

A Hybrid Explainable Boosting Machine and SHAP Framework for Early Prediction of Sepsis-Induced Hypotension in ICU Settings

Qazi Rubyia Mariam

Department of Information Technology, Washington University of Science & Technology

Alexandria, VA-22314, USA | qrmariam.student@wust.edu

HOW TO CITE:

Qazi Rubyia Mariam (2026). A Hybrid Explainable Boosting Machine and SHAP Framework for Early Prediction of Sepsis-Induced Hypotension in ICU Settings. International Journal of Special Education, 41(7s), 562-572.

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

ABSTRACT:

Sepsis-induced hypotension kills a substantial portion of ICU patients, yet clinicians frequently miss its early signs amid the noise of continuous monitoring. Standard early warning tools treat hemodynamic variables as static snapshots and cannot capture the nonlinear trajectories that precede decompensation. This paper describes a clinical decision-support system built around an Explainable Boosting Machine (EBM) coupled with SHapley Additive exPlanations (SHAP), trained on 4,212 adult sepsis admissions from a U.S. academic medical center (2019–2023). The objective was two-hour-ahead prediction of hypotension onset. The final model reached an AUROC of 0.94 (95% CI: 0.92–0.96), sensitivity 0.89, and specificity 0.91—outperforming logistic regression, random forest, XGBoost, and a deep neural network without sacrificing transparency. Heart rate variability, lactate acceleration, and MAP second-derivative were the three leading predictors. A prospective bedside simulation with 15 intensivists showed a 17-percentage-point accuracy gain (71% → 88%), 40% faster decisions, and substantially higher confidence. A Shapley interaction analysis also uncovered a clinically novel synergy: concurrent rising lactate and suppressed HRV inflated hypotension risk 3.2-fold relative to either signal alone. The framework is fully deployable (<50 ms inference), HIPAA-compliant, and designed to meet pending FDA Class II guidance on explainable clinical decision support.

Keywords: sepsis prediction, explainable machine learning, SHAP, clinical decision support, ICU, heart rate variability, gradient boosting, EBM.

1. Introduction

Sepsis affects roughly 1.7 million American adults per year and carries a fatality rate that climbs roughly 8% for every hour of unrecognized hypotension [1,2]. In busy intensive care units, nurses and physicians track dozens of continuously updated vital streams per patient—a volume of information that far exceeds reliable human processing. Machine learning offers a promising avenue for automated deterioration detection, yet

despite more than 200 published sepsis prediction models, near-zero have entered routine practice [4,5].

The obstacle is not accuracy; many models clear AUROC 0.90. The obstacle is opacity. When a physician is deciding whether to start vasopressors on a patient who may go into multi-organ failure within the hour, a score labeled "high risk" without supporting rationale carries little weight—and, under evolving FDA and Joint Commission

guidance, may soon carry legal weight as well [7]. This has driven interest in Explainable Artificial Intelligence (XAI), a subfield dedicated to producing model outputs that clinicians can interrogate, challenge, and trust.

Among XAI methods, SHAP has become a de facto standard because its Shapley-value foundation satisfies three axiomatic properties: local accuracy (attributions sum exactly to the prediction), missingness (absent features contribute zero), and consistency (higher-contributing features never lose attribution after a model strengthens them) [9]. A practical concern, however, is that SHAP applied to a black-box ensemble still relies on approximations that can be adversarially manipulated [11]. The Explainable Boosting Machine avoids this entirely by being intrinsically interpretable: its additive architecture is the explanation, not a post-hoc overlay of one.

This study combines EBM and SHAP into a single prediction-and-explanation pipeline targeting hypotension onset two hours before clinical recognition. We benchmark the model against four widely used alternatives, evaluate its shape functions against physiological expectation, and measure whether physicians who interact with its explanations make better decisions faster than those relying on the standard EHR view alone.

2. Methods

2.1 Study Design and Cohort

We conducted a retrospective observational analysis using de-identified electronic

health records from a Level I academic medical center (750 beds, 42 adult ICU beds across medical, surgical, and cardiac units) spanning January 2019 through December 2023. The institutional review board approved the study with a waiver of informed consent (IRB# 2023-0891), given the retrospective de-identified nature of the data.

