Early-Horizon Multimodal ICU Mortality Prediction Without Retraining

Alexander Bakumenko[1]* ABAKUME@CLEMSON.EDU
D. Hudson Smith[2] DANE2@CLEMSON.EDU
Janine Hoelscher[3] JANINEH@CLEMSON.EDU
[1] *School of Computing, Clemson University, North Charleston, SC, USA*
[2] *Department of Mathematical and Statistical Sciences, Clemson University, Clemson, SC, USA*
[3] *Department of Bioengineering, Clemson University, Clemson, SC, USA*

Abstract

Earlier ICU mortality prediction is more clinically useful because it can identify high-risk patients while treatment decisions can still change. Yet most models are trained on data from a fixed time window, so it is unclear whether a model trained on the first 48 hours of ICU data remains reliable when used earlier in the ICU stay. We evaluated a multimodal ICU mortality model trained once at 48 hours and then applied unchanged at 6, 12, 24, and 48 hours on MIMIC-III. The model combines an LSTM for physiological time-series data, a finetuned ClinicalModernBERT model for clinical notes, and a logistic regression fusion layer. Performance remained strong at earlier time points, suggesting that useful mortality prediction is possible earlier in the ICU stay even without retraining. At 6 hours, the model achieved AUROC 0.777 and remained well-calibrated (expected calibration error, ECE 0.038) without any recalibration, and it outperformed both single-modality models at every horizon. The benefit of combining both modalities was most evident at earlier horizons, when physiological data were sparse: agreement between the two specialists dropped by more than half from 48 to 6 hours, while the median contribution from clinical notes increased from 37% to 49%. A Bayesian version of the fusion layer showed that uncertainty decreased for survivors as more data accumulated but remained high for non-survivors; the most uncertain cases were up to 4.9 times more likely to be non-surviving patients. Continuous hourly analyses further showed that clinical notes provide stable context between documentation events. Simply carrying forward the most recent note matched or outperformed note-decay and documentation-gap alternatives. These results suggest that a multimodal ICU mortality model trained on 48 hours of data can provide trustworthy earlier predictions without retraining, while also identifying the cases that remain hardest to interpret.

Keywords: ICU mortality prediction, multimodal fusion, early-horizon evaluation, epistemic uncertainty, Bayesian stacking, NUTS, clinical notes, calibration, zero-shot transfer, logistic meta-learner

1. Introduction

Early ICU mortality prediction is most useful when intervention is still possible. Machine learning ICU mortality predictors are overwhelmingly trained and evaluated at a single fixed

* Corresponding author.

horizon, typically 24 or 48 hours, yet the window where intervention is most actionable is earlier Harutyunyan et al. (2019); Topol (2019). When such a model is applied before its training horizon without adaptation, its trustworthiness, here meaning reliable behavior with calibrated discrimination and quantified per-case uncertainty, cannot be assumed Subbaswamy and Saria (2020). For multimodal models that combine structured physiological time-series with clinical text, the mismatch is even greater: a model trained at 48 h encounters systematically different input distributions at, for instance 6 or 12 h, such as shorter physiological trajectories and fewer available clinical notes, both of which alter the evidence presented to the model Khadanga et al. (2019). This mismatch is not uniform: clinical notes arrive sparsely but often capture broader, longer-lived context, whereas physiological time-series update frequently and primarily reflect current state Zhang et al. (2022). Whether such a model remains trustworthy when applied earlier than its training horizon has not been systematically characterized.

A clinical prediction cannot be trusted on discrimination alone; three properties matter jointly: *discrimination*, so that higher-risk patients can be ranked above lower-risk patients; *calibration*, so that predicted probabilities reflect observed outcome frequencies; and *per-case uncertainty*, so that clinicians can distinguish stable predictions from those that warrant additional scrutiny Combi et al. (2022); Tonekaboni et al. (2019). These properties need not degrade in parallel as horizon shortens. A model may retain rank-order performance while losing calibration, or remain accurate in aggregate while being unreliable for a clinically important subset of episodes. This distinction matters concretely: calibration errors directly distort decision thresholds and risk communication Van Calster et al. (2019), and principled per-case uncertainty has been identified as an essential requirement for machine-assisted medical decision making Begoli et al. (2019). Neither property is well-documented in the multimodal ICU mortality literature. A recent scoping review of multimodal ICU in-hospital mortality studies found that only one-third reported any calibration metric, and trustworthiness analysis across different time horizons was essentially absent Bakumenko et al. (2026). From a pre-clinical standpoint, this is a substantive gap: predictable behavior, calibration, and principled uncertainty quantification have been identified as required properties for any predictive system before clinical evaluation is contemplated Bakumenko et al. (2025b).

We address this gap through a *zero-shot horizon transfer* evaluation paradigm. A multimodal ensemble whose 48-hour behavior has already been established Bakumenko et al. (2025a) is evaluated at 6, 12, 24, and 48 hours without retraining any specialist model and without re-estimating the fusion rule combining the specialist model outcomes. We characterize early-horizon trustworthiness along four axes: (1) discrimination and calibration across horizons; (2) changes in modality complementarity and per-case contribution structure under frozen late fusion; (3) per-case epistemic uncertainty (uncertainty in the model's own fusion weights) via a Bayesian No-U-Turn Sampler (NUTS) extension of the logistic regression (LR) stacker Hoffman et al. (2014); and (4) continuous hourly sensitivity of predictions to the note imputation strategy under irregular note updates. This paper makes four contributions:

(i) We provide a framework for systematically evaluating performance across earlier prediction horizons (6–48 h) without retraining, and show that a frozen multimodal en-

semble retains discrimination and calibration while consistently outperforming both unimodal specialists.

(ii) We show that modality complementarity is highest at the earliest horizons: clinical notes account for a larger relative share of the fused prediction, while the two specialist outputs are largely uncorrelated. Together, these trends provide a model-grounded explanation for the preserved fusion advantage when physiological evidence is most limited.

(iii) We introduce a Bayesian NUTS extension of the LR meta-stacker that preserves point-estimate discrimination while adding per-case epistemic uncertainty, revealing a class-level asymmetry in which survivors stabilize but non-survivors remain persistently uncertain across horizons.

(iv) We show that clinical notes act as stable background information: explicitly modeling how notes age over time did not improve predictions, suggesting that notes provide steady contextual signal that remains relevant over time.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work offers several insights relevant to multimodal clinical prediction beyond the ICU mortality setting. First, we share a methodology for zero-shot trustworthiness profiling: a frozen model can be characterized across earlier horizons without retraining, identifying where performance, calibration, and per-case reliability remain acceptable before external validation. Second, modalities can be most complementary when information is sparse early on: as the observation window extends and physiological evidence accumulates, the two modalities become progressively more redundant. Multimodal systems should therefore be evaluated across plausible deployment horizons, not only at the training horizon. Third, transparent linear fusion preserves an important interpretability property: a logistic meta-learner yields algebraically exact per-case modality attribution, avoiding reliance on post-hoc approximation methods for fusion-level interpretation. Finally, clinical notes can encode stable patient context throughout the ICU stay. Rather than assuming notes become stale over time, this should be evaluated empirically: notes may carry general admission-level information that remains relevant well beyond when they were written.

2. Methods

2.1. Data, Cohort Construction, and Prediction Task

We use the MIMIC-III clinical database Johnson et al. (2016). Structured physiological time-series are prepared following the benchmark pipeline of Harutyunyan et al. (2019): 17 clinical variables are resampled into hourly bins, missing values imputed via last-observation-carried-forward (LOCF) with per-variable normal-value fallback, and each hour is represented as a 76-dimensional feature vector. Clinical notes are preprocessed following Khadanga et al. (2019): notes without a valid chart time are excluded, and notes are concatenated chronologically. The 48-hour baseline architecture is inherited from Bakumenko et al. (2025a), which reports full training details. This baseline fuses physiological time-series and clinical notes, two of the most commonly combined modalities in multimodal ICU mortality

Table 1: Horizon-specific cohort sizes and in-hospital mortality prevalence. Prevalence is stable (13.01%–12.73%), indicating no horizon-dependent outcome bias.

Horizon	N	Positive (y=1)	Prevalence
6 h	11,881	1,546	13.01%
12 h	14,888	1,917	12.88%
24 h	15,554	1,995	12.83%
48 h	15,861	2,019	12.73%

prediction Bakumenko et al. (2026). Here, we focus on horizon-transfer characterization, cross-horizon calibration, fusion-layer uncertainty, and note-persistence analysis.

The prediction task is binary in-hospital mortality (IHM) fixed to post-48 h through end of stay, evaluated independently using data available at $h \in \{6, 12, 24, 48\}$ hours from ICU admission. For a given episode at horizon h, the time-series input is formed by retaining only the first h hourly vectors and zero-padding the remaining $48 - h$ positions (end-padding). The notes input is the concatenation of all notes charted within h hours; no notes from beyond the horizon are used. For each horizon, an ICU stay is included if and only if (i) at least h hours of structured measurements are available and (ii) at least one clinical note is charted within the first h hours. The train/validation/test partition follows the benchmark split of Harutyunyan et al. (2019); each split is then filtered independently by the horizon-specific criteria. Cohort sizes and IHM prevalence across horizons are summarized in Table 1.

