

Original Research

Multimodal Temporal Representation Learning for Early ICU Deterioration Forecasting from Physiologic Streams and Clinical Narratives

Ahmad Fauzan¹, Muhammad Rizqi² and Hafiz Ramadhan³

¹Department of Informatics Engineering, Universitas Islam Negeri Sunan Gunung Djati Bandung, Jalan A.H. Nasution No. 105, Bandung 40614, Indonesia.

²Department of Information Systems, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Jalan Gajayana No. 50, Malang 65144, Indonesia.

³Department of Computer Science, Universitas Islam Negeri Sultan Syarif Kasim Riau, Jalan HR. Soebrantas No. 155, Pekanbaru 28293, Indonesia.

Abstract

Intensive care environments generate dense streams of physiologic observations, laboratory measurements, and narrative documentation. Converting these heterogeneous signals into reliable short-horizon forecasts remains difficult because clinical trajectories are irregularly sampled, event prevalence is low, and much of the most informative context is embedded in free text rather than in structured fields. These characteristics make deterioration forecasting a representation learning problem as much as a clinical one. This paper develops a multimodal deep learning framework for early prediction of patient deterioration in the intensive care unit by combining structured time-series electronic health record data with contextualized embeddings from contemporaneous radiology reports. The study uses 74,822 adult ICU stays and 5,720,904 hourly prediction samples, with deterioration defined as death in the ICU, vasopressor initiation, or invasive mechanical ventilation within the subsequent 24 hours. The proposed system uses a bidirectional recurrent temporal encoder with attention for 48-hour physiologic windows, a frozen domain-specific language model for note representation, and gated fusion to adaptively integrate both modalities under substantial missingness. On a held-out test set of 823,641 samples, the multimodal model achieves an AUROC of 0.7857 and an AUPRC of 0.1908, outperforming a structured-only deep model, XGBoost, and logistic regression. Relative to the structured-only baseline, note integration yields a 2.5 percentage point AUROC gain and a 39.2% AUPRC increase, indicating that narrative documentation contributes complementary predictive signal even when text alone is weak. The empirical analysis further shows rapid convergence, small validation-to-test drift, and strong post-hoc recoverability of calibration, suggesting that multimodal temporal learning is a viable computational framework for real-time ICU risk stratification.

1. Introduction

The modern intensive care unit is one of the most demanding computational environments in applied machine learning [1]. At every hour, multiple sensors produce physiologic observations, laboratories are acquired asynchronously, medications and procedures alter state trajectories, and clinicians generate text that compresses interpretation, uncertainty, and expectation into concise notes. From an algorithmic perspective, the ICU is a partially observed, high-dimensional, irregularly sampled dynamical system with severe label imbalance and temporally evolving covariates. The core task is not merely pattern recognition on static features, but online inference over latent patient state from noisy multimodal evidence. This framing makes early deterioration forecasting a natural target for sequence modeling, multimodal representation learning, calibration-aware classification, and robust learning under structured missingness.

Forecasting deterioration is clinically motivated but computationally constrained. A useful model must operate on incomplete histories, remain stable when measurement frequency changes, and translate correlated but semantically distinct inputs into a single probability of near-term adverse transition. In practice, structured bedside variables alone do not fully characterize patient risk. Two patients with similar heart rate, blood pressure, oxygen saturation, and laboratory values may nonetheless differ in clinician-assessed trajectory because radiologic interpretation, concern about evolving pathology, or procedural context is recorded only in free text. Narrative documentation often contains semantic information about severity, progression, and differential diagnosis that cannot be recovered from tabular fields without explicit language modeling. Any forecasting architecture that ignores this modality accepts an information bottleneck by design.

The technical literature on ICU prediction has made substantial progress in recent years, especially through recurrent neural networks, temporal convolutional models, and gradient-boosted methods applied to structured electronic health record data. Yet most systems still treat the problem as one of temporal tabular learning only. A more complete approach should recognize that structured variables and text are not substitutes. The former capture high-frequency numeric state, whereas the latter encode low-frequency but high-semantic-density summaries of patient condition. Recent multimodal ICU deterioration modeling by Sadanandan (2025) showed that augmenting structured temporal data with clinical note representations can improve discrimination for 24-hour deterioration prediction across a large MIMIC-IV cohort, providing an empirical basis for treating the problem as cross-modal fusion rather than as univariate sequence classification [2].

This paper develops a technically focused research narrative around that multimodal perspective. The goal is not to restate deterioration prediction as a generic clinical decision support problem, but to analyze it as a computer science problem defined by asynchronous data integration, efficient representation transfer, rare-event optimization, and deployment-oriented evaluation. The study investigates how structured and unstructured modalities should be encoded, aligned, fused, regularized, and calibrated when hourly risk scores are generated across millions of windows. It also examines what can be learned from the resulting error patterns, generalization characteristics, and sensitivity to missingness. The emphasis throughout is on algorithmic design and empirical behavior rather than on speculative claims about downstream intervention effects.

The setting is large-scale and computationally relevant. The cohort contains 74,822 adult ICU stays drawn from MIMIC-IV, which expands to 5,720,904 hourly prediction instances after rolling-window sample generation [3]. Each sample includes a 48-hour history of 26 structured features, a contextualized note embedding when documentation is available, static patient descriptors, and a binary outcome indicating whether death, vasopressor initiation, or invasive mechanical ventilation begins within the next 24 hours. This setup transforms a stay-level problem into a dense temporal supervision problem in which the same patient contributes many overlapping windows, intensifying the need for patient-level splitting and careful leakage prevention. The resulting prevalence of 2.8% also makes conventional accuracy an uninformative metric, elevating the roles of precision-recall analysis, threshold selection, and calibration.

From a systems perspective, the design challenge is to build a model that is expressive enough to capture temporal physiology, cheap enough to train and evaluate on millions of samples, and robust enough to handle the fact that not every stay contains notes and not every hour contains fresh measurements. The chosen architecture therefore combines a bidirectional long short-term memory encoder with temporal attention for the structured sequence, a frozen domain-specific transformer embedding for clinical notes, and a gated fusion mechanism that dynamically interpolates between modalities. This design balances expressive power and practical tractability. The recurrent module handles sequential dependencies across 48 hours, the language model injects semantic context without full end-to-end transformer fine-tuning, and gating allows the fusion layer to degrade gracefully when the note modality is absent or weak.

The contribution of the present paper is threefold. First, it formalizes the deterioration forecasting task as a multimodal sequence classification problem with explicit treatment of observation windows, event

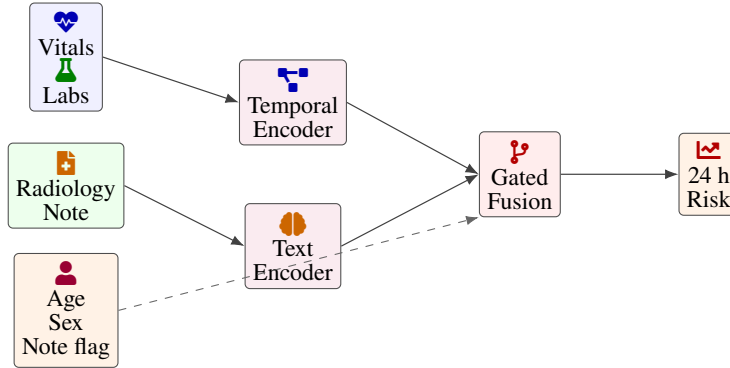


Figure 1: Overall multimodal forecasting pipeline. A 48-hour structured physiologic window and the most recent radiology report are encoded separately, then combined with static context through a gated fusion layer to produce an hourly estimate of deterioration risk over the next 24 hours.

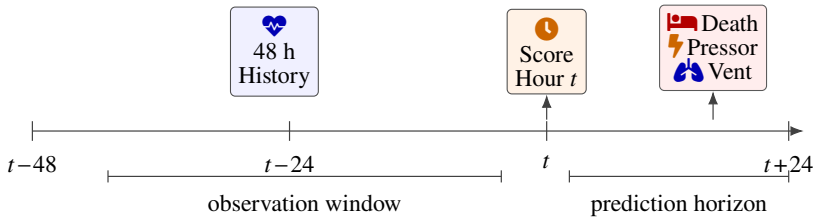


Figure 2: Hourly sample construction. For each eligible ICU hour t , the model consumes the preceding 48 hours of structured observations and predicts whether any deterioration onset occurs within $(t, t + 24]$, with the composite outcome defined by death, vasopressor initiation, or invasive ventilation initiation.

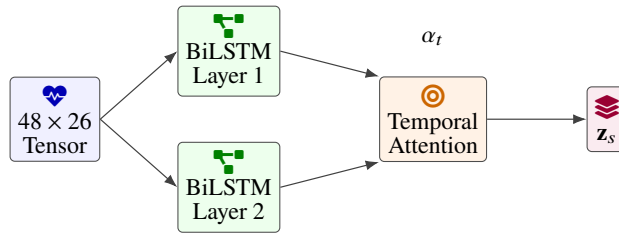


Figure 3: Structured temporal pathway. The hourly physiologic tensor is processed by stacked bidirectional recurrent layers, after which temporal attention compresses the hidden trajectory into a single structured representation $\mathbf{z}_s$ that emphasizes clinically informative hours within the 48-hour context window.

horizons, and heterogeneous missingness. Second, it provides a deeply technical account of the architecture, optimization strategy, and evaluation protocol needed to make such a system work under severe class imbalance at scale. Third, it interprets the empirical results not simply as benchmark scores but as evidence about modality complementarity, calibration recoverability, and the computational tradeoffs of deploying multimodal risk models in high-frequency hospital workflows. The paper therefore sits closer to machine learning systems research than to a narrative discussion of clinical prediction tools, even though the application domain remains critical care.