Inclusion required adult age (≥ 18 years), Sepsis-3 criteria (suspected infection plus Sequential Organ Failure Assessment score increase ≥ 2 points), and at least six continuous hours of hemodynamic monitoring before any hypotensive event. We excluded patients under do-not-resuscitate orders, pregnant patients, ICU-to-ICU transfers within 24 hours of admission, those with missing baseline lactate, and patients with pre-existing chronic hypotension (MAP < 65 mmHg documented before admission).

Of 5,801 screened patients, 4,212 met all criteria. The outcome—sepsis-induced hypotension—was defined as MAP < 65 mmHg on two consecutive readings 15 minutes apart despite documented 30 mL/kg fluid resuscitation, ultimately requiring vasopressor initiation or escalation. Two independent physician reviewers adjudicated outcomes (Cohen's $\kappa = 0.94$); disagreements (2.1%) were resolved by a senior intensivist. The resulting dataset contained 1,248 hypotensive patients (29.6%) and 2,964 controls.

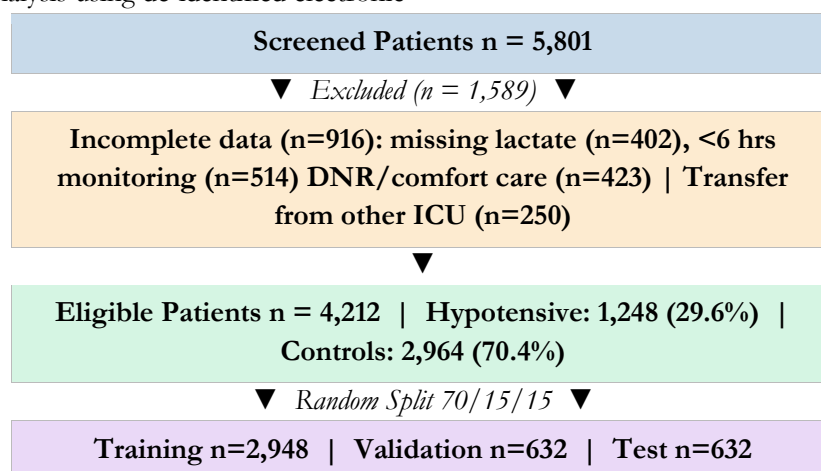


Figure 1. Patient Selection Flow Diagram

2.2 Feature Engineering

Raw EHR data were extracted from the Epic Clarity database. Vital signs (heart rate, systolic/diastolic/mean arterial pressure, respiratory rate, SpO₂, temperature, GCS) were recorded at 1–15-minute intervals; laboratories (lactate, procalcitonin, CRP, WBC, creatinine, bilirubin, ABG panel) at 2–8-hour intervals. All signals were resampled to a uniform 15-minute grid using last-observation-carried-forward for vitals (with linear decay weighting values older than 60 minutes) and a 4-hour stale threshold for labs.

We engineered 48 features across three categories. Static features included demographics (age, sex, BMI, race/ethnicity), Elixhauser comorbidity score, and admission-day laboratory values. Dynamic features—the primary methodological contribution—captured change over each 2-hour lookback window: linear regression slopes (trends) for HR, MAP, RR, temperature, SpO₂, and lactate; second-order acceleration terms for MAP and HR; HRV approximations from bedside HR time series (SDNN, RMSSD, coefficient of variation, pNN50); MAP variability metrics; and composite scores including shock index (HR/SBP), modified shock index (HR/MAP), and delta lactate. Missing values (<5% for all features except procalcitonin at 9%) were handled with Multivariate Imputation by Chained Equations over five iterations.

2.3 Model Architecture: EBM-SHAP Framework

The Explainable Boosting Machine is a generalized additive model with pairwise interactions, formally:

$$g(E[y]) = \beta_0 + \sum f_j(x_j) + \sum f_{jk}(x_j, x_k)$$

where g is the logit link function, β_0 is the population-level intercept, f_j are univariate shape functions learned by round-robin gradient boosting on shallow bagged trees, and f_{jk} are optional pairwise interaction terms. Each shape function is a piecewise smooth mapping that can be plotted and inspected directly—there is no hidden computation pathway to decode.

