

Original Research

Stochastic-Process Representations for Informative Missingness in Multimodal ICU Forecasting under Distribution Shift and Causal Masking Constraints

Prakash Adhikari¹, Manoj Gautam² and Suman Bhandari³

¹Department of Computer Science and Engineering, Purbanchal University School of Science and Technology, Biratnagar–Itahari Road, Biratnagar 56613, Nepal.

²Department of Information and Communication Technology, Lumbini Technological University, Siddhartha Highway, Butwal 32907, Nepal.

³Department of Software Engineering, Nepal Open University, Man Bhawan–Jawalakhel Road, Lalitpur 44700, Nepal.

Abstract

This paper develops a framework for risk forecasting that treats irregular sampling and missingness as first-class stochastic objects rather than as artifacts to be imputed away. We formalize electronic health record inputs as marked point processes and derive a risk estimator that conditions on both observed values and the endogenous measurement policy. The model couples a latent continuous-time state space for patient dynamics with a separately parameterized observation-intensity model, enabling explicit reasoning about informative missingness, temporal uncertainty, and asynchrony between numeric streams and text-derived context. We propose modality-stable fusion operators that remain well-defined when notes are absent or temporally ambiguous, and we analyze learning objectives that preserve discrimination under extreme class imbalance while allowing post-hoc calibration to recover probability semantics. Finally, we outline evaluation protocols that separate gains due to physiologic signal from gains due to measurement-policy shortcuts, including stratified testing across missingness regimes, shift-sensitive calibration checks, and event-level operational metrics. The resulting treatment provides a principled path from irregular clinical logs to reliable probabilistic forecasting models that are compatible with streaming deployment and retrospective auditing.

1. Introduction

Forecasting clinical deterioration in intensive care is often framed as a supervised learning task in which a model maps a patient’s recorded history to a probability of an adverse event in a future horizon [1]. While this framing is operationally useful, it obscures an essential computational fact: the recorded history is not an unbiased sample of an underlying physiologic trajectory. It is a censored, workflow-mediated trace, composed of measurements taken when devices are attached, when nurses chart, when clinicians order labs, and when documentation is finalized. From a machine learning perspective, the input is a joint realization of at least two coupled processes: a latent patient state evolution and an observation process that decides which features become visible at which times. Standard pipelines that discretize time, forward-fill values, and treat missingness as a mask can succeed empirically, but they typically conflate these processes in ways that are hard to diagnose under shift and hard to justify when one cares about stable probability semantics.

Informative missingness is the most visible symptom of this coupling. In a simple sensor network, missingness might reflect random packet loss. In the ICU, missingness often reflects clinical reasoning, resource constraints, and patient stability. A stable patient may have fewer arterial blood gas measurements, fewer lactate measurements, and fewer interventions, producing sparser observation traces. Conversely, a highly unstable patient may have fewer recorded values in certain channels because of

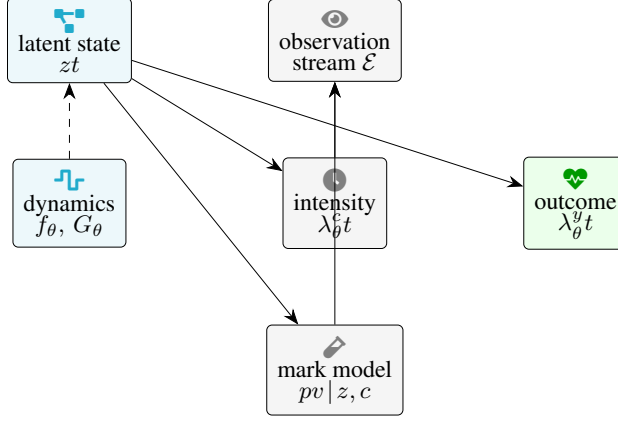


Figure 1: Generative separation of physiology and measurement: a continuous-time latent state drives (i) channel-wise measurement propensity and (ii) observed values, while an outcome intensity is conditioned on the same latent trajectory.

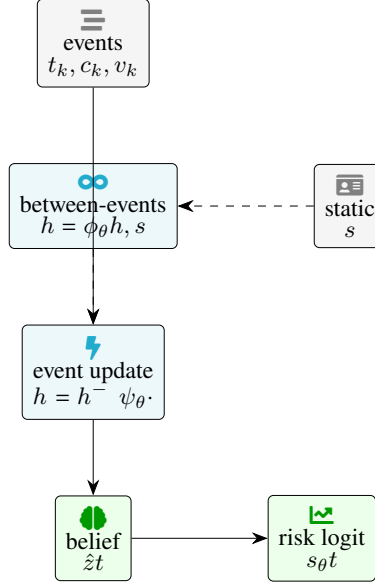


Figure 2: Event-driven encoder: continuous evolution propagates hidden state across irregular gaps, with sparse jump updates at measurement arrivals to produce a queryable belief state for downstream risk scoring.

emergent transport, device changes, or documentation delay, even while the patient is clinically worsening. Thus, missingness patterns can encode acuity, but not in a way that is globally monotone or institution-invariant [2]. A model that learns to use missingness as a proxy for severity may obtain good ranking on one dataset while failing under another measurement policy, and it may produce probabilities that are miscalibrated when the base rate changes or when measurement intensity shifts.