The 6-hour cohort (N=11,881; Table 1) is the most restrictive, excluding stays shorter than 6 hours and stays without an early note. This cohort serves a dual role: as the discrete 6h evaluation split, and as the fixed cohort for all continuous hourly analyses (Section 2.7).

2.2. Multimodal Architecture

The system comprises two independently trained specialist models and a lightweight meta-learner. The *time-series specialist* is a stacked bidirectional LSTM Schuster and Paliwal (1997) (8 units) followed by a unidirectional LSTM Hochreiter and Schmidhuber (1997) (16 units), dropout, and a dense sigmoid output. It processes a 48×76 matrix of 17 clinical variables sampled hourly, each hour represented as a 76-dimensional feature vector following Harutyunyan et al. (2019). A `Masking(mask_value=0)` layer is included in the 48 h base architecture; its role during early-horizon evaluation is described in Section 2.3. The *clinical notes specialist* is ClinicalModernBERT Lee et al. (2025) finetuned end-to-end on in-hospital mortality labels, with an 8192-token context window and a linear classification head on the `[CLS]` token. It processes all notes charted within h hours, concatenated chronologically. Both specialists output a scalar mortality probability $p \in (0, 1)$ and were trained once on 48-hour MIMIC-III data.

Late fusion is performed by a meta-learner that combines the two specialist outputs (Section 2.4). All model parameters are frozen at evaluation time: the specialists generate predictions from truncated inputs, and the meta-learner combines them using the same parameters learned at 48 h. No component is adapted to any earlier horizon. This is the design constraint that makes zero-shot horizon transfer a meaningful evaluation paradigm.

2.3. Zero-Shot Horizon Transfer

Evaluating the ensemble at horizon $h < 48$ h is not a simple input truncation: the LSTM was trained exclusively on 48-timestep sequences, so shorter inputs alter the temporal context it was trained to read; and the clinical notes input must be restricted to notes charted within h hours, so no future documentation leaks in.

Time-series truncation. For horizon h, the time-series input is formed by retaining the first h hourly vectors and zero-padding the remaining $48 - h$ positions at the end (end-padding). The `Masking` layer then causes the LSTM to skip zero-padded timesteps entirely; they are excluded from recurrent computation rather than processed as zero-valued inputs, preventing contamination of the hidden state with non-physiological signal. A consequence is that predictions are generated from the hidden state at position $h-1$ rather than position 47: the model operates on only the observed physiology, but from an earlier position in the sequence than it encountered during training. End-padding preserves the temporal alignment learned during training; front-padding breaks it.

Clinical notes truncation. The notes input at horizon h is the concatenation of all notes charted within the first h hours. No notes from beyond the horizon are used. For the discrete evaluation horizons (6, 12, 24, 48 h), notes are extracted at the corresponding cutoff. For the continuous hourly evaluation, the most recent available note is carried forward (Section 2.7).

Meta-learner. Fusion weights are fixed at their 48h-trained values. No horizon-specific re-estimation is performed at any stage. Horizon evaluation thus tests both the specialists' ability to predict from incomplete evidence and the meta-learner's robustness to the positional shift in temporal context.

2.4. Stacking Algorithm Selection

We evaluated the generalization of six meta-learner algorithms on the held-out test set: logistic regression (LOGREG), uniform averaging (AVG), gradient boosted trees (GBM), random forests (RF), multilayer perceptron (MLP), and XGBoost (XGB) Wolpert (1992); Breiman (1996); Bakumenko et al. (2025a). Each candidate used the same two standardized specialist logits, evaluated independently at all four horizons. LOGREG and AVG were the top tier at every horizon, while GBM, RF, MLP, and XGB showed consistently lower AUPRC (Appendix Figure 6); with only two meta-features, higher-capacity stackers fail to improve discrimination.

While LOGREG and AVG showed similar discrimination across horizons, we focus on LOGREG for two architectural advantages: (i) exact additive per-case modality decomposition (Section 2.6); and (ii) direct Bayesian extension via NUTS for per-case epistemic uncertainty without changing point predictions (Section 2.5).

2.5. NUTS Meta-Stacker and Per-Case Epistemic Uncertainty

Transition from logistic regression to Bayesian No-U-Turn Sampler. The logistic meta-learner described in Section 2.4 operates deterministically: given fixed weights w^*

learned at 48 h, it maps specialist logits to a single mortality probability per episode. However, it provides no information about how confident that prediction is for a given episode. To characterize fusion-level epistemic uncertainty without altering the primary model, we replace the point-estimate w^* with a posterior distribution $p(w \mid \mathcal{D})$, which propagates weight uncertainty to a per-case predictive interval. All other design choices (specialist architecture, fusion in logit space, fixed 48h training) are unchanged. All performance results in Sections 3.1 and 3.2 are based on the deterministic LR stacker.

Base specialists and logit-space meta-features. We extend the ensemble from Sections 2.2–2.3 with a fusion-level uncertainty component, keeping all specialist parameters fixed. At each prediction horizon $h \in \{6, 12, 24, 48\}$ h, each specialist outputs a probability $p_{\mathrm{TS},h} \in (0,1)$ and $p_{\mathrm{CN},h} \in (0,1)$ for in-hospital mortality. To preserve the interpretable additive structure of the stacker, we operate in logit space. Let $\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right)$. We define the 2D meta-feature vector $x_h = [\mathrm{logit}(p_{\mathrm{TS},h}), \mathrm{logit}(p_{\mathrm{CN},h})]^\top$ and standardize it to z_h using the empirical mean μ and element-wise standard deviation s computed from the 48-hour training meta-features.

Bayesian logistic meta-learner trained once at 48h. We replace the deterministic meta-learner with a Bayesian logistic regression. For the 48-hour training episodes $\{(z_{48}^{(i)}, y^{(i)})\}_{i=1}^N$, we place weakly informative Gaussian priors on the modality weights $w \sim \mathcal{N}(0, 1.5^2 I)$ and intercept $b \sim \mathcal{N}(0, 1.0^2)$, following standard weak-prior practice for logistic regression Gelman et al. (2008). The likelihood for the binary mortality outcome $y^{(i)} \in \{0,1\}$ is:

$$y^{(i)} \mid z_{48}^{(i)}, w, b \sim \mathrm{Bernoulli}\left(\sigma\left(b + w^\top z_{48}^{(i)}\right)\right), \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function. Posterior inference is performed with the No-U-Turn Sampler (NUTS), yielding $S = 1000$ posterior samples $\{(w^{(s)}, b^{(s)})\}_{s=1}^S$. Convergence diagnostics are reported in Appendix Table 13.

The posterior is learned *once at 48h* and then *reused across horizons*. Horizon trends in uncertainty therefore reflect changes in input evidence (specialist logits under truncation), not horizon-specific re-estimation of fusion weights. This preserves the zero-shot transfer property of the full system.

Posterior predictive inference. For a test episode at horizon h with standardized logits z_h, we compute $p_h^{(s)} = \sigma\left(b^{(s)} + w^{(s)\top} z_h\right)$ for $s = 1, \ldots, S$, and use the posterior predictive mean $\hat{p}_h = \frac{1}{S}\sum_{s=1}^S p_h^{(s)}$ as the point estimate.

Epistemic uncertainty metric. We quantify epistemic uncertainty per episode as the width of a central posterior predictive interval on the probability scale:

$$u_h = Q_{0.95}\left(\{p_h^{(s)}\}_{s=1}^S\right) - Q_{0.05}\left(\{p_h^{(s)}\}_{s=1}^S\right), \tag{2}$$

where $Q_q(\cdot)$ denotes the empirical q-quantile across posterior samples. The distribution of u_h is reported by horizon overall and stratified by true label $y \in \{0,1\}$ using quantiles and violin plots.

Note that two interval types appear throughout: bootstrap 95% CIs on AUROC and AUPRC (B=1000 resamples) quantify evaluation-sample variance, reflecting how metrics

would vary across test sets drawn from the same population. Posterior predictive intervals u_h (Eq. 2) are distinct; they quantify parameter-level uncertainty per episode, characterizing how much the fused probability for a given patient varies across the posterior over fusion weights.

Performance evaluation. For each horizon, we evaluate AUROC and AUPRC on the test set using $\hat{p}_h$, with 95% confidence intervals via nonparametric bootstrap ($B = 1000$ resamples) Tibshirani and Efron (1993).

2.6. Modality Attribution Decomposition

The LR meta-learner's additive linear structure enables an exact, per-case decomposition of the fused prediction into modality-specific contributions: not a post-hoc approximation, but an algebraic identity (Eq. 5 in Section 3.2). The effective raw-logit weights w_{TS} and w_{CN} are obtained by dividing the standardized-space coefficients β by the corresponding training standard deviations s; the fused log-odds then decompose exactly as $b_{\text{eff}} + c_{\text{TS}} + c_{\text{CN}}$, where $c_i = w_i\, x_i^{\text{raw}}$ is the signed contribution of modality i in raw log-odds units, and $x_{\text{TS}}, x_{\text{CN}}$ denote the un-standardized specialist logits. Positive values push the prediction toward higher mortality odds; negative values push toward lower.

