Principled Multimodal Risk Estimation for ICU Deterioration: Integrating Structured Sequences and Unstructured Notes with Robust Fusion

Thanawat Boonmee[1] **and Kittipong Saelim**[2]

[1]Department of Computer Engineering, Uttaradit Rajabhat University, 27 Injaimee Road, Tha It, Mueang Uttaradit, Uttaradit 53000, Thailand
[2]School of Information Technology, Sakon Nakhon Rajabhat University, 680 Nittayo Road, That Choeng Chum, Mueang Sakon Nakhon, Sakon Nakhon 47000, Thailand

ABSTRACT Intensive care units continuously generate heterogeneous data streams that encode evolving physiologic state and clinician interpretation. Despite dense monitoring, clinically meaningful deterioration events often emerge from subtle, multivariate trajectories that are difficult to recognize early using threshold-based scoring rules. Computational approaches that treat prediction as a sequential inference problem can exploit temporal dependencies, but they are frequently constrained by irregular sampling, missingness, and the fact that many clinically salient cues appear only in free-text documentation. This paper develops a for early deterioration forecasting that integrates structured time-series measurements with unstructured clinical text under a unified representation-learning perspective. We formalize ICU forecasting as conditional risk estimation over partially observed multichannel sequences, introduce a principled multimodal encoding strategy that maps both modalities into a shared latent space, and propose a fusion mechanism that remains well-defined when text is absent or temporally misaligned. The modeling components are analyzed in terms of identifiability under informative missingness, robustness to label sparsity, and calibration under extreme class imbalance. We emphasize objectives that control both ranking performance and probability quality, including imbalance-aware proper losses and post-hoc calibration operators that preserve discrimination. The resulting methodology is positioned as a general computational template rather than a dataset-specific recipe. We discuss design tradeoffs among recurrent, convolutional, and attention-based temporal encoders; contrast early, late, and gated fusion from an information bottleneck viewpoint; and derive operational metrics appropriate for low-prevalence event forecasting. The overall contribution is a cohesive technical treatment of multimodal ICU risk modeling with explicit attention to irregular sampling, fusion stability, and calibrated decision support.

KEYWORDS

1. Introduction

Forecasting adverse clinical trajectories in the ICU is fundamentally a problem of sequential inference under partial observability [1]. At any moment, the patient state is not directly measured; instead, it is glimpsed through sporadic laboratories, quasi-continuous vital signs, clinician interventions, and narrative documentation that summarizes evolving hypotheses. A deterioration event such as the onset of respiratory failure or shock is rarely an instantaneous discontinuity; it is more often the endpoint of interacting processes that manifest as modest drifts, increasing variance, changing correlations, or discordance between physiologic channels and clinician concern. From a computational standpoint, these characteristics imply that prediction models must learn from temporal structure, handle irregular sampling without introducing spurious signal, and represent heterogeneous modalities without collapsing them into lossy aggregates.

Historically, operational systems in critical care have favored

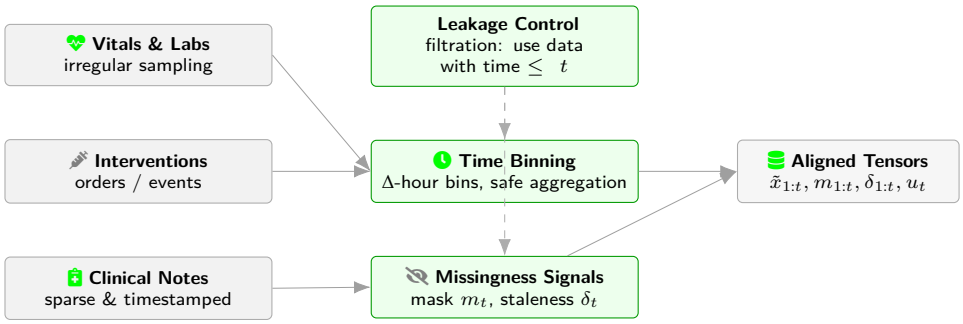

Figure 1 Multistream ICU signals are aligned into discrete-time bins while preserving observation patterns. Masks and per-channel staleness are retained as first-class inputs, and conservative filtration prevents temporally invalid features from entering the forecasting context.

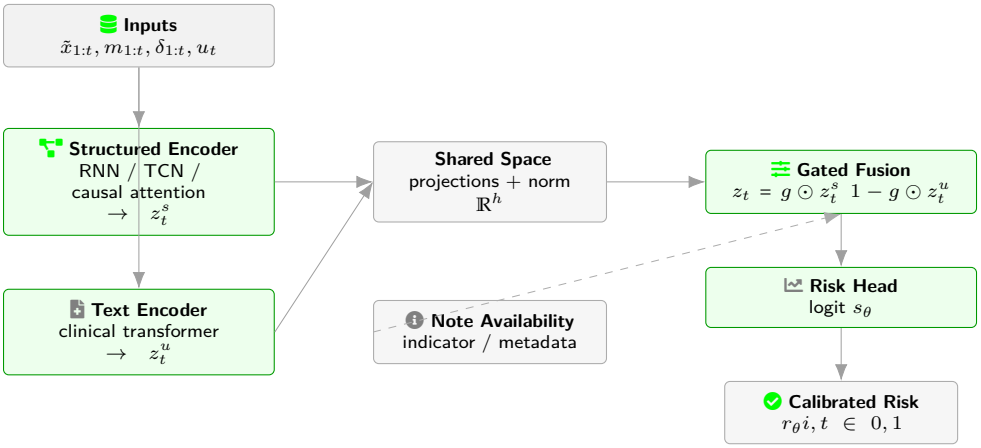

Figure 2 Unified multimodal forecaster: structured and textual encoders map irregular EHR and clinical notes into a shared latent space. A gated fusion operator yields a stable representation that degrades gracefully when text is absent or misaligned, enabling calibrated near-term deterioration risk estimation.

interpretability and simplicity, often using fixed thresholds and additive scoring rules. Such systems are attractive because they are easy to audit, cheap to deploy, and relatively stable under distribution shift. However, their functional form severely limits expressivity. A linear additive rule over a small set of variables cannot represent interactions such as compensatory tachycardia masking hypotension, divergent trends in lactate and oxygenation, or the clinically common scenario where a single value is uninformative but a pattern of change is crucial [2]. Moreover, rigid thresholds are fragile to baseline heterogeneity. An absolute respiratory rate of 24 breaths/min may be less important than a rise from 14 to 24 over a short window, yet trend sensitivity is difficult to encode manually without proliferating rules.

Modern machine learning reframes deterioration forecasting as the estimation of a conditional probability of future outcomes given the observed history. This reframing is computationally appealing because it supports continuous risk scoring, integrates multivariate covariates, and naturally accounts for tradeoffs among sensitivity, specificity, and alert burden. Yet applying it in the ICU introduces technical obstacles that are less central in canonical sequence modeling. The first obstacle is irregular observation. Clinical data are event-driven; clinicians measure more when concerned and less when stable, leading to informative missingness. The second obstacle is delayed and heterogeneous labeling [3]. Many relevant outcomes are composite or proxy events that reflect clinical actions rather than direct physiological thresholds, complicating causal interpretation. The third obstacle is modality mismatch. Structured channels capture numeric measurements, but narrative notes contain contextual information such as differential diagnoses, evolving goals, imaging interpretations, and informal severity statements that may not appear in any structured field.

