SOFA-Reconstruction Deep Markov Model (SR-DMM) for Dynamic ICU Mortality Prediction with Clinical Notes

Anonymous Author(s)
Affiliation
Address
email

Abstract

Timely risk updates are critical in the ICU, where rapid physiologic changes and sparse, irregular documentation make multimodal early warning challenging. We propose SR-DMM, a SOFA-reconstruction multimodal Deep Markov Model that fuses noisy structured time series with asynchronous clinical notes in a probabilistic state-space framework to update 24-hour mortality risk hourly using information available up to time t. SR-DMM reconstructs SOFA-defining raw indicators (Raw6) via a generative emission objective, encouraging physiologically meaningful latent severity trajectories while integrating notes at the encoder/posterior to directly refine state inference without heuristic decay or late fusion. In 5-fold nested cross-validation, SR-DMM improves performance over common baselines.

1 Introduction

Triage and the allocation of limited medical resources across the emergency department–to–ICU continuum are closely linked to patient outcomes. The ICU, in particular, requires high-frequency monitoring and timely interventions, motivating risk prediction tools that can identify impending deterioration early enough to guide monitoring intensity and therapeutic resources [8].

ICU prediction is challenging because clinical modalities differ in temporal characteristics: device-derived vital signs and some laboratory measurements are relatively regular time series, whereas clinical notes are free-text documents recorded irregularly. Although notes can contain condensed expert assessments, integrating them with time series raises key issues in temporal alignment and aggregation [3]. Moreover, real-world decision making often depends on near-term risk; accordingly, early-warning benchmarks formulate the task as predicting deterioration or death within the next 24 hours at each time point [2].

Purely discriminative training on a binary outcome may overemphasize predictive cues without representing the underlying evolution of physiology. Generative state-space approaches instead model observations as arising from latent clinical-state dynamics. In this context, the Deep Markov Model (DMM) parameterizes transition and emission distributions with neural networks and uses variational inference to learn nonlinear state-space models [9].

The Sequential Organ Failure Assessment (SOFA) score summarizes organ dysfunction, and serial SOFA changes have been associated with prognosis; Sepsis-3 operationalizes sepsis as infection with organ dysfunction defined by an increase in SOFA (≥ 2) [5–7]. These foundations suggest that ICU prognosis should be interpreted not only through mortality risk, but also along a clinically meaningful organ-failure severity axis.

In this study, we propose a dynamic prediction model that integrates hourly aligned structured time series (vital signs, laboratory results, and interventions) with irregular clinical notes in an MIMIC-IV ICU cohort to estimate, at each time point t, the probability of death within the subsequent 24 hours. The model is designed to reconstruct raw physiologic indicators underlying SOFA, encouraging latent states to align with organ-dysfunction severity while supporting early-warning mortality prediction.

We align notes to the hourly grid and incorporate them directly into the inference encoder so that narrative context can influence latent-state estimation, rather than injecting text via fixed heuristics such as time-decay functions. By integrating multimodal evidence within a coherent state-space perspective, we aim to improve the reliability and clinical meaningfulness of dynamically updated mortality risk estimates in noisy, missingness-prone ICU settings. We provide an open-source implementation of SR-DMM to support reproducible ICU research at (https://anonymous.4open.science/r/sr-dmm-36BA/).

2 Related Work

2.1 ICU Prediction and Research Gap

Dynamic ICU early-warning has been benchmarked as an hourly prediction task that estimates acute deterioration or death within the next 24 hours, enabling standardized evaluation of time-varying risk models [2]. Alongside risk prediction, severity scoring has been widely operationalized using SOFA, a standardized index of organ dysfunction [5], with serial SOFA trajectories associated with prognosis [6]. DeepSOFA further demonstrated that feeding the *SOFA-defining raw physiologic variables* directly (rather than discrete SOFA scores) can reduce discretization-related information loss and better capture temporal deterioration patterns for hourly mortality risk estimation [8]. Our work shares the motivation of tracking severity over time, while extending it by (i) learning latent states under an explicit **Raw6 reconstruction** objective and (ii) integrating clinical text into the same state-inference process.

In parallel, clinical notes provide clinician interpretations and context that are difficult to infer from structured signals alone, but must be mapped into fixed-dimensional representations to be used by predictive models. A common fusion pipeline embeds each note and aligns/aggregates note embeddings into per-timepoint vectors along the ICU timeline; for example, Khadanga et al. used CNN-based note embeddings and constructed time-specific text representations via an exponentially decayed weighted sum of past notes, reporting improved ICU prediction performance [3]. To address irregular sampling and missingness in ICU time series more broadly, latent-variable generative models—including DMM-family approaches—represent patient status as a latent trajectory and learn it via amortized variational inference; Krishnan et al. introduced structured inference networks with a **smoothing-style variational posterior** that leverages future observations during training to stabilize latent-state estimation in nonlinear state-space models [9].