Two derived metrics characterize fusion behavior across horizons. *Signed contributions* c_{TS} and c_{CN} capture the direction and patient-level heterogeneity of each modality's influence. *Unsigned share* $|c_{\text{CN}}|/(|c_{\text{TS}}| + |c_{\text{CN}}|)$, summarized as a per-episode median, captures relative modality dominance. Specialist logit correlation $r(x_{\text{TS}}, x_{\text{CN}})$, computed on raw specialist logits rather than contributions, quantifies information redundancy between specialists independently of fusion reweighting. Together, these three quantities provide the evidence base for the modality complementarity analysis (Section 3.2).

2.7. Continuous Hourly Evaluation Under Documentation Sparsity

We hypothesize that clinical notes may encode stable admission-level context, including presenting diagnosis, initial clinician assessment, and chronic conditions, that remains predictive between documentation events rather than behaving as a rapidly decaying signal. To test this, we evaluated the frozen multimodal ensemble hourly from hour 6 through hour 48 on the fixed 6h cohort (N=11,881), holding cohort composition constant to isolate evidence accumulation from population shifts.

Method 1: Last-Observation-Carried-Forward (LOCF, null model). Notes are cumulative documents: the 12h note contains all text from the first 12 hours, not only hour 12. We exploit this structure directly: the 6h note is used for hours 6–11, the 12h note for hours 12–23, the 24h note for hours 24–47, and the 48h note for hour 48. LOCF is the null model: if notes persist as stable contextual inputs, this simple strategy should suffice.

Method 2: Exponential decay (post-hoc rescaling, no retraining). To test whether note contributions should diminish monotonically as documentation age grows, we modulate the CN logit at inference time:

$$\text{logit}(p_{\text{meta}}) = b_{\text{eff}} + w_{\text{TS}}\, x_{\text{TS}} + w_{\text{CN}}\, e^{-\alpha \Delta h}\, x_{\text{CN}}, \tag{3}$$

Table 2: Discrimination and calibration performance by time horizon for the multimodal ensemble and both modality specialists. All metrics are computed on the held-out test set with 95% bootstrap CIs ($B = 1000$). ECE Pre/Post: expected calibration error before/after recalibration. Brier scores are reported in Appendix Table 12.

Model	Horiz.	AUROC	AUPRC	ECE (Pre)	ECE (Post)
Ensemble	48h	0.891 [0.872,0.909]	0.565 [0.503,0.625]	0.022	0.022
	24h	0.853 [0.831,0.875]	0.481 [0.420,0.538]	0.026	0.026
	12h	0.805 [0.778,0.831]	0.382 [0.318,0.447]	0.034	0.034
	6h	0.777 [0.741,0.809]	0.306 [0.243,0.367]	0.038	0.038
TS (LSTM)	48h	0.855 [0.832,0.877]	0.485 [0.425,0.545]	0.029	0.024
	24h	0.807 [0.778,0.832]	0.380 [0.319,0.441]	0.048	0.030
	12h	0.751 [0.719,0.782]	0.319 [0.261,0.376]	0.073	0.027
	6h	0.693 [0.654,0.732]	0.259 [0.201,0.320]	0.101	0.035
CN (ft-CMB)	48h	0.876 [0.856,0.895]	0.526 [0.462,0.586]	0.133	0.024
	24h	0.839 [0.816,0.862]	0.460 [0.402,0.519]	0.126	0.027
	12h	0.782 [0.753,0.810]	0.344 [0.287,0.403]	0.105	0.035
	6h	0.759 [0.724,0.790]	0.271 [0.213,0.327]	0.083	0.037

where Δh is hours elapsed since the most recent note and $\alpha = \ln 2/t_{1/2}$ is the decay rate corresponding to half-life $t_{1/2}$. We test three half-lives ($t_{1/2} \in \{6, 12, 24\}$ h). This is post-hoc rescaling only; no meta-learner parameters are re-estimated.

Method 3: Additive staleness feature (stacker retrained). Documentation timing could carry independent predictive signal, as a proxy for monitoring intensity or care-escalation patterns, even if note content itself does not decay. We test this by adding Δh as a third meta-feature and retraining the stacker on validation predictions:

$$\text{logit}(p_{\text{meta}}) = b_{\text{eff}} + w_{\text{TS}}\, x_{\text{TS}} + w_{\text{CN}}\, x_{\text{CN}} + w_{\Delta}\, \Delta h. \tag{4}$$

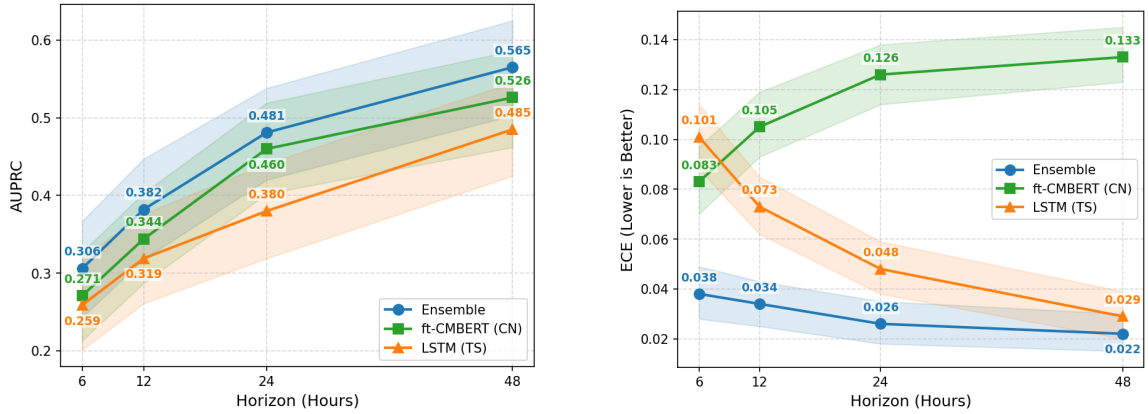
The learned coefficient w_{Δ} directly quantifies whether documentation gaps carry predictive signal independent of note content. Results are reported in Section 3.4.

3. Results

3.1. Discrimination and Calibration Are Preserved Without Retraining

Stacking algorithm selection. A comparison of six stacking algorithms (Section 2.4) confirms that LOGREG and AVG dominate at every horizon (Appendix Figure 6); LOGREG is selected over AVG for its exact per-case modality decomposition. LOGREG's two-weight-plus-intercept structure also motivates the Bayesian extension in Section 3.3: placing a posterior over this parsimonious model adds per-case epistemic uncertainty without changing the predictive structure.

Discrimination. We evaluated whether a logistic-regression stacker trained on 48-hour multimodal data retains meaningful discrimination at earlier horizons under a strict zero-adaptation constraint: no retraining, no recalibration, and no horizon-specific tuning. The ensemble achieves AUROC 0.891 and AUPRC 0.565 at 48h, and AUROC 0.777 and AUPRC 0.306 at 6h (Table 2), outperforming both specialists at every horizon by an average of



(a) AUPRC across horizons

(b) ECE across horizons (lower is better)

Figure 1: AUPRC (left) and ECE (right) across horizons for the multimodal ensemble and both specialists. Shaded bands show 95% bootstrap confidence intervals ($B = 1000$). Left: the ensemble (blue, circle) outperforms ft-CMB (CN, green) and LSTM (TS, orange) at every horizon; ft-CMB degrades faster than LSTM in AUPRC (48h→6h: ft-CMB −0.255, LSTM −0.226). Right: the LR ensemble maintains ECE ≤ 0.038 with identical pre- and post-calibration values; both specialists require recalibration at early horizons. AUROC values by horizon are in Table 2; precision–recall and ROC curves are in Appendix Figure 5.

+0.018 AUROC and +0.033 AUPRC. Figure 1 shows AUPRC and ECE confidence intervals across horizons. Performance degrades smoothly (AUROC drop 48h→6h: ensemble −0.114, LSTM −0.162, ft-CMB −0.117), with LSTM losing more AUROC and ft-CMB losing more AUPRC (LSTM: −0.226; ft-CMB: −0.255), an asymmetric degradation examined mechanistically in Section 3.2. Horizon-specific baselines showed that zero-shot performance remained close to retraining while substantially exceeding SOFA-h at every horizon (Appendix L).

Calibration. The LR ensemble requires no post-hoc recalibration at any horizon: ECE is 0.038 at 6h, 0.034 at 12h, 0.026 at 24h, and 0.022 at 48h, with pre- and post-calibration values identical throughout (Table 2). Both specialists require post-hoc recalibration. LSTM is most miscalibrated at early horizons (pre-calibration ECE 0.101 at 6h, 2.7× the ensemble at the same horizon); ft-CMB shows the opposite pattern (ECE 0.133 at 48h), reflecting systematic overconfidence that recalibration corrects substantially. These results indicate that ensemble outputs are directly interpretable as mortality probabilities across 6–48h without additional recalibration. The smallest early-horizon agreement-defined subgroup (6h Agree High, $n = 33$) showed unstable subgroup-level calibration, consistent with limited sample size and distribution shift in rare high-risk consensus cases (see Appendix Table 7).