From the perspective of representation learning, integrating text into ICU forecasting is attractive because it can inject high-level semantic context that is otherwise unavailable. Text, however, brings its own difficulties. Notes are sparse and irregular, their timestamps may reflect documentation time rather than observation time, and their content may encode future knowledge if written after a decision is made. Computationally, this raises alignment and leakage concerns that must be addressed through careful temporal conditioning and defensible feature construction. Even when temporally valid, the information in text can be redundant or noisy, and naive concatenation can lead to overfitting, especially under low event prevalence [4].

This paper focuses on the computer-science aspects of building a multimodal deterioration forecaster that integrates structured time-series and clinical text while remaining robust to

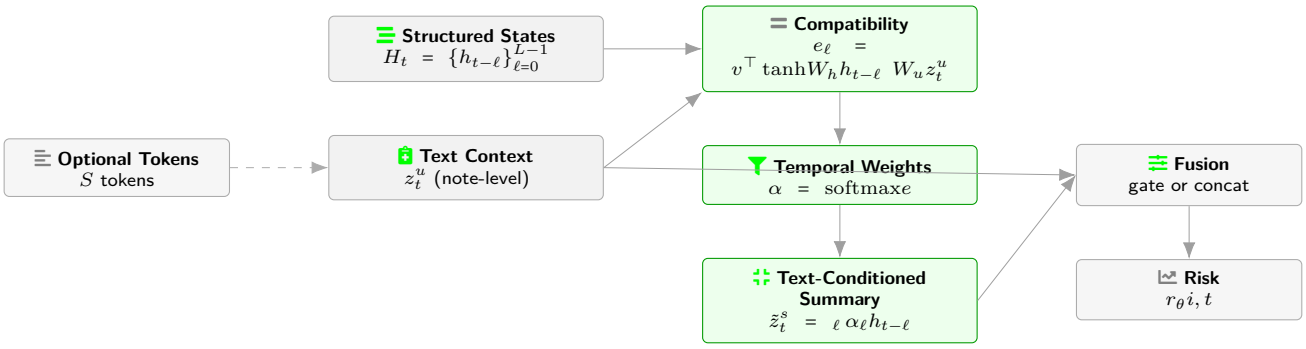

Figure 3 Cross-modal sequence-to-context coupling: text provides a query-like context that shapes attention over the structured lookback window, yielding a text-conditioned physiologic summary. The resulting representation can be fused with text via a stable operator to produce near-term deterioration risk.

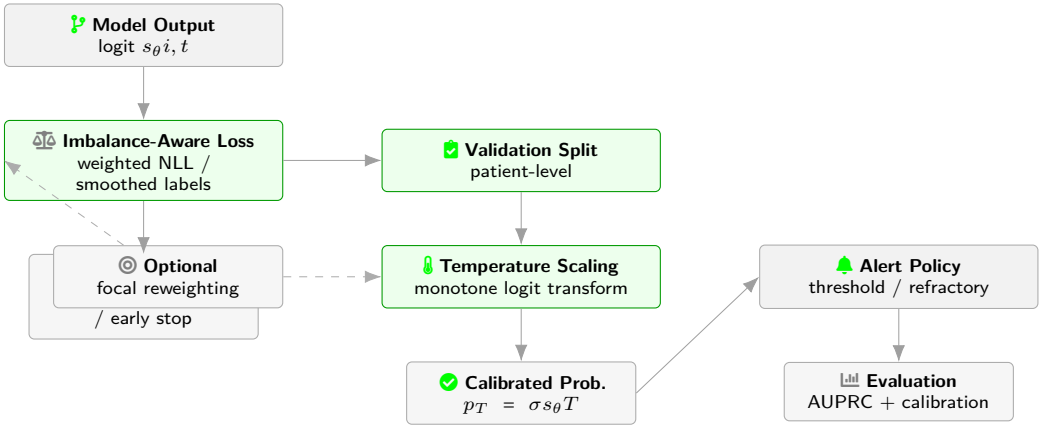

Figure 4 Training-to-decision pathway under low prevalence: an imbalance-aware objective is paired with patient-level validation and post-hoc calibration to preserve ranking while improving probability quality. A deployment-aligned alert policy is evaluated using precision–recall behavior and calibration diagnostics.

missingness and distributional artifacts. We adopt a formal sequence model view, introduce a multimodal embedding pipeline that is stable under absent notes, and analyze training objectives suited to severe class imbalance. The discussion emphasizes algorithmic choices, computational complexity, and evaluation methodology. The overall aim is to present a coherent technical blueprint that clarifies what must be specified to make multimodal ICU forecasting well-posed, reproducible, and diagnostically testable.

2. Problem Setup and Data Model

Let a patient episode be indexed by i and time be discretized into uniform bins of duration Δ, commonly one hour in ICU applications. At each discrete time $t \in \{1, \dots, T_i\}$, structured measurements are represented by a vector $x_{i,t} \in \mathbb{R}^d$, where d aggregates vitals, laboratory values, and optionally derived features such as recent slopes. Because ICU measurements are irregular, many components of $x_{i,t}$ are unobserved. We represent this with a binary mask $m_{i,t} \in \{0,1\}^d$ indicating which components are observed in bin t, and optionally a time-since-last-observation vector $\delta_{i,t} \in \mathbb{R}^d$ capturing per-channel staleness. The unstructured modality is a collection of notes $\mathcal{N}_i = \{n_{i,j}, \tau_{i,j}\}_{j=1}^{J_i}$ where $n_{i,j}$ is text and $\tau_{i,j}$ is the associated timestamp. A prediction query is issued at time t and concerns an outcome $y_{i,t} \in \{0,1\}$ indicating whether a deterioration event occurs within a horizon H bins into the future.

To prevent ill-defined conditioning, the model must satisfy a strict filtration constraint: when predicting at time t, only information available up to and including time t may be used. Formally, the predictor is a measurable function of the sigma-algebra generated by $\{x_{i,s}, m_{i,s}, \delta_{i,s}\}_{s \le t}$ and $\{n_{i,j}, \tau_{i,j} : \tau_{i,j} \le t\}$. This constraint appears trivial, but in practice it must be enforced during preprocessing to avoid label leakage through charting delays, copy-forward notes, or features computed using future bins [5]. The constraint also implies that, if a note is present, its representation should be tied to its timestamp and only included for predictions after it is available.

The modeling goal is to estimate a conditional risk function

$$r_\theta i, t = \mathbb{P}\big(y_{i,t} = 1 \,|\, \mathcal{H}_{i,t}\big)$$
$$\mathcal{H}_{i,t} = \big(\{x_{i,s}, m_{i,s}, \delta_{i,s}\}_{s \le t}, \; \{n_{i,j} : \tau_{i,j} \le t\}\big), \qquad (2.1)$$

where θ parameterizes the predictor. The evaluation is performed over a set of prediction times, typically excluding early bins where insufficient history exists. The horizon H is chosen to balance clinical lead time and predictability; larger horizons reduce leakage risk but increase uncertainty.

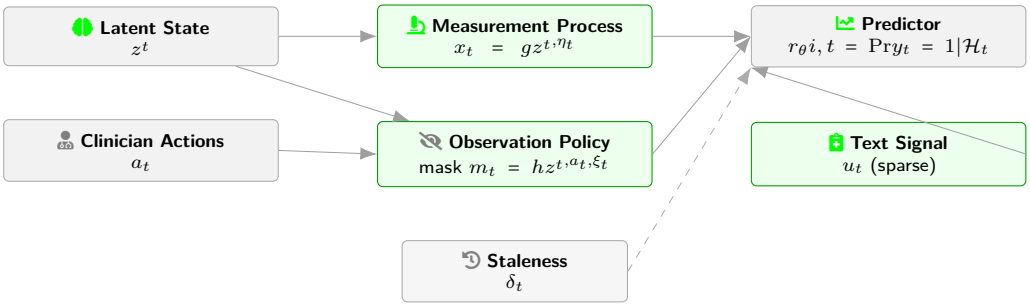

Figure 5 Predictive semantics under informative missingness: observation patterns are generated by a policy that depends on latent state and actions, making masks predictive but non-causal. The forecaster conditions jointly on values, masks, staleness, and sparse textual context to approximate the Bayes-optimal risk under the observed data process.