Despite reported gains from time series–note fusion, existing methods often implement the temporal contribution of notes through heuristics. In particular, decay-based aggregation specifies a monotonic recency assumption and introduces sensitivity to the tuned decay parameter λ [3]. Moreover, when text is appended only at the prediction stage, the mechanism by which notes—as clinician interpretations of patient state—modify **state inference** and representation learning remains structurally ambiguous. This limitation is amplified by the sparsity and heterogeneity of documentation: discontinuous note injection can induce artifacts, encouraging over-reliance on note-present time points or leading the model to ignore text altogether. Finally, analyses that explicitly validate how text alters inferred states (e.g., physiologic consistency or severity-aligned trends) remain limited. Consequently, the underlying mechanism through which clinical text refines latent-state inference—and whether improvements reflect true state refinement versus documentation-driven artifacts—remains insufficiently characterized.

3 Data and Preprocessing

We constructed a multimodal ICU time-series cohort from MIMIC-IV and MIMIC-IV-Note by integrating structured signals (vitals, labs, interventions) and unstructured clinical notes. We adopt an hourly **dynamic prediction** setting that updates the risk of death within the next 24 hours using information available up to time t. The pipeline includes cohort filtering, hourly discretization, Raw6

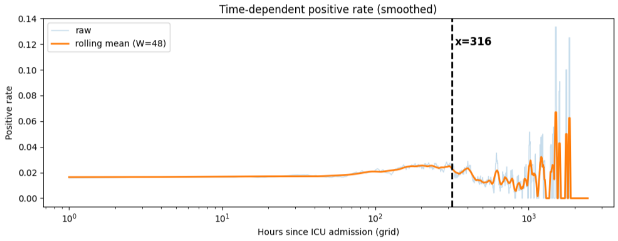

Figure 1: **Temporal changes in positive rate.** Raw time-specific positive rate (light-blue) and a 48-hour rolling mean (orange). After 316 hours (vertical dashed line), fluctuations increase due to the sharp reduction in remaining stays.

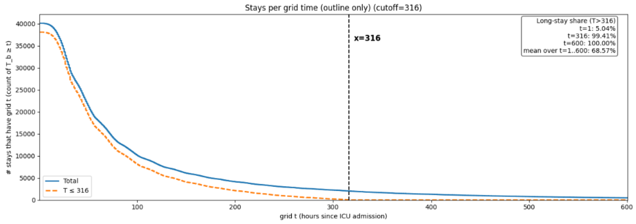

Figure 2: **Distribution of remaining ICU patients over time.** The number of patients still in the ICU decreases substantially by 316 hours (vertical dashed line), motivating the LOS cap for stability.

definition/alignment (SOFA components), label/mask construction, and missingness + multimodal alignment [2, 8].

3.1 Cohort Construction, Hourly Grid, Labels, and LOS Cut-off

Following benchmark and prior multimodal ICU work [2, 3], we retained only each patient's first ICU stay (leakage control), excluded ICU LOS <4 hours (early-window label instability), restricted clinical text to **radiology reports in MIMIC-IV-Note**, and retained only stays with **at least one non-UNK token** after tokenization (i.e., **non-UNK token count** ≥ 1 at the stay level). We also capped LOS at 316 hours to avoid a high-variance sparsity regime. The final cohort comprised 40,098 stays (3,385 deaths; 8.44% stay-level). On the hourly grid, we obtained 3,712,542 time steps, of which 70,331 (1.89%) were positive for "death within the next 24 hours."

We discretized each ICU stay into a fixed **hourly grid** anchored at ICU admission, consistent with the decompensation benchmark setup [2]. For continuous variables, we use the last measurement within each hour; for event variables (procedures/medications), we encode interval-level indicators or summary counts. We define $y_{b,t}$ as whether death occurs within the next 24 hours and use a validity mask $\nu_{b,t}$ to exclude time points without a full 24-hour horizon from loss/evaluation. Given the low base rate among valid time steps ($\approx 1.9\%$), we report PRC/AUPRC in addition to AUROC [14].

We cap ICU LOS at 316 hours to avoid a high-variance regime where few patients remain at risk. Beyond this point, the number of ongoing stays drops sharply (Figure 2) and time-specific positive-rate estimates become unstable (Figure 1), increasing sensitivity to a small long-stay subgroup and degrading training stability.

3.2 Raw6 (SOFA Components), Clinical Text, and Missingness

We define **Raw6** as the six physiologic indicators underlying SOFA and use them as reconstruction-aligned targets: respiration (PaO_2/FiO_2), coagulation (platelets), liver (bilirubin), cardiovascular (MAP/vasopressors), CNS (GCS), and renal (creatinine/urine output). Raw6 was generated using SOFA-related queries from the **MIT-LCP mimic-code** pipeline and mapped to the hourly grid [16]. This follows DeepSOFA's motivation to avoid information loss from SOFA score discretization, while differing in that Raw6 is modeled as observations to be generatively explained rather than only inputs [8].