3.2. Modality Complementarity Increases at Shorter Horizons

Table 2 shows that the LSTM specialist performs substantially worse than the notes specialist at earlier horizons: AUROC falls from 0.855 to 0.693 for TS versus 0.876 to 0.759 for

CN, yet the LR ensemble consistently exceeds both. This asymmetric degradation cannot be explained by either specialist alone; we therefore examine the fusion dynamics.

We hypothesized that modality complementarity, the degree to which specialists capture distinct, non-redundant information, would increase at shorter horizons, providing a model-grounded account of the preserved fusion advantage when physiological evidence is most limited. The unsigned CN share, $|c_{\mathrm{CN}}|/(|c_{\mathrm{TS}}| + |c_{\mathrm{CN}}|)$ per episode, increases monotonically as horizon shortens: median share is 0.370 at 48h, 0.394 at 24h, 0.438 at 12h, and 0.486 at 6h (Figure 2). At 6h, clinical notes and physiological signals contribute approximately equally to the fused prediction. This shift is a structural consequence of the late-fusion architecture under fixed meta-learner weights, not horizon-specific reweighting: as horizon shortens, the LSTM's output logit compresses toward zero while LR weights remain fixed, so CN gains relative influence. This contribution trajectory was unchanged after stripping explicit code-status and goals-of-care terms from the clinical notes (Appendix Table 8).

The Pearson correlation between specialist logits falls from $r = 0.696$ at 48h to $r = 0.308$ at 6h (Figure 2), with variance explained dropping from 48.4% to 9.5%, indicating a transition from largely overlapping to largely distinct signals as fewer physiological hours are available. This declining correlation provides a model-grounded explanation for the preserved fusion margin observed in Section 3.1: decorrelated specialists produce more diverse error patterns, making their combination more effective Kuncheva and Whitaker (2003). Figure 2 shows the joint trend directly: CN share rises as specialist correlation falls, with the two curves crossing near 12h. Both modalities should therefore be retained for early-horizon evaluation; dropping either would incur the largest complementarity cost precisely when physiological data are scarcest. Episode-level dominance patterns are consistent with this aggregate trend: balanced cases (CN share 45–55%) increase from 11.5% at 48h to 19.0% at 6h, TS-dominant cases fall from 16.7% to 10.0% (Appendix Figure 8), and the proportion of both-high specialist consensus cases (Agree High: both predict $p > 0.5$) declines from 5.1% to 1.8% as fewer physiological hours are available to drive both specialists into high-mortality agreement (Appendix Table 3).

Signed contributions. Modality share and logit correlation describe aggregate complementarity; signed contributions reveal the directional mechanism: time-series pushes most patients toward low-risk with growing force as data accumulate, while clinical notes maintain a distinctive positive tail that drives high-risk identifications at all horizons.

To evaluate late-fusion behavior across horizons, we analyzed signed modality contributions from the logistic-regression stacker at 6, 12, 24, and 48 hours. Because coefficients were converted from standardized-logit space to effective raw-logit coefficients, fused predictions decompose exactly as:

$$\mathrm{logit}(p_{\mathrm{meta}}) = b_{\mathrm{eff}} + c_{\mathrm{TS}} + c_{\mathrm{CN}}, \quad c_{\mathrm{TS}} = w_{\mathrm{TS}}\, x_{\mathrm{TS}}, \quad c_{\mathrm{CN}} = w_{\mathrm{CN}}\, x_{\mathrm{CN}}. \tag{5}$$

Appendix Figure 7 shows the patient-level contribution distributions at each horizon. With ≈11% in-hospital mortality, central tendency is below zero for both modalities throughout. For time-series, contributions grow more negative and more dispersed across all percentiles as physiology accumulates; the p95 remains near zero at all horizons, confirming the LSTM rarely contributes strong high-risk signal. For clinical notes, the positive tail strengthens: the fraction above zero rises from 10.2% at 6h to 19.8% at 48h.

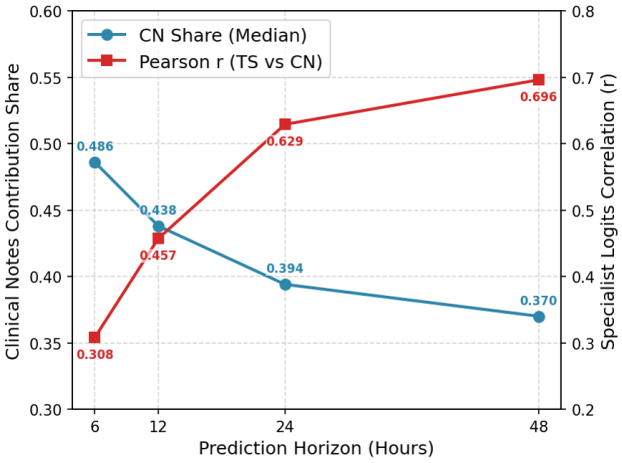


Figure 2: Modality complementarity across horizons. Median clinical-note contribution share (blue circles, left axis) increases as horizon shortens, from 0.370 at 48 h to 0.486 at 6 h, while Pearson correlation between specialist logits (red squares, right axis) decreases from 0.696 to 0.308. Together, these trends indicate increasing complementarity under temporal truncation: as physiological evidence becomes sparser, clinical notes contribute a larger relative share and the two specialists become less redundant.

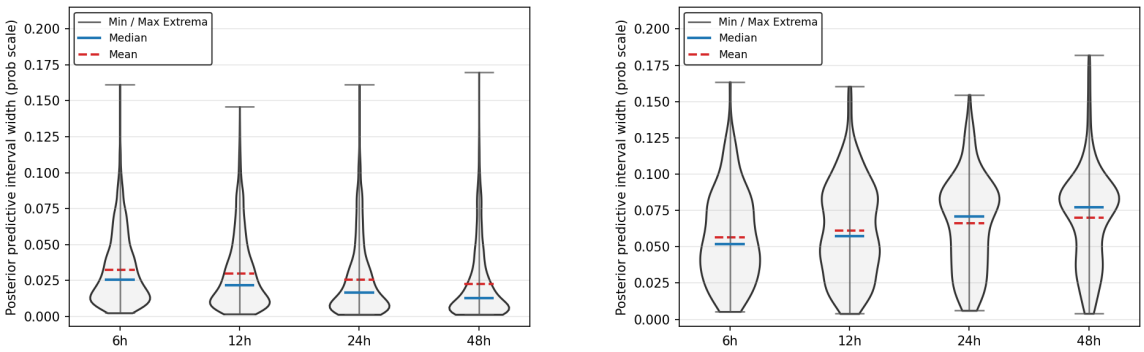
TS is attenuated at early horizons, CN remains anchored throughout; this divergence is the per-episode manifestation of the complementarity pattern in Section 3.2. Full per-quantile statistics are in Appendix Table 4.

3.3. Per-Case Epistemic Uncertainty and Its Class Asymmetry

The preceding sections establish that the frozen LR-based fusion remains effective and interpretable across early horizons. We next ask a different question: how much confidence should be placed in any individual fused prediction? The NUTS extension of the LR stacker provides this: by treating fusion weights as random variables, it produces a posterior predictive distribution per episode whose width (u_h, Eq. 2) quantifies how much the predicted probability varies under uncertainty about the fusion parameters.

NUTS recovers identical LR discrimination. The Bayesian (NUTS) meta-stacker yields essentially identical discrimination to the deterministic LR meta-stacker at every horizon (AUROC and AUPRC differences ≤ 0.001; full comparison in Appendix Table 5), confirming that uncertainty quantification is added without any performance trade-off. This is expected for a 3-parameter model, where the posterior predictive mean closely tracks the maximum-likelihood solution while the posterior over weights enables per-case interval estimation.

Overall uncertainty. Using Eq. (2) with a fixed posterior learned at 48 h, we quantified horizon-wise posterior predictive interval widths (q95–q5) on the probability scale. Median posterior predictive width increases from 0.0156 at 48 h to 0.0282 at 6 h (1.81×); corresponding ratios at 12h and 24h are 1.59× and 1.26× respectively (see Appendix Figure 9).