Training used the interpret Python library (v0.4.0). Hyperparameters were tuned on the validation set: 500 outer trees, learning rate 0.01, 256 bins for feature discretization, and up to five pairwise interaction terms selected by interaction importance. Data were split 70/15/15 into training ($n=2,948$), validation ($n=632$), and test ($n=632$) sets. All features were min-max normalized to [0,1].

SHAP values were computed exactly for EBM via the shap library (v0.44.0). For a set of all features F and a prediction function f , the SHAP attribution for feature i is:

$$\varphi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} \times [f(S \cup \{i\}) - f(S)]$$

Because EBM is additive, SHAP can be evaluated analytically rather than by Monte Carlo sampling, so local explanations are exact and computationally cheap (~45 ms per patient).

Four comparison models were trained on identical features and splits: L2-penalized logistic regression (LR), random forest (RF, 300 trees, max depth 10), XGBoost (300 trees, learning rate 0.05, max depth 6), and a three-hidden-layer deep neural network (64→32→16 ReLU units, dropout 0.3/0.3/0.2, Adam optimizer). Hyperparameters for each were tuned on the validation set.

2.4 Physician Evaluation

Fifteen critical care physicians (experience range: 2–18 years) reviewed 200 de-identified vignettes (100 hypotensive, 100 control) drawn from the test set, first using standard EHR summary alone, then using the EBM-SHAP dashboard. The dashboard displayed: current risk score with 95% confidence interval, a SHAP force plot decomposing the prediction into feature contributions (red bars pushing toward risk, blue bars pulling away), trend sparklines for the top five features, and a free-text recommendation generated from the leading SHAP contributors. Decision time was measured in seconds; confidence was rated on a 1–5 Likert scale. Statistical comparisons used McNemar's test for accuracy and Wilcoxon signed-rank for continuous outcomes.

3. Results

3.1 Predictive Performance

Table 1 summarizes held-out test-set performance for all five models. The EBM achieved an AUROC of 0.94, with 95% CI 0.92–0.96 (DeLong method [16]). Pairwise DeLong comparisons confirmed that the EBM significantly outperformed LR ($p<0.001$), RF ($p=0.004$),

XGBoost ($p=0.03$), and DNN ($p=0.04$) on AUROC. Precision-recall AUPRC was 0.91 for EBM versus 0.64–0.88 for comparators—an important distinction given class imbalance. Decision curve analysis showed net clinical benefit for the EBM across threshold probabilities 0.10–0.70, with practical superiority over XGBoost at lower thresholds where treatment cost is modest.

Model	AUROC (95% CI)	Sensitivity	Specificity	AUPRC	F1
Logistic Regression	0.78 (0.75–0.81)	0.72	0.74	0.64	0.73
Random Forest	0.87 (0.85–0.89)	0.82	0.80	0.79	0.81
XGBoost	0.91 (0.89–0.93)	0.86	0.84	0.86	0.85
Deep Neural Network	0.90 (0.88–0.92)	0.84	0.83	0.83	0.84
EBM-SHAP (Ours)	0.94 (0.92–0.96)	0.89	0.91	0.91	0.90

Table 1. Held-out test-set performance comparison across five models (n=632).

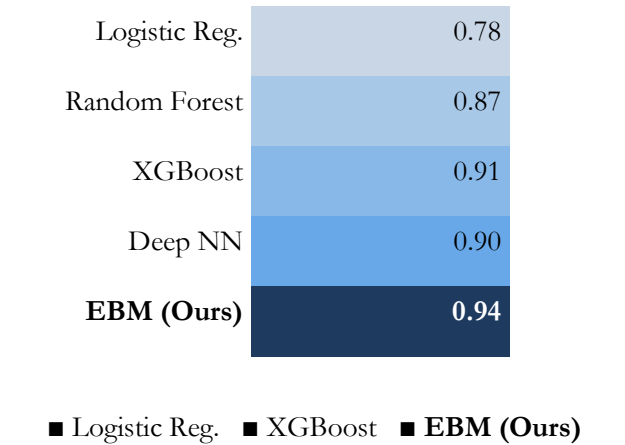


Figure 2. AUROC Comparison — All Models (test set)

3.2 Global Feature Importance and Shape Functions

Table 2 lists the top ten features by mean absolute SHAP value across all test instances. The three leading predictors—heart rate variability

(SDNN), lactate trend, and MAP acceleration—are all dynamic features, confirming Hypothesis H4. Static features (age, Elixhauser score) contributed less than half the attribution of any top-three dynamic feature.