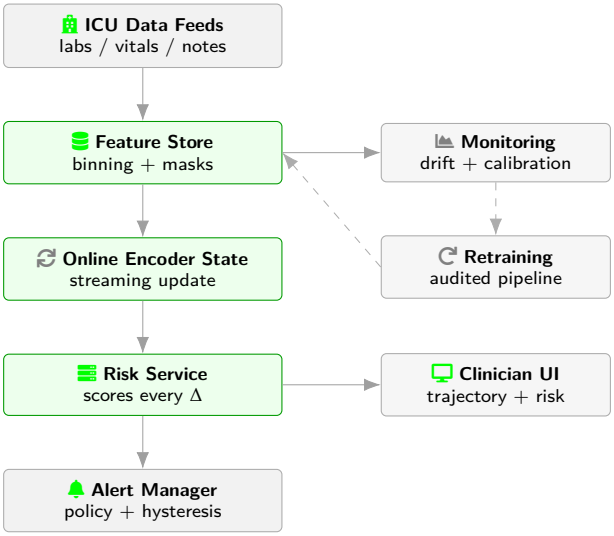

Figure 6 Streaming deployment blueprint: real-time ingestion and feature alignment feed an online-capable encoder and a scoring service. Alerts are produced via an explicit policy layer, while monitoring tracks discrimination and calibration drift, enabling audited retraining without disrupting bedside workflows.

Irregular sampling introduces a subtle statistical issue: missingness may be informative, and the pattern of observation itself can be predictive. Treating missing values as zeros without additional structure implicitly assumes missing-at-random after conditioning on observed values, which is rarely defensible in the ICU. A more accurate computational stance is to regard $x_{i,t}, m_{i,t}, \delta_{i,t}$ jointly as the observation, allowing the model to learn associations between measurement patterns and outcome. This does not solve causal confounding, but it improves predictive adequacy by acknowledging that measurement decisions reflect clinician concern [6]. The key technical requirement is that the model should not conflate imputation artifacts with physiologic signal; this motivates architectures that process masks and staleness explicitly rather than relying solely on imputed values.

The unstructured modality can be treated as a sequence of documents or as a time-varying textual context. In many datasets, note frequency is low relative to vitals, and a pragmatic approach is to define a time-indexed text representation $u_{i,t}$ as the embedding of the most recent note with timestamp $\leq t$, or a pooled embedding over all prior notes with a decay kernel. This yields a discrete-time multimodal input $x_{i,1:t}, m_{i,1:t}, \delta_{i,1:t}, u_{i,t}$ suitable for standard neural encoders while respecting temporal availability.

3. Preprocessing, Alignment, and Leakage Control

Preprocessing choices effectively define the statistical problem, particularly when the model is evaluated on millions of time-indexed samples derived from fewer patient episodes. A core decision is temporal binning. Continuous measurements and irregular labs are mapped into fixed bins via aggregation operators such as mean, median, last-observation, or clinically motivated summaries. Aggregation must be defined per variable to avoid injecting future information; for example, using the maximum within a bin that includes a post-query measurement would violate filtration. A safe operator is to include only measurements with timestamps $\leq$ the bin endpoint and compute summaries over the interval, but the bin endpoint itself must coincide with the prediction query time.

Imputation is often unavoidable because deep sequence models benefit from dense tensors [7]. Yet imputation is not innocu-

Table 1 Key preprocessing and alignment decisions for ICU forecasting.

Component	Typical choices	Design constraint
Temporal binning	Fixed-width bins of size Δ (e.g., 1 hour) shared across all channels	Query time aligned with bin endpoint
Aggregation	Mean, median, last-value, or clinically motivated summaries within each bin	Summaries must exclude measurements after the query time
Imputation	Forward-fill, interpolation, model-based or constant imputation of missing values	Preserve uncertainty via masks; avoid creating pseudo-future signal
Staleness encoding	Per-channel time-since-last-observation vector $\delta_{i,t}$	Enables model to discount stale information explicitly
Text alignment	Notes aligned by filing timestamp, optionally with a safety buffer before the query	Only notes strictly prior to the query time are included
Dataset splitting	Train/validation/test splits at the patient level rather than the sample level	Prevent leakage across episodes and correlated contexts

ous. Forward-fill can encode clinician measurement frequency into the imputed values, effectively extending an observation into future bins even when it may no longer reflect the current state. A staleness vector $\delta_{i,t}$ partially mitigates this by allowing the model to discount old values. Another strategy is to maintain both an imputed value and a mask, and to constrain early layers to treat the mask as a first-class input. From a computer-science standpoint, the goal is not to recover the true unobserved value but to provide a stable computational representation of partial information.

Text alignment is more fragile. Notes may be written after an event has occurred but filed with a chart time that is ambiguous. A conservative approach is to treat note availability at the filing timestamp and include only notes whose filing time is strictly less than the query time. When documentation delays are substantial, an additional safety margin can be imposed by requiring that notes precede the query by at least a small buffer [8]. The cost is reduced text coverage, but the benefit is reduced leakage risk. Because this paper emphasizes methodological defensibility, such conservative alignment is preferable when the goal is to build a reliable predictive system rather than to maximize apparent performance.

Dataset splitting must be performed at the patient level rather than the sample level to avoid leakage through repeated episodes or correlated histories. When generating hourly samples from each episode, the number of samples per patient can be large, and random sample-level splitting yields overly optimistic estimates by placing near-duplicate contexts in train and test sets. A patient-level split ensures that the model generalizes across individuals rather than memorizing idiosyncratic measurement patterns.

Class imbalance is intrinsic to ICU deterioration forecasting when posed as hourly risk prediction. Even if a large fraction of patients experience some adverse event during an ICU stay, the per-hour event rate can be low because the horizon window is short and only new events count [9]. This has algorithmic consequences. Loss functions that treat positives and negatives symmetrically can be dominated by negatives, producing a model that ranks well but fails to identify rare events at useful precision. Additionally, evaluation metrics must reflect the operational regime; the area under the precision-recall curve is often more informative than the area under the ROC curve when prevalence is low.

A subtle preprocessing issue is label definition. If the outcome is a composite of multiple events, the label is a union over event indicators within the horizon window. This union operation can introduce heterogeneity: distinct mechanisms may lead to different event types, and the model may learn a mixture predictor. This is acceptable for a composite alarm but complicates interpretability. A computationally principled compromise is to treat the composite label as the target for ranking but to maintain auxiliary labels for each component to support error analysis and to detect cases where the model is effectively predicting only one component [10].

Finally, the preprocessing pipeline should be treated as executable specification. Many failures in clinical ML arise from preprocessing ambiguity rather than from model design. From a systems perspective, reproducibility requires deterministic binning, deterministic imputation rules, explicit timestamp handling, and clear split logic. These elements are as important as architectural novelty because they define the information boundary between history and future.

Table 2 Structured temporal encoder families and their properties.

