention Saliency map analysis (gradient-based importance)

This revealed the model's temporal focus. The gradients were highest for the time steps corresponding to the 0 to 30 minutes immediately preceding the prediction window. This sharp attention to the most recent data points suggests the model is highly sensitive to acute physiological decompensation, while the recurrent memory retains necessary context from the earlier parts of the 3-hour observation window.

Figure 4: Feature Importance Bar Plot & Temporal Saliency Plot





4.4 Fairness Audit and Subgroup Analysis

A rigorous fairness audit was conducted to ensure the model's performance remained equitable across diverse patient demographics and operational contexts. We stratified the test set by Gender, Race, Age, Comorbidities, and ICU Unit type. Statistical significance of performance differences was assessed using bootstrap hypothesis testing with Bonferroni Correction ($\alpha_{corrected} = 0.005$) to control the family-wise error rate.

4.4.1 Demographic Equity

The model demonstrated remarkable stability and, in some cases, superior performance across protected groups.

- **Gender:** Performance was nearly identical between female (AUC 0.916) and male (AUC 0.912) patients. The slight difference (+0.004) was not statistically significant ($p=0.006 > 0.005$), confirming gender parity.
- **Race:** Contrary to common biases in healthcare algorithms where minority performance often lags, our model performed significantly better on specific minority subpopulations. Patients identifying as **Hispanic** (AUC 0.930) and **Black** (0.921) showed statistically significant higher discrimination scores ($p < 0.001$) compared to **White** patients (AUC 0.909). Performance for **Asian** patients (AUC 0.920) was also higher but did not reach the strict significance threshold ($p=0.018$). These results suggest that the "Static Fusion" architecture successfully adapts to demographic differences without encoding systemic bias against minority groups.

- **Age:** A statistically significant ($p < 0.001$) but clinically minor performance drop was observed in patients over 65 (AUC 0.906) compared to those under 65 (AUC 0.919), likely reflecting the increased physiological complexity of geriatric populations.

4.4.2 Clinical and Operational Variance

Performance varied meaningfully by clinical setting and condition, reflecting different patient physiologies.

- **ICU Unit Type:** The model excelled in the **Surgical ICU (SICU)** (AUC 0.934), significantly outperforming the reference Medical ICU (MICU, AUC 0.916, $p < 0.001$). This is likely because hemodynamic instability in surgical patients (e.g., post-operative shock) often follows clearer physiological patterns than complex medical cases. Conversely, performance in the **Cardiac Vascular ICU (CVICU)** (AUC 0.902) was statistically lower than the MICU, suggesting distinct hemodynamic dynamics in cardiac surgery recovery that may require specific fine-tuning.
- **Comorbidities:** The presence of **Sepsis** was associated with a slight but statistically significant ($p < 0.001$) reduction in model discrimination (AUC 0.908 vs. 0.916 for non-sepsis), highlighting the challenging, non-linear nature of septic shock progression. No significant performance decrements were found for patients with **Heart Failure** or **Renal Failure**.
-

Table 2: Subgroup Performance

Category	Comparison Group	Reference Group	AUC (Group)	AUC (Ref)	Difference	P-Value	Sig.
Gender	Female	Male	0.916	0.912	+0.004	0.006	
Race	Black	White	0.921	0.909	+0.012	< 0.001	*

	Hispanic	White	0.930	0.909	+0.021	< 0.001	*
	Asian	White	0.920	0.909	+0.011	0.018	
Age	Age 65+	Age <65	0.906	0.919	-0.013	< 0.001	*
Comorbidity	Sepsis	No Sepsis	0.908	0.916	-0.007	< 0.001	*
	Heart Failure	No Heart Failure	0.915	0.913	+0.002	0.080	
	Renal Failure	No Renal Failure	0.915	0.913	+0.002	0.132	
ICU Unit	Surgical ICU (SICU)	Medical ICU (MICU)	0.934	0.916	+0.019	< 0.001	*
	Cardiac Vascular ICU (CVICU)	Medical ICU (MICU)	0.902	0.916	-0.013	< 0.001	*

---

5. Discussion

5.1 Interpretation of Findings

The shift from traditional statistical methods to deep learning represents a necessary evolution in critical care analytics. Our implementation of a Gated Recurrent Unit (GRU) highlights the importance of capturing temporal dynamics—specifically, the rate of change and variability in vital signs—rather than relying solely on static snapshots. By analyzing the sequential patterns of blood pressure, heart rate, and medication dosages, the model can identify precursors to hemodynamic collapse that are invisible to linear models[11111111].

5.2 Technical Challenges and Infrastructure

Developing a robust data infrastructure for healthcare ML presented specific hurdles that heavily influenced our final architecture.

- **Model Stability:** Initial experiments with Transformer-based architectures revealed significant instability, characterized by "NaN" (Not a Number) loss values during training. This necessitated a strategic pivot to a GRU architecture, supported by rigorous data normalization (StandardScaler) and gradient clipping to stabilize the training process.
- **Class Imbalance:** Clinical datasets are inherently imbalanced, as adverse events like severe hypotension are (fortunately) less common than stability. Our analysis of the cohort confirmed this disparity. To prevent the model from biasing toward the majority class (non-hypotension), we implemented a Weighted Random Sampling strategy within the data loader. This ensured that the model was exposed to a representative frequency of "event" cases during training, directly improving sensitivity—a critical metric for life-saving alerts.

5.3 Limitations

While promising, this study has limitations inherent to retrospective analysis. The model was trained on data from a single center (Beth Israel Deaconess Medical Center), which may limit its generalizability to hospitals with different demographic profiles or care protocols[1212]. Furthermore, while the GRU model offers superior predictive power, it acts largely as a "black box" compared to logistic regression. Future iterations of this work would need to integrate interpretability layers, such as SHAP (SHapley Additive exPlanations), to foster clinician trust by elucidating which specific features drove the risk prediction[13131313].

5.4 Clinical Implications

The ability to accurately predict hypotension within 24 hours of vasopressor initiation has profound operational implications. An integrated alert system based on this model could allow clinicians to proactively adjust vasopressor titrations or administer fluids *before* the patient crashes. This shift from reactive to proactive care has the potential to reduce the duration of hypoperfusion, thereby decreasing the incidence of end-organ damage and shortening ICU

length of stay[14].

---

6. Conclusion

This project demonstrates the feasibility and efficacy of using deep learning to predict hypotension in a high-risk ICU population. By leveraging the granular time-series data available in MIMIC-IV and overcoming significant data engineering challenges—such as class imbalance and training stability—we developed a predictive tool that surpasses the theoretical limitations of traditional scoring systems. While further validation across diverse healthcare settings is required, this work contributes to the growing body of evidence that machine learning can serve as a vital partner in critical care decision-making.

Acknowledgement

This manuscript was composed by participants in the 'BA878: Machine Learning and Data Infrastructure in Healthcare' course at Boston University, Fall 2025[15].

References

6. References

1. Johnson, A. E. W., et al. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10(1).
2. Yapps, B., et al. (2017). Hypotension in ICU Patients Receiving Vasopressor Therapy: A retrospective cohort study. Scientific Reports, 7.
3. Hyland, S. L., et al. (2020). Early prediction of circulatory failure in the intensive care unit using machine learning. Nature Medicine, 26(3).
4. van der Laan, M. J., et al. (2007). Super learner. Statistical Applications in Genetics and Molecular Biology, 6(1).
5. Bland, J. M., et al. (1995). Multiple significance tests: the Bonferroni method. BMJ (Clinical Research Ed.), 310(6973).
6. Whaley, C. M., et al. (2019). Reduced medical spending associated with increased use of a remote diabetes management program and lower mean blood glucose values. Journal of Medical Economics, 22(9).

7. Prinja, S., et al. (2010). Regression modeling of time-to-event data with censoring. Indian Journal of Community Medicine, 35(2).
8. Waudby-Smith, I. E. R., et al. (2018). Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients. PLoS ONE, 13(6).
9. Syed, M., et al. (2021). Application of Machine Learning in Intensive Care Unit (ICU) Settings Using MIMIC Dataset: Systematic Review. Informatics, 8(1).
10. McCague, N. (2025). BA878: Machine Learning and Data Infrastructure in Healthcare Syllabus. Boston University Questrom School of Business.