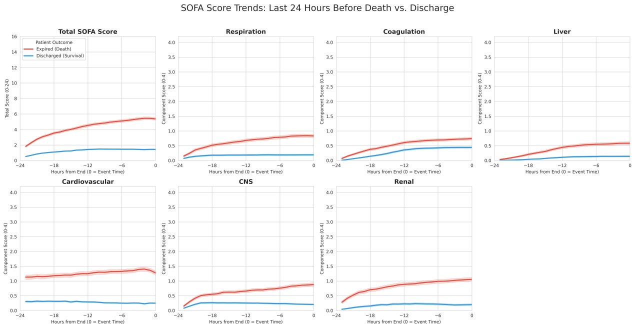

Figure 3: **Temporal trends of total and component SOFA scores in the last 24 hours.** Event-aligned mean trajectories ($t \in [-24, 0]$) show increasing severity toward death, while remaining low before discharge.

Table 1: **Input composition (grouped) and observation rates.**

Group	**Obs. rate (%)**	**Examples (type)**
Static demo.	90–100	age/sex/insurance (cat/cont)
Vitals (hourly)	30–95	HR, BP, RR, SpO_2, Temp (cont)
Labs (non-Raw6)	2–7	lactate, WBC, pH (cont)
Raw6 (SOFA comps.)	100 (binned)	P/F, PLT, bili, MAP/vaso, GCS, Cr/UO (cont/cat)
Interventions/events	0.1–40	fluids/meds/vent/proc (cat)
Note features	3–5	severity/uncertainty/abnormality (cat/cont)
Note tokens/meta	3–5	token bag + modality/region bags (emb)

For clinical text, we use **radiology reports from MIMIC-IV-Note** and align note timestamps to elapsed hours since ICU admission, mapping them onto the hourly grid with row-level identifiers [17]. We define **text existence at the stay level** as having **at least one non-UNK token** after preprocessing (non-UNK token count ≥ 1). Notes are structured with a local LLM (Ollama; `qwen2.5-coder:7b`) to produce (i) fixed-dimensional structured features and (ii) lexical tokens with negation markers and `UNK` handling. Token sequences are serialized into EmbeddingBag-compatible arrays/offsets. Exam metadata (modality/region) are normalized using the **LOINC–RSNA Radiology Playbook** and encoded as separate meta-bags to preserve imaging context [19, 20]. Text inputs are aggregated per hour and injected only when notes are present.

To handle irregular sampling, we do not forward-fill vitals/labs; instead, we provide explicit **mask variables** so models can exploit both values and availability patterns. We additionally include a text-availability indicator η_t to distinguish note-observed from note-absent intervals, following benchmark and latent-variable modeling conventions where missingness itself can be informative [2, 9].

4 Proposed Method: SOFA Reconstruction Deep Markov Model (SR-DMM)

We propose the **SOFA-Reconstruction Deep Markov Model (SR-DMM)**, a probabilistic state-space model for noisy/irregular ICU time series and sparse, asynchronous clinical notes. Given an hourly ICU trajectory for patient b with $t = 1, \dots, T_b$, SR-DMM represents the evolving condition using latent states $z_{b,t}$ and predicts the risk of death within the next 24 hours, $\hat{p}_{b,t}$. A key design is to **reconstruct Raw6** (SOFA component variables) via a generative emission model, constraining the latent space to preserve physiologic severity structure rather than collapsing into purely discriminative shortcuts.

4.1 Formulation

Structured inputs at time t are Raw6 $x_{b,t}$ with mask $m_{b,t}$, additional continuous summaries $u_{b,t}$, event features $e_{b,t}$, and static covariates s_b. Text inputs are represented by structured features $r_{b,t}$, token/meta-bag embedding $g_{b,t}$, and availability indicator $\eta_{b,t}$. The time-dependent label is

$$y_{b,t} = 1 \iff \text{death occurs in } (t,\, t + 24\text{h}], \tag{1}$$

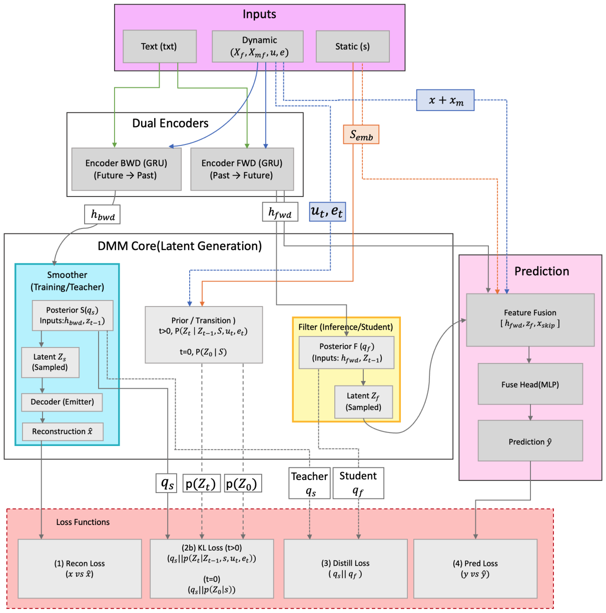

Figure 4: **Overview of SR-DMM.** SR-DMM reconstructs Raw6 and predicts 24-hour mortality risk on an hourly grid. Text is aligned to the grid and fused into the inference encoder to influence latent-state estimation.

with a validity mask $\nu_{b,t}$ to exclude time points without a full 24-hour horizon.