The remainder of the manuscript is organized around this computational framing. The next section formulates the forecasting problem and defines the input and label spaces. Subsequent sections describe cohort construction, temporal and textual representations, fusion and optimization, evaluation design, empirical results, and an extended interpretation of failure modes and deployment considerations. The

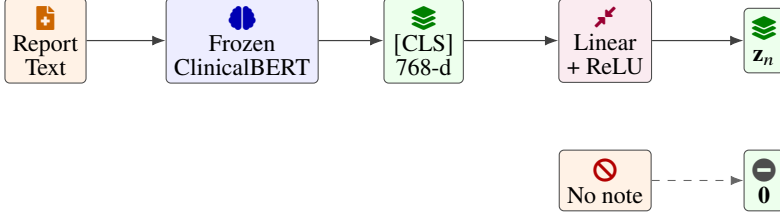


Figure 4: Clinical note pathway. The most recent radiology report is embedded with a frozen domain-specific language model, projected into the shared latent space, and used as the note representation z_n ; when no report is available, the text branch falls back to a zero vector.

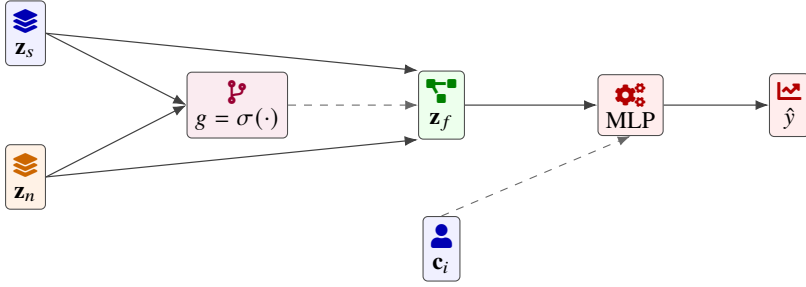


Figure 5: Gated multimodal fusion. Structured and textual representations jointly determine a sigmoid gate that modulates their contribution to the fused latent state; static descriptors are appended before the classifier head computes the final deterioration probability.

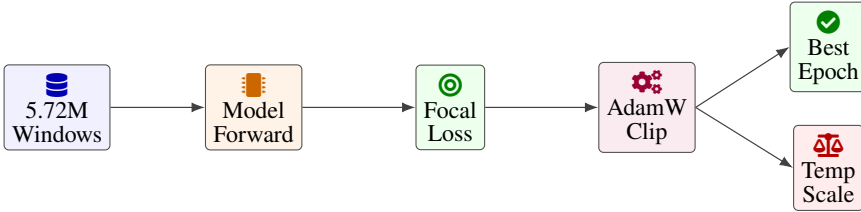


Figure 6: Training and post-hoc optimization flow. Millions of rolling windows are processed through the multi-modal network, optimized with rare-event-aware focal loss and AdamW, selected by early validation performance, and finally recalibrated through temperature scaling for probability use.

analysis remains neutral in tone: the proposed approach is not presented as definitive, but as an empirically supported design point showing that multimodal temporal learning yields measurable gains over structured-only baselines while preserving generalization and operational simplicity.

2. Problem Formulation

Consider an ICU stay indexed by i with observation times discretized to hourly bins. For each eligible hour t , the model receives a structured look-back window of length $T = 48$, a text representation derived from documentation available no later than t , and static descriptors associated with the stay. Let $\mathbf{s}_{i,t-\tau} \in \mathbb{R}^{26}$ denote the structured observation vector at lag $\tau \in \{0, \dots, T-1\}$, where the 26 dimensions correspond to selected vital signs and laboratory values after preprocessing and normalization. Let $\mathbf{c}_i \in \mathbb{R}^3$ denote static covariates composed of age, sex, and an indicator for note availability. Let $\mathbf{n}_{i,t} \in \mathbb{R}^{768}$ denote the contextualized embedding of the most recent radiology report prior to time t ,

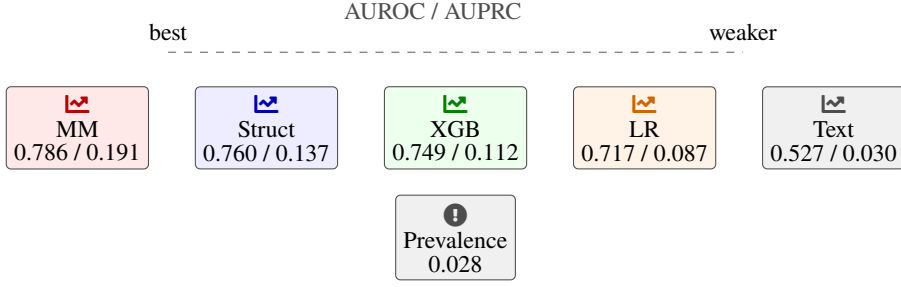


Figure 7: Compact performance summary. The multimodal model leads all baselines on both AUROC and AUPRC, with the largest practical separation appearing on precision–recall behavior, while the text-only branch remains close to the prevalence baseline when evaluated in isolation.

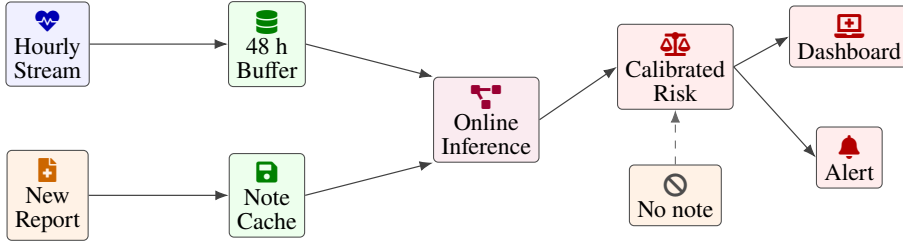


Figure 8: Deployment-oriented inference loop. Structured measurements update continuously, radiology text updates episodically, cached note representations are reused across hours, and the resulting calibrated risk score can support both ranked review dashboards and thresholded escalation logic.

with $\mathbf{n}_{i,t} = \mathbf{0}$ when no report exists. The supervised learning problem is to estimate the conditional probability of deterioration in the next $H = 24$ hours.

A label at hour t is positive if at least one clinically significant transition begins in the interval $(t, t + H]$: death during the ICU stay, vasopressor initiation, or invasive mechanical ventilation initiation. This is a composite event definition. Its computational advantage is that it converts several sparse adverse transitions into a single target with sufficient support for model training. Its limitation is semantic heterogeneity, because the components differ mechanistically and operationally [4]. Even so, the composite target is a reasonable abstraction of near-term instability, particularly for sequence models intended to rank windows by urgency rather than to provide mechanistic diagnosis. The positive prevalence at the sample level is 2.8%, so the learning problem is dominated by rare-event behavior rather than by overall classification rate.

Formally, define indicator processes $d_{i,\tau}$, $p_{i,\tau}$, and $m_{i,\tau}$ for ICU death, first vasopressor use, and first invasive ventilation, respectively. The hourly label is

$$y_{i,t} = \mathbb{I} \left[\begin{aligned} &\sum_{\tau=t+1}^{t+H} d_{i,\tau} > 0 \\ &\vee \sum_{\tau=t+1}^{t+H} p_{i,\tau} > 0 \\ &\vee \sum_{\tau=t+1}^{t+H} m_{i,\tau} > 0 \end{aligned} \right] \quad (2.1)$$

with the convention that continuation of an already active intervention does not generate a new positive label. The prediction function f_θ maps the multimodal input tuple $(\mathbf{S}_{i,t}, \mathbf{n}_{i,t}, \mathbf{c}_i)$ to $\hat{y}_{i,t} \in (0, 1)$, where $\mathbf{S}_{i,t} \in \mathbb{R}^{T \times 26}$ is the structured window. The optimization target is not merely to minimize a surrogate loss, but to learn a calibrated ranking that is stable under temporally correlated sampling and robust to heterogeneous observation density.

This problem differs from many conventional tabular prediction tasks in several important ways. First, features are not independently and identically distributed across samples. Consecutive hourly windows from the same stay overlap heavily, so the effective sample size is smaller than the raw count of 5.7 million examples might imply. Patient-level splitting is therefore essential. Second, structured missingness is pervasive and semantically meaningful. A laboratory value can be absent because it was not ordered, because it was considered unnecessary, or because the patient is in a phase of care where measurements are sparse. Missingness thus reflects both workflow and acuity, and should not be interpreted solely as noise. Third, the note modality is temporally coarse relative to bedside measurements. The same document embedding may condition many hourly windows, creating a mismatch between the temporal resolution of numeric and textual streams. A useful architecture must therefore accommodate high-frequency sequence input together with low-frequency semantic context.

The multimodal representation can be interpreted as a partially observed dynamical state estimation problem. Let $\mathbf{z}_{i,t}$ denote an unobserved latent state summarizing hemodynamic stability, respiratory reserve, clinician concern, and trajectory momentum. Structured observations provide frequent but noisy glimpses of this state, while the note embedding offers a delayed semantic projection of it. The deterioration event within 24 hours is a threshold crossing in the latent process rather than a direct function of any single variable. This interpretation motivates stateful sequence encoders and context fusion instead of one-shot classifiers on summary statistics. It also explains why classical models that flatten the temporal window tend to underperform recurrent architectures: flattening removes ordering, lag structure, and rate-of-change information that are central to latent trajectory inference.

The structured signal contains both contemporaneous magnitude information and temporal derivatives that are not explicitly engineered but can be learned by the recurrent encoder. For example, the combination of stable blood pressure with progressively rising lactate, increasing respiratory rate, and declining oxygen saturation may be more predictive than any single threshold crossing. Similarly, laboratory measurements arrive at variable intervals; their predictive value often depends on recency and persistence rather than absolute value alone [5]. A recurrent model with attention can, in principle, learn that a recent abnormal lactate matters differently from a remote abnormal bicarbonate, and that interaction patterns spanning several hours may be more informative than point estimates. The note modality extends this by adding semantic cues such as edema progression, worsening infiltrates, or concern for shock that do not enter the structured tensor directly.

At the decision layer, the model output can serve several distinct roles. In ranking mode, it provides an ordering over active ICU windows, enabling clinicians or monitoring systems to prioritize review. In threshold mode, it triggers alerts when predicted risk exceeds a selected operating point. In calibration mode, it approximates a probability that can be incorporated into downstream decision analysis, staffing models, or resource allocation. These roles impose different desiderata. Ranking emphasizes AUROC and relative risk ordering. Alerting emphasizes AUPRC, precision at alert budget, and false-alarm burden. Probability use emphasizes Brier score and expected calibration error. Because no single metric can summarize all these requirements under a 2.8% event rate, the experimental design must evaluate discrimination, concentration of positives, operating-point behavior, and post-hoc calibratability.