Multimodal inputs compound this issue. The unstructured modality, typically clinical notes or reports, is not sampled continuously. Notes are event-driven and can be temporally ambiguous: a note may summarize observations over a period, be written after a decision has been made, or be copied forward with minor edits. A naive use of text embeddings risks introducing future information through documentation lag. A conservative use of text reduces leakage but also yields piecewise-constant context signals that may align poorly with physiologic changes. The computational challenge is to represent text in a way

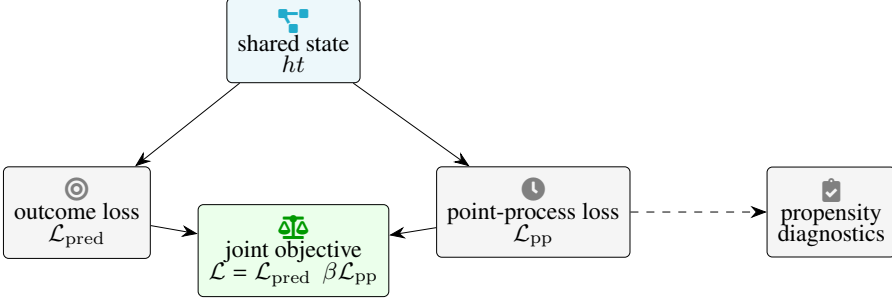


Figure 3: Coupled training signals: the representation is constrained to explain both outcome discrimination and observed measurement timing, enabling controlled incorporation of sampling-policy information.

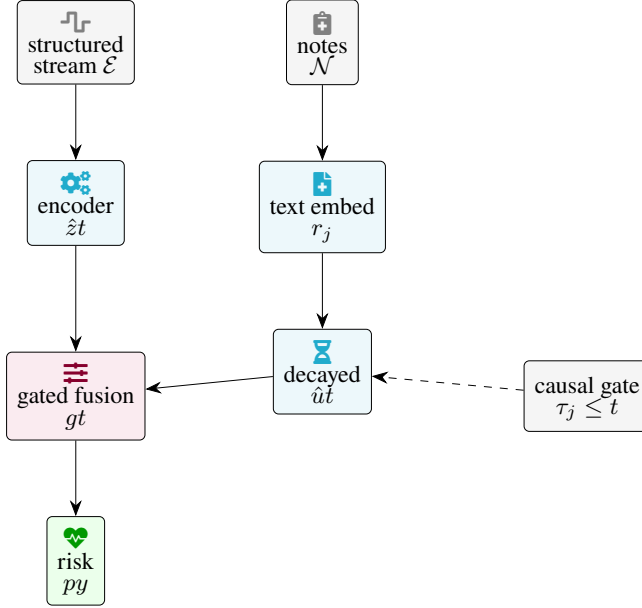


Figure 4: Modality-stable fusion: structured and text pathways remain independently well-defined, while a gated operator blends them using a decayed text context and a conservative availability-time constraint.

that respects causality constraints, leverages semantic context when available, and degrades gracefully when it is absent, without turning note availability into a shortcut feature.

This paper argues that a principled treatment of ICU forecasting should explicitly model the observation process and treat missingness and sampling intensity as stochastic objects. The goal is not to claim a fully causal interpretation of risk, but to construct estimators that remain stable under realistic perturbations to measurement policy. We therefore propose a modeling framework with three tightly coupled components. First, we represent structured measurements as a marked temporal point process whose marks include observed values and missingness masks, and we introduce an observation-intensity model that captures the endogenous likelihood that a measurement is taken [3]. Second, we model the latent patient state as a continuous-time stochastic process that can be queried at discrete decision times, thereby avoiding implicit assumptions induced by fixed binning. Third, we fuse text-derived context using modality-stable operators that are defined even when text is missing or delayed, ensuring that the model has an explicit fallback mode.