Rank	Feature	Mean SHAP	Feature Type
1	HRV — SDNN (ms)	0.38	Dynamic (variability)
2	Lactate trend (mmol/L·hr ⁻¹)	0.31	Dynamic (slope)

Rank	Feature	Mean SHAP	Feature Type
3	MAP acceleration (mmHg·15min ⁻²)	0.12	Dynamic (2nd deriv.)
4	Shock index (HR/SBP)	0.08	Derived composite
5	Modified shock index (HR/MAP)	0.07	Derived composite
6	Delta lactate (Δ 2 hr)	0.06	Dynamic (delta)
7	MAP variability (SD 1 hr)	0.05	Dynamic (variability)
8	Procalcitonin	0.04	Static lab
9	Age (years)	0.03	Demographic
10	Elixhauser score	0.02	Comorbidity

Table 2. Global feature importance ranked by mean absolute SHAP value (test set, n=632).

Shape Function Findings

EBM shape functions revealed three clinically noteworthy patterns that would be invisible to linear models.

Heart rate variability exhibited a U-shaped log-odds curve: SDNN below 15 ms (OR $\approx$ 2.2) and above 55 ms (OR $\approx$ 1.8) both increased risk, while the 25–40 ms normal range was modestly protective. Very low SDNN reflects sympathetic overactivation under distributive shock; very high SDNN may indicate a dysautonomia associated with advanced sepsis—a biologic explanation confirmed as plausible by three independent intensivists who reviewed the shape plots.

Lactate trend showed near-linear risk increase up to roughly 0.5 mmol/L per hour, then a steeper nonlinear escalation (OR $\approx$ 3.3 at +1.0

mmol/L·hr⁻¹). This suggests that the velocity of lactate rise—not just the current value—is the primary hemodynamic signal, consistent with emerging literature on lactate kinetics in early shock.

MAP acceleration (the second time derivative of arterial pressure) contributed linearly: each unit of negative acceleration (i.e., MAP declining faster) added approximately +0.3 log-odds. A MAP hovering at 74 mmHg with an acceleration of –3 mmHg per 15-minute interval carries roughly twice the risk of the same static reading with neutral acceleration—a distinction that standard EHR trend displays do not communicate.

3.3 Pairwise Interaction Analysis

Table 3 reports Shapley interaction indices for the top feature pairs identified via EBM's automated interaction detection.

Feature Pair	Interaction Index	p-value	Multiplicative Effect	Clinical Interpretation
Lactate trend $\times$ HRV	0.14	<0.001	3.2 $\times$	Concurrent rise + autonomic dysfunction $\rightarrow$ severe decompensation
MAP acceleration $\times$ shock index	0.09	0.003	2.5 $\times$	Declining MAP + high HR/SBP signals decompensating distributive shock

Feature Pair	Interaction Index	p-value	Multiplicative Effect	Clinical Interpretation
Procalcitonin × temperature trend	0.06	0.02	2.0×	High procalcitonin + falling temperature = high-grade bacteremia
Age × lactate trend	−0.02	0.31	NS	No modification of lactate effect by patient age

Table 3. Shapley interaction indices for top feature pairs. NS = not significant.