4.2 SR-DMM: Generative Model, Inference, Prediction, and Objective

SR-DMM defines a latent state-space model with Gaussian conditionals (diagonal covariance) parameterized by neural networks. The initial latent state depends on static covariates,

$$p(z_{b,1} \mid s_b) = \mathcal{N}\Big(\mu_{p0}(s_b),\ \mathrm{diag}(\sigma^2_{p0}(s_b))\Big), \tag{2}$$

and for $t \geq 2$ the transition depends on the previous state and covariates,

$$p(z_{b,t} \mid z_{b,t-1}, u_{b,t}, e_{b,t}, s_b) = \mathcal{N}\Big(\mu_p(\cdot),\ \mathrm{diag}(\sigma^2_p(\cdot))\Big). \tag{3}$$

Raw6 is generated from the latent state via an emission model,

$$p(x_{b,t} \mid z_{b,t}) = \mathcal{N}\Big(\mu_x(z_{b,t}),\ \mathrm{diag}(\sigma^2_x(z_{b,t}))\Big), \tag{4}$$

which induces an explicit reconstruction objective that aligns latent states with physiologic severity structure.

Because exact posteriors are intractable, SR-DMM uses variational inference with a *smoothing* posterior for training and a *filtering* posterior for online prediction. Both are parameterized by GRU encoders operating on structured inputs and projected text. The projected text vector is

$$\sqcup_{b,t} = \phi\big([r_{b,t}; g_{b,t}; \eta_{b,t}]\big), \tag{5}$$

where $\phi(\cdot)$ is a lightweight projection module, enabling end-to-end learning of how notes affect latent-state inference. The forward (filter) and backward (smoother; training only) encoders update

$$h_f(b,t) = \mathrm{GRU}_{\mathrm{fwd}}\Big(h_f(b,t-1),\ [x_{b,t}, m_{b,t}, u_{b,t}, e_{b,t}, \sqcup_{b,t}]\Big), \tag{6}$$

$$h_s(b,t) = \mathrm{GRU}_{\mathrm{bwd}}\Big(h_s(b,t+1),\ [x_{b,t}, m_{b,t}, u_{b,t}, e_{b,t}, \sqcup_{b,t}]\Big), \tag{7}$$

and define Gaussian posteriors

$$q_f\big(z_{b,t} \mid z_{b,t-1}, h_f(b,t), s_b\big) = \mathcal{N}\Big(\mu_{qf}(\cdot),\ \mathrm{diag}(\sigma^2_{qf}(\cdot))\Big), \tag{8}$$

$$q_s\big(z_{b,t} \mid z_{b,t-1}, h_s(b,t), s_b\big) = \mathcal{N}\Big(\mu_{qs}(\cdot),\ \mathrm{diag}(\sigma^2_{qs}(\cdot))\Big). \tag{9}$$

Online mortality risk prediction uses the filtering path. We form

$$\psi_{b,t} = [h_f(b,t);\ z_f(b,t);\ s_{\mathrm{emb}}(s_b)], \qquad \tilde{x}_{b,t} = g_{\mathrm{in}}([x_{b,t}; m_{b,t}]), \tag{10}$$

and predict

$$\hat{p}_{b,t} = \sigma\Big(f\big([\psi_{b,t}; \tilde{x}_{b,t}]\big)\Big), \tag{11}$$

where $f(\cdot)$ is an MLP and $\sigma(\cdot)$ is the sigmoid.

Training jointly optimizes Raw6 reconstruction, KL regularization, mortality prediction, and distillation from the smoother (teacher) to the filter (student):

$$\mathcal{L} = \mathcal{L}_{\mathrm{recon}} + \beta\,\mathcal{L}_{\mathrm{KL}} + \lambda_{\mathrm{death}}\,\mathcal{L}_{\mathrm{death}} + \lambda_{\mathrm{dist}}\,\mathcal{L}_{\mathrm{distill}}. \tag{12}$$

Reconstruction is computed only over observed Raw6 dimensions using $m_{b,t}$, mortality BCE is computed only on valid time points using $\nu_{b,t}$, and distillation matches q_f to q_s while stopping gradients into the teacher.