A central theme throughout is identifiability and evaluation. If the observation process carries information about clinician concern, conditioning on it can improve prediction, but it also risks learning

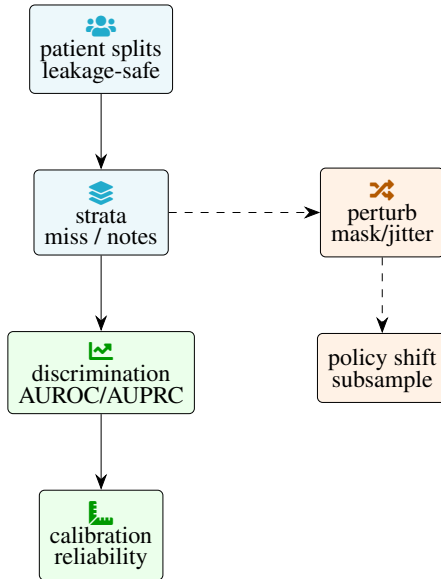


Figure 5: Shift-oriented evaluation: report metrics within missingness and modality strata, and add controlled stress tests that probe sensitivity to observation-pattern perturbations and measurement-rate changes.

Element	Notation	Description	Example in ICU setting
Patient state	zt	Latent continuous-time physiology and context	Multi-organ stability, underlying trajectory
Structured events	$\mathcal{E}$	Timestamped numeric measurements	Heart rate, lactate, arterial blood gas
Note events	$\mathcal{N}$	Asynchronous text documents	Progress notes, consults, radiology reports
Decision times	t	Times at which risk is scored	Hourly deterioration risk estimates
Outcome label	y_t	Event in a prediction horizon	Deterioration within next H hours

Table 1: Key elements of the multimodal ICU risk forecasting setup.

dataset-specific shortcuts. Distinguishing these regimes requires both mathematical analysis and careful evaluation design. We therefore develop objectives that separate discrimination from probability calibration, and we emphasize stratified evaluation across missingness regimes and shift conditions. The aim is to provide a technically grounded blueprint for building models whose performance improvements can be attributed to robust signal extraction rather than to brittle correlations in documentation patterns.

2. Related Computational Foundations

The computational literature relevant to this problem spans several areas: sequence modeling with irregular sampling, probabilistic modeling of missing data, point process learning, multimodal representation learning, and calibration under class imbalance. A key idea across these areas is that when time stamps are irregular and the observation pattern is informative, the correct object to model is not a regularly sampled vector sequence but an event stream. In computer science terms, the input is closer to a log of events than to a fixed-rate signal, and therefore the right inductive bias often involves continuous-time models or event-based encoders.

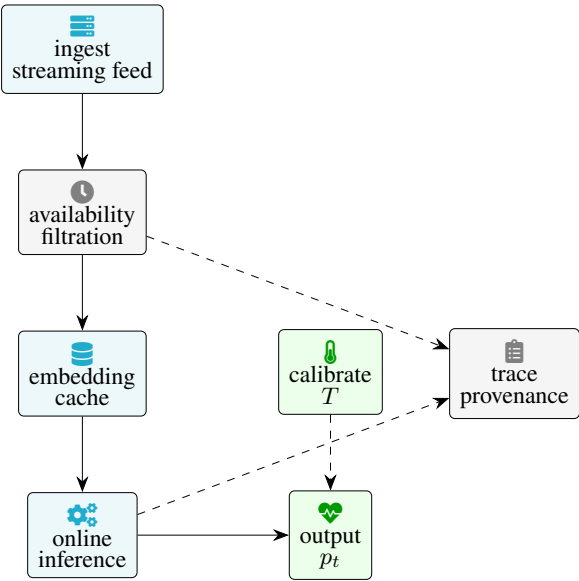


Figure 6: Deployment path with auditability: streaming ingestion is filtered by availability time, embeddings are cached at arrival, inference runs event-driven, calibration is applied as a lightweight post-layer, and each prediction is reproducible via a stored trace.

Source	Mechanism	Effect on data	Risk if mis-modeled
Clinical concern	Extra tests when instability suspected	Denser labs near deterioration	Shortcut on “more labs $\Rightarrow$ higher risk”
Resource limits	Staff and device constraints	Fewer measurements in busy periods	Apparent lower acuity from sparse data
Device availability	Device attach / detach patterns	On–off monitoring windows	Policy-specific sampling frequencies
Transport / procedures	Gaps during imaging or OR	Missing vitals during high risk	False impression of stability in gaps
Documentation lag	Notes written after events	Text summarizing past and future	Label leakage via note content and timing

Table 2: Illustrative sources of informative missingness in ICU records.