Encoder	Strengths	Limitations
Causal RNN (GRU/LSTM)	Handles variable-length sequences and supports streaming state updates	Sequential computation; gradients can degrade on very long histories
Bidirectional RNN on past window	Uses both past and near-future within a fixed lookback window for richer context	Requires fixed window; not naturally streaming; window choice is a hyperparameter
Temporal CNN	Parallelizable convolutions with controllable receptive field, good for local pattern detection	Limited ability to model very long-range dependencies without deep stacks
Temporal transformer	Global attention over the lookback window, flexible sequence-to-sequence modeling	Quadratic cost in window length and more sensitive to hyperparameter choices

4. Multimodal Representation Learning Architecture

A multimodal ICU forecaster can be decomposed into three algorithmic modules: a structured temporal encoder, a text encoder, and a fusion operator. The structured encoder maps a length-L window of multivariate observations into a fixed-dimensional latent vector $z^s_{i,t} \in \mathbb{R}^h$ that summarizes the recent physiologic trajectory. The text encoder maps the available notes into a latent vector $z^u_{i,t} \in \mathbb{R}^h$ in the same representational space. The fusion operator produces a combined representation $z_{i,t}$ from which a probabilistic classifier outputs $r_\theta i, t$.

For structured sequences, recurrent neural networks remain a competitive baseline because they handle variable length and preserve temporal order. A bidirectional recurrent encoder is typically inappropriate for real-time forecasting because it uses future context, but it can be used on a fixed lookback window if the window is entirely in the past relative to the query time. In that setting, bidirectionality is an algorithmic device for richer within-window modeling, not a violation of temporal causality [11]. Alternatively, causal temporal convolutions provide parallelizable computation with controlled receptive fields, and transformer-based temporal attention can model long-range dependencies at higher computational cost.

A general structured encoder can be written as

$$\begin{aligned} h_{t-\ell} &= \phi_\theta\big(\tilde{x}_{t-\ell}, m_{t-\ell}, \delta_{t-\ell}, h_{t-\ell-1}\big) \\ z^s_t &= \mathrm{Pool}\Big(\{h_{t-\ell}\}_{\ell=0}^{L-1}\Big), \end{aligned} \tag{4.1}$$

where $\tilde{x}$ denotes an imputed or normalized value vector and Pool may be attention pooling, last-state selection, or a learned set function. Attention pooling is appealing because it can focus on salient moments, such as abrupt changes, and it yields a weight distribution that can be inspected for debugging. However, attention weights are not inherently causal explanations; they are parameters of a learned aggregation and must be interpreted cautiously.

For text, transformer encoders provide contextual embeddings that capture domain-specific semantics when pretrained on clinical corpora. In ICU settings, text length is often truncated for computational reasons. A typical approach is to extract a document embedding from a special classification token or to pool token embeddings. Because note frequency is low relative to vitals, the text embedding is frequently reused across multiple prediction times until a new note appears [12]. This induces a piecewise-constant text context, which is a limitation but can still provide useful high-level signal.

The key representational design problem is to align the structured and text latents. If z^s and z^u are learned independently, their scales and geometries may differ, making fusion unstable. A pragmatic approach is to project both into a shared space via learnable linear maps with normalization. One can view this as learning modality-specific encoders followed by a common latent bottleneck. The shared space encourages the classifier to treat both modalities as alternative sources of evidence rather than as unrelated feature blocks.

Fusion can be performed by concatenation, additive combination, mixture-of-experts gating, or cross-attention. Concatenation is simple but often brittle under missing text because the classifier may overfit to patterns in the presence indicator [13]. A gated fusion operator is attractive because it yields a well-defined limit when text is absent. Let $g_t \in 0, 1^h$ be a gating vector and define

$$\begin{aligned} g_t &= \sigma\big(W_g\, z^s_t; z^u_t; s_t\ \ b_g\big) \\ z_t &= g_t \odot z^s_t \quad 1 - g_t \odot z^u_t, \end{aligned} \tag{4.2}$$

where s_t may include static covariates and note-availability indicators, σ is the logistic function, and $\odot$ is elementwise

Table 3 Text representation strategies for clinical notes.

Strategy	Granularity	Trade-offs
Most recent note embedding	Single document embedding from the last note with $\tau_{i,j} \leq t$ reused until a new note arrives	Simple alignment; may ignore earlier but still relevant narrative context
Pooled history with decay	Aggregation over all prior notes using recency-aware pooling or decay kernels	Captures longitudinal semantic context at higher computational and memory cost
Token-level contextualization	Full token sequence passed through a transformer encoder per note or group of notes	High fidelity representation; difficult to align hourly and computationally expensive
Frozen vs fine-tuned encoder	Either keep pretrained text encoder fixed or update during supervised training	Freezing improves stability; fine-tuning adapts to task but can overfit and amplify leakage

product. This structure ensures that if $z_t^u = 0$ for missing text, the model can learn $g_t \approx \mathbf{1}$ and fall back to structured features. Conversely, if structured channels are extremely sparse, the gate can shift weight toward text, though the utility of text-only prediction depends on note quality and temporal validity.

From a computer-science perspective, the gating vector can be interpreted as a learned soft routing mechanism. It is not equivalent to attention over time; instead it mediates between modalities. The gating network itself can be regularized to prevent degenerate solutions such as always selecting one modality. For example, an entropy penalty on the distribution of gate values across samples can encourage nontrivial usage, but such regularization must be applied carefully to avoid forcing the model to use noisy text.

5. Cross-Modal Attention and Sequence-to-Context Coupling

While gating provides a stable fusion primitive, it does not explicitly model interactions between specific temporal events and specific textual cues [14]. Cross-modal attention addresses this by allowing the structured sequence representation to query the text representation, or vice versa, producing context-aware latents. The design challenge is that the text modality may be represented as a single vector when notes are sparse, whereas structured data retains temporal resolution. A flexible approach is to treat the structured encoder as producing a sequence of hidden states $\{h_{t-\ell}\}_{\ell=0}^{L-1}$ and to treat text as either a document vector or a token sequence. Cross-attention can then compute a weighted aggregation of structured states conditioned on text, effectively selecting temporal patterns relevant to the textual context.

Consider a simplified form where text is a single vector z_t^u and structured states are $H_t \in \mathbb{R}^{L \times h}$. We can define a compatibility score between each structured state and the text vector, yielding attention weights over time:

$$
\begin{aligned}
e_\ell &= v^\top \tanh\big(W_h h_{t-\ell} \; W_u z_t^u \; b\big) \\
\alpha_\ell &= \frac{\exp e_\ell}{\sum_{k=0}^{L-1} \exp e_k} \\
\tilde{z}_t^s &= \sum_{\ell=0}^{L-1} \alpha_\ell \, h_{t-\ell}.
\end{aligned}
\tag{5.1}
$$

The resulting $\tilde{z}_t^s$ is a text-conditioned summary of the structured window. This mechanism can be used in conjunction with gating, where gating operates on $\tilde{z}_t^s$ and z_t^u. The computational complexity is OLh per sample for the attention calculation, which is typically acceptable for moderate L such as 48 bins, but can become costly for large windows or high throughput.

If token-level text representations are used, cross-attention can be extended to attend over both time and tokens, but the cost grows as $OL \cdot S \cdot h$ where S is the number of tokens, making it more demanding. In many clinical settings, the marginal gain of token-level fusion may not justify the cost, especially when text is sparse and repeated across many time samples. An alternative is to compute a compact text representation with a small number of learned prototypes, effectively compressing tokens into a few vectors that can be queried efficiently [15].

A systems-oriented concern is that cross-modal attention can amplify leakage if text contains explicit statements about imminent interventions. Even if the note precedes the query time, it may encode plans that reflect clinician decisions, which can turn the model into a predictor of clinician action rather than patient physiology. This is not necessarily undesirable for operational triage, but it changes the semantics of the risk score. The safest computational stance is to treat the model as

Table 4 Multimodal fusion operators for combining structured and text latents.