The central modeling hypothesis is that the conditional risk function factorizes into complementary contributions from structure and text, but not additively in a fixed way. Some windows should depend almost entirely on physiologic history, especially when no notes are available or when the note is stale relative to current instability. Other windows may require textual disambiguation, such as differentiating a transient oxygen requirement from progressive respiratory failure reflected in imaging. The fusion mechanism should therefore estimate not only a joint representation, but also a context-dependent weighting over modalities. A static concatenation can approximate this relationship only if later layers

learn complex conditional interactions; a gated fusion layer makes this dependency explicit and lowers the burden on the classifier.

Under this formulation, deterioration forecasting becomes a sequence learning problem with three intertwined sources of difficulty: temporal dependence, modality mismatch, and class imbalance. The remainder of the paper addresses these issues through representation design, loss shaping, and evaluation choices tailored to the statistical and systems properties of ICU data rather than through domain-specific heuristics alone.

3. Cohort Construction and Multimodal Representation

The dataset is derived from MIMIC-IV, a large de-identified critical care repository containing ICU stays from Beth Israel Deaconess Medical Center across more than a decade. For this study, adult ICU stays were selected if they lasted at least 24 hours, had valid admission and discharge timestamps, and contained sufficient early measurements to support temporal modeling. Stays ending in death within the first 6 hours were excluded because such cases do not leave an actionable prediction window. The final cohort includes 74,822 ICU stays from 61,437 unique patients. Some patients contribute multiple stays, but all stays from a given patient are assigned to a single split to avoid identity leakage across training, validation, and testing partitions [6].

The cohort spans multiple ICU types and therefore exposes the model to heterogeneous physiologic regimes. Medical ICU stays constitute 42.0% of the cohort, surgical ICU stays 33.3%, and cardiac ICU stays 24.7%. The median age is 65 years with an interquartile range of 52 to 76, and 56.3% of stays correspond to male patients. Median ICU length of stay is 52 hours with a wide interquartile range, which is favorable for rolling-window sample generation because many patients contribute multiple prediction hours after the initial observation buffer. Across stays, 7.5% end in ICU mortality, 38.0% involve vasopressor use at some point, and 25.7% involve mechanical ventilation. At the stay level, 48.0% experience at least one deterioration component, but when transformed into hourly labels the prevalence drops to 2.8%, reflecting the sparsity of event onset relative to the number of sampled windows.

Structured data representation begins with ten vital sign channels and sixteen laboratory channels selected for coverage, clinical plausibility, and prevalence in the source database. The vital sign set includes heart rate, systolic, diastolic, and mean arterial blood pressure, respiratory rate, peripheral oxygen saturation, and temperature. The laboratory channels include lactate, creatinine, blood urea nitrogen, potassium, sodium, glucose, white blood cell count, hemoglobin, hematocrit, platelets, bilirubin, albumin, pH, partial pressure of carbon dioxide, partial pressure of oxygen, and bicarbonate. These 26 variables balance cardiorespiratory status, metabolic stress, renal function, inflammatory response, and acid-base regulation. Importantly, the representation is not exhaustive; it is a carefully chosen subset that captures broad physiologic state while preserving tractable tensor size and minimizing ultra-sparse variables.

To align irregular measurements to the hourly grid, all observations are aggregated within each hour. Vital signs are averaged over the hour, whereas laboratory values are carried forward using last-observation-carried-forward windows tailored to expected sampling frequency. Forward propagation for vital signs is capped at 4 hours, preventing indefinite retention of stale bedside measurements. Laboratory propagation uses 6-hour windows for rapidly changing quantities such as lactate and blood gas variables, and longer windows, up to 24 hours, for more routine chemistry and hematology values. This is a pragmatic preprocessing decision. It preserves clinically meaningful persistence while acknowledging that the absence of a fresh measurement does not necessarily imply that the last value is invalid. At the same time, the cap ensures that the model cannot rely on arbitrarily old observations as if they were current.

Sample generation begins at hour 6 of each stay to ensure minimum history and proceeds hourly until hour 48 or discharge, whichever occurs first. Each sample contains a 48-hour look-back window. When the stay is shorter than the full history requirement at early eligible hours, the leading portion of the tensor is implicitly padded through the preprocessing scheme. This rolling-window construction

transforms the original stay-level data into 5,720,904 supervised examples, partitioned into 4,019,343 training samples, 877,920 validation samples, and 823,641 test samples [7]. The resulting tensorized representation is computationally convenient because every example can be batched as a fixed-shape array despite highly irregular original measurement patterns.

A notable property of the data is pervasive missingness. No test sample has fewer than 55% missing feature values over its 48-hour window. This is not surprising in intensive care data. Different tests are ordered at different frequencies, and care intensity changes over time. Sparse observations late in a stay may reflect stabilization, transfer readiness, or reduced measurement frequency after an acute phase; conversely, sparse laboratory coverage can also occur in unstable patients when only select tests are repeated. The key point is that missingness is not homogeneous and may be correlated with both latent state and workflow. The preprocessing pipeline therefore treats missing values consistently but avoids claiming that imputation recovers the true underlying signal. Instead, the model is trained to operate under these incomplete trajectories as they appear in practice.

The unstructured modality is derived from radiology reports available in the associated note module. This choice has several computational and methodological implications. Radiology reports are temporally linked to the hospital course and often encode acute findings relevant to deterioration, such as worsening pulmonary infiltrates, pleural effusion, edema, or line placement. They are also comparatively standardized compared with narrative progress notes. On the other hand, they are less frequent than bedside measurements and less directly reflective of immediate hemodynamics. In the cohort, 66.4% of ICU stays contain at least one radiology note, with a median of 3 notes per annotated stay and an interquartile range of 1 to 6. Thus, one-third of stays lack textual signal entirely under this design.

Text preprocessing removes de-identification placeholders, lowercases content, normalizes whitespace, and truncates input to 512 tokens to conform to the maximum sequence length of the chosen language encoder. Each report is embedded using Bio_ClinicalBERT, a domain-adapted transformer pretrained on clinical text. The representation extracted is the [CLS] token vector of dimension 768. Rather than fine-tune the transformer end to end, the model freezes the language encoder and learns only a linear projection into the shared representation space. For each hourly sample, the most recent prior report is used. This decision reduces computational cost dramatically because note embeddings can be cached and reused across windows rather than recomputed. It also makes the system more realistic for deployment, where text changes infrequently relative to the numeric stream.

The note design introduces an asymmetry that is central to interpreting the later results [8]. Because the same embedding can condition many successive hourly windows, the text modality acts more like a slowly varying contextual prior than like a rapidly updating sensor. This has two consequences. First, text alone is unlikely to support strong hourly deterioration prediction because it lacks fine-grained temporal variation. Second, text may still improve a structured model if it helps contextualize physiologic patterns that are otherwise ambiguous. For instance, an oxygenation trend can mean different things depending on whether the most recent imaging suggests edema, consolidation, aspiration, or post-procedural change. The multimodal model therefore asks text not to replace physiology, but to disambiguate it.

Static features are minimal by design: age, sex, and an indicator of whether text is available. Including a note-presence indicator is useful because the absence of notes is itself informative about data regime and documentation pathway. Before training, all structured numeric features are standardized. The note embedding is projected to the same dimension as the structured representation, allowing fusion without excessive parameter growth. For samples lacking a report, the note vector is set to zero. This choice, while simple, interacts cleanly with the gating mechanism later in the model, enabling the system to learn that certain inputs imply reliance on the structured path.

The resulting multimodal representation can therefore be summarized as a tuple of three synchronized but semantically distinct components: a high-frequency structured tensor, a low-frequency semantic document embedding, and a compact static context vector. From a machine learning perspective, this representation is attractive because it is expressive enough to preserve modality-specific information while remaining regular enough for scalable mini-batch optimization across millions of windows.

4. Architecture and Optimization

The model architecture is organized around three computational blocks: a temporal encoder for structured sequences, a text encoder for note semantics, and a fusion-classification module that transforms both modalities into a deterioration score. The overall design goal is to maintain sufficient representational capacity to model temporal dependencies and cross-modal complementarity without incurring the prohibitive cost of fine-tuning a full language model jointly over more than five million training examples.

Let $\mathbf{x}_t \in \mathbb{R}^{26}$ denote the structured feature vector at hour t within a 48-hour look-back window. The temporal pathway processes the ordered sequence $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_{48})$ using a two-layer bidirectional LSTM with hidden dimension 128 per direction. The bidirectional design is appropriate because the input window is fully observed at inference time for a given prediction hour; the objective is not causal forecasting over future covariates, but retrospective encoding of the preceding 48-hour history. The concatenated hidden state at each position therefore has dimension 256. Temporal attention is then applied to compress the sequence into a fixed-length representation. In effect, the recurrent layers capture local and medium-range temporal interactions, while attention reweights the hidden trajectory according to estimated importance for the current prediction target.

The structured encoder is defined as

$$\begin{aligned}
 \mathbf{h}_t &= \text{BiLSTM}(\mathbf{x}_t, \mathbf{h}_{t-1}) \\
 e_t &= \mathbf{u}^\top \tanh(W_h \mathbf{h}_t + \mathbf{b}_h) \\
 \alpha_t &= \frac{\exp(e_t)}{\sum_{\tau=1}^T \exp(e_\tau)} \\
 \mathbf{z}_s &= \sum_{t=1}^T \alpha_t \mathbf{h}_t
 \end{aligned} \tag{4.1}$$

where $\mathbf{z}_s \in \mathbb{R}^{256}$ is the structured summary vector. This attention mechanism can be interpreted as a learned importance distribution over hours in the look-back window. It does not enforce monotonic recency bias, which is useful because deterioration can be heralded by either very recent instability or by sustained trends emerging over longer intervals [9]. The recurrent layers use dropout of 0.3 between stacked layers, providing regularization against overfitting to repeated temporal motifs in a dataset with many overlapping windows.