Irregular time-series modeling has evolved from classical state space models to neural formulations such as recurrent networks with time gates, temporal convolutions with dilation, and attention mechanisms with relative time embeddings [4]. Many practical ICU models discretize time into hour bins and rely on forward filling. This approach yields a computationally convenient tensor but introduces implicit assumptions: forward filling assumes persistence of values over time, and binning assumes that the order within a bin is irrelevant. These assumptions can be partially mitigated by staleness features that encode time since last observation, but the representational geometry still reflects binning artifacts. In contrast, continuous-time latent variable models treat observations as events and evolve the latent state between observations according to a learned dynamics, which can better represent sparse measurements and variable intervals.

Missing data modeling has a long history in statistics, often categorized into missing completely at random, missing at random, and missing not at random. In the ICU, missingness is frequently missing not at random because it depends on unobserved patient state and clinician action. For prediction, one

Component	Symbol / equation	Role	Notes
Latent dynamics	$dz_t = f_\theta z_t, u_t dt$	Evolves patient state in continuous time	Captures physiology and interventions
Diffusion noise	$G_\theta z_t dW_t$	Stochastic variability in state	Models uncertainty between events
Measurement intensity	$\lambda_\theta^c t$	Rate of observing channel c	Encodes observation policy and concern
Value model	$p_\theta v \mid z_t, c$	Distribution of observed value v	Separates value generation from timing
Outcome intensity	$\lambda_\theta^y t$	Hazard of deterioration event	Drives risk forecast over horizons

Table 3: Core pieces of the continuous-time generative model for measurements and outcomes.

Stage	Operation	Inputs	Purpose
Initialization	h_0 from static covariates	Demographics, admission data	Set starting belief over z_0
Between events	$\frac{dh_t}{dt} = \phi_\theta h_t, s$	Hidden state, static context	Evolve state over irregular gaps
Event update	$ht_k = ht_k^-$ $\psi_\theta ht_k^-, c_k, v_k$	Channel and value at t_k	Incorporate new measurement event
Belief projection	$\hat{z}t_k = \rho_\theta ht_k$	Updated hidden state	Low-dimensional summary of history
Prediction heads	$\lambda_\theta^c t, s_\theta t$	$\hat{z}t$ and features	Model intensities and risk jointly

Table 4: Event-driven encoder architecture for irregular ICU measurement streams.

Strategy	Definition	Behavior when notes are absent	Leakage / shortcut risk
Structured-only	Use $\hat{z}t$ only	No change; baseline model	Lower leakage, misses text signal
Early concatenation	Concatenate z_s and z_u	Undefined if z_u missing	High risk from note timing and content
Gated fusion	$zt = g \odot z_s + 1 - g \odot z_u$	Learns $g \approx 1$ when no text	Controlled influence of sparse notes
FiLM modulation	Text scales / shifts z_s	Reverts to identity modulation	Text reweights physiological features
Late fusion	Combine separate scores	Falls back to structured score	Less interaction, more robust to dropouts

Table 5: Comparison of modality fusion mechanisms under asynchronous text and causal masking.

need not recover the counterfactual fully observed trajectory; however, one does need an estimator that remains stable under changes in the missingness mechanism. In this sense, robust prediction resembles domain generalization where the domain shift is induced by a change in observation policy rather than a change in latent physiology. Computer-science methods for domain invariance, adversarial learning, and invariant risk minimization are therefore conceptually relevant, though their assumptions must be interpreted carefully in clinical settings where shifts are multifactorial.

Objective	Loss	Handles imbalance?	Effect on probability semantics
Standard cross-entropy	ℓ_{CE}	No	Probabilities calibrated to base rate if model is well specified
Class-weighted CE	$w\ell$ $w-\ell-$	Yes, via reweighting	Changes implicit prior, scores not directly calibrated
Focal loss	$1 - p^\gamma \ell$	Yes, focuses on hard examples	Improves recall, distorts probability scale
Point-process likelihood	likeli- NLL of $\lambda_\theta^c t$	Models observation times	Encourages informative hidden state about sampling
Auxiliary supervision	self- Reconstruction / forecasting losses	Uses unlabeled data	Shapes representation; probability meaning set by main loss

Table 6: Learning objectives used to train representation, intensity models, and risk predictors.

Tool / metric	Level	Quantity measured	Example use
Temperature scaling	Logit space	Global calibration of scores	Adjust probabilities after imbalance-aware training
Isotonic / spline maps	Logit or score space	Flexible monotone calibration	Capture non-linear reliability curves
AUROC	Sample level	Discrimination over all thresholds	Compare ranking across model variants
AUPRC	Sample level	Precision–recall under class imbalance	Reflect burden when positives are rare
ECE / reliability plots	Sample and stratum level	Gap between predicted and empirical risk	Monitor calibration across missingness regimes

Table 7: Calibration tools and evaluation metrics for probabilistic ICU forecasting.