5 Experimental Setup and Evaluation

We evaluate ICU mortality forecasting as an hourly **dynamic prediction** task. At each time t, the model estimates

$$\hat{p}_{b,t} = \Pr\big(y_{b,t} = 1 \mid \mathcal{I}_{b,t}\big), \tag{13}$$

where $y_{b,t} = 1$ indicates death within the next 24 hours and $\mathcal{I}_{b,t}$ denotes the information available up to t [2]. We use a validity mask $\nu_{b,t}$ to exclude time points without a full 24-hour horizon and compute losses/metrics only over $\{(b,t) : \nu_{b,t} = 1\}$.

To reduce model-selection bias, we employ **nested cross-validation**: 5-fold patient-level *outer* splits for final testing and 3-fold *inner* splits for hyperparameter selection via Optuna [13]. Hyperparameters are selected by maximizing inner-validation **AUPRC**, after which the model is refit on the full outer training set and evaluated on the held-out outer test set. Results are reported as mean $\pm$ SD across outer folds. All experiments were conducted on a high-performance Linux computing environment equipped with an **Intel Xeon Gold 6334 CPU (3.60GHz), 1TB RAM, and NVIDIA A30 GPUs** to ensure reproducibility and computational efficiency.

Because positive events are rare, **AUPRC** is the primary metric, with **AUROC** reported as a complementary measure [14]. All metrics are computed over valid time points only.

We compare SR-DMM against multiple baselines and ablations under identical preprocessing, inputs, masking, and evaluation. Specifically, we include (i) a heuristic-decay late-fusion baseline that aggregates past note embeddings with a tuned exponential decay and fuses them with an LSTM encoding of structured time series [3], (ii) a deterministic-fusion GRU baseline that concatenates a per-timepoint projected text representation with the GRU hidden state for prediction, and (iii) a DMM ablation without text that preserves Raw6 reconstruction, latent transition/emission structure, inference encoders, prediction head, and objective. The proposed SR-DMM injects text directly into the inference encoder (filter/smoother) at each time point without heuristic decay and reconstructs Raw6 to constrain latent severity states; the prediction head and objective follow Section 4.

6 Experimental Results

We summarize quantitative performance, calibration, trajectory behavior, and attribution/representation analyses in a single unified view. Table 2 reports test performance under 5-fold nested cross-validation: the proposed **Multimodal DMM** achieves the highest mean **PR-AUC** with relatively low fold-to-fold variability, suggesting a stable gain under severe class imbalance.

Table 2: **Quantitative comparison of dynamic mortality prediction performance on the test set under 5-fold nested cross-validation (mean ± standard deviation).**

Model	Architecture	Text Integration Strategy	PR-AUC	ROC-AUC
Baseline 1	LSTM	Heuristic Decay [3]	0.385 ± 0.013	0.913 ± 0.005
Baseline 2	GRU	Encoder Input	0.387 ± 0.020	0.914 ± 0.008
Ablation	DMM	None	0.398 ± 0.020	0.917 ± 0.007
Proposed	DMM	Encoder Input	$\mathbf{0.410 \pm 0.013}$	$\mathbf{0.923 \pm 0.003}$

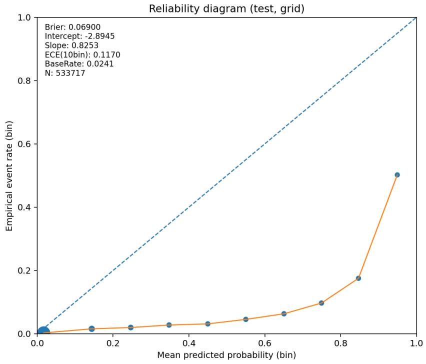

Figure 5: **Reliability diagram on the test set (grid-level).** In certain probability ranges, empirical event rates are lower than mean predicted probabilities, indicating mild **overestimation** (often associated with $\beta < 1$).

Beyond scalar metrics, we analyze **dynamic risk trajectories** aligned to terminal events over the last 72 hours (Figure 6). The death cohort exhibits a clear upward trend toward the event, whereas the discharge cohort remains low and tends to decrease, supporting the model output as a time-varying early-warning signal.

We additionally evaluate **probability calibration** with a grid-level reliability diagram (Figure 5) and summarize calibration via the logistic calibration model

$$\text{logit}(y) = \alpha + \beta \cdot \text{logit}(p). \tag{14}$$

Calibration is computed at the **grid level**, so long-stay patients contribute more time points; thus calibration should be interpreted alongside intended threshold-based clinical use.