(a) Survivors (y=0): uncertainty decreases with horizon

(b) Non-survivors (y=1): uncertainty remains persistently high

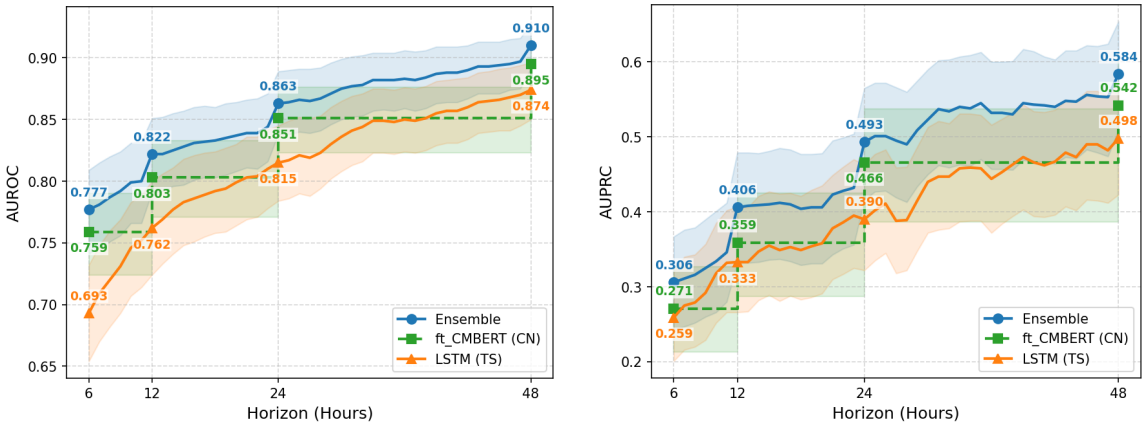
Figure 3: Class-stratified epistemic uncertainty by horizon (posterior predictive width q95–q5, probability scale). Survivors stabilize as physiological evidence accumulates; non-survivors do not. The class asymmetry becomes more pronounced at later horizons (median ratio 2.0 at 6h $\rightarrow$ 6.0 at 48h), flagging the cases the model finds hardest to classify: those whose fused specialist signals do not converge to a confident prediction regardless of observation window. Full per-class quantiles are in Appendix Table 6; overall horizon trend is in Appendix Figure 9.

Notably, the upper tail (q90–q95) remains comparatively stable across horizons, suggesting that a subset of episodes sustain high epistemic uncertainty regardless of available data, motivating the class-stratified analysis below.

Survivors stabilize and non-survivors remain uncertain. Label-stratified distributions reveal that the overall horizon trend is driven primarily by the negative class. For $y = 0$ (survivors), median uncertainty decreases as horizon increases. For $y = 1$ (non-survivors), uncertainty is consistently higher and its median increases with horizon, producing a class asymmetry that becomes more pronounced at later horizons.

Appendix Table 6 gives full quantiles by class and horizon. The ratio of median uncertainty for non-survivors to survivors increases from 2.02 at 6 h to 5.96 at 48 h. By the time maximum physiological evidence is available, the model is substantially more certain about survivors while its uncertainty about non-survivors has barely moved from its 6h level. While probability-scale widths are also shaped by sigmoid geometry, our primary claim concerns their horizon-wise evolution within each class, rather than geometry-independent cross-class magnitudes.

This asymmetry has a specific model-level reading. For survivors, additional ICU monitoring provides increasingly consistent physiological evidence that the specialist logits integrate into a confident low-risk prediction. For non-survivors, the fused specialist signals do not converge: the posterior remains wide because no stable configuration of the two weights produces a consistently high-risk output for these episodes; this is not a model failure, but an accurate signal of inherent classification difficulty. What the Bayesian stacker makes explicit, and the LR point estimate conceals, is that difficulty, per episode and per horizon.



(a) Hourly AUROC (6h–48h)

(b) Hourly AUPRC (6h–48h)

Figure 4: Continuous hourly discrimination under LOCF on the fixed 6h cohort (N=11,881). Shaded bands show 95% bootstrap confidence intervals. Ensemble (blue), ft-CMB specialist (green, dashed), LSTM specialist (orange). Performance rises at documentation events (hours 12, 24, and 48; vertical step increases) and accumulates gradually between events from physiological data alone. The ensemble consistently exceeds both specialists at every hour. Carrying the most recent note forward does not degrade prediction between documentation events, consistent with the note-persistence hypothesis.

Operability annotation. The practical question is whether high epistemic uncertainty merely reflects data scarcity, which would make it redundant with the point estimate, or whether it preferentially concentrates on episodes the model finds hardest to classify. Across all horizons, high-uncertainty episodes are substantially enriched for non-surviving patients. At 48 h, the top 10% most-uncertain episodes contain non-surviving patients at 4.90× the base rate; the pattern holds consistently from 6h (3.05×) through 48h. This confirms that epistemic uncertainty carries information beyond the point estimate: it concentrates on the cases where the model's fused specialist signals fail to converge, rather than on cases that are simply data-sparse.

However, this also cautions against naive uncertainty-gated abstention in a highly imbalanced setting: deferring on high-uncertainty cases would disproportionately defer non-surviving patients. We therefore treat epistemic uncertainty as an operability annotation, a flag that a given prediction warrants closer scrutiny, rather than as a deployed deferral or triage policy.

3.4. Clinical Notes Persistence as Stable Contextual Inputs

We hypothesized that clinical notes encode stable admission-level context, including presenting diagnosis, clinician assessment, and comorbidities, that remains predictive between documentation events. If so, LOCF should suffice without staleness penalties. We tested this against two decay-assuming alternatives (Section 2.7).

3.4.1. LOCF Baseline: Stepwise Gains at Documentation Events

Using the fixed 6h cohort (N=11,881) evaluated hourly from hour 6 through hour 48, the LOCF ensemble achieves AUROC 0.777 at 6h and rises to 0.910 at 48h; AUPRC rises from 0.306 to 0.584 over the same range (Figure 4). The trajectories display a characteristic step pattern: performance accelerates at documentation events (hours 12, 24, and 48, when a new cumulative note is incorporated), and accumulates gradually between events from time-series accumulation alone. Carrying a note forward across an inter-event interval does not harm prediction; the ensemble maintains or improves at every between-event hour, confirming the LOCF assumption: a note written at hour 6 remains informative at hour 11.

Exponential decay. To test whether note contributions should diminish as time since documentation grows, we applied exponential decay to the CN logit at inference time: $\text{logit}(p_{\text{meta}}) = b_{\text{eff}} + w_{\text{TS}}x_{\text{TS}} + w_{\text{CN}} \cdot e^{-\alpha\Delta h} \cdot x_{\text{CN}}$, where Δh is hours since the most recent note. Three decay rates were tested (half-life 6h, 12h, 24h) as post-hoc rescaling without retraining. All three variants perform within or below the LOCF confidence band at all evaluation hours; stronger decay consistently reduces AUROC and AUPRC at hours distant from documentation events, precisely the intervals where the staleness hypothesis predicts the largest benefit.

Time-since-note. Documentation timing could still be informative independent of note content, as a proxy for monitoring intensity or care escalation patterns. To test this, we extended the stacker with Δh as a third meta-feature and retrained on validation predictions: $\text{logit}(p_{\text{meta}}) = b_{\text{eff}} + w_{\text{TS}}x_{\text{TS}} + w_{\text{CN}}x_{\text{CN}} + w_{\Delta}\Delta h$. The learned coefficient $w_{\Delta} \approx 0.01$ is effectively zero, and performance is indistinguishable from the two-feature LOCF baseline. Across all three tests, LOCF is sufficient: clinical notes encode admission-level context that operates on a different informational timescale from hourly physiology, and prediction quality was not sensitive to documentation timing within the tested window.

4. Discussion

In this work, we examined whether a transparent multimodal ICU mortality ensemble trained once at 48 h can be used earlier without retraining, and whether its predictions remain trustworthy under that zero-shot horizon shift. The main finding is that early-horizon use does not reduce trustworthiness to discrimination alone: the LR-based ensemble retained both discrimination and strong calibration across 6–48 h while consistently outperforming both unimodal specialists. The calibration result is particularly important because it is empirical under horizon shift rather than guaranteed by design. The LR stacker was fitted on the 48h training distribution; applying the same frozen fusion rule at earlier horizons changes the specialist logits it receives, so calibration could have degraded substantially. Instead, ECE remained ≤ 0.038 at all horizons, whereas both specialists required post-hoc recalibration to approach comparable reliability. While calibration remains underreported in the multimodal ICU mortality prediction literature, this shows that probabilistic reliability can be preserved under zero-shot horizon transfer in a transparent late-fusion architecture.

The preserved fusion advantage at earlier horizons has a model-grounded explanation in changing modality complementarity. As horizon shortens, the time-series specialist degrades faster than the notes specialist, while the ensemble remains above both. This is accompanied

by two consistent shifts in fusion behavior: the median clinical-notes contribution share rises from 0.370 at 48 h to 0.486 at 6 h, and the Pearson correlation between specialist logits falls from $r = 0.696$ to $r = 0.308$, reducing shared variance from 48.4% to 9.5%. The two specialists therefore become less redundant when physiological evidence is sparse; lower redundancy implies more diverse error structure and a greater fusion benefit Kuncheva and Whitaker (2003). The signed-contribution analysis confirms this at episode level: the time-series specialist (TS) increasingly resolves the low-risk majority, while the clinical notes specialist (CN) maintains a persistent positive tail, an attribution that is algebraically exact because the LR stacker is linear.

The Bayesian NUTS extension shows that population-level performance is an incomplete account of trustworthiness. What it contributes is per-case epistemic reliability. The key finding is a class-level asymmetry: survivors stabilize as evidence accumulates while non-survivors remain persistently uncertain, a pattern not explained by data sparsity, because it persists when input evidence is maximal. The model becomes progressively more certain about episodes whose fused signals support survival, while remaining uncertain about episodes whose specialist signals do not converge. Epistemic uncertainty therefore functions as an operability annotation, a per-case flag that a prediction warrants closer scrutiny, not a justification for uncertainty-gated abstention, which in an imbalanced setting would disproportionately defer non-surviving patients.