Point process models provide a natural formalism for event streams [5]. A temporal point process represents event times as random variables with intensity functions that may depend on history. Marks can represent event types, measured values, or metadata such as which variable was measured. In the ICU, each measurement can be represented as an event with a variable identifier and a value, and the history dependence captures both physiologic dynamics and workflow patterns. Neural point processes extend this by using recurrent networks or attention to parameterize the intensity function. Importantly, the intensity can be interpreted as a model of measurement propensity, which can be used as an input to prediction or as a component of a joint model that disentangles observation from state.

Multimodal learning introduces additional challenges. Fusion strategies can be broadly understood as early fusion, late fusion, and intermediate fusion. Early fusion concatenates raw features or low-level embeddings, which is simple but brittle under missing modalities. Late fusion combines independently trained predictors, which is stable but can fail to exploit cross-modal interactions. Intermediate fusion, including gating and cross-attention, can represent interactions but introduces risks of shortcut learning, particularly when one modality is sparse and its presence is itself informative. A computationally careful fusion operator should therefore be explicitly defined under modality absence, and the training procedure should encourage robustness to modality dropout [6].

Axis / test	Description	What is varied	Insight provided
Missingness fraction	Group by proportion of unobserved values	Dense vs sparse windows	Separates physiologic signal from observation-policy signal
Note availability	Stratify by presence of recent notes	Text-rich vs text-poor periods	Measures dependence on unstructured modality
ICU unit type	Evaluate per unit or service line	Medical, surgical, mixed units	Detect domain shift across local policies
Synthetic event subsampling	Randomly drop measurement events	Channel-specific or global rates	Stress-test robustness to policy change
Synthetic value masking	Mask observed values, keep masks	Observation pattern vs value content	Reveal shortcutting on missingness alone

Table 8: Stratifications and stress tests for understanding model behavior under missingness and shift.

Component	Requirement	Design choice	Audit artifact
Data ingestion	Low-latency streaming	Event-driven updates from EHR feeds	Raw event logs with identifiers
Timestamp handling	Causal availability	Distinguish chart vs availability time	Per-prediction list of visible events
Note embedding cache	Efficient text handling	Encode once at note arrival	Cache of note embeddings with times
Calibration maintenance	Stable probabilities	Periodic refitting of temperature T	Versioned calibration parameters
Alerting policy	Operationally usable outputs	Thresholds, hysteresis, refractory periods	Log of alerts and triggering scores

Table 9: Systems-level considerations for deploying streaming ICU risk models with auditability.

Calibration under imbalance is another foundational topic. Neural models trained with imbalance-aware losses often produce miscalibrated probabilities, even when ranking is strong. Post-hoc calibration methods such as temperature scaling and isotonic regression can recover probability semantics without changing ranking if the calibration map is monotone. For ICU decision support, calibration is not optional if probabilities are used to set alert thresholds, allocate scarce resources, or interpret risk trajectories. Moreover, calibration can drift under prevalence shift and measurement-policy shift, which motivates modular pipelines in which calibration can be refreshed without retraining the entire encoder.

These strands motivate the framework developed in this paper: treat observations as events, model measurement propensity explicitly, fuse modalities with stability guarantees under absence, and separate discrimination learning from calibration recovery so that probability semantics can be maintained under realistic shift.

3. Generative Modeling of Measurement and Missingness

We formalize the structured ICU record as a marked event stream. Let $\mathcal{E} = \{t_k, c_k, v_k\}_{k=1}^K$ denote a sequence of measurement events up to a decision time, where $t_k \in \mathbb{R}$ is the event time, $c_k \in \{1, \dots, d\}$ is the channel identifier, and $v_k \in \mathbb{R}$ is the observed value. Channels can represent vitals, labs, or derived measurements. Missingness at a decision time is then not a primitive; it is the complement of event

occurrence in a recent interval. This reframing is computationally useful because it allows the model to encode the observation pattern through intensity functions rather than through ad hoc masks.

We posit an underlying latent state process $zt \in \mathbb{R}^h$ that represents patient physiology and context. The latent state evolves in continuous time according to a stochastic differential equation or a learned drift-diffusion model [7]. Observations are generated as marks of a point process whose intensity depends on the latent state, capturing informative sampling. A minimal generative specification is

$$\begin{aligned} dz_t &= f_\theta z_t, ut \, dt \\ &\quad + G_\theta z_t dW_t \\ \lambda_{\theta c}^c t &= \exp(a_c^\top z_t \, b_c) \\ p_{\theta v} \mid z_t, c &= \mathcal{N}(\mu_{\theta, c} z_t, \sigma_{\theta, c}^2 z_t), \end{aligned} \tag{3.1}$$

where ut is an exogenous input process capturing interventions or static covariates, W_t is a Wiener process, $\lambda_{\theta c}^c t$ is the intensity of measurement events for channel c , and the mark distribution models observed values conditional on state. This specification explicitly separates two mechanisms: the dynamics of the latent state and the dynamics of observation. Importantly, it allows the model to represent the fact that measurement frequency varies with state and channel. The exponential link for intensity yields a log-linear dependence on latent state, but other parameterizations can be used.