To understand model reliance on input pathways, we use IG/SHAP-style attribution (Figure 7). IG indicates that the hidden state **H** contributes most strongly, consistent with recurrent accumulation of longitudinal structured signals and note-derived context; static variables (S) and the skip connection (X_FUSE) also contribute. The *direct* contribution of the latent variable **Z** is smaller, aligning with the interpretation that z_t primarily regularizes/structures representations through Raw6 reconstruction rather than serving as a direct discriminative shortcut. Finally, we quantify how notes alter state inference via a counterfactual **representation shift** under text removal,

$$\Delta h_t = \left\| h_t^{\text{text}} - h_t^{\text{no-text}} \right\|_2, \tag{15}$$

and the associated prediction shift $\Delta p_t = |p_t^{\text{text}} - p_t^{\text{no-text}}|$ (Figure 8). Across alignments, shifts are larger in the death cohort, indicating stronger text-driven state updates along higher-risk trajectories. The relationship between Δh_t and Δp_t is further visualized with hexbin plots (Figure 9), showing a positive association that is stronger in the death cohort. A case study (Figure 10) illustrates that Δh_t spikes at note times and then decays, consistent with note-triggered encoder updates; text can shift risk trajectories earlier/more sharply in deterioration cases while leaving low-risk cases largely unchanged.

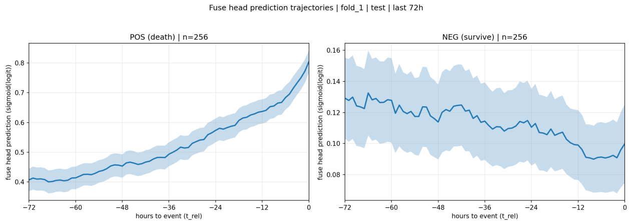

Figure 6: **Dynamic mortality risk prediction trajectories in the last 72 hours prior to event.** Trajectories are aligned such that the event time is $t = 0$ and visualized over $t_{\text{rel}} \in [-72, 0]$. Solid curves denote means and shaded regions indicate confidence intervals.

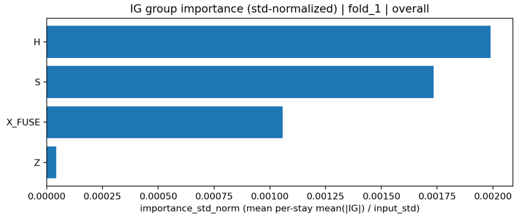

Figure 7: **Overall Integrated Gradients importance by input group.** IG importance is aggregated by pathway (H, S, X_FUSE, Z) using per-stay mean $|\text{IG}|$ normalized by input standard deviation. Larger values indicate higher direct output sensitivity to that pathway.

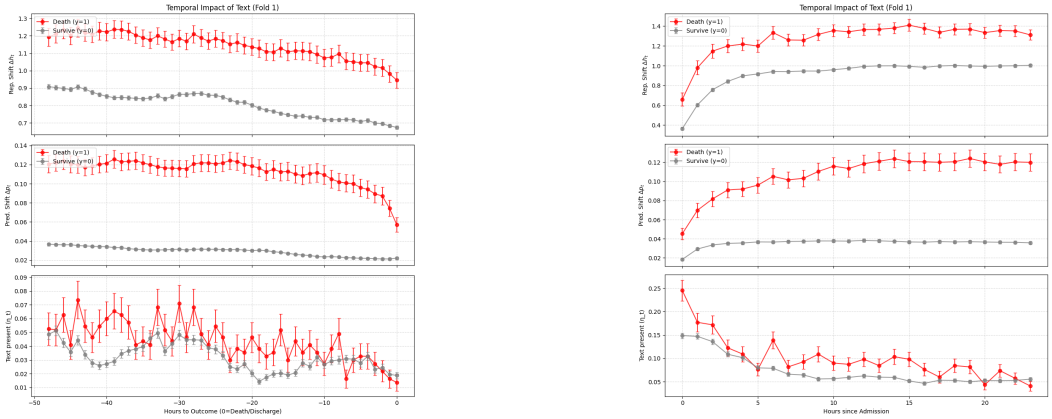

Figure 8: **Temporal impact of text.** (a) Outcome-aligned summaries over $[-48, 0]$ hours relative to death/discharge. (b) Admission-aligned summaries over $[0, 24]$ hours (LOS $\geq$ 48h). Top: Δh_t. Middle: Δp_t. Bottom: text availability η_t. Curves are shown for death ($y = 1$) and survival ($y = 0$).

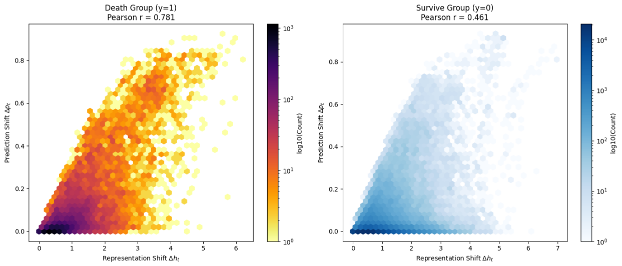

Figure 9: **Relationship between representation shift and prediction shift (Δh–Δp correlation).** Hexbin plots show the association between Δh_t and Δp_t under counterfactual removal of text (color indicates $\log_{10}$ frequency). Panels are stratified by outcome and Pearson correlation coefficients are reported.