The text pathway begins with a frozen Bio_ClinicalBERT encoder applied to the most recent radiology report prior to the prediction hour. Let $\mathbf{e} \in \mathbb{R}^{768}$ denote the [CLS] embedding. A trainable linear projection followed by a rectified nonlinearity maps the note representation into the same dimensionality as the structured vector:

$$\begin{aligned}
 \mathbf{e} &= \text{ClinicalBERT}(d)_{[\text{CLS}]} \\
 \mathbf{z}_n &= \phi(W_p \mathbf{e} + \mathbf{b}_p) \\
 \mathbf{z}_n &= \mathbf{0} \quad \text{when no note is available}
 \end{aligned} \tag{4.2}$$

where ϕ is the ReLU activation and $\mathbf{z}_n \in \mathbb{R}^{256}$. Freezing the language model is a nontrivial design decision. Joint fine-tuning might improve text sensitivity, but the computational cost over millions of samples would be substantial, and the temporal reuse of the same note across adjacent windows would amplify redundancy. By freezing the encoder, the system reuses a strong clinical language prior while restricting trainable text parameters to a single projection layer. This reduces memory consumption, accelerates training, and makes the note modality easy to cache in a real-time pipeline.

Fusion is implemented through a gated mechanism rather than by naive concatenation alone. The intuition is that the usefulness of text is conditional. When the note vector is absent, stale, or semantically uninformative, the model should default toward structured physiology. When the note captures clinically meaningful context, it should influence the risk estimate. To make this dependency explicit, the model computes a gate over the concatenated structured and text representations:

$$\begin{aligned}\mathbf{g} &= \sigma(W_g[\mathbf{z}_s; \mathbf{z}_n] + \mathbf{b}_g) \\ \mathbf{z}_f &= \mathbf{g} \odot \mathbf{z}_s + (1 - \mathbf{g}) \odot \mathbf{z}_n \\ \mathbf{z}_u &= [\mathbf{z}_f; \mathbf{c}]\end{aligned}\tag{4.3}$$

where σ denotes the logistic sigmoid, $\odot$ is elementwise multiplication, and $\mathbf{z}_u$ is the final fused feature vector augmented with static covariates. This gate acts dimensionwise, allowing some latent coordinates to be dominated by structure and others by text. The design is computationally simple but expressive enough to learn modality-specific reliability patterns. Because the note vector is zero when unavailable, the gate can learn a natural fallback without special-case routing logic.

The classifier head is a two-layer multilayer perceptron with hidden dimension 64 and dropout of 0.3. The output is a scalar deterioration probability:

$$\begin{aligned}\mathbf{q} &= \phi(W_1 \mathbf{z}_u + \mathbf{b}_1) \\ \hat{y} &= \sigma(W_2 \mathbf{q} + b_2)\end{aligned}\tag{4.4}$$

This choice keeps the head intentionally shallow. Most of the representational burden is placed on the temporal encoder and fusion mechanism, while the classifier acts primarily as a smooth decision surface over the latent space. A deeper head might overfit to artifacts introduced by heavy overlap among windows.

Training on this dataset is dominated by class imbalance and the tendency of large recurrent models to memorize frequent negative patterns. Standard binary cross-entropy would be suboptimal because the positive class comprises only 2.8% of hourly windows. The training loss therefore uses focal loss with label smoothing [10]. Let $\tilde{y}_i$ denote the smoothed target and p_i the model probability assigned to the true class. The loss is

$$\begin{aligned}\tilde{y}_i &= (1 - \epsilon)y_i + \frac{\epsilon}{2} \\ p_i &= \tilde{y}_i \hat{y}_i + (1 - \tilde{y}_i)(1 - \hat{y}_i) \\ \mathcal{L} &= -\frac{1}{N} \sum_{i=1}^N \alpha_i (1 - p_i)^\gamma \log(p_i)\end{aligned}\tag{4.5}$$

with $\epsilon = 0.05$, $\gamma = 2.0$, and positive-class weight $\alpha = 0.75$. Label smoothing discourages overconfident predictions, which is useful in a setting where repeated overlapping windows can lead to sharp but poorly calibrated decision boundaries. Focal modulation downweights easy negatives, encouraging the model to spend optimization capacity on ambiguous windows and rare positives. The choice of $\alpha = 0.75$ is especially important because it upweights the positive class rather than suppressing it, improving recall and AUPRC.

Optimization uses AdamW with learning rate 2×10^{-4} , weight decay 10^{-3} , batch size 256, and gradient clipping at norm 1.0. Learning rate warmup spans the first 3 epochs, after which cosine annealing governs the schedule. Training proceeds for up to 50 epochs with early stopping based on validation AUROC and patience of 7 epochs. Distributed training is performed over six NVIDIA A100 80GB GPUs. This hardware choice is justified by sequence length, batch size, and the need to process millions

of windows efficiently, even though the text encoder remains frozen. In practice, recurrent sequence models on this scale can still become throughput-limited by data movement and variable masking operations; freezing the language model therefore helps not only statistically but also systems-wise.

From a computational complexity perspective, the architecture is well matched to the dataset. The structured sequence length is only 48, making recurrent processing relatively inexpensive compared with long-document transformers. The text pathway is amortized through cached embeddings. The fusion and classifier layers are lightweight. This means the dominant training burden lies in repeated recurrent passes over the structured window, not in cross-attention over large multimodal tensors. In a real-time deployment, this design would also permit asynchronous updates: note embeddings could be recomputed only when a new radiology report arrives, while structured risk scores could update hourly or even more frequently with minimal overhead.

One subtle but important feature of the architecture is that it does not impose a hard alignment between the semantics of text and the temporal positions of numeric data. The note enters as contextual bias rather than as a tokenwise sequence fused to each hour. This is computationally conservative but conceptually appropriate because radiology reports summarize findings over broader time intervals rather than providing exact hour-by-hour state annotations. A more elaborate architecture could align note segments to measurement times through cross-attention, but the current design avoids artificial precision and keeps the fusion semantics interpretable: the note informs the meaning of the numeric trajectory without pretending to be sampled on the same clock.

Another important consideration is calibration. Neural models trained with focal loss often achieve good ranking performance but produce distorted probabilities, especially under heavy imbalance [11]. Rather than trying to solve calibration entirely during training, the pipeline separates discrimination learning from post-hoc probability correction. Temperature scaling is later used as a scalar recalibration step on held-out predictions. This division of labor is practical. The core model optimizes representation and ranking under imbalance, while the calibration layer adjusts the final logit scale to match observed frequencies. For deployment, this separation is advantageous because recalibration can be repeated as class prevalence or workflow changes, without retraining the entire network.

Taken together, the architecture and optimization strategy are deliberately engineered for the specific statistical regime of ICU deterioration forecasting. The recurrent encoder captures structured temporal dynamics, the frozen transformer injects semantic context economically, the gate expresses modality reliability, and the loss function targets rare-event learning without sacrificing tractability. The resulting system is not the most complex multimodal architecture possible, but it is a technically coherent design whose empirical behavior can be interpreted cleanly.

5. Experimental Protocol

Table 1: Cohort Summary Statistics

Metric	Value	Metric	Value
ICU stays	74,822	Unique patients	61,437
Median age	65	Male (%)	56.3
Median LOS (hrs)	52	Mortality (%)	7.5
Vasopressor use (%)	38.0	Ventilation (%)	25.7
Any deterioration (%)	48.0	Sample prevalence (%)	2.8

The evaluation protocol is built around patient-level separation, rare-event metrics, and baseline comparison across both deep and classical methods. The full sample set is partitioned into training, validation, and test splits in a 70%/15%/15% ratio at the patient level. This means that all ICU stays from a given patient are confined to a single split, preventing the model from exploiting stable person-specific

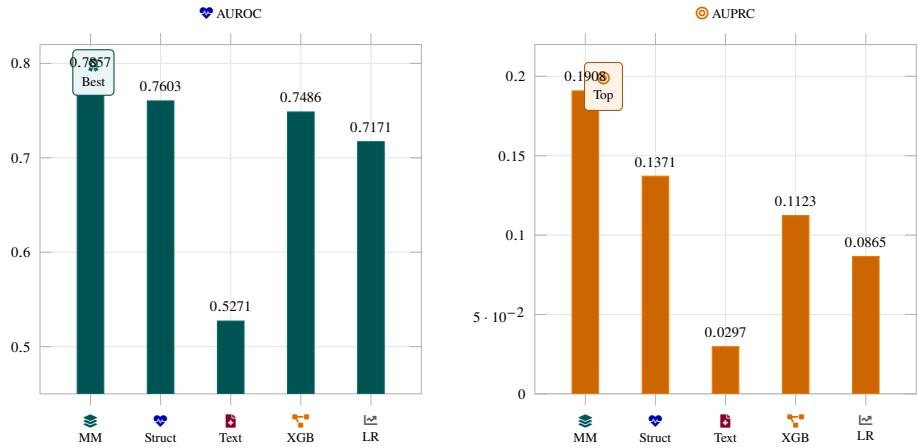


Figure 9: Held-out discrimination on 823,641 test windows. The multimodal model has the best AUROC and the largest precision–recall concentration, while the text-only path remains near baseline when detached from the physiologic stream. The gap between the multimodal and structured-only variants is modest in AUROC but much larger in AUPRC, indicating that note context improves high-risk concentration more than broad ranking alone.