The risk forecasting objective concerns an outcome event in a future horizon, which we can represent as another point process with intensity $\lambda_{\theta}^y t$ that depends on latent state. Alternatively, in a discrete-time decision system, one can define a binary label for whether an event occurs in the next H hours. In either case, the central object is the conditional probability of outcome given observed history. Under the generative model, this is obtained by filtering: computing the posterior distribution of zt given the event history and then integrating outcome intensity over the horizon. In practice, exact filtering is intractable, so one uses amortized variational inference or neural filters.

A computationally pragmatic discriminative approximation is to build an encoder that maps the observed event stream to a belief state $\hat{z}t$, which approximates sufficient statistics of the latent posterior. One can implement the encoder as a continuous-time recurrent update that processes events as they arrive [8]. Between events, the hidden state evolves according to a learned ordinary differential equation, and at each event it receives an update conditioned on channel and value. A typical update structure is

$$\begin{aligned} \frac{dht}{dt} &= \phi_\theta ht, s, \quad t \in t_{k-1}, t_k \\ ht_k &= ht_k^- \, \psi_\theta(ht_k^-, c_k, v_k) \\ \hat{z}t_k &= \rho_\theta(ht_k), \end{aligned} \tag{3.2}$$

where ht is a continuous-time hidden state, s are static covariates, and the superscripts $-$ and $^+$ denote pre- and post-event states. This event-driven encoder naturally captures irregular intervals because the ODE evolution depends on elapsed time. It also allows channel-specific updates, which is critical when different variables have different clinical meaning and sampling frequency.

To incorporate informative sampling explicitly, the encoder can also model the intensity of events. One can augment the loss with a term that fits the observed event times under a learned intensity, thereby forcing the hidden state to carry information about measurement propensity. If one only trains on outcome labels, the model may still use missingness patterns, but it will do so implicitly and potentially in a brittle way. By explicitly fitting event intensities, one can regularize the representation and obtain diagnostic signals about what the model believes drives measurement. A joint objective can be written

as

$$\begin{aligned} \mathcal{L}\theta = \mathcal{L}_{\text{pred}}\theta \\ \beta \int_0^T \lambda_{\theta}^c t \, dt \\ - \sum_{k: c_k=c} \log \lambda_{\theta}^c t_k), \end{aligned} \quad (3.3)$$

where the bracketed term is the negative log-likelihood of event times under a point process intensity and β trades off prediction and observation modeling [9]. This does not provide causal disentanglement, but it provides a structured inductive bias: the hidden state must support both predicting deterioration and predicting measurement propensity, which can reduce reliance on spurious correlations that do not generalize.

When discrete-time binning is required for compatibility with existing pipelines, the event-stream framework still offers benefits. One can derive bin-level features such as counts of events per channel, time since last event, and imputed values based on last observation, but treat these as derived from an underlying point process. This suggests a representation that includes both values and intensity proxies, rather than only values and masks. In computational terms, the input becomes a tuple of feature tensors and intensity tensors, enabling the model to learn different weights for physiologic magnitudes and for observation dynamics.

4. Modality Fusion under Asynchrony and Causal Masking

Unstructured text introduces a second asynchronous event stream. Let $\mathcal{N} = \{\tau_j, r_j\}_{j=1}^J$ denote note events, where τ_j is the availability time and r_j is a text representation produced by a pretrained encoder. The availability time must be treated conservatively to respect causality constraints. In many ICU datasets, documentation time can lag observation time, and note content can summarize future interventions. A causal masking constraint therefore requires that at decision time t , the model may only condition on notes with $\tau_j \leq t$, and even then, it should be designed to be robust if τ_j is an imperfect proxy for when the information became available.

A fusion operator must satisfy two technical properties. The first is definability under missing modalities: if no notes are available, the fused representation should reduce to a structured-only representation without requiring special-case handling that introduces distributional artifacts [10]. The second is stability under temporal asynchrony: if notes are sparse and shared across many decision times, the fusion operator should not induce spurious oscillations in risk purely because a note is repeated.