The continuous hourly analyses support the note-persistence hypothesis. All three staleness tests converge on the same conclusion: clinical notes encode admission-level context that operates on a different informational timescale from hourly physiology and is not superseded by vital sign evolution within the first 48 h. Within the tested window, prediction quality was not sensitive to documentation timing, which may simplify operational requirements for systems of this type.

Limitations. First, evaluation uses single-institution MIMIC-III data collected from 2001–2012; generalizability to contemporary practice, other hospitals, electronic health record systems, and patient populations is unverified, and demographic subgroup behavior remains uncharacterized. Second, decision-curve analysis is not reported and net clinical benefit has not been characterized; these results should not be interpreted as evidence of clinical utility. Third, this study characterizes a late-fusion architecture, while horizon-transfer behavior may differ for other fusion paradigms. Fourth, epistemic uncertainty is quantified only at the fusion layer; aleatoric and specialist-level uncertainty, as well as comparisons with alternative uncertainty-quantification methods, are outside the present scope. Fifth, cohort construction requires at least one clinical note within the first h hours; settings with sparser early documentation may exhibit different note-persistence behavior.

5. Conclusion

This paper provides an answer to a practical translational question: whether a transparent multimodal ICU mortality model trained at 48 h can be meaningfully characterized at earlier horizons without retraining. Within that zero-shot setting, trustworthiness proved characterizable across four measurable dimensions: discrimination, calibration, fusion behavior, and per-case uncertainty, without any retraining.

Taken together, the frozen ensemble preserved calibration and discrimination across 6–48 h; its fusion advantage was accompanied by increasing modality complementarity; its Bayesian extension revealed a persistent uncertainty asymmetry concentrated in hard non-survivor cases; and clinical notes behaved as a stable contextual input rather than a decaying signal. Rather than treating early prediction as an accuracy-only problem, this paper shows that early-horizon multimodal prediction can be characterized as a trustworthiness problem, with explicit evidence on when reliable signal emerges, how fusion behaves when physiology is sparse, and which cases remain intrinsically hard to classify.

Acknowledgments

This research was supported by the National Science Foundation (NSF) under Award No. OIA-2242812. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF.

References

Alexander Bakumenko, Janine Hoelscher, and Hudson Smith. Transparent early icu mortality prediction with clinical transformer and per-case modality attribution. *arXiv preprint arXiv:2511.15847*, 2025a.

Alexander Bakumenko, Aaron J Masino, and Janine Hoelscher. Before the clinic: Transparent and operable design principles for healthcare ai. *arXiv preprint arXiv:2511.01902*, 2025b.

Alexander Bakumenko, Janine Hoelscher, D. Hudson Smith, Svetlana Bakumenko, Aaron J. Masino, and Jihad S. Obeid. Multimodal icu in-hospital mortality prediction: A scoping review of transparency, calibration, and translation readiness. *Research Square*, March 2026. doi: 10.21203/rs.3.rs-9217081/v1.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

Carlo Combi, Beatrice Amico, Riccardo Bellazzi, Andreas Holzinger, Jason H Moore, Marinka Zitnik, and John H Holmes. A manifesto on explainability for artificial intelligence in medicine. *Artificial Intelligence in Medicine*, 133:102423, 2022.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008. doi: 10.1214/08-AOAS191.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

Alistair E W Johnson, David J Stone, Leo A Celi, and Tom J Pollard. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39, 2018.

Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. Using clinical notes with time series data for icu management. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6432–6437, 2019.

William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. APACHE II: a severity of disease classification system. *Critical care medicine*, 13(10): 818–829, 1985.

Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.

Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (SAPS II) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963, 1993.

Simon A Lee, Anthony Wu, and Jeffrey N Chiang. Clinical modernbert: An efficient and long context encoder for biomedical text. *arXiv preprint arXiv:2504.03964*, 2025.

Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020.

Robert J Tibshirani and Bradley Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1):1–436, 1993.

Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pages 359–380. PMLR, 2019.

Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

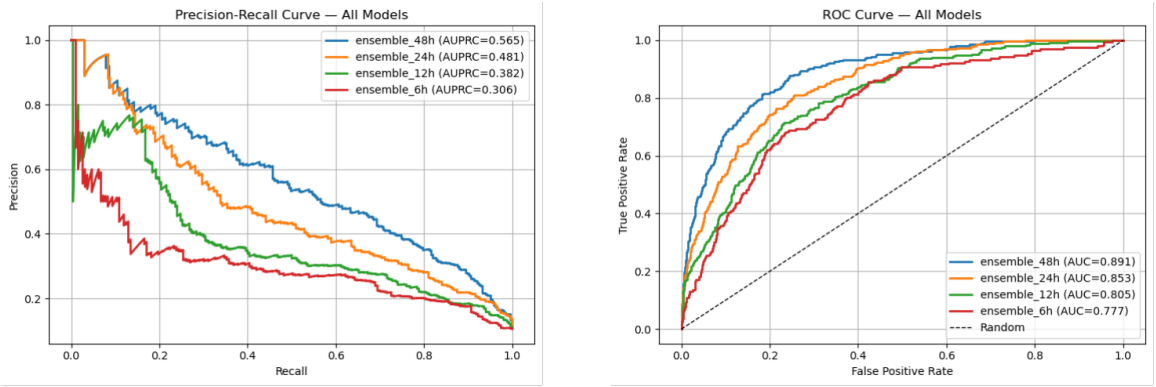
Ben Van Calster, David J McLernon, Maarten Van Smeden, Laure Wynants, and Ewout W Steyerberg. Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1): 230, 2019.

J-L Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, C Kathryn Reinhart, PeterM Suter, and Lambertius G Thijs. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure: On behalf of the working group on sepsis-related problems of the european society of intensive care medicine (see contributors to the project in the appendix). *Intensive care medicine*, 22(7):707–710, 1996.

David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

Jingqing Zhang, Luis Daniel Bolanos Trujillo, Ashwani Tanwar, Julia Ive, Vibhor Gupta, and Yike Guo. Clinical utility of automatic phenotype annotation in unstructured clinical notes: intensive care unit use. *BMJ Health & Care Informatics*, 29(1):e100519, 2022.

Appendix A. Ensemble Discrimination Curves



(a) Precision–recall curves

(b) ROC curves

Figure 5: Precision–recall (left) and ROC (right) curves for the multimodal LR ensemble at 6, 12, 24, and 48 hours on the held-out test set. AUPRC and AUROC values match Table 2.

Appendix B. Stacking Algorithm Comparison

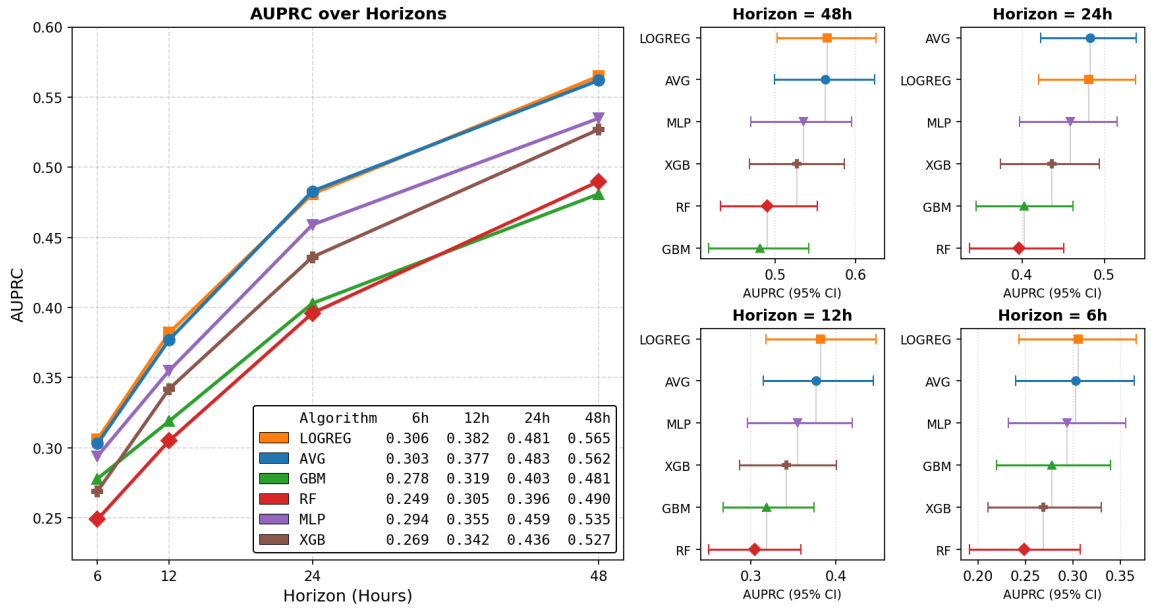


Figure 6: Stacking algorithm comparison on held-out test AUPRC. Left: AUPRC trajectories across horizons for six meta-learner candidates (LOGREG, AVG, GBM, RF, MLP, XGB). Right: per-horizon forest plots with 95% bootstrap confidence intervals. LOGREG and AVG consistently match or outperform all other candidates at every horizon; tree-based and neural alternatives show lower point estimates and wider intervals, consistent with variance induction from an underdetermined 2-feature input.