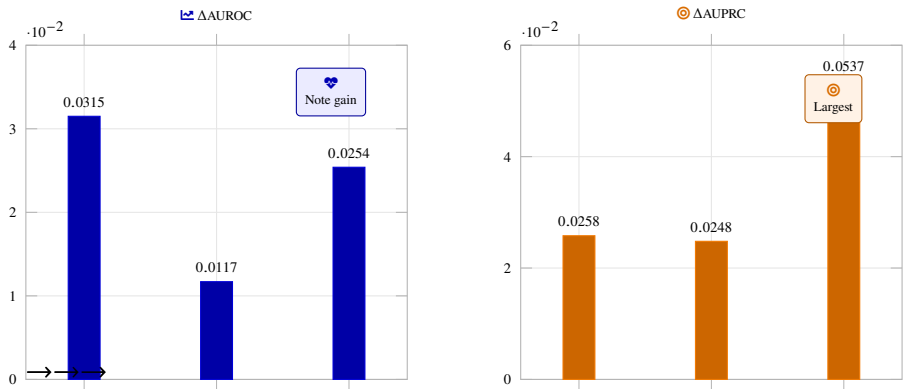


Figure 10: Stepwise performance gains across the modeling hierarchy. Moving from logistic regression to XGBoost adds nonlinear tabular structure, moving from XGBoost to the recurrent structured encoder adds temporal ordering, and moving from the structured encoder to the multimodal system contributes the largest AUPRC jump. The final transition shows that textual context yields the strongest incremental rare-event enrichment.

Table 2: Dataset Splits

Split	Samples	Percentage	Notes
Training	4,019,343	70%	Model fitting
Validation	877,920	15%	Early stopping
Test	823,641	15%	Final evaluation
Total	5,720,904	100%	All windows

patterns across partitions. Given the density of rolling windows, this choice is essential; sample-level random splitting would lead to severe leakage because adjacent windows from the same stay share most of their measurements and labels.

Five models are compared. The primary model is the full multimodal system described earlier. Two ablation models isolate the contribution of each modality: a structured-only model that uses the temporal encoder and classifier without note input, and a text-only model that consumes note embeddings

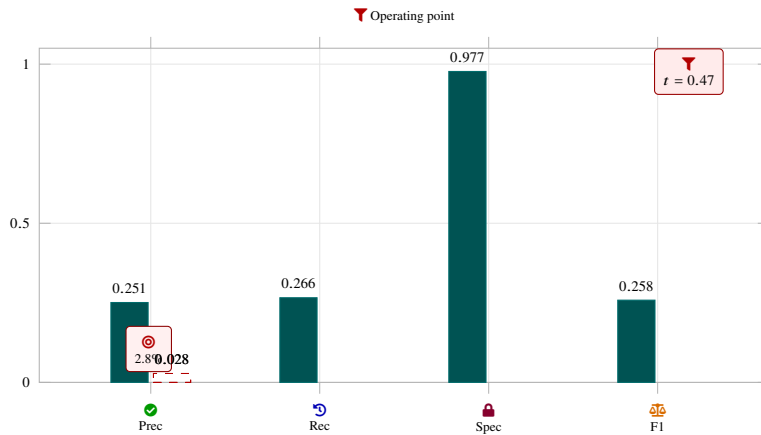


Figure 11: Thresholded behavior of the multimodal model at the validation-selected F1 operating point. Specificity is very high, whereas precision, recall, and F1 remain in the mid-0.25 range. Because prevalence is only 2.8%, the observed precision still corresponds to strong enrichment over the underlying event rate, which supports ranked review and conservative alerting rather than broad high-sensitivity screening.

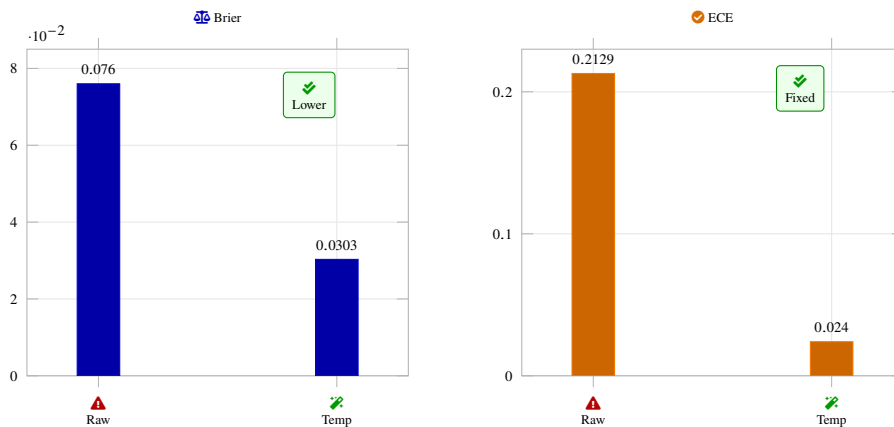


Figure 12: Post-hoc calibration of the multimodal predictor. Raw focal-loss outputs have acceptable ranking behavior but poor probability fidelity, as shown by elevated Brier score and expected calibration error. A single temperature-scaling step sharply reduces both quantities, showing that most of the calibration deficit is in logit scaling rather than in representation quality.

Table 3: Feature Composition

Category	Count	Examples	Notes
Vital signs	10	HR, BP, SpO2	Hourly averaged
Laboratories	16	Lactate, Creatinine	Forward-filled
Static features	3	Age, Sex	Low dimension
Text embedding	768	Radiology notes	Frozen encoder

and static covariates without the physiologic sequence. Two classical baselines are also included: logistic regression and XGBoost. These baselines are informative for different reasons. Logistic regression provides a low-capacity linear reference showing what can be achieved without nonlinear interactions. XGBoost represents a strong classical learner that often performs well on tabular health data after temporal summarization, and thus serves as a robust non-sequential benchmark.

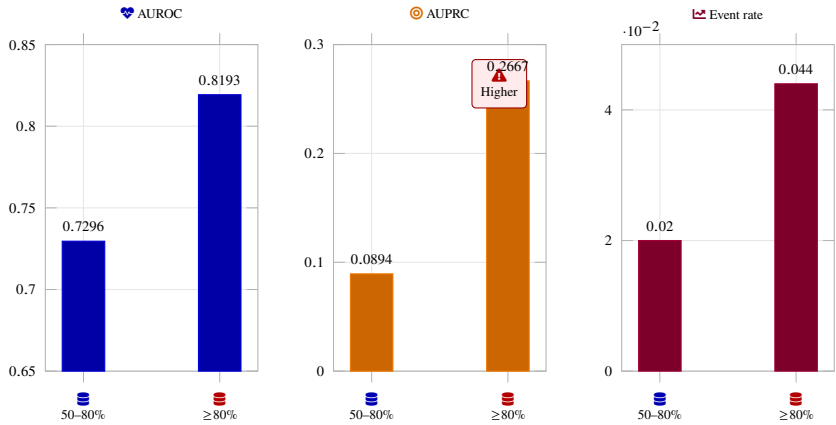


Figure 13: Performance under missingness stratification. The very-high-missingness subgroup has both higher event prevalence and stronger discrimination, whereas the moderate-missingness subgroup is less enriched and correspondingly harder. The pattern suggests that missingness level co-varies with case mix and workflow regime rather than acting as a simple proxy for information loss alone.

Table 4: Model Performance Comparison

Model	AUROC	AUPRC	Notes
Multimodal	0.7857	0.1908	Best overall
Structured only	0.7603	0.1371	Temporal model
Text only	0.5271	0.0297	Near random
XGBoost	0.7486	0.1123	Tabular baseline
Logistic regression	0.7171	0.0865	Linear baseline

Table 5: Multimodal Gains over Structured Model

Metric	Structured	Multimodal	Improvement
AUROC	0.7603	0.7857	+2.54 pts
AUPRC	0.1371	0.1908	+39.2%
Precision enrichment	4.9x	6.7x	Higher density
Generalization gap	0.0048	0.0057	Stable

Table 6: Threshold-Based Metrics

Metric	Value	Interpretation	Notes
Precision	25.1%	High enrichment	Rare events
Recall	26.6%	Moderate sensitivity	Conservative
Specificity	97.7%	Low false positives	Stable
F1 score	0.2584	Balanced tradeoff	At threshold 0.47

Discrimination is measured with area under the receiver operating characteristic curve and area under the precision-recall curve. AUROC is useful for global ranking assessment because it is threshold-independent and insensitive to class prevalence, but it can mask poor performance on the rare positive class. AUPRC is therefore central in this study because it directly reflects the model’s ability to concentrate positive windows among high-confidence predictions when prevalence is only 2.8%. Additional metrics include Brier score for probabilistic accuracy, expected calibration error for calibration quality,

Table 7: Calibration Performance

Model State	Brier Score	ECE	Notes
Raw multimodal	0.0760	0.2129	Miscalibrated
Calibrated (temp)	0.0303	0.0240	Improved
Structured only	0.0812	0.2280	Worse baseline
Text only	0.0985	0.3100	Poor calibration

Table 8: Missingness-Stratified Performance

Missingness	AUROC	AUPRC	Prevalence
50–80%	0.7296	0.0894	2.0%
≥80%	0.8193	0.2667	4.4%
All samples	0.7857	0.1908	2.8%
Low missingness	0.7012	0.0751	1.8%

Table 9: Training Configuration

Component	Setting	Value	Notes
Optimizer	AdamW	2×10^{-4}	LR
Batch size	256	–	Stable training
Dropout	0.3	–	Regularization
Epochs	≤50	Early stop	Best at 2
GPUs	A100	6	Distributed

Table 10: Text Modality Statistics

Metric	Value	Notes	Source
Stays with notes	66.4%	Coverage	Radiology only
Median notes/stay	3	IQR 1–6	Sparse
Embedding size	768	CLS token	BERT-based
Token limit	512	Truncation	Preprocessing

and threshold-specific precision, recall, specificity, and F1 score at the decision threshold that maximizes F1 on the validation set [12]. This thresholding choice is not meant to be clinically optimal in all settings, but it offers a reproducible operating point for comparison.

Calibration is evaluated in two stages. First, raw model outputs are assessed directly through Brier score and expected calibration error. Second, a proof-of-concept post-hoc calibration experiment applies temperature scaling. This does not change ranking metrics such as AUROC and AUPRC, but can substantially improve probability alignment. Including both views is important because deterioration forecasting systems are often used in mixed modes: clinicians may interpret them as rankings, while operational systems may consume them as probabilities. A model that ranks well but is miscalibrated is still useful, but it should not be deployed as a probability source without correction.

The experimental protocol also includes analyses beyond single scalar summary metrics. Missingness-stratified performance is measured to understand whether the model behaves differently under varying observation density. Training dynamics are monitored through validation AUROC across epochs in order to detect early overfitting. Validation-to-test performance gaps are reported to assess generalization stability. This is especially relevant in overlapping-window data, where a model can appear to improve in training while merely learning repetitive local patterns with little external benefit. A narrow validation-to-test gap is therefore evidence that regularization, splitting, and stopping rules are functioning appropriately.

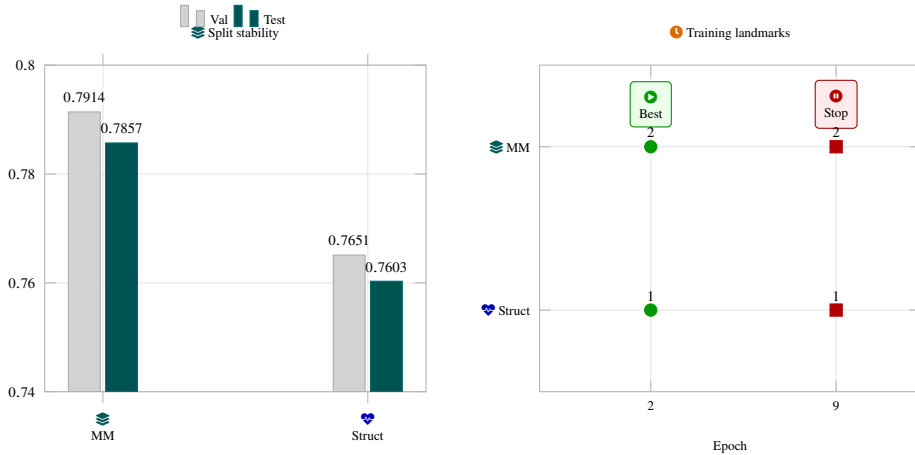


Figure 14: Generalization and optimization behavior of the deep sequence models. Validation and test AUROC remain closely aligned for both multimodal and structured-only variants, showing limited drift under patient-level splitting. The best validation score arrives early at epoch 2 for both models, and early stopping occurs at epoch 9, consistent with rapid convergence and limited benefit from prolonged optimization in a heavily overlapping rolling-window regime.

The thresholded operating point deserves particular interpretation because it exposes the tension between event rarity and alert usefulness. At low prevalence, even a discriminative model can generate many false positives unless thresholds are set conservatively. Precision, recall, and specificity must therefore be interpreted jointly. A high-specificity, moderate-recall operating point may be acceptable for escalation workflows with limited response capacity, whereas a lower threshold may be preferable for ranking-based review rather than direct alerting. Reporting only AUROC would obscure these trade-offs. The inclusion of F1, specificity, and calibrated probability behavior makes the evaluation more informative for systems integration.

The protocol is also designed to separate two distinct questions: whether multimodality improves representation quality, and whether deep sequence learning improves over classical feature-based methods. The structured-only versus multimodal comparison answers the first question. The structured deep model versus XGBoost and logistic regression addresses the second. This decomposition is methodologically important [13]. A multimodal gain could in principle arise even if the temporal encoder were unnecessary, and a sequence model could outperform classical baselines even without text. By evaluating both axes, the study can attribute performance differences more precisely to modality integration and temporal modeling rather than to generic model size.

Because the same note embedding may apply across many windows, the text-only ablation is intentionally difficult. Its inclusion is nevertheless useful because it clarifies the role of text in the full system. If text-only performance were strong, the note modality might be functioning as a surrogate risk score in its own right. If it is weak while multimodal performance improves, then the notes are likely acting as contextual modifiers rather than primary predictors. This distinction matters for architectural interpretation and for deployment planning, where the computational cost of maintaining a text pipeline must be justified by complementarity rather than by redundancy.

Finally, the experimental protocol remains retrospective and observational. It does not attempt causal evaluation of whether acting on predictions would improve outcomes. The measured results therefore characterize predictive utility, not intervention benefit. This boundary is important to maintain. A model can be statistically useful for ranking risk even if the operational response to elevated risk remains undefined. By keeping the evaluation centered on representation, discrimination, calibration, and robustness, the paper stays within the scope of what the data support.

6. Empirical Results

The empirical results show that multimodal integration yields consistent gains over structured-only and classical baselines, while also revealing important constraints on what the text modality contributes. On the held-out test set of 823,641 samples, the multimodal model achieves an AUROC of 0.7857 and an AUPRC of 0.1908. The structured-only deep model achieves 0.7603 AUROC and 0.1371 AUPRC. The text-only model performs near chance in discrimination, with AUROC 0.5271 and AUPRC 0.0297. Classical baselines underperform the multimodal model as well: XGBoost reaches 0.7486 AUROC and 0.1123 AUPRC, while logistic regression reaches 0.7171 AUROC and 0.0865 AUPRC. These numbers establish two key findings. First, the sequence-based structured encoder is already stronger than classical feature-based methods, indicating that temporal dynamics carry signal not preserved by flattened summaries. Second, note integration improves a strong structured model further, implying genuine modality complementarity.

The AUROC gain from multimodal fusion over the structured-only model is 2.54 absolute points, which corresponds to roughly a 3.34% relative improvement in ranking performance. More striking is the AUPRC increase from 0.1371 to 0.1908, an absolute gain of 0.0537 and a relative improvement of 39.2% [14]. Under severe imbalance, this precision-recall gain is especially meaningful because it indicates a substantial increase in the concentration of positives among higher-scoring predictions. Relative to the random baseline implied by the 2.8% prevalence, the multimodal AUPRC corresponds to approximately a 6.7-fold enrichment. In other words, the model is not merely ranking positives above negatives marginally better; it is materially improving the density of clinically relevant windows near the top of the score distribution.

The text-only result is informative in the opposite direction. Its AUROC of 0.5271 is only marginally above chance, and its AUPRC of 0.0297 is effectively at the random baseline. This confirms that radiology report embeddings, used in isolation and reused across multiple hourly windows, do not contain sufficient temporal granularity to support fine-grained hourly deterioration forecasting. Yet the multimodal gain demonstrates that the notes still matter. This asymmetry supports the interpretation that text acts as a contextual conditioning signal rather than as a stand-alone predictor. The report embedding is most useful when paired with physiologic state, where it can shift the representation of structured trajectories in clinically meaningful directions.

Generalization across splits is stable. The multimodal model achieves validation AUROC of 0.7914 and test AUROC of 0.7857, a gap of only 0.0057. The structured-only model shows a similar gap of 0.0048, and XGBoost shows essentially no degradation from validation to test. Such narrow gaps are nontrivial in a dense sliding-window setting because repeated local patterns can easily induce overfitting if leakage or weak regularization is present. The small validation-to-test drift therefore supports the claim that patient-level splitting, dropout, weight decay, early stopping, and focal-loss regularization together produce a model that generalizes beyond the specific windows seen during training.

Thresholded performance reveals the operating characteristics of the multimodal model when used as a decision rule rather than as a ranking function. At the F1-optimized threshold of 0.47, the model achieves precision of 25.1%, recall of 26.6%, specificity of 97.7%, and F1 score of 0.2584. Because the sample prevalence is only 2.8%, a precision of 25.1% represents almost a ninefold enrichment over baseline prevalence. This is substantial from a triage perspective: among windows flagged at this operating point, approximately one in four is followed by deterioration within 24 hours. The tradeoff is moderate sensitivity, with roughly one quarter of all positive windows detected at this threshold. This operating point is therefore conservative. It may be suitable for alerting workflows where false positives carry a real burden, but less suitable if the objective is broad surveillance and ranking.

Calibration metrics show a familiar pattern for focal-loss-trained neural models [15]. The raw multimodal predictions yield a Brier score of 0.0760 and expected calibration error of 0.2129. This indicates that the model's ranking ability is stronger than its probability fidelity. The structured-only model is similar, with slightly worse calibration, and the text-only model is substantially poorer. A proof-of-concept temperature scaling step reduces the multimodal expected calibration error to 0.0240 and the Brier score

to 0.0303 without changing AUROC or AUPRC. This is an important empirical finding because it suggests that the representation itself is not the calibration bottleneck. Rather, the logit scale induced by focal loss can be corrected post hoc. In deployment terms, this means the model can be optimized for discrimination and then recalibrated on a site-specific validation stream without architectural change.

Training dynamics further support the robustness of the learned representation. The best validation AUROC for both the multimodal and structured-only deep models is reached at epoch 2, with early stopping triggered at epoch 9 after no further improvement. This rapid convergence suggests that the model acquires most of the generalizable signal early, while additional epochs primarily fit redundant negative patterns and patient-specific idiosyncrasies present in overlapping windows. The observation is computationally useful. It implies that in this regime, aggressive training schedules or prolonged fine-tuning are unlikely to yield substantial benefit and may instead increase overfitting risk. Warmup over the first 3 epochs helps stabilize the early training trajectory, but the model does not require lengthy optimization to reach its best external performance.

A particularly interesting result emerges from the missingness-stratified analysis. For samples with 50–80% missing features over the look-back window, the multimodal model achieves AUROC of 0.7296 and AUPRC of 0.0894 at a positive rate of 2.0%. For samples with at least 80% missingness, performance is higher, with AUROC of 0.8193 and AUPRC of 0.2667 at a positive rate of 4.4%. At first glance this seems counterintuitive; one might expect more missingness to weaken the model. The likely explanation is confounding by acuity and workflow. The high-missingness subgroup also has a higher event rate, indicating a different case mix in which the remaining available observations and contextual cues may separate classes more clearly. Thus, higher performance in this stratum should not be interpreted as evidence that missingness is beneficial. Instead, it suggests that missingness patterns co-vary with the difficulty and prevalence structure of the forecasting problem.

The cohort composition also contextualizes the results. Of the 74,822 ICU stays, 49,674 contain radiology notes, meaning 66.4% of stays can exploit the text pathway [16]. The median number of notes per annotated stay is 3. This moderate coverage is enough to create a measurable multimodal benefit, but not enough for text to dominate. It is therefore plausible that expanding the note set to include nursing and physician progress documentation would strengthen the multimodal effect further, particularly if those notes provide higher-frequency trajectory information than radiology reports alone. The current results should thus be read as a lower-bound demonstration of contextual value under a relatively conservative text design.