We define a structured belief state $\hat{z}t$ from the event-driven encoder and a text context state $\hat{u}t$ from note events. The text context can be updated only at note times and held constant otherwise, or it can be evolved with a decay dynamics to discount older notes. A simple continuous-time update is

$$\begin{aligned} \frac{d\hat{u}t}{dt} &= -\kappa \hat{u}t \\ \hat{u}\tau_j &= \hat{u}\tau_j^- \xi_{\theta} r_j, \end{aligned} \quad (4.1)$$

where κ is a decay rate and ξ_{θ} is a learned projection of the note embedding. This yields a text context that gradually decays unless refreshed by new notes, which can reduce the impact of stale documentation.

For fusion, we consider a gated operator that computes a mixture between structured and text representations. Let $z_s t = P_s \hat{z}t$ and $z_u t = P_u \hat{u}t$ be projected into a shared space. The fused state

is

$$\begin{aligned} gt &= \sigma(W_g z_s t \ U_g z_u t \ b_g) \\ zt &= gt \odot z_s t \\ 1 - gt &\odot z_u t, \end{aligned} \tag{4.2}$$

which ensures that if text is absent and $z_u t = 0$, the model can learn $gt \approx 1$ and behave like a structured model. To avoid shortcut learning through text availability, the training procedure can include modality dropout that randomly zeros $z_u t$ with some probability even when notes exist, thereby encouraging the structured pathway to remain predictive.

Cross-modal attention can be used to condition the structured summary on text, but in ICU settings it introduces a risk of overfitting and leakage amplification [11]. A computationally conservative alternative is to allow text to modulate the structured pathway through feature-wise affine transformations. One can implement this as a learned FiLM-style modulation where text produces a scale and shift applied to structured features:

$$\begin{aligned} \alpha t &= A_\theta z_u t \\ \beta t &= B_\theta z_u t \\ \tilde{z}_s t &= \alpha t \odot z_s t \ \beta t. \end{aligned} \tag{4.3}$$

The fused state can then be computed using $\tilde{z}_s t$ instead of $z_s t$. This approach allows text to reweight physiologic features without requiring token-level alignment, and it reduces computational cost relative to cross-attention.

Causal masking constraints also apply within the structured stream when binning is used. If measurements arrive late and are backfilled, an online system must decide whether to incorporate them into past states. For real-time alerting, the causally correct approach is to ignore backfilled values for past decisions, while possibly using them for retrospective auditing. This implies that the encoder’s state update must be driven by availability time rather than by chart time. In practice, systems often have multiple timestamps, and the choice among them determines the effective filtration. A robust design logs the exact data available at each decision time, making the model’s conditioning explicit and auditable.

Finally, fusion must be compatible with calibration objectives. If text strongly shifts logits when a note arrives, the probability scale can change abruptly, undermining stable thresholds [12]. Gated fusion and decayed text context mitigate this by smoothing the influence of text. In addition, one can add a temporal smoothness regularizer on logits to discourage abrupt changes absent strong evidence in the structured stream, though this must be balanced against responsiveness to real deterioration.

5. Learning Objectives, Identifiability, and Calibration under Extreme Imbalance

The forecasting task is typically defined at discrete decision times, such as hourly intervals, with a label indicating whether deterioration occurs within a horizon. Let $y_t \in \{0, 1\}$ denote the label at decision time t and let $p_\theta t = \sigma s_\theta t$ be the predicted probability derived from a logit $s_\theta t$ computed from the fused representation. Because the positive rate is often low, naive cross-entropy training can yield conservative models with poor recall. Imbalance-aware objectives improve detection but can distort probability semantics.

A widely used imbalance-aware objective is focal loss, which downweights easy negatives and emphasizes hard examples. Another approach is class-weighted cross-entropy, which reweights positives. Both change the effective target distribution and therefore alter the relationship between the learned probability and the true conditional probability under the original prevalence. This is not a problem if one uses the output purely as a score, but it is a problem if one interprets it as a probability. A technically clean approach is to separate discrimination and calibration. The encoder and classifier are trained to optimize

ranking and rare-event detection, and a post-hoc calibration map is learned on held-out data to recover probability semantics [13].

Let $\mathcal{D}$ be a training set of decision-time samples. A discriminative training objective can be written as

$$\mathcal{L}_{\text{disc}} \theta = \sum_{t,y \in \mathcal{D}} \ell_{\text{imb}}(p_{\theta} t, y) + \lambda \|\theta\|_2^2, \quad (5.1)$$

where ℓ_{imb} is an imbalance-aware loss and λ is weight decay. The calibration stage then learns a monotone map c so that $\hat{p}t = c p_{\theta} t$ is calibrated on a calibration set. Temperature scaling in logit space is a common choice because it preserves ranking:

$$\hat{p}t = \sigma\left(\frac{s_{\theta} t}{T}\right) \\ T^* = \arg \min_{T > 0} \sum_{t,y \in \mathcal{C}} \ell_{\text{CE}}(\hat{p}t, y), \quad (5.2)$$

where $\mathcal{C}$ is a held-out calibration set and ℓ_{CE} is standard cross-entropy. This yields calibrated probabilities without changing the ordering induced by logits.