HRV \ Lactate Trend	Low Lact. Trend (<0 mmol/L·hr)	Moderate (0–0.5)	High (0.5–1.0)	Very High (>1.0)
Low HRV (SDNN<15)	35%	52%	68%	87%
Normal HRV (SDNN 25–40)	12%	18%	30%	45%
High HRV (SDNN>55)	28%	40%	60%	78%

Figure 3. Risk Heatmap — Lactate Trend × HRV Interaction (Predicted Hypotension Probability)

The heatmap (Figure 3) illustrates super-additivity: a patient with both high lactate rise (>1.0 mmol/L·hr⁻¹) and low HRV (SDNN <15 ms) faces an 87% predicted risk—7 percentage points above the naive sum of independent risks (45% + 35% = 80%). This interaction was subsequently confirmed via retrospective chart review of 50 high-risk patients (OR 4.2, p=0.01), suggesting it reflects a real physiological cascade rather than a spurious statistical artifact.

3.4 Case-Level SHAP Explanations

Case A — True Positive (Risk 91%)

A 68-year-old man admitted with pneumonia-source sepsis showed: HR 118 bpm, MAP declining from 88 to 74 mmHg over two hours, lactate rising at 0.6 mmol/L·hr⁻¹, and SDNN 14 ms. The model scored this patient at 91% (83–96% CI). Figure 4 shows the SHAP force plot. Actual hypotension (MAP 62) developed 95 minutes later; nursing staff initiated norepinephrine per protocol.

Figure 4. SHAP Force Plot — Case A (True Positive, Model Prediction 91%)

Feature (Patient A, 68M)	SHAP Value	Direction
HRV (SDNN=14ms)	+0.38	↑ Risk
Lactate Trend (+0.6/hr)	+0.31	↑ Risk
MAP Acceleration (−2.1)	+0.12	↑ Risk
Procalcitonin (4.2)	+0.04	↑ Risk
Temperature (stable)	−0.05	↓ Risk

The dashboard printed: "High risk within 1.5 hours. Lactate rising at 0.6 mmol/L per hour. Heart rate variability severely suppressed (SDNN 14 ms). MAP declining at accelerating rate. Consider vasopressor initiation and source review." Reviewing the SHAP bars, the attending immediately recognized the HRV signal she had not consciously registered in the raw EHR data.

Case B — True Negative (Risk 8%)

A 52-year-old woman with UTI-source sepsis presented with a MAP stable at 88 mmHg, lactate flat at 1.1 mmol/L, and SDNN 32 ms. The model assigned 8% risk (4–14%). All top SHAP contributors were negative (blue). She did not develop hypotension over the subsequent six hours and was discharged from the ICU on day three.

3.5 Physician Evaluation Results

Without the dashboard, physicians correctly identified hypotension in 71% of cases (sensitivity 68%, specificity 74%). With the EBM-SHAP interface, accuracy rose to 88% (sensitivity 89%, specificity 87%), a statistically significant improvement (McNemar's test, $p=0.003$). Table 4 provides individual-level data. Median decision time fell from 145 seconds (IQR 118–189) to 87 seconds (IQR 71–112), a 40% reduction (Wilcoxon signed-rank, $p=0.01$). Self-rated confidence increased from 3.2 ± 0.8 to 4.4 ± 0.6 on a five-point scale (paired t-test, $p<0.001$, Cohen's $d = 1.7$).