Fusion type	Description	Behavior with missing text	Computational aspects
Early concatenation	Concatenate z_t^s and z_t^u and feed into a shared classifier	Classifier may implicitly treat note presence as a severity proxy	Minimal overhead; simplest multimodal baseline
Late averaging	Independently map each modality to logits and average or weight them	Remains defined if one modality default score is used when absent	Requires duplicate classifier heads and weight tuning
Gated fusion	Learn gate g_t and set $z_t = g_t \odot z_t^s \;\; 1 - g_t \odot z_t^u$	Naturally falls back to structured-only when z_t^u is zeroed out	Adds a small gating network; modest extra parameters
Cross-modal attention	Condition structured summary on text or text on structured states via attention	Robust if temporal alignment is valid; can amplify subtle leakage cues	Higher cost, especially with token-level text representations

Table 5 Strategies for handling informative missingness in structured data.

Technique	Description	Advantages	Risks
Zero imputation only	Replace missing values with a constant after normalization without explicit mask usage	Very simple representation compatible with most architectures	Confounds missingness with true values; assumes missing-at-random
Forward fill + staleness	Propagate last observed value and track time since observation in $\delta_{i,t}$	Encodes measurement timing and value persistence explicitly	May propagate stale information far into the future if not downweighted
Value/mask projections	Apply separate linear maps to values and masks and sum before nonlinear layers	Encourages separation between measurement magnitude and observation pattern	Slightly more complex parameterization and tuning choices
Measurement-process modeling	Treat mask m_t as generated by a process dependent on latent state and actions	Better reflects informative missingness and clinician measurement behavior	Increased complexity and potential identifiability issues

forecasting a composite event that includes clinical actions, and to evaluate it accordingly, while avoiding causal claims.

Another important coupling is the relationship between missingness and attention. If structured states contain imputed values, attention may focus on artifacts such as repeated imputed constants. Incorporating masks into the attention score mitigates this by allowing the compatibility function to downweight low-information bins. One can define an effective state $\bar{h}_{t-\ell}$ that concatenates the hidden state with a summary of missingness in that bin, or incorporate missingness directly into the recurrent update so that hidden states reflect observation reliability.

From an optimization viewpoint, cross-modal attention introduces additional parameters that can overfit under low prevalence [16]. Regularization methods such as dropout on attention weights, weight decay, and early stopping are often necessary. Additionally, because attention is a softmax, it can become sharply peaked early in training, leading to poor gradient flow. Temperature scaling of attention logits or entropic regularization can stabilize training, though such techniques should be tuned with care.

6. Mathematical Modeling of Risk, Missingness, and Generalization

To clarify what the model is estimating, it is useful to adopt a latent-state view. Let $z_{i,t}^*$ denote an unobserved physiologic state and let observations be generated by a process that depends on $z_{i,t}^*$ and clinician actions. A simplified generative picture is

$$\begin{aligned} z_{t1}^* &= f\big(z_t^*, a_t, \varepsilon_t\big) \\ x_t &= g\big(z_t^*, \eta_t\big) \odot m_t \quad \mathrm{NA} \odot 1 - m_t \\ m_t &= h\big(z_t^*, a_t, \xi_t\big)\,, \end{aligned} \tag{6.1}$$

where a_t denotes latent clinician action variables, $\varepsilon_t, \eta_t, \xi_t$ are noise terms, and NA denotes unobserved components. The

Table 6 Learning objectives and calibration mechanisms for low-prevalence risk prediction.

Component	Purpose	Notes
Class-weighted cross-entropy	Counter class imbalance by upweighting positive examples during training	Retains proper scoring; weights tuned using validation prevalence and performance
Focal-style loss	Emphasize hard or misclassified examples and reduce influence of easy negatives	Improves rare-event recall but can degrade probability calibration quality
Label smoothing	Replace hard labels with slightly softened targets to prevent overconfident fits	Helps regularize under label noise and extreme imbalance
Temperature scaling	Post-hoc rescaling of logits s_θ via $p_T = \sigma s_\theta T$	Preserves ranking while improving calibration on a held-out validation set

missingness mask m_t depends on state and actions, making it informative. The prediction target y_t is a function of future states and actions within horizon H [17]. Under this model, a predictor that conditions on $x_{1:t}, m_{1:t}$ is estimating a risk that marginalizes over the unobserved state and actions, not the intrinsic physiologic risk independent of measurement policy.

This distinction matters for interpretation but does not invalidate prediction. From a computational standpoint, the goal is to approximate the Bayes optimal predictor under the observed data-generating process. A multimodal model that includes text implicitly conditions on clinician reasoning, potentially reducing uncertainty about a_t and thereby improving predictive performance. This is consistent with the view that notes provide a noisy but informative observation of latent clinical state that is not captured numerically.

To formalize the discriminative learning objective, consider a model $r_\theta x_{1:t}, m_{1:t}, \delta_{1:t}, u_t \in 0, 1$ trained by minimizing an expected loss. For imbalanced binary outcomes, a common choice is a weighted proper loss or a focal-style loss that emphasizes hard positives. A generic imbalance-aware loss with label smoothing can be written as

$$\begin{aligned} \tilde{y} &= 1 - \epsilon\, y \quad \frac{\epsilon}{2} \\ \ell\theta &= -w_y \left(\tilde{y} \log r_\theta \quad 1 - \tilde{y} \log 1 - r_\theta\right), \end{aligned} \tag{6.2}$$

where w_y upweights positive samples and ϵ reduces overconfidence. Focal modifications multiply the loss by a factor that downweights easy examples; this can improve precision-recall behavior but may harm calibration because it is not a strictly proper scoring rule [18]. Consequently, when focal-style objectives are used, post-hoc calibration becomes more important if the probabilities are to be interpreted quantitatively.

Calibration can be treated as a learned monotone transformation of logits. Let s_θ be the logit output and define a temperature-scaled probability

$$\begin{aligned} p_T &= \sigma\left(\frac{s_\theta}{T}\right) \\ T^\star &= \arg\min_{T>0} \sum_{i,t \in \mathcal{V}} \ell_{\mathrm{NLL}}\left(p_T,\ y_{i,t}\right), \end{aligned} \tag{6.3}$$

where $\mathcal{V}$ is a validation set. This preserves ranking under $T > 0$ while adjusting probability sharpness. In deployment, calibration should be monitored because the mapping can drift under changes in patient population or clinical practice.

Generalization in this setting is complicated by the dependence among samples derived from the same patient episode. Even with patient-level splitting, within-episode samples are correlated, and standard i.i.d. assumptions used in learning theory do not directly apply. A pragmatic computational stance is to treat each patient as a unit and consider the model as learning a function over patient histories [19]. This suggests evaluating uncertainty at the patient level rather than the sample level, for example by bootstrapping patients. It also implies that regularization and early stopping should be tuned based on patient-level splits, not on random sample splits.

Another mathematically salient issue is the relationship between AUROC, AUPRC, and decision thresholds. AUROC measures the probability that a random positive receives a higher score than a random negative, which is prevalence-invariant, while AUPRC depends on prevalence and reflects performance in the low-false-positive regime. In low-prevalence tasks, a model can have high AUROC but low AUPRC if its score distributions overlap substantially. Therefore, a technically complete evaluation must report both ranking metrics and thresholded operating points. It should also report calibration metrics such as the Brier score or expected calibration error when probabilities are used for decision support [20].

Table 7 Evaluation protocol elements for ICU deterioration forecasting.