Identifiability issues arise because the model conditions on both values and observation patterns. If measurement propensity is correlated with outcome due to clinician behavior, a predictor that uses propensity may generalize poorly when that behavior changes. One can address this by regularizing the model to reduce reliance on observation features, but doing so blindly can reduce performance because propensity may be legitimately predictive in a given workflow. A more nuanced approach is to design the representation so that physiologic evidence and propensity evidence are separable, enabling evaluation to quantify how much each contributes.

In the event-stream framework, one can represent measurement propensity via learned intensities $\lambda_{\theta}^c t$ and include them as inputs to the risk predictor. To reduce shortcut learning, one can apply an information bottleneck that limits the mutual information between propensity features and the final logit, while still allowing some contribution. Operationally, this can be implemented by adding noise to propensity features or by penalizing the norm of weights connecting propensity features to the classifier. Another approach is to train with data augmentation that perturbs measurement patterns, such as randomly dropping measurement events or jittering event times, forcing the model to rely more on robust patterns rather than exact missingness configurations [14]. Such augmentation must be designed carefully to avoid generating unrealistic sequences, but mild perturbations can improve stability under small policy shifts.

A complementary identifiability perspective concerns the text modality. Note embeddings can encode clinician recognition of deterioration, which may be correlated with imminent interventions that define the label. If the model uses text to predict clinician action rather than physiologic deterioration, it can achieve high performance but change the semantics of prediction. This is not necessarily undesirable for triage, but it must be acknowledged. A computational strategy to control this is causal masking and note-type restriction, such as using only note categories that are less likely to contain explicit future decisions. Another strategy is to introduce a delay buffer so that a note influences predictions only after some time has elapsed, reducing the likelihood that it contains information about events within the forecast horizon. This trades off information content for leakage control.

Calibration also interacts with missingness. If high missingness correlates with higher base rate of events, the model’s probability outputs can be systematically different across missingness regimes. A single global calibration map may be insufficient, because the score-to-probability relationship can differ by regime [15]. One can address this by calibrating with covariates, learning a calibration function that depends on summary features such as missingness fraction or ICU type. However, covariate-dependent calibration risks overfitting and can be unstable when subgroup sizes are small. A practical compromise

is to calibrate globally but monitor calibration stratified by missingness regime, refreshing calibration as needed.

Finally, probability semantics depend on the label definition and horizon. If the horizon changes, the conditional probability changes. Therefore, calibration must be tied to a fixed label semantics. In systems that provide multiple horizons, one can train a multi-horizon predictor that outputs a vector of probabilities, each calibrated separately. This can be done by sharing the encoder and using separate output heads, with calibration maps per horizon.

6. Evaluation under Missingness Stratification and Distribution Shift

Evaluation in ICU forecasting often reports AUROC and AUPRC computed over all decision-time samples in a held-out set. While these metrics are informative, they can obscure the role of missingness and observation policy. If a model performs better on high-missingness samples, that could indicate that the latent clinical signal is stronger in those samples, or it could indicate that missingness itself is serving as a proxy for severity [16]. Distinguishing these interpretations requires stratified evaluation and careful reporting of base rates.

A principled evaluation protocol begins with patient-level splits to prevent leakage across correlated samples. It then defines a set of stratification variables, such as missingness fraction in the lookback window, note availability, ICU unit type, and time since admission. For each stratum, one reports discrimination metrics and calibration metrics along with the stratum’s base rate. Because AUPRC depends on prevalence, reporting base rate is essential to interpret differences. In addition, one should evaluate event-level detection performance, such as whether an event is preceded by at least one alert within a lead-time window, because sample-level metrics can be inflated by repeated high scores around an event.

Missingness stratification is particularly revealing. In practice, one often finds that high-missingness strata have different base rates and different score distributions. A concrete example of this phenomenon is that performance can appear higher in the high-missingness stratum, with both AUROC and AUPRC improving, even though missingness is generally viewed as harmful. This can occur because high-missingness strata correspond to later ICU phases or different care pathways with stronger signal separation [17]. In one reported stratification, the AUROC increased from 0.7296 in a 50–80% missingness group to 0.8193 in a $\geq 80\%$ missingness group, alongside a base-rate increase from 2.0% to 4.4%, illustrating how missingness can be confounded with acuity and prevalence in realistic cohorts, and why stratified reporting is necessary to interpret aggregate metrics, as described by Sadanