Orchestrating LLMs to Explain Shock Predictions by DL Model for ICU Care: A Reasoning Scorecard Approach

Bhanu Pratap Singh
University of Texas at Austin
Austin, TX, USA
bps1418@utexas.edu

Abstract

Shock prediction in critical care settings remains a major challenge, requiring both accurate early detection and explainable reasoning to support clinicians. We present a novel pipeline where a Transformer-based deep learning (DL) model predicts shock probability based on ICU vitals and labs, followed by explanation generation using multiple large language models (LLMs) — GPT-4, Gemini 1.5 Pro, and Mistral. To evaluate explanation quality, we introduce an orchestration strategy with DeepSeek R1 to score reasoning outputs on *Transparency*, *Consistency*, *Clarity*, and *Completeness*. Our scorecard results show strengths and weaknesses of each LLM, providing insights for developing more reliable explainable AI in healthcare.

Keywords

Shock Prediction, ICU Care, Explainable AI, Large Language Models, Reasoning Scorecard, Deep Learning, Healthcare AI

ACM Reference Format:

Bhanu Pratap Singh. 2025. Orchestrating LLMs to Explain Shock Predictions by DL Model for ICU Care: A Reasoning Scorecard Approach. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Shock is a life-threatening condition in ICU settings that demands early detection and intervention [5]. Predictive models, especially deep learning-based, have shown potential for identifying shock onset [3]. However, clinical adoption remains hindered by a lack of transparent and trustworthy explanations [1].

We propose a two-stage system: (1) a Transformer-based DL model trained on MIMIC-III data to predict shock probability, and (2) an explainability layer where GPT-4, Gemini 1.5 Pro, and Mistral generate clinical reasoning based on the model's output. We further design a **reasoning scorecard** evaluated by DeepSeek R1 LLM to systematically benchmark the quality of generated explanations.

Contributions:

- A shock prediction model achieving AUC of 0.8226 on ICU patient data.
- An orchestration framework combining GPT-4, Gemini, Mistral for explanation generation.
- A novel Reasoning Scorecard evaluating *Transparency*, *Consistency*, *Clarity*, and *Completeness*.
- A comparative analysis identifying LLM strengths and limitations in medical explainability.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, Washington, DC, USA
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Related Work

Early works such as [5] demonstrated the feasibility of shock prediction using vital signs. Explainability in healthcare has gained traction with methods like SHAP [2] and LIME [4], yet generating coherent clinical reasoning remains challenging [9]. Recent studies [8] have explored LLMs for medical reasoning but lack quantitative scorecard-based evaluations.

3 Methodology

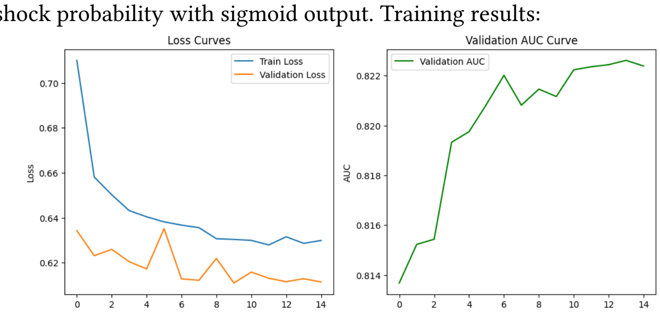
3.1 Data Processing

We extract ICU vitals and lab measurements from MIMIC-III database for the first 12 hours after admission. Records missing essential vitals/labs were filtered to ensure meaningful prediction. Final dataset: **4,957 patients**.

The generated patient summaries [6] and model explanation reasoning dataset [7] have been made publicly available to support reproducibility and further research.

3.2 Transformer Shock Prediction Model

Our DL model projects input features using a linear layer, applies 2-layer Transformer encoding (4 heads, 32-d embedding), and predicts shock probability with sigmoid output. Training results:



Final model achieves **AUC = 0.8226**. Confusion matrix and SHAP global feature importance are shown in Figures 1 and 2.

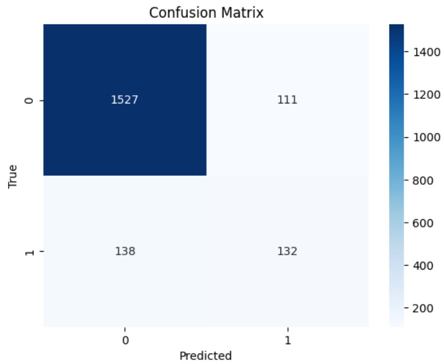


Figure 1: Confusion Matrix of Shock Prediction Model

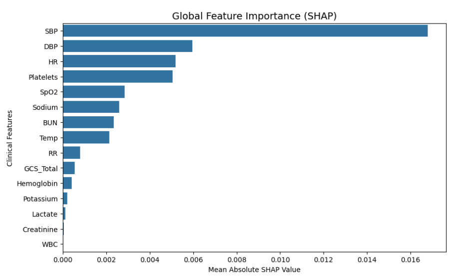


Figure 2: Global Feature Importance using SHAP

Note: SHAP values indicate the impact of each feature on the model's output. Higher absolute SHAP values signify greater influence.

3.3 Reasoning Generation by LLMs

Using top features + model prediction + probability, we generate prompts per patient. Three models: (1) GPT-4 via OpenAI API, (2) Gemini 1.5 Pro API, (3) Mistral 7B locally (Ollama).

3.4 Reasoning Scorecard

Each explanation was scored on 4 axes:

- **Transparency**: Are assumptions, risks discussed?
- **Consistency**: No contradictions or factual errors.
- **Clarity**: Easy for clinician to understand.
- **Completeness**: All key vitals/labs interpreted.

Scores were generated by DeepSeek R1 model prompting a formal rubric.

4 Results

4.1 LLM Performance

Radar plots and bar charts summarize average scores across models (Figures 3, 4, 5).

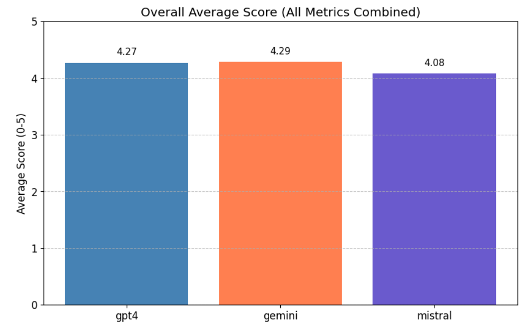


Figure 3: Overall Average Reasoning Score per LLM. Average scores across all axes for each LLM. Higher scores indicate better performance.

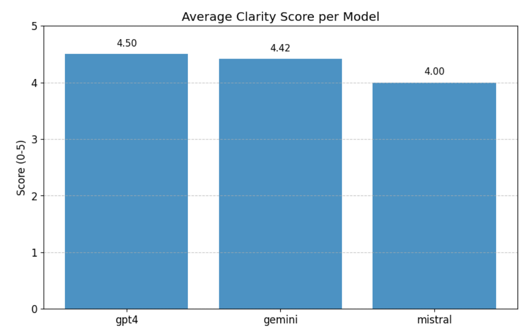


Figure 4: Clarity Score Comparison

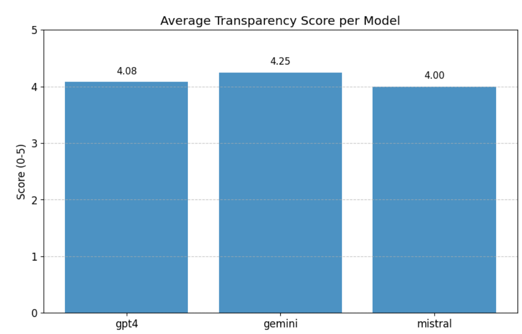


Figure 5: Transparency Score Comparison

4.2 Key Insights

- GPT-4 produced the most *Consistent* explanations.
- Gemini showed high *Transparency* but lower *Clarity*.
- Mistral explanations were more *Concise*, but missed detailed reasoning in some cases.
- DeepSeek R1 orchestration provided robust automated evaluation.

5 Conclusion

This work demonstrates that LLMs, when orchestrated carefully, can effectively generate clinical reasoning for shock prediction in ICU patients. Our scorecard approach surfaces model-specific tendencies and quality gaps. Future work will explore using ensembles of explanations, further fine-tuning LLMs for ICU context, and validating impact with real clinicians.

Acknowledgment

This project was completed as part of "AI in Healthcare" coursework at University of Texas at Austin.

References

[1] David Gunning. 2017. Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)* (2017). DARPA XAI Program.

[2] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*. 4765–4774.

[3] Michael Moor, Bastian Rieck, Max Horn, et al. 2019. SepsisNet: Early detection of sepsis from clinical data using deep learning. In *NeurIPS Workshop on Machine Learning for Health*.

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.

[5] Vivek Shankar et al. 2020. Early Prediction of Hemodynamic Shock in the ICU with Deep Learning on High-Resolution Vital Sign Time Series. *Critical Care Explorations* 2, 10 (2020), e0249.

[6] Bhanu Pratap Singh. 2024. Patient Summary Dataset for Shock Prediction. https://github.com/bps1418/AI_in_Healthcare/blob/master/HighRiskProject/data/patient_summary.csv. Accessed: 2024-05-01.

[7] Bhanu Pratap Singh. 2024. Shock Reasoning Scorecard Dataset. https://github.com/bps1418/AI_in_Healthcare/blob/master/HighRiskProject/data/shock_reasoning_scorecard.csv. Accessed: 2024-05-01.

[8] Youzhi Yang et al. 2023. Evaluations and Benchmarks for LLMs in Medicine: A Survey. *arXiv preprint arXiv:2306.00983* (2023).

[9] Haoyang Zhang et al. 2023. Can Large Language Models Be Good Clinical Reasoners? *arXiv preprint arXiv:2305.09617* (2023).