Aspect	Metric or design choice	Rationale
Ranking performance	AUROC and average precision (area under the precision-recall curve)	Capture global discrimination, with AUPRC more informative at low prevalence
Alert precision	Precision-recall curves and F1 at a validation-chosen operating threshold	Reflects utility in high-alert-cost regimes and supports threshold selection
Calibration	Brier score, expected calibration error, and reliability diagrams	Ensure that predicted probabilities match observed event frequencies
Event-based detection	Episode-level sensitivity given a minimal lead-time and refractory periods	Aligns evaluation with how alarms would be triggered in practice
Robustness strata	Metrics stratified by note availability, missingness level, or ICU hour	Diagnose modality dependence and shortcut learning on observation patterns

7. Optimization Under Extreme Imbalance and Sparse Text

Training multimodal models on ICU data is a large-scale optimization problem with specific pathologies. The first pathology is gradient domination by negatives. In hourly forecasting with rare positives, the majority of minibatch examples are negative, and the gradient signal may push the model to reduce false positives at the expense of missing events. Class weighting partially addresses this, but naive weighting can lead to unstable training when positive examples are noisy or heterogeneous. Focal-style objectives provide a dynamic reweighting that emphasizes misclassified positives, but they can create sharp decision boundaries that are poorly calibrated.

The second pathology is modality collapse. If text is present only for a subset of episodes, the model may learn to use note availability as a proxy for severity or care pathway. This can occur even when note availability is not causally related to deterioration, because it may correlate with imaging frequency or documentation patterns [21]. A computational mitigation is to include an explicit note-availability indicator as an input and to apply regularization that discourages the classifier from relying solely on that indicator. Another mitigation is to train with modality dropout, randomly masking the text representation during training to force the model to remain competent with structured data alone. This can be understood as a form of data augmentation over modalities.

A third pathology is temporal shortcut learning. If the dataset includes later ICU hours where interventions are more likely, the model may use time-since-admission as a predictor rather than physiologic evidence. Including time features can be useful, but it risks encoding care processes rather than patient state. A cautious approach is to include time features but to test sensitivity by evaluating performance on restricted windows or by reporting stratified metrics by ICU hour. From an engineering standpoint, such diagnostics are essential to determine whether the model is learning clinically meaningful patterns or exploiting dataset artifacts [22].

Optimization stability is also affected by the computational treatment of missingness. If imputed values are set to zero after normalization, the zero vector may be a plausible physiologic value for some variables but not others. The model can misinterpret missingness as a real signal unless the mask is provided. Furthermore, if the mask is provided but the first layer is unconstrained, the model can still entangle mask and value in undesirable ways. One practical strategy is to use separate linear projections for values and masks, summing them before nonlinearity, which encourages a clearer separation between observed magnitudes and observation indicators.

Text embeddings can be either frozen or fine-tuned. Freezing reduces compute and may reduce overfitting when labeled data are limited, but it can limit task-specific adaptation. Fine-tuning improves expressivity but increases risk of leakage amplification if the text contains subtle future information [23]. A compromise is to use frozen embeddings with a learnable projection head, or to fine-tune with strong regularization and limited learning rates. In addition, truncation length and note selection strategy matter. Using the most recent note simplifies alignment but may discard earlier clinically important context. Pooling multiple notes with temporal decay can increase robustness, but it may also increase the chance of including irrelevant or redundant text.

Table 8 Deployment and robustness considerations for multimodal ICU forecasters.

Dimension	Challenge	Representative design choice
Streaming inference	Updating risk scores as new vitals and labs arrive in real time	Use causal encoders or cached hidden states for incremental updates
Alarm policy	Limiting repeated alerts and alert fatigue for staff	Apply refractory periods, hysteresis thresholds, or episode-level alert aggregation
Distribution shift	Changes in patient mix, practice patterns, or documentation templates	Perform temporal validation, monitor performance, and recalibrate as needed
Text privacy	Sensitive information embedded in free-text clinical notes	Prefer on-premises encoders, limited logging, and storage of embeddings rather than raw text
Model interpretability	Providing usable insight into predictions for monitoring and debugging	Use attention summaries, ablation tests, and perturbation-based sensitivity analyses

In the imbalanced regime, evaluation-time threshold selection should be treated as a separate optimization problem. Choosing a threshold to maximize F1 can yield a balance between precision and recall, but F1 implicitly weights false positives and false negatives equally, which may not match operational costs. In practice, one may choose a threshold that meets a target precision to manage alert fatigue, or that maximizes expected utility under a cost model [24]. The computational contribution here is not to specify a universal threshold but to require that threshold selection be performed on a validation set and reported transparently along with the resulting operating characteristics.

A concrete empirical finding reported in prior work is that incorporating clinical-note representations alongside structured time-series features can materially improve precision-recall performance for deterioration forecasting; for example, Sadanandan (2025) reports a relative improvement of 39.2% in AUPRC when adding note information to a structured-only baseline under a 24-hour horizon formulation [25]. This type of result is consistent with the information-theoretic view that text provides complementary context that reduces uncertainty about latent state and impending interventions, even when text alone is insufficient for accurate hourly prediction.

8. Evaluation Protocols and Metric Design for Low-Prevalence Forecasting

A technically defensible evaluation protocol must address dependence structure, prevalence effects, calibration, and operational constraints. Patient-level splits are necessary but not sufficient. Because a single patient can contribute many hourly samples, reporting sample-level confidence intervals can be misleadingly narrow. A more appropriate approach is to compute metrics per patient episode and then aggregate, or to bootstrap at the patient level. This yields uncertainty estimates that reflect variability across individuals rather than across correlated timepoints [26].

Metric selection should reflect the intended use of the model. If the model is used primarily for ranking patients by near-term risk, AUROC and AUPRC are relevant. AUROC assesses global ranking, while AUPRC emphasizes performance on the positive class. In low-prevalence settings, AUPRC is particularly important because it captures the precision attainable at useful recall levels. However, AUPRC can vary with prevalence across evaluation subsets, so comparisons across institutions or time periods should account for baseline rate differences. Reporting the positive rate alongside AUPRC is therefore essential.

Thresholded metrics such as precision, recall, specificity, and F1 should be reported at a clearly defined threshold selection rule. A common rule is to select the threshold that maximizes F1 on validation data, but this should be interpreted as an engineering choice, not as an intrinsic property of the model [27]. In many deployments, the acceptable alert rate is constrained, which corresponds to selecting a threshold that yields a target predicted positive fraction. Computationally, this is equivalent to choosing a quantile of the risk score distribution, which can be more stable than optimizing a metric under noisy validation estimates.

Calibration assessment is necessary whenever probabilities

are used to support decisions, allocate resources, or communicate risk. Brier score provides a proper measure of mean squared error of probability predictions. Expected calibration error summarizes deviations between predicted and observed frequencies across bins but depends on binning choice. Reliability diagrams provide more granular insight. When focal losses or class-weighted objectives are used, raw probabilities are often miscalibrated, making post-hoc calibration beneficial. Importantly, calibration should be performed on a separate calibration set or via cross-validation to avoid optimistic bias [28].

A critical but sometimes overlooked aspect is temporal evaluation. Because risk is predicted at multiple timepoints, a model can achieve high sample-level metrics while being operationally suboptimal if it triggers repeated alarms for the same episode. A more operationally meaningful metric is event-based detection: whether the model provides at least one high-risk warning within a specified lead-time window before the event, possibly with constraints on alert frequency. Implementing event-based metrics requires defining how to map time-indexed probabilities into episode-level alerts, for example via maximum risk in the last K hours or via a refractory period after an alert. This moves evaluation closer to how models would be used in practice.

Another evaluation dimension is robustness to missingness and modality availability. If text is missing for a subset of episodes, metrics should be reported separately for episodes with and without notes [29]. Similarly, performance can be stratified by missingness level in structured features, because missingness can correlate with acuity and care processes. Such stratified evaluation is technically informative because it reveals whether the model is robust across observation regimes or whether its performance depends on a particular documentation pattern.