Appendix C. Signed Modality Contribution Distributions

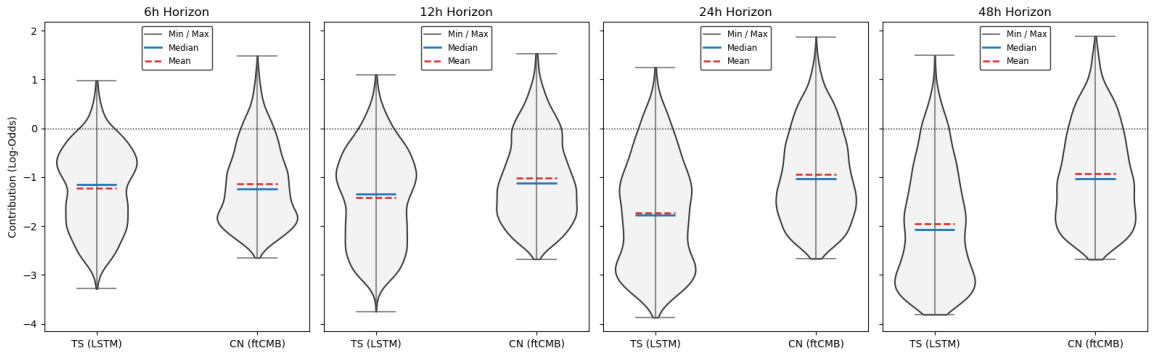


Figure 7: Signed modality contribution distributions across decision horizons (LR stacker). Each violin shows patient-level weighted contributions in raw log-odds units for the LSTM (TS) and ClinicalModernBERT (CN) specialists. The dotted line at zero marks no directional effect. Values below zero push the ensemble toward lower mortality odds; values above zero push toward higher mortality odds. Blue solid = median; red dashed = mean. TS contributions grow more negative and more dispersed with time; CN contributions maintain a distinctive positive tail at all horizons.

Appendix D. Modality Dominance Patterns

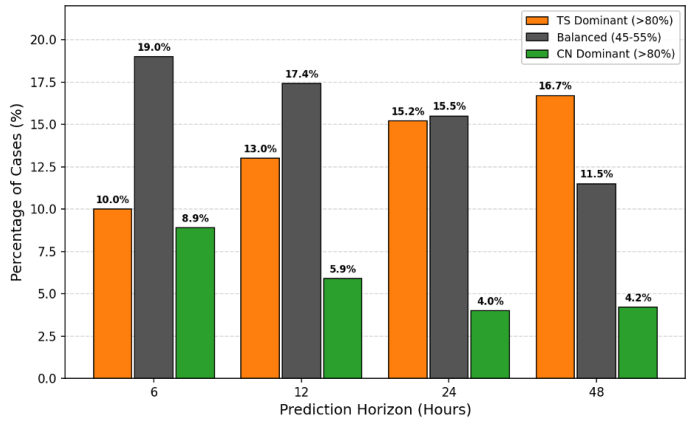


Figure 8: Episode-level modality dominance patterns by horizon. TS Dominant: CN share $< 20\%$; Balanced: CN share 45–55%; CN Dominant: CN share $> 80\%$. At earlier horizons, TS-dominant cases decrease (16.7%→10.0%) and balanced cases increase (11.5%→19.0%), consistent with increasing two-modality fusion at early horizons.

Appendix E. Specialist Agreement Patterns

Table 3: Specialist agreement patterns across horizons (test set). Agree High: both specialists predict $p > 0.5$; Conflict: one above, one at or below 0.5; Agree Low: both predict $p \leq 0.5$. Counts and row percentages shown.

Horizon	Agree High (↑↑)	Conflict (↑↓)	Agree Low (↓↓)
48h	124 (5.1%)	394 (16.1%)	1930 (78.8%)
24h	105 (4.4%)	399 (16.7%)	1892 (79.0%)
12h	72 (3.1%)	346 (15.1%)	1874 (81.8%)
6h	33 (1.8%)	213 (11.6%)	1591 (86.6%)

Appendix F. Signed Modality Contribution Statistics

Table 4: Summary statistics for signed modality contribution distributions (LR stacker). All values in raw log-odds units. Positive fraction: proportion of episodes where the modality contributes a positive (mortality-increasing) log-odds term.

Horizon	Modality	Mean	Median	SD	p5	p25	p75	p95	Pos. fraction
6h	TS (LSTM)	−1.220	−1.153	0.830	−2.570	−1.899	−0.578	−0.002	5.0%
	CN (ft-CMB)	−1.132	−1.243	0.804	−2.233	−1.809	−0.568	0.295	10.2%
12h	TS (LSTM)	−1.413	−1.345	0.952	−2.940	−2.208	−0.678	0.114	6.3%
	CN (ft-CMB)	−1.013	−1.121	0.866	−2.250	−1.712	−0.367	0.493	15.1%
24h	TS (LSTM)	−1.731	−1.780	1.097	−3.311	−2.714	−0.896	0.154	6.9%
	CN (ft-CMB)	−0.940	−1.038	0.926	−2.282	−1.685	−0.240	0.636	18.5%
48h	TS (LSTM)	−1.960	−2.069	1.160	−3.518	−2.994	−1.085	0.101	6.5%
	CN (ft-CMB)	−0.936	−1.038	0.975	−2.314	−1.754	−0.194	0.720	19.8%

Appendix G. NUTS vs. LR Discrimination Parity

Table 5: Discrimination performance: Bayesian NUTS vs. deterministic LR meta-stacker. Differences are ≤0.001 at all horizons, confirming the Bayesian extension is non-destructive.

Metric	NUTS	LR	Δ
6h horizon			
AUROC	0.776 [0.741,0.809]	0.777 [0.741,0.809]	−0.001
AUPRC	0.306 [0.243,0.367]	0.306 [0.243,0.367]	0.000
12h horizon			
AUROC	0.805 [0.778,0.830]	0.805 [0.778,0.831]	0.000
AUPRC	0.382 [0.317,0.448]	0.382 [0.318,0.447]	0.000
24h horizon			
AUROC	0.853 [0.831,0.875]	0.853 [0.831,0.875]	0.000
AUPRC	0.481 [0.420,0.538]	0.481 [0.420,0.538]	0.000
48h horizon			
AUROC	0.891 [0.872,0.909]	0.891 [0.872,0.909]	0.000
AUPRC	0.564 [0.503,0.624]	0.565 [0.503,0.625]	−0.001

Appendix H. Overall Epistemic Uncertainty by Horizon

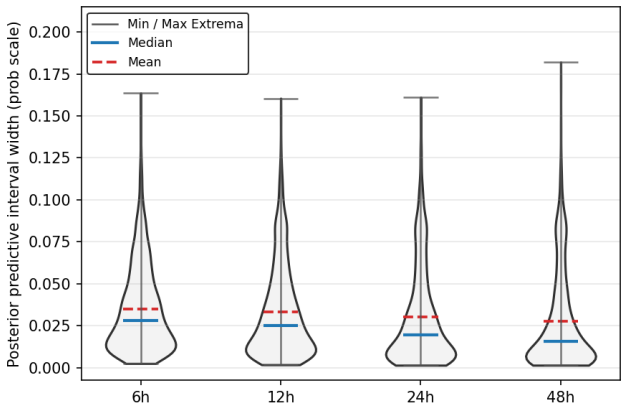


Figure 9: Epistemic uncertainty by horizon (all episodes). Posterior predictive interval width (q95–q5) on the probability scale, under a fixed posterior learned at 48 h. Median width increases 1.81× at 6h relative to 48h; the upper tail (q90–q95) remains stable, indicating a subset of persistently uncertain episodes characterized by the class-stratified analysis in Figure 3.

Appendix I. Epistemic Uncertainty by Class and Horizon

Table 6: Posterior predictive interval width (q95–q5) by horizon and outcome class.

Class	Horizon	Mean	Median	q90	q95
y=0	6h	0.032567	0.025689	0.068168	0.079425
y=0	12h	0.029884	0.021966	0.068452	0.083485
y=0	24h	0.025799	0.016441	0.065811	0.083452
y=0	48h	0.022511	0.012959	0.060253	0.078685
y=1	6h	0.056440	0.051876	0.097837	0.111306
y=1	12h	0.061233	0.057353	0.103024	0.112199
y=1	24h	0.066291	0.071038	0.103363	0.120587
y=1	48h	0.070139	0.077181	0.101046	0.114001

Appendix J. Calibration by Specialist Agreement Category

Table 7: Ensemble calibration stratified by specialist agreement category and horizon. Actual Rate: observed mortality rate. Predicted Prob: mean ensemble predicted probability. Calibration Error: absolute difference. The 6h Agree High subgroup ($n = 33$) shows the largest calibration error (0.231), consistent with sample size instability.