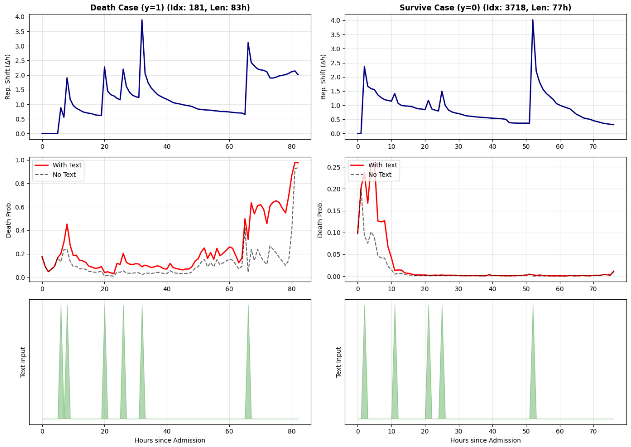

Figure 10: **Case study: text-driven latent-state updates and risk trajectory under counterfactual text removal.** Each panel shows (top) Δh_t, (middle) predicted mortality risk (with text vs no text), and (bottom) text input times for (a) a death case and (b) a discharge case.

7 Discussion

Prior multimodal ICU time-series studies often align clinical notes to structured signals using **decay-weighted aggregation** with a tuned decay parameter λ. While straightforward, such heuristics do not fully reflect the **irregular and sparse** occurrence of notes or the **heterogeneous documentation patterns** across clinicians and settings, and they typically inject text only at the prediction stage. In contrast, SR-DMM integrates text at the **encoder level**, so that notes directly influence **latent-state inference** within the probabilistic state-space model. This design enables end-to-end learning of how textual context calibrates the evolving patient state, avoiding fixed monotonic decay assumptions.

Beyond fusion, SR-DMM introduces an auxiliary objective that **reconstructs Raw6**—the physiologic variables underlying SOFA—to encourage the latent states to align with clinically meaningful **severity axes** and to reduce reliance on purely discriminative shortcuts. Consistent with this motivation, predicted risk trajectories exhibit clear separation between non-survivors and survivors, with non-survivors showing sharper risk escalation as endpoints approach, supporting the model output as a time-varying early-warning signal.

In summary, we proposed **SR-DMM**, a SOFA-reconstruction–based multimodal Deep Markov Model for hourly dynamic prediction of 24-hour mortality risk using information available up to time t. The model improves performance over comparative baselines while unifying irregular structured signals and sparse clinical notes in a coherent state-space framework. Key methodological elements include **encoder-level text fusion** (instead of late fusion with decay heuristics [3]) and a **distillation** term that transfers the advantages of smoothing-based training to filtering-time inference. Future work includes external validation beyond MIMIC-IV and leveraging posterior uncertainty (e.g., variance) for uncertainty-aware early warning and downstream clinical decision support.

References

[1] Johnson AEW, Pollard TJ, Shen L, et al. **MIMIC-IV, a freely accessible electronic health record dataset.** *Scientific Data*. 2023;10:1.

[2] Harutyunyan H, Khachatrian H, Kale DC, et al. **Multitask learning and benchmarking with clinical time series data.** *Scientific Data*. 2019;6:96.

[3] Khadanga S, Aggarwal K, Joty S, et al. **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)*. 2019. p. 6432–6437.

[4] Thorsen-Meyer H-C, Nielsen AB, Nielsen AP, et al. **Dynamic and explainable machine learning prediction of mortality in intensive care: a retrospective study of time-series data and ensemble learning.** *The Lancet Digital Health*. 2020;2(6):e179–e191.

[5] Vincent J-L, Moreno R, Takala J, et al. **The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure.** *Intensive Care Medicine*. 1996;22(7):707–710.

[6] Ferreira FL, Bota DP, Bross A, Mélot C, Vincent J-L. **Serial evaluation of the SOFA score to predict outcome in critically ill patients.** *JAMA*. 2001;286(14):1754–1758.

[7] Singer M, Deutschman CS, Seymour CW, et al. **The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3).** *JAMA*. 2016;315(8):801–810.

[8] Shickel B, Tighe PJ, Bihorac A, Rashidi P. **DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning.** *Scientific Reports*. 2019;9:1879.

[9] Krishnan RG, Shalit U, Sontag D. **Structured inference networks for nonlinear state space models.** In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*. 2017. p. 2101–2109.

[10] Kingma DP, Welling M. **Auto-Encoding Variational Bayes.** In: *International Conference on Learning Representations (ICLR)*. 2014. (arXiv:1312.6114).

[11] Hinton GE, Vinyals O, Dean J. **Distilling the knowledge in a neural network.** arXiv preprint arXiv:1503.02531. 2015.

[12] Lundberg SM, Lee S-I. **A unified approach to interpreting model predictions.** In: *Advances in Neural Information Processing Systems (NeurIPS) 30*. 2017. p. 4765–4774.

[13] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. **Optuna: A next-generation hyperparameter optimization framework.** In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*. 2019. p. 2623–2631.