Finally, the evaluation protocol should incorporate ablations that isolate the contribution of each modality and of the fusion mechanism. A structured-only model, a text-only model, and a multimodal model form a minimal ablation suite. Additional ablations such as removing masks, removing staleness, or replacing cross-modal attention with concatenation can diagnose whether observed gains are due to principled multimodal integration or due to incidental proxies. From a scientific computing standpoint, ablations are not merely explanatory; they are essential debugging tools that can reveal leakage, shortcut learning, or modality collapse.

9. Robustness, Interpretability, and Deployment Considerations

Robustness in ICU forecasting concerns both statistical shift and systems reliability. Statistical shift can arise from changes in patient mix, clinical protocols, documentation practices, or measurement devices [30]. A multimodal model is particularly exposed to documentation shift because note templates and clinical language evolve. From a computer-science perspective, robustness can be improved by designing representations that are less sensitive to superficial lexical variation, for example by relying on domain-pretrained encoders and by applying normalization steps that remove de-identification artifacts. However, robustness is ultimately an empirical property that must be measured through temporal validation and external testing.

An important robustness concept is invariance under measurement policy changes. Because missingness is informative, a model trained in one institution may implicitly encode that institution's measurement patterns, which may not transfer. Including masks and staleness helps, but it also allows the model to depend on those patterns. One approach to improve portability is to constrain the model to be less sensitive to masks by applying adversarial objectives that encourage invariance, or by explicitly modeling measurement processes. Such approaches, while conceptually appealing, increase complexity and require careful tuning to avoid harming predictive performance in the source domain [31].

Interpretability is multifaceted. For structured sequences, attention pooling weights can be inspected to see which times were emphasized. Gradient-based attribution can identify which variables most influence the output. For text, token-level saliency can highlight influential phrases. These tools are useful for debugging, but they do not provide causal explanations. Moreover, in multimodal models, attributions can be unstable because small changes in one modality can change gating and attention patterns. A cautious interpretability strategy is to use attributions as sanity checks rather than as clinical rationales, and to prioritize counterfactual stress tests such as masking a variable channel or perturbing a note to see whether predictions behave plausibly.

Deployment introduces additional algorithmic constraints [32]. Real-time systems must handle streaming data, late-arriving labs, and occasional data outages. This suggests using causal encoders and maintaining a state that can be updated incrementally rather than recomputing over a full window at each step. Recurrent encoders support incremental updates naturally, while windowed transformers may require caching key-value states or using sliding-window attention. Text integration must handle new notes as they arrive; a piecewise-constant text embedding updated on note events is computationally efficient. The fusion operator should be stable when text is absent, which favors gated fusion over naive concatenation.

Another deployment issue is alarm management. If the model outputs a continuous risk at every hour, naive thresholding can generate repeated alerts [33]. A refractory period, hysteresis thresholding, or aggregation into event-level alerts can reduce alert fatigue. These mechanisms should be considered part of the system design rather than afterthoughts, because they affect clinical workflow and the effective operating point. From a formal standpoint, one can treat the alerting policy as a function π mapping risk trajectories to discrete alerts, and evaluate the joint system r_θ, π on event-based outcomes.

Privacy and governance are also relevant. Text contains sensitive information and may include details not captured in structured fields. Even when de-identified, text can increase re-identification risk. Secure deployment therefore requires access control, logging, and potentially differential privacy techniques if models are trained on sensitive data. From a computational perspective, minimizing raw text retention by storing only embeddings can reduce exposure, though embeddings themselves

can leak information under adversarial attacks [34]. These considerations shape system architecture choices such as whether to run text encoders on-premises and whether to log intermediate representations.

Finally, model monitoring is essential. Performance can degrade over time due to shift, and calibration can drift, causing over- or underestimation of risk. Monitoring should track both discrimination (e.g., AUROC over recent months) and calibration (e.g., reliability in recent windows), while acknowledging that outcome labels may arrive with delay. Drift detection methods can also monitor feature distributions, missingness patterns, and text embedding statistics. Because ICU care is high stakes, any deployment should include rollback mechanisms, human-in-the-loop oversight, and clear documentation of model limitations, particularly regarding what the risk score represents and what it does not.

10. Conclusion

Multimodal ICU deterioration forecasting is a technically rich sequence modeling problem characterized by irregular sampling, informative missingness, sparse text context, and extreme class imbalance. A computationally sound approach requires explicit temporal conditioning, robust handling of missingness through masks and staleness signals, and fusion operators that remain stable when one modality is absent [35]. Gated fusion provides a well-defined mechanism for combining structured and textual representations while supporting graceful degradation, and cross-modal attention offers a way to couple temporal patterns with semantic context when alignment is defensible.

The learning objective and evaluation protocol are as important as the architecture. In low-prevalence settings, ranking metrics alone are insufficient; precision-recall behavior, thresholded operating points, and probability calibration must be assessed to understand operational utility. Imbalance-aware losses can improve rare-event detection but often necessitate post-hoc calibration if probabilities are used quantitatively. Reproducible preprocessing, patient-level splitting, and ablation suites are necessary to diagnose leakage and shortcut learning, particularly when text is incorporated.

From a computer-science viewpoint, the central methodological challenge is not merely to add an extra modality, but to define a coherent representation and optimization pipeline that respects information boundaries and remains robust under real-world data imperfections. Future work in this area is likely to benefit from tighter modeling of measurement processes, more efficient streaming-friendly encoders, and evaluation frameworks that align more closely with event-based clinical workflows. The resulting systems should be treated as decision-support components whose outputs require careful calibration, monitoring, and governance when integrated into critical care environments [36].

Acknowledgments

We would like to thank the reviewers of this research for their helpful comments and suggestions.

References

[1] L. Harding, S. Bekaert, and J. V. Appleton, "Exploring the challenges of using electronic health record systems in nursing research.," *Nurse researcher*, vol. 28, pp. 14–19, 4 2020.

[2] E. V. Bernstam, J. L. Warner, J. C. Krauss, E. P. Ambinder, W. S. Rubinstein, G. A. Komatsoulis, R. S. Miller, and J. L. Chen, "Quantifying interoperability: An analysis of oncology practice electronic health record data variability.," *Journal of Clinical Oncology*, vol. 37, pp. e18080–e18080, 5 2019.

[3] T. Adedeji, H. Fraser, and P. Scott, "Implementing electronic health records in primary care using the theory of change: A nigerian case study (preprint)," 9 2021.

[4] J. P. Drozda, A. Zeringue, B. Dummitt, B. Yount, and F. S. Resnic, "How real-world evidence can really deliver: a case study of data source development and use," *BMJ surgery, interventions, & health technologies*, vol. 2, pp. e000024–, 3 2020.

[5] A. Gohar, S. AbdelGaber, and M. Salah, "A proposed patient-centric healthcare framework for better semantic interoperability using blockchain," *Zenodo (CERN European Organization for Nuclear Research)*, 1 2022.

[6] P. Y. Chan, S. E. Perlman, D. C. Lee, J. R. Smolen, and S. Lim, "Neighborhood-level chronic disease surveillance: Utility of primary care electronic health records and emergency department claims data.," *Journal of public health management and practice : JPHMP*, vol. 28, pp. E109–E118, 6 2020.

[7] T. K. Dang, X. Lan, J. Weng, and M. Feng, "Federated learning for electronic health records," *ACM Transactions on Intelligent Systems and Technology*, vol. 13, pp. 1–17, 6 2022.