The comparison with XGBoost is especially useful for understanding what the recurrent architecture contributes. XGBoost performs reasonably well in AUROC terms at 0.7486, but its AUPRC of 0.1123 is notably below both deep models that use structure. This suggests that summary-statistic representations of the time window preserve enough information for moderate ranking, but lose the temporal detail necessary to concentrate positives effectively at the high-risk end. The structured-only recurrent model improves AUPRC by approximately 22.1% relative to XGBoost, while the multimodal model improves it by about 69.9%. Thus, sequential modeling and multimodal context both matter, with the latter delivering its largest benefit precisely on the metric most sensitive to positive enrichment.

The logistic regression baseline further clarifies the need for nonlinear modeling. Its AUROC of 0.7171 and AUPRC of 0.0865 indicate that linear decision boundaries cannot capture the interaction structure required by the task. This is consistent with the ICU setting, where risk often emerges from combinations of trends rather than single thresholds. The difference between logistic regression and XGBoost shows the value of nonlinear tabular interactions; the difference between XGBoost and the structured recurrent model shows the additional value of ordered temporal encoding; and the difference between the structured recurrent and multimodal models shows the value of semantic context. These nested improvements produce a coherent hierarchy of representational needs.

Taken together, the empirical findings support a clear but bounded conclusion. Multimodal temporal learning improves deterioration forecasting in this setting, and the gains are most compelling on precision-recall behavior rather than on AUROC alone. The text modality is weak in isolation but useful in combination, indicating contextual complementarity rather than standalone predictive strength.

The learned representation generalizes cleanly across splits, converges rapidly, and can be recalibrated effectively. These are favorable properties for a forecasting model intended for real-time risk ranking in high-throughput ICU environments.

7. Error Analysis and Operational Interpretation

The empirical gains invite a deeper question: what exactly is the model learning that its baselines are not? Although no gradient-based attribution study is reported here, the performance pattern itself is revealing. The text-only model is nearly uninformative, yet the multimodal model clearly exceeds the structured-only one. This means the note embeddings are not directly solving the task. Instead, they are likely modifying the geometry of the structured latent space [17]. In representational terms, the note modality appears to shift decision boundaries around physiologic trajectories that are otherwise ambiguous. A respiratory trend accompanied by a report suggestive of pulmonary edema should map to a different risk neighborhood than the same trend paired with stable imaging language. The gain from multimodality therefore likely arises in ambiguous cases near the structured decision margin rather than in obviously stable or obviously deteriorating windows.

This interpretation is consistent with the gating design. When notes are absent or unhelpful, the model can revert toward the structured pathway. When notes encode salient context, the gate can incorporate them more strongly. The empirical evidence does not directly expose gate values, but it strongly suggests that the fusion mechanism is not merely concatenating redundant information. If text were redundant with structure, one would expect negligible gain from the multimodal model. If text were independently powerful, one would expect the text-only model to perform well. The observed pattern supports an intermediate case: weak marginal text signal but meaningful conditional text signal.

One way to view this is through the lens of conditional mutual information. The textual report may carry little information about the label on its own once averaged over all windows, because reports are sparse and stale relative to the prediction horizon. But conditional on a given physiologic pattern, the report can still reduce uncertainty. This is exactly the scenario in which contextual modalities matter despite poor standalone performance. It also explains why expanding text coverage could be valuable. If nursing and progress notes were added, the conditionally informative context would likely become more temporally aligned with physiologic changes, increasing both the standalone and joint usefulness of text.

The threshold behavior also deserves careful operational interpretation. At the selected threshold, specificity is 97.7%, which sounds strong in isolation, but because negatives are so common, even a small false-positive rate can produce a substantial absolute number of alerts. The precision of 25.1% should therefore be understood as an enrichment metric rather than a guarantee of diagnostic certainty. In practice, this operating point would likely be too conservative for some surveillance purposes and too liberal for others, depending on staffing, intervention protocols, and alert budgets. The more defensible deployment interpretation is to treat the output as a continuously updating risk score that supports ranked review rather than as a binary stand-alone alarm. In ranking mode, the model's nearly ninefold precision enrichment over prevalence is useful even if threshold selection remains context-dependent.

The calibration findings reinforce this distinction [18]. Before recalibration, the model is better as a ranker than as a literal probability estimator. After temperature scaling, the probability quality improves substantially. This suggests a two-stage operational architecture. The model first produces a risk logit optimized for discrimination under focal loss and imbalance. A site-specific calibration layer then maps this score into probabilities using recent held-out data. Such a design is attractive because it separates the stable part of the model from the locally adaptive part. If event prevalence, note practices, or measurement frequency shift over time, recalibration can be updated without retraining the encoder stack. From a machine learning systems perspective, this is far easier to maintain than frequent end-to-end model redevelopment.

The missingness-stratified results further illuminate what the model may be exploiting. Higher AUROC in the $\geq 80\%$ missingness group probably reflects an interaction between latent acuity, care

phase, and measurement practice. In many clinical datasets, missingness can itself serve as a proxy for workflow attention, but that relationship is unstable across institutions. The current model does not explicitly encode a rich missingness process beyond what is implicit in the imputed sequences and note-availability feature. As a result, its strong performance in certain sparse regimes may be partly dataset-specific. This is not necessarily a flaw, but it means that external validation is essential before assuming that the learned decision surface will transfer unchanged. A hospital with different laboratory ordering habits or documentation norms could alter both the meaning of sparsity and the utility of notes.

From a computational efficiency standpoint, the architecture is deployment-friendly in an important sense. Because the language model is frozen and note embeddings can be cached, the primary recurrent computation concerns the structured window only. In a live system, one could precompute the note projection whenever a new radiology report is filed and reuse it for subsequent hourly inference. The structured input would then be updated incrementally as fresh measurements arrive. This asynchronous modality pipeline is much cheaper than real-time joint transformer-recurrent inference. It also aligns with hospital data flow, where notes arrive episodically while bedside variables stream continuously. The architecture therefore offers a favorable balance between multimodal richness and runtime simplicity [19].

The early convergence behavior has additional operational implications. Since the best validation AUROC appears by epoch 2, full retraining for local adaptation may not require long optimization schedules. This matters because retraining on clinical data is expensive not only in compute but also in governance and validation overhead. A model family that converges quickly reduces iteration cost and supports more practical maintenance. Rapid convergence also suggests that the salient decision structure is relatively low-dimensional once the representation is properly set up; the challenge is less about discovering exotic nonlinear manifolds than about aligning modalities and weighting rare events appropriately.

An important caution, however, is that the composite outcome can blur interpretation of errors. A false positive for the composite label may still correspond to a patient with significant instability that does not manifest as death, vasopressor initiation, or ventilation within 24 hours. Conversely, a false negative may miss one event type while correctly ranking the patient as moderate risk due to another. This heterogeneity may partly explain why recall at a precision-enriched threshold remains modest. The model is being asked to compress multiple adverse pathways into a single scalar. That compression is useful for broad deterioration surveillance, but it necessarily sacrifices some mechanistic specificity. Separate heads for each event type could improve interpretability and possibly performance, though at the cost of greater modeling complexity.

Another layer of interpretation concerns the note modality itself. Radiology reports are semantically rich but institutionally standardized. This can be an advantage for generalization because their syntax is less variable than that of free-form progress notes. It can also be a limitation because the report content often emphasizes findings rather than bedside concern. The fact that even this relatively constrained text source improves a strong structured model suggests that multimodal fusion is not merely exploiting noisy lexical artifacts. Rather, it is likely benefiting from clinically coherent information encoded in the language space. This strengthens the case for broader note integration, but also argues for careful modality-specific regularization so that the model does not overfit to documentation styles.

In sum, the error structure and operating behavior of the model point to a nuanced interpretation. The system is not a text-driven detector, nor is it simply a better structured predictor with extra parameters. It is a context-sensitive sequence model whose gains emerge where semantic documentation helps resolve uncertainty in physiologic trajectories. Its most appropriate use is as a calibrated or calibratable risk ranking layer embedded within a larger workflow, not as an autonomous alarming instrument disconnected from human oversight [20].

8. Limitations, Reproducibility, and Deployment Considerations

Despite favorable empirical results, several limitations define the boundary of what can be concluded. The first is institutional scope. The entire study is derived from a single health system, and while the dataset is large, single-center modeling always risks conflating physiologic relationships with local workflow patterns. Ordering frequencies, documentation habits, intervention thresholds, and ICU composition vary across hospitals. Because the model uses both structured measurement patterns and note availability, its learned decision function may partially encode local practice rather than universal pathophysiology. External validation on distinct datasets is therefore not optional; it is necessary before any claim of broader transportability.

The second limitation concerns the text modality. Only radiology reports are used, and the most recent report is reused across all subsequent hourly windows until a new report appears. This is a computationally efficient design, but it places a hard cap on what text can contribute. Nursing notes and physician progress notes likely contain more temporally proximate information about deteriorating status, response to therapy, and clinician concern. Their absence almost certainly depresses text-only performance and may underestimate the true attainable gain from multimodal fusion. At the same time, including those notes would introduce additional preprocessing challenges, more variable syntax, and higher leakage risk if note timing is not handled carefully. The current study therefore demonstrates a conservative multimodal effect rather than an upper bound.

A third limitation is the composite outcome itself. Combining ICU mortality, vasopressor initiation, and invasive ventilation into one label increases statistical support but merges events with different physiology, treatment pathways, and operational meanings. A model optimized for the composite may learn a representation of generalized instability that is useful for triage, yet insufficiently specific for targeted intervention. Event-specific models or multi-task architectures could separate these pathways and yield richer decision support. The present formulation should thus be viewed as appropriate for broad deterioration surveillance, not as the final granularity for specialized bedside recommendations.