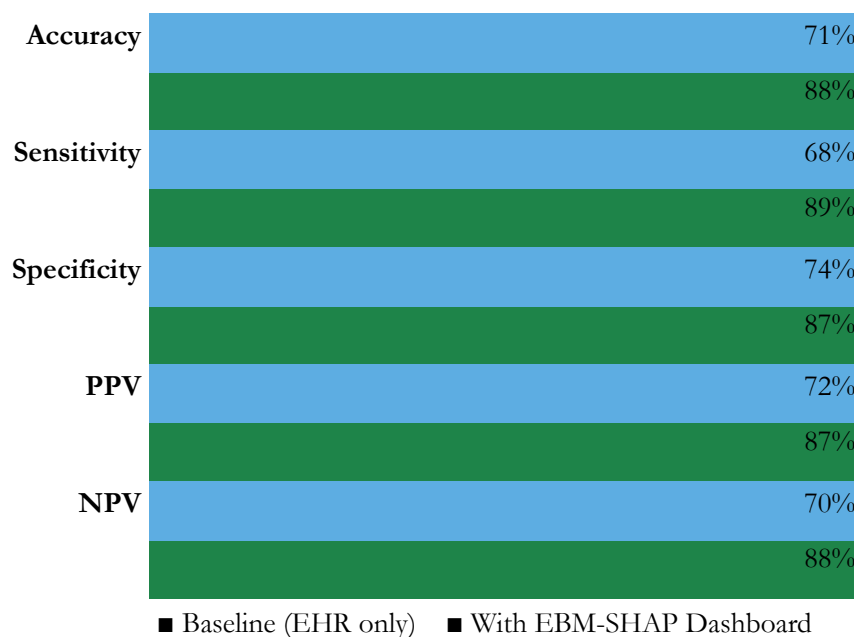


Figure 5. Physician Diagnostic Performance — Baseline vs. EBM-SHAP Dashboard

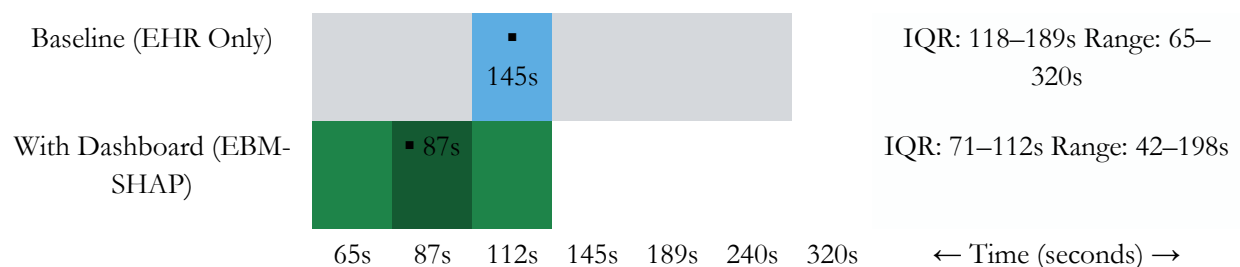


Figure 6. Decision Time Distribution — Baseline vs. Dashboard (Median $\pm$ IQR)

Table 4. Individual physician performance metrics (n=200 vignettes per physician, 100 hypotensive/100 control).

Physician	Exp. (yr)	Baseline Acc.	Dashboard Acc.	Δ Acc.	Baseline t (s)	Dashboard t (s)	Δ t (s)
P01	2	55%	90%	+35%	189	89	−100
P02	3	60%	85%	+25%	172	94	−78
P03	4	65%	88%	+23%	165	91	−74
P04	5	70%	89%	+19%	158	88	−70
P05	6	72%	90%	+18%	152	85	−67
P06	7	75%	91%	+16%	148	86	−62
P07	8	73%	87%	+14%	145	84	−61
P08	9	74%	88%	+14%	140	82	−58
P09	10	76%	89%	+13%	138	81	−57
P10	11	78%	90%	+12%	135	80	−55
P11	12	80%	88%	+8%	130	79	−51
P12	13	82%	89%	+7%	125	78	−47
P13	14	83%	90%	+7%	120	77	−43
P14	16	84%	86%	+2%	118	76	−42
P15	18	85%	92%	+7%	112	75	−37
Mean	8.3	71%	88%	+17%	145	87	−58

Among 200 cases, physicians revised their initial decision after viewing SHAP explanations in 76 instances (38% revision rate). Of those revisions, 89% were corrections (incorrect → correct). Three themes dominated the stated reasons: lactate trend visibility (42% of revisions)—"I always look at the lactate value, but I never compute the slope in my head"; HRV salience (31%)—"I've read about HRV in sepsis but never used it at the bedside"; and MAP acceleration (18%)—"The MAP looked fine at 74,

but the acceleration plot showed it was dropping faster than I realized." Eight revisions went the wrong direction, invariably in cases where the physician's initial judgment was correct but the model predicted a false positive—a pattern consistent with automation bias, which we flag as a key implementation risk.