[8] X. Shen, S. Ma, P. Vemuri, M. R. Castro, P. J. Caraballo, and G. J. Simon, "A novel method for causal structure discovery from ehr data and its application to type-2 diabetes mellitus," *Scientific reports*, vol. 11, pp. 21025–, 10 2021.

[9] M. Noaeen, S. Amini, S. Bhasker, Z. Ghezelsefli, A. Ahmed, O. Jafarinezhad, and Z. S. H. Abad, "Unlocking the power of ehrs: Harnessing unstructured data for machine learning-based outcome predictions.," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2023, pp. 1–4, 7 2023.

[10] J. A. Goldstein, J. S. Weinstock, L. Bastarache, D. B. Larach, L. G. Fritsche, E. M. Schmidt, C. M. Brummett, S. Kheterpal, G. R. Abecasis, J. C. Denny, and M. Zawistowski, "Labwas: novel findings and study design recommendations from a meta-analysis of clinical labs in two independent biobanks," 4 2020.

[11] A. Blanco, A. Pérez, and A. Casillas, "Exploiting icd hierarchy for classification of ehrs in spanish through multi-task transformers.," *IEEE journal of biomedical and health informatics*, vol. 26, pp. 1–1, 3 2022.

[12] A. Calvitti, H. Hochheiser, S. Ashfaq, K. Bell, Y. Chen, R. E. Kareh, M. T. Gabuzda, L. Liu, S. Mortensen, B. Pandey, S. Rick, R. L. Street, N. Weibel, C. R. Weir, and Z. Agha, "Physician activity during outpatient visits and subjective workload," *Journal of biomedical informatics*, vol. 69, pp. 135–149, 3 2017.

[13] C. Craven, *Improved Population Management Requires Management: Care Coordination*, pp. 189–192. Productivity Press, 5 2019.

[14] M. Casey, I. Moscovice, and J. S. McCullough, "Rural primary care practices and meaningful use of electronic health records: The role of regional extension centers," *The Journal of rural health : official journal of the American Rural Health Association and the National Rural Health Care Association*, vol. 30, pp. 244–251, 9 2013.

[15] B. M. Hollister, N. A. Restrepo, E. Farber-Eger, D. C. Crawford, M. C. Aldrich, and A. L. Non, "Psb - development and performance of text-mining algorithms to extract socioeconomic status from de-identified electronic health records.," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 22, pp. 230–241, 11 2016.

[16] Z. Diao and F. Sun, "A deep-learning neural network approach for secure wireless communication in the surveillance of electronic health records," *Processes*, vol. 11, pp. 1329–1329, 4 2023.

[17] M. A. Alanezi, "A novel methodology for providing security in electronic health record using fuzzy based multi agent system," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 17, pp. 93–102, 11 2021.

[18] L. Tong, H. Wu, M. D. Wang, and G. Wang, "Introduction of medical genomics and clinical informatics integration for p-health care.," *Progress in molecular biology and translational science*, vol. 190, pp. 1–37, 7 2022.

[19] J. H. Holmes, J. Beinlich, M. R. Boland, K. H. Bowles, Y. Chen, T. S. Cook, G. Demiris, M. Draugelis, L. Fluharty, P. Gabriel, R. W. Grundmeier, C. W. Hanson, D. S. Herman, B. E. Himes, R. A. Hubbard, C. E. Kahn, D. Kim, R. Koppel, Q. Long, N. Mirkovic, J. S. Morris, D. L. Mowery, M. D. Ritchie, R. J. Urbanowicz, and J. H. Moore, "Why is the electronic health record so challenging for research and clinical care," *Methods of information in medicine*, vol. 60, pp. 32–48, 7 2021.

[20] F. Severac, E.-A. Sauleau, N. Meyer, H. Lefevre, G. Nisand, and N. Jay, "Non-redundant association rules between diseases and medications: an automated method for knowledge base construction.," *BMC medical informatics and decision making*, vol. 15, pp. 29–29, 4 2015.

[21] D. Yoon, M. Y. Park, N. K. Choi, B. J. Park, J. H. Kim, and R. W. Park, "Detection of adverse drug reaction signals using an electronic health records database: Comparison of the laboratory extreme abnormality ratio (clear) algorithm.," *Clinical pharmacology and therapeutics*, vol. 91, pp. 467–474, 1 2012.

[22] M. Marshall, P. Campbell, J. Bailey, C. A. Chew-Graham, P. Croft, M. Frisher, R. Hayward, R. Negi, T. Rathod-Mistry, S. Singh, L. Robinson, A. Sumathipala, N. Thein, K. Walters, S. Weich, and K. P. Jordan, "Feasibility of linking markers of dementia-related health in primary care medical records to cognitive function assessed in a specialist dementia service," 10 2022.

[23] M. M, N. B. J, S. A. R, N. A. R, and J. P. A, "Securing patient health record in blockchain with abe access control," *International Research Journal on Advanced Science Hub*, vol. 2, pp. 145–149, 9 2020.

[24] S. Lee, E. Kim, and K. A. Monsen, "Public health nurse perceptions of omaha system data visualization," *International journal of medical informatics*, vol. 84, pp. 826–834, 7 2015.

[25] 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.

[26] A. Kormiltsyn and A. Norta, *Dynamically Integrating Electronic - With Personal Health Records for Ad-hoc Healthcare Quality Improvements*, pp. 385–399. Germany: Springer International Publishing, 11 2017.

[27] H. Ren, J. Wang, W. X. Zhao, and N. Wu, "Kdd - rapt: Pre-training of time-aware transformer for learning robust healthcare representation," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3503–3511, ACM, 8 2021.

[28] R. Zhang, W. Zhang, N. Liu, and J. Wang, *DASFAA (3) - Susceptible Temporal Patterns Discovery for Electronic Health Records via Adversarial Attack*, pp. 429–444. Germany: Springer International Publishing, 4 2021.

[29] L. N. Tengeh and M. Bimerew, "Primary healthcare nurses' perceptions about their ability to utilise electronic health records," *Obzornik zdravstvene nege*, vol. 56, pp. 176–183, 9 2022.

[30] N. A. Dreyer, C. D. Mack, R. B. Anderson, E. M. Wojtys, E. B. Hershman, and A. K. Sills, "Lessons on data collection and curation from the nfl injury surveillance program.," *Sports health*, vol. 11, pp. 440–445, 7 2019.

[31] C. Roth, R. E. Foraker, P. R. O. Payne, and P. J. Embi, "Community-level determinants of obesity: harnessing the power of electronic health records for retrospective data analysis," *BMC medical informatics and decision making*, vol. 14, pp. 36–36, 5 2014.

[32] A. Boonstra, T. L. Jonker, M. van Offenbeek, and J. F. Vos, "Persisting workarounds in electronic health record system use: types, risks and benefits.," *BMC medical informatics and decision making*, vol. 21, pp. 1–14, 6 2021.

[33] A. G. Hall, G. K. Davlyatov, G. N. Orewa, T. S. Mehta, and S. S. Feldman, "Multiple electronic health record-based measures of social determinants of health to predict return to the emergency department following discharge.," *Population health management*, vol. 25, pp. 771–780, 10 2022.

[34] M. Jane, "Towards the development of metamed: A blockchain-based ehr system integrated with a browser-based web application for interhospital communication and patient data ownership empowerment," in *Proceedings of the International Conference on Industrial Engineering and Operations Management*, pp. 360–372, IEOM Society International, 5 2023.

[35] "Ehrs are a tool, not a solution: The nyc primary care information project," 3 2013.

[36] D. X. Yang and Y. A. Kim, "Patient-physician interactions and electronic health records," *JAMA*, vol. 310, pp. 1857–1857, 11 2013.