Horizon	Category	N	Actual Rate	Predicted Prob	Calib. Error
48h	Agree High	124	0.669	0.674	0.005
	Conflict	394	0.279	0.311	0.032
	Agree Low	1930	0.038	0.044	0.006
24h	Agree High	105	0.543	0.648	0.105
	Conflict	399	0.306	0.324	0.018
	Agree Low	1892	0.044	0.052	0.008
12h	Agree High	72	0.528	0.620	0.092
	Conflict	346	0.249	0.331	0.082
	Agree Low	1874	0.067	0.065	0.002
6h	Agree High	33	0.394	0.625	0.231
	Conflict	213	0.272	0.326	0.054
	Agree Low	1591	0.077	0.072	0.005

Appendix K. Sensitivity of Clinical-Notes Contribution Share to Explicit Goals-of-Care Language

To assess whether the horizon-wise modality-contribution pattern was driven by explicit code-status or goals-of-care language, we removed the terms *DNR*, *DNI*, *do not resuscitate*, *do not intubate*, *comfort care*, *palliative*, *hospice*, and *goals of care* from each horizon-specific clinical-notes document using case-insensitive phrase matching. We then reran the frozen clinical-notes specialist and applied the original frozen 48h LR stacker using unchanged

time-series predictions and fusion parameters. Table 8 compares the original and stripped-note modality-complementarity results on the test sets.

Table 8: Sensitivity of modality complementarity to explicit code-status and goals-of-care language. Flagged episodes contain at least one target term. CN Share is the median per-episode clinical-notes contribution share. Pearson r is the correlation between time-series and clinical-notes specialist logits. Stripping produced negligible changes and preserved the higher relative notes contribution and lower specialist correlation at earlier horizons.

Horizon	Flagged episodes N (%)	CN Share Original	CN Share Stripped	Δ Share	Pearson r Original	Pearson r Stripped
6h	70 (3.81)	0.486	0.487	+0.001	0.308	0.311
12h	107 (4.67)	0.438	0.438	+0.000	0.457	0.458
24h	146 (6.09)	0.394	0.393	−0.001	0.629	0.629
48h	194 (7.92)	0.370	0.371	+0.001	0.696	0.696

Appendix L. Additional Baseline Comparisons

L.1. Horizon-Specific Retraining

To quantify the performance trade-off associated with zero-shot horizon transfer, we retrained both specialists and the LR fusion layer independently at 6h, 12h, and 24h using the corresponding horizon-specific training and validation data. Table 9 reports held-out test discrimination for the retrained time-series specialist, clinical-notes specialist, and multimodal ensemble. All results are shown as mean [95% confidence interval] from 1,000 bootstrap resamples.

Table 9: Discrimination of horizon-specific retrained models on the held-out test sets. All specialists and the LR fusion layer were retrained independently at each indicated horizon. Results are shown as mean [95% bootstrap confidence interval].

Horizon	Model	AUROC	AUPRC
6h	TS (LSTM)	0.763 [0.731,0.799]	0.307 [0.250,0.372]
	CN (ft-CMB)	0.746 [0.713,0.780]	0.248 [0.199,0.301]
	Ensemble	0.795 [0.764,0.826]	0.336 [0.273,0.406]
12h	TS (LSTM)	0.798 [0.768,0.824]	0.367 [0.307,0.434]
	CN (ft-CMB)	0.786 [0.755,0.813]	0.354 [0.294,0.412]
	Ensemble	0.828 [0.801,0.852]	0.413 [0.346,0.475]
24h	TS (LSTM)	0.809 [0.783,0.835]	0.375 [0.319,0.436]
	CN (ft-CMB)	0.841 [0.819,0.864]	0.458 [0.399,0.515]
	Ensemble	0.861 [0.840,0.883]	0.486 [0.424,0.547]

Relative to the frozen zero-shot ensemble reported in Table 2, horizon-specific retraining improved AUROC by 0.018, 0.023, and 0.008 and AUPRC by 0.030, 0.031, and 0.005 at 6h, 12h, and 24h, respectively. The differences were modest, and by 24h the retrained and zero-shot ensembles were nearly equivalent.

L.2. Horizon-Specific SOFA-h Clinical Baseline

Traditional ICU severity and organ-dysfunction scores include APACHE II, SAPS II, and SOFA Knaus et al. (1985); Le Gall et al. (1993); Vincent et al. (1996). As a clinical reference, we evaluated horizon-specific complete SOFA-h. Unlike the reduced 17-variable benchmark representation used by the time-series specialist, SOFA-h was computed from all six standard SOFA organ-system subscores derived from available MIMIC-III physiologic, laboratory, and vasopressor data: respiratory (PaO_2/FiO_2), coagulation (platelets), liver (bilirubin), cardiovascular (vasopressor requirements), neurologic (Glasgow Coma Scale), and renal (creatinine/urine output). The derivation followed the organ-system definitions used in the MIT-LCP MIMIC severity-score concept implementation Johnson et al. (2018). For each horizon h, SOFA-h used the maximum severity observed during the first h hours.

Table 10: Discrimination of the horizon-specific SOFA-h clinical baseline on the held-out test sets. SOFA-h was computed using all six organ-system subscores and only information available within the indicated horizon. Results are shown as mean [95% bootstrap confidence interval].

Horizon	AUROC	AUPRC
6h	0.597 [0.549,0.646]	0.200 [0.153,0.251]
12h	0.641 [0.601,0.681]	0.235 [0.190,0.285]
24h	0.652 [0.613,0.691]	0.266 [0.216,0.324]
48h	0.690 [0.650,0.729]	0.305 [0.251,0.365]

The zero-shot multimodal ensemble exceeded SOFA-h by 0.180, 0.164, 0.201, and 0.201 AUROC at 6h, 12h, 24h, and 48h, respectively, and also achieved higher AUPRC at every horizon. The AUROC confidence intervals did not overlap at any horizon.

L.2.1. Availability of SOFA-h Component Measurements

Table 11 summarizes physical measurement missingness for the variables required to calculate the SOFA-h subscores. Each episode was counted as missing when no corresponding measurement was available within the indicated horizon. Laboratory availability was the primary early-horizon constraint, particularly for bilirubin, creatinine, and platelets.

Table 11: Physical missingness of measurements required for SOFA-h calculation. Cells report the number of episodes without an available measurement divided by the horizon-specific cohort size, with the corresponding percentage in parentheses.

Measurement	6h ($N = 11{,}881$)	12h ($N = 14{,}888$)	24h ($N = 15{,}554$)	48h ($N = 15{,}861$)
GCS (neurologic)	204 (1.7%)	134 (0.9%)	72 (0.5%)	37 (0.2%)
PaO_2/FiO_2 (respiratory)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
MAP (cardiovascular)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Bilirubin (liver)	9,346 (78.7%)	10,543 (70.8%)	9,881 (63.5%)	9,050 (57.1%)
Creatinine (renal)	3,716 (31.3%)	2,039 (13.7%)	197 (1.3%)	38 (0.2%)
Platelets (coagulation)	3,773 (31.8%)	2,121 (14.2%)	302 (1.9%)	53 (0.3%)

At 6h, bilirubin was unavailable in 78.7% of episodes, while creatinine and platelets were unavailable in 31.3% and 31.8%, respectively. Their availability improved as the observation window increased, whereas respiratory and cardiovascular measurements were available for all episodes at every evaluated horizon.

L.3. Previously Reported MIMIC-III Benchmarks

For additional context, the 48h ensemble achieved AUROC 0.891 and AUPRC 0.565, compared with previously reported MIMIC-III in-hospital mortality results of AUROC 0.870 and AUPRC 0.533 Harutyunyan et al. (2019) and AUROC 0.865 and AUPRC 0.525 Khadanga et al. (2019). These comparisons are descriptive because differences in modality availability and cohort construction may affect the evaluated populations.

Appendix M. Brier Scores by Horizon

Table 12 reports probabilistic accuracy across horizons. Pre and Post denote values before and after the same post-hoc recalibration procedure used for ECE.

Table 12: Brier scores before and after recalibration for the multimodal ensemble and both modality specialists. Lower values indicate better probabilistic accuracy.

Model	Horizon	Brier (Pre)	Brier (Post)
Ensemble	48h	0.068	0.068
	24h	0.076	0.076
	12h	0.084	0.084
	6h	0.084	0.084
TS (LSTM)	48h	0.075	0.075
	24h	0.086	0.083
	12h	0.097	0.088
	6h	0.104	0.089
CN (ft-CMB)	48h	0.105	0.072
	24h	0.110	0.078
	12h	0.114	0.087
	6h	0.105	0.087

Appendix N. MCMC Convergence Diagnostics

We reran the Bayesian fusion-layer model using three chains, each with 500 warmup iterations followed by 1,000 retained draws, and a target acceptance probability of 0.8. Table 13 reports parameter-level convergence diagnostics. All $\widehat{R}$ values were below 1.01, bulk effective sample sizes exceeded 1,555, and no divergent transitions occurred, supporting stable posterior estimation.

Table 13: NUTS convergence diagnostics for the Bayesian fusion layer. Bulk ESS denotes bulk effective sample size. Three chains were run with 500 warmup iterations and 1,000 retained draws per chain; no divergent transitions were observed.

Parameter	$\widehat{R}$	Bulk ESS
TS weight	1.0006	1555.7
CN weight	1.0004	1570.6
Intercept	1.0000	1680.5

Appendix O. Code Availability

Source code supporting the study-specific evaluations is available at https://github.com/Alexbav0304/early-horizon-multimodal-icu.