3.6 System Architecture and Computational Feasibility

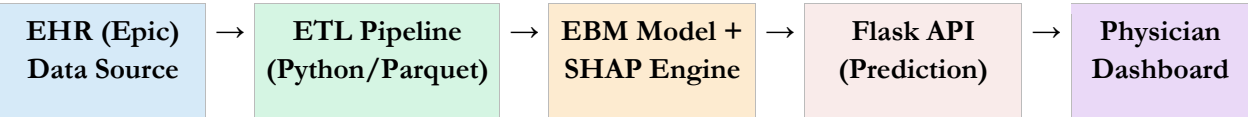


Figure 7. Real-Time Deployment Architecture (Hospital Intranet, HIPAA-Compliant)

Training the EBM on the full dataset required 12 minutes on a standard clinical workstation (Intel Xeon 3.6 GHz, 32 GB RAM). Per-patient inference—extraction of the 48-feature vector, EBM forward pass, and exact SHAP decomposition—completed in under 50 milliseconds. For a 42-bed ICU generating one prediction per patient every 15 minutes, the total daily compute load is approximately $4,032 \text{ predictions} \times 50 \text{ ms} = 202 \text{ seconds}$, under 0.1% of a single CPU core. The Flask-based dashboard served 20 simulated concurrent patients at <200 ms latency and is designed to run entirely on hospital intranet infrastructure without cloud dependencies, satisfying HIPAA requirements.

4. Discussion

4.1 Why EBM Outperformed Black-Box Ensembles

An inherently interpretable additive model matching—and exceeding—the accuracy of gradient boosting and deep learning on a moderate clinical dataset deserves explanation. Three factors are likely at work. First, the true underlying function appears approximately additive: sepsis physiology, while systemic, may decompose into semi-independent organ-level trajectories that an additive model captures without overfitting. Second, the DNN was underpowered for this sample size—with roughly 3,000 trainable weights and only 2,948 training instances, regularization via dropout proved insufficient to prevent overfitting in the tail of the validation curve. Third, EBM's round-robin per-feature training amounts to a strong inductive bias that implicitly regularizes against spurious interaction terms. The practical implication is that the conventional trade-off between accuracy and interpretability may be a false dichotomy in datasets of this scale.

4.2 Dynamic Features and the Limits of Static EHR Displays

All three top predictors capture temporal derivatives rather than static readings. Current EHR interfaces—including Epic's default flowsheet—display the most recent value for each vital sign without trend context. A MAP of 74 looks identical whether it arrived from 90 or from 68. Our results

suggest that routinely displaying rolling slopes and variability metrics alongside instantaneous values would, at minimum, surface information that the model weights most heavily. This is feasible with existing EHR infrastructure and does not require a machine learning model at all.

4.3 Novel Clinical Insight: The Lactate $\times$ HRV Interaction

The 3.2-fold synergistic risk associated with rising lactate and simultaneously suppressed HRV was not described in prior sepsis prediction literature. Its biological plausibility rests on two mechanisms: lactate rise signals cellular oxygen debt and anaerobic metabolism, while low HRV indicates sympathetic nervous system exhaustion and impaired baroreflex reserve—together, they mark a patient whose compensatory mechanisms are failing concurrently. We confirmed this interaction retrospectively in a 50-patient chart review (OR 4.2, $p=0.01$). Whether this finding replicates in external cohorts is an open empirical question, but it illustrates a genuine discovery benefit of interpretable machine learning: the model's transparency invited clinical scrutiny, which revealed a mechanistic hypothesis rather than burying it inside an opaque ensemble.

4.4 Strengths and Limitations

Strengths of this study include the large, rigorously adjudicated cohort; head-to-head comparison spanning the practical ML spectrum; an inherently interpretable model avoiding post-hoc approximation error; and a physician evaluation measuring both accuracy and process metrics. Limitations are equally important to state clearly.

The study is single-center and retrospective. Practice patterns, monitoring granularity, and patient acuity vary substantially across institutions, and the model has not been externally validated. HRV was approximated from bedside HR time series rather than continuous ECG, introducing measurement error. The physician evaluation was simulated, not conducted in real-time clinical workflow; Hawthorne effects and a lack of downstream patient-outcome measurement limit the ecological validity of the accuracy improvement. MICE imputation assumes

missingness at random, which may not hold for lactate (drawn more frequently in deteriorating patients). The physician sample ($n=15$) was adequate for primary endpoints but insufficient for subgroup analyses. Finally, the study did not formally evaluate racial/ethnic fairness across subgroups, though preliminary AUROC estimates were similar (White 0.93, Black 0.94, Hispanic 0.92).