[14] Davis J, Goadrich M. **The relationship between Precision-Recall and ROC curves.** In: *Proceedings of the 23rd International Conference on Machine Learning (ICML)*. 2006. p. 233–240.

[15] Che Z, Purushotham S, Cho K, Sontag D, Liu Y. **Recurrent neural networks for multivariate time series with missing values.** *Scientific Reports*. 2018;8:6085.

[16] MIT Laboratory for Computational Physiology. **mimic-code: MIMIC Code Repository** [Internet]. GitHub repository; Accessed 2026-01-30.

[17] van Veen D, et al. **Adapted large language models can outperform medical experts in clinical text summarization.** *Nature Medicine*. 2024.

[18] Yang X, et al. **GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records.** *npj Digital Medicine*. 2022.

[19] Vreeman DJ, Abhyankar S, Wang KC, Carr C, Collins B, Rubin DL, Langlotz CP. **The LOINC RSNA radiology playbook – a unified terminology for radiology procedures.** *Journal of the American Medical Informatics Association*. 2018;25(7):885–893.

[20] Regenstrief Institute (LOINC). **LOINC–RSNA Radiology Playbook (RSNA collaboration)** [Internet]. Web resource; Accessed 2026-01-30.

[21] Özyurt Y, Kraus M, Hatt T, Feuerriegel S. **AttDMM: An Attentive Deep Markov Model for Risk Scoring in Intensive Care Units.** In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2021. p. 3452–3462.

AI Co-Scientist Challenge Korea Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the proposed SR-DMM, the target task (hourly 24-hour mortality prediction), multimodal fusion with clinical notes, and the auxiliary Raw6 reconstruction objective, which match the contributions evaluated in Sections 5–6.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 (Discussion) acknowledges limitations including single-center evaluation on MIMIC-IV and points to future work such as external validation and uncertainty-aware extensions (e.g., leveraging posterior uncertainty).

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical theorems requiring proofs; it provides methodological and model formulations instead.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: The paper describes the cohort definition, hourly discretization, label/mask setup, and the nested cross-validation evaluation protocol (Sections 3 and 5). However, it does not fully specify the complete implementation details needed to faithfully reproduce the reported numbers (e.g., exact preprocessing scripts and configurations, full hyperparameter search space and selected values, optimizer/training schedules, and random seeds).

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: While the MIMIC-IV and MIMIC-IV-Note datasets are available through credentialed access via PhysioNet , we provide our full codebase for data preprocessing and model training at (https://anonymous.4open.science/r/sr-dmm-36BA/) to ensure the reproducibility of our findings.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [No]

   Justification: The paper specifies the nested cross-validation design (5-fold outer, 3-fold inner), the use of Optuna for model selection, and the primary evaluation metrics (AUPRC/AUROC) in Section 5. However, key training details required to fully understand and replicate training are not explicitly listed (e.g., optimizer type, learning rate schedule, batch size, number of epochs/early stopping, regularization settings, and the Optuna search space and selected hyperparameter values).

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 2 reports mean ± standard deviation across outer cross-validation folds, which captures split-to-split variability for the main reported metrics. Figures 6 and 8 also visualize trajectory summaries with shaded uncertainty bands (described as confidence intervals in the captions), although the exact computation procedure is not detailed.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Section 5 reports the compute environment (Intel Xeon Gold 6334 CPU @ 3.60GHz, 1TB RAM, NVIDIA A30 GPUs), but does not provide execution time per run or total compute.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://nips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses de-identified ICU EHR data accessed under standard credentialed procedures and focuses on clinical risk prediction/interpretability. It does not involve new data collection, subject recruitment, or deployment, and aims to support safer clinical decision-making rather than harmful applications.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper motivates potential positive impacts such as improved early warning and resource allocation in ICUs (Sections 1 and 7). However, it does not explicitly discuss potential negative societal impacts (e.g., bias/fairness concerns, misuse, over-reliance risks, deployment harms) and mitigation strategies.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk generative model or a newly scraped dataset; it uses existing credentialed-access clinical data and reports an experimental study.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The manuscript credits key external assets via citations (e.g., MIMIC-IV / MIMIC-IV-Note and mimic-code in the References). However, it does not explicitly list the license names/versions or the full terms-of-use for each asset in the paper (beyond general credentialed access expectations), so the checklist requirement of explicitly stating licenses/terms is not fully met.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release an anonymized code repository for SR-DMM (linked in the main paper) that includes the implementation and supporting scripts. The repository provides documentation (e.g., README) describing prerequisites, data access requirements for MIMIC-IV (no redistribution), and example commands/configuration needed to run training and evaluation.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work is a secondary analysis of de-identified EHR data and does not involve crowdsourcing or new human-subject recruitment.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This is a secondary analysis of de-identified MIMIC-IV data accessed under the PhysioNet data use agreement process; the study does not involve new participant recruitment or intervention.