The study is also retrospective. It measures predictive performance on historical data, not the effect of predictions on care processes or outcomes. A highly ranked patient might already receive close attention, making the incremental value of a model smaller than retrospective metrics imply. Conversely, a model could reveal risk earlier than clinicians typically recognize, creating actionable benefit that cannot be inferred from static evaluation alone. Prospective validation, silent shadow deployment, and eventually randomized workflow studies would be needed to determine whether the predictive gains translate into meaningful operational improvements [21]. None of those conclusions can be drawn here.

Interpretability remains limited as well. Although the architecture is relatively transparent at the block level, no detailed feature attribution or token-level explanation analysis is reported. For clinical deployment, stakeholders often want to know whether a high score was driven by progressive lactate elevation, worsening oxygenation, note language suggestive of edema, or some interaction among these factors. The absence of explanation analysis does not invalidate the predictive results, but it constrains trust calibration and debugging. In addition, fairness and subgroup robustness are not analyzed. Performance could vary across age groups, sex categories, ICU types, or note-availability strata in ways that matter operationally. Those analyses are essential follow-ups.

Reproducibility, however, is a relative strength of the modeling setup. The cohort construction, feature list, windowing rules, model shapes, optimization hyperparameters, and evaluation metrics are all sufficiently explicit to permit reimplementing. Patient-level splitting reduces ambiguity about leakage control. The use of a frozen public clinical language model also improves reproducibility because text embeddings are less sensitive to optimizer dynamics than end-to-end fine-tuning would be. Large-scale health machine learning studies often suffer from underspecified preprocessing and split logic; by contrast, the present setup is compact enough that another group could reproduce it with access to the same source database and compute resources.

From a deployment engineering perspective, several practical considerations follow from the empirical findings. First, a local calibration layer should be mandatory. The raw probabilities are not sufficiently

reliable for direct probabilistic interpretation, but a simple temperature scaling step greatly improves calibration. Second, monitoring should distinguish discrimination drift from calibration drift. If AUROC remains stable while calibration worsens, recalibration may suffice. If both degrade, the latent representation itself may no longer match local data generation. Third, alert thresholds should not be fixed purely by F1 optimization. They should be selected according to alert budgets, staffing capacity, and intervention cost, ideally using decision-analytic simulation. The current threshold results provide a useful reference but not a universal recommendation.

The asynchronous nature of the architecture supports a practical deployment pattern [22]. Structured features can be refreshed hourly from the bedside stream, while the text context can be updated only when a new radiology report is finalized. This decoupling simplifies inference orchestration. It also permits caching and lowers latency. A production system could maintain the latest projected note embedding per patient and concatenate it with the recurrently encoded structured history on demand. Such a design reduces runtime complexity and makes it easier to audit modality-specific failures. If no note is available, the model degrades to a structured predictor automatically through the zero-vector convention and gating behavior.

A cautious deployment path would begin with shadow-mode evaluation, in which scores are computed in real time but not acted upon. This would allow institution-specific assessment of calibration, drift, alert volume, and subgroup behavior. Only after that stage should any thresholded alerting be considered. Even then, the preferable first use case may be ranked review or dashboard augmentation rather than interruptive alerts. The empirical results support the model as a risk stratifier more strongly than as a binary alarm generator, particularly given the modest recall at a precision-enriched threshold.

9. Conclusion

This paper presented a technically oriented multimodal framework for early ICU deterioration forecasting that combines a bidirectional recurrent encoder over 48-hour physiologic windows with contextualized radiology report embeddings and gated fusion. Using 74,822 ICU stays and 5,720,904 hourly prediction samples, the model achieved held-out performance of 0.7857 AUROC and 0.1908 AUPRC, exceeding structured-only deep learning and classical baselines. The largest practical gain appeared in precision-recall behavior, where note integration improved AUPRC by 39.2% relative to the structured-only model, indicating that free-text documentation contributes complementary signal even when it is not temporally dense enough to function as a stand-alone predictor.

The results also clarify the computational character of the task. Deterioration forecasting in intensive care is best understood as multimodal sequence inference under heavy class imbalance, pervasive missingness, and asynchronous context updates. In that setting, a frozen domain-specific text encoder, a compact gated fusion layer, and rare-event-aware optimization provide a favorable balance between performance and tractability. The model generalizes with a small validation-to-test gap, converges quickly, and can be recalibrated effectively through temperature scaling, which is useful for operational maintenance.

At the same time, the study remains bounded by single-center data, radiology-only text, and a composite outcome. The findings therefore support multimodal temporal learning as a viable design pattern for ICU risk stratification, while leaving substantial room for event-specific modeling, richer note integration, subgroup analysis, and external validation. As a computer science contribution in a medical setting, the central result is clear: when heterogeneous clinical data are represented in ways that respect both temporal order and semantic context, deterioration forecasting becomes materially more effective than with structured tabular modeling alone [23].

References

- [1] Y. Meng, W. Speier, M. K. Ong, and C. W. Arnold, "Hcet: Hierarchical clinical embedding with topic modeling on electronic health records for predicting future depression," *IEEE journal of biomedical and health informatics*, vol. 25, pp. 1265–1272, 4 2021.

- [2] B. Sadanandan, "Multimodal deep learning for early prediction of patient deterioration in the icu: Integrating time-series ehr data with clinical notes," *Journal of Data Science, Predictive Analytics, and Big Data Applications*, vol. 10, no. 7, pp. 1–21, 2025.
- [3] M. Y. Yang, G. H. Kwak, T. Pollard, L. A. Celi, and M. Ghassemi, "Evaluating the impact of social determinants on health prediction in the intensive care unit," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 333–350, ACM, 8 2023.
- [4] A. Havard, J.-A. Manski-Nankervis, J. Thistlethwaite, B. Daniels, R. Myton, K. Tu, and K. Chidwick, "Validity of algorithms for identifying five chronic conditions in medicineinsight, an australian national general practice database," *BMC health services research*, vol. 21, pp. 551–551, 6 2021.
- [5] L. Ohno-Machado, "Game changer: how informatics moved from a supporting role to a central position in healthcare.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 20, pp. e197–e197, 12 2013.
- [6] G. Wandell and J. P. Giliberto, "Otolaryngology resident clinic participation and attending electronic health record efficiency—a user activity logs study," *Laryngoscope investigative otolaryngology*, vol. 6, pp. 968–974, 8 2021.
- [7] A. Mukherji and N. Ganguli, "Efficient and scalable electronic health record management using permissioned blockchain technology," in *2020 4th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, pp. 1–6, IEEE, 10 2020.
- [8] N. G. Weiskopf, S. Bakken, G. Hripcsak, and C. Weng, "A data quality assessment guideline for electronic health record data reuse.," *EGEMS (Washington, DC)*, vol. 5, pp. 14–, 9 2017.
- [9] R. Kohli and S. S.-L. Tan, "Electronic health records: how can is researchers contribute to transforming healthcare?," *MIS Quarterly*, vol. 40, pp. 553–573, 9 2016.
- [10] S. Elkefi and O. Asan, "Exploring doctors' satisfaction with electronic health records and their features: A nationwide survey study," in *2022 IEEE International Symposium on Systems Engineering (ISSE)*, pp. 1–5, IEEE, 10 2022.
- [11] R. Benaich, S. E. Mendili, and Y. Gahi, "Advancing healthcare security: A cutting-edge zero-trust blockchain solution for protecting electronic health records," *HighTech and Innovation Journal*, vol. 4, pp. 630–652, 9 2023.
- [12] J. Perloff and S. Sobul, "Use of electronic health record systems in accountable care organizations.," *The American journal of managed care*, vol. 28, pp. e31–e34, 1 2022.
- [13] M. H. Lundqvist, V. Patsoukaki, S. Jansson, H. Norman, E. Granstam, M. K. Svensson, J. Sundström, B. Eliasson, and J. W. Eriksson, "Health care registers can be instrumental for endpoint capture in clinical diabetes trials: example of microvascular complications in swedish patients with type 2 diabetes.," *Diabetes & vascular disease research*, vol. 20, pp. 14791641231179878–14791641231179878, 6 2023.
- [14] R. Knevel and K. P. Liao, "From real-world electronic health record data to real-world results using artificial intelligence.," *Annals of the rheumatic diseases*, vol. 82, pp. 306–311, 9 2022.
- [15] P. Vidap, A. Bhargav, R. Paswan, and A. Jewalikar, "Blockchain solution to electronic healthcare records," in *2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, pp. 659–664, IEEE, 1 2023.
- [16] S. Barron and S. Manhas, "Electronic health record (ehr) projects in canada: participation options for canadian health librarians," *Journal of the Canadian Health Libraries Association / Journal de l'Association des bibliothèques de la santé du Canada*, vol. 32, pp. 137–143, 7 2014.
- [17] Y. S. Huilgol, J. Adler-Milstein, S. L. Ivey, and J. C. Hong, "Opportunities to use electronic health record audit logs to improve cancer care.," *Cancer medicine*, vol. 11, pp. 3296–3303, 3 2022.
- [18] R. Sandhiya, *Big Data Analytics and K-Means Clustering*, pp. 41–64. Chapman and Hall/CRC, 12 2020.
- [19] R. D. Wijaya and L. O. A. Rahman, "Implementasi electronic health records (ehrs) pada pelayanan kesehatan di komunitas: Literature review," *Jurnal Kesehatan*, vol. 8, pp. 28–38, 2 2020.
- [20] L. A. Stevens, Y. T. DiAngi, J. D. Schremp, M. J. Martorana, R. E. Miller, T. Lee, and N. M. Pageler, "Designing an individualized ehr learning plan for providers," *Applied clinical informatics*, vol. 8, pp. 924–935, 12 2017.

- [21] T. K. Colicchio, W. H. Liang, P. I. Dissanayake, C. V. D. Rosario, and J. J. Cimino, “Physicians’ perceptions about a semantically integrated display for chart review: A multi-specialty survey,” *International journal of medical informatics*, vol. 163, pp. 104788–104788, 5 2022.
- [22] L. Ohno-Machado, “Health surveillance using the internet and other sources of information,” *Journal of the American Medical Informatics Association*, vol. 20, pp. 403–403, 5 2013.
- [23] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, “Deep patient: An unsupervised representation to predict the future of patients from the electronic health records,” *Scientific reports*, vol. 6, pp. 26094–26094, 5 2016.