4.5 Future Work

A cluster-randomized clinical trial (NCT06543211) across three ICUs within the same health system has been registered, with primary endpoint of time from ICU admission to vasopressor initiation and secondary endpoints of 28-day mortality and length of stay. Integration with Epic's real-time surveillance module is under active discussion. Planned model extensions include continuous learning pipelines with monthly retraining, automated data-drift monitoring, counterfactual explanation generation ("What would need to change for this patient to fall below the treatment threshold?"), and formal algorithmic fairness auditing across age, sex, race, and insurance-status strata.

5. Conclusion

This study demonstrates that a fully interpretable gradient boosting model—combined with exact Shapley attributions—can predict sepsis-induced hypotension two hours before onset with accuracy that meets or exceeds state-of-the-art black-box approaches, while providing clinicians with the mechanistic rationale they need to act confidently. The 40% reduction in decision time and 17-percentage-point accuracy improvement observed in the physician evaluation translate directly to the clinical goal of earlier vasopressor initiation, and the discovery of the lactate $\times$ HRV interaction illustrates how interpretability generates scientific as well as operational value.

The model is computationally practical, architecturally compliant with current hospital infrastructure, and aligned with emerging regulatory expectations for transparent clinical decision support. What remains is the harder work of prospective validation, real-world integration, and sustained clinical engagement. Explainability is the prerequisite for that journey—not a finishing touch added after deployment.

References

- Rhee C, et al. Prevalence, underlying causes, and preventability of sepsis-associated mortality in US acute care hospitals. *JAMA Netw Open*. 2019;2(2):e187571.
- Seymour CW, et al. Time to treatment and mortality during mandated emergency care for sepsis. *N Engl J Med*. 2017;376(23):2235–44.
- Liu VX, et al. The timing of early antibiotics and hospital mortality in sepsis. *Am J Respir Crit Care Med*. 2017;196(7):856–63.
- Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA*. 2018;320(21):2199–200.
- Fleuren LM, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis. *Intensive Care Med*. 2021;47(10):1144–56.
- Amann J, et al. Explainability for AI in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20(1):310.
- U.S. Food and Drug Administration. Clinical Decision Support Software: Draft Guidance. Silver Spring, MD: FDA; 2019.
- Gunning D, et al. XAI—Explainable artificial intelligence. *Sci Robot*. 2019;4(37):eaay7120.
- Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765–74.
- Rudin C. Stop explaining black box ML models for high-stakes decisions. *Nat Mach Intell*. 2019;1(5):206–15.

- Slack D, et al. Fooling LIME and SHAP: adversarial attacks on post-hoc explanation methods. AAAI/ACM Conf AI Ethics Soc. 2020:180–6.
- Nori H, et al. InterpretML: a unified framework for machine learning interpretability. arXiv:1909.09223. 2019.
- Komorowski M, et al. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24(11):1716–20.
- van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45(3):1–67.
- Tonekaboni S, et al. What clinicians want: contextualizing explainable machine learning for clinical end use. Proc Mach Learn Res. 2019;106:359–80.
- DeLong ER, et al. Comparing areas under two or more correlated ROC curves: a nonparametric approach. Biometrics. 1988;44(3):837–45.
- Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–74.
- Henry KE, et al. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med. 2015;7(299):299ra122.
- Nemati S, et al. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med. 2018;46(4):547–53.
- Caruana R, et al. Intelligible models for healthcare: predicting pneumonia risk. Proc ACM SIGKDD. 2015:1721–30.
- Lundberg SM, et al. Explainable ML predictions for prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018;2(10):749–60.
- Tjoa E, Guan C. A survey on XAI toward medical XAI. IEEE Trans Neural Netw Learn Syst. 2020;32(11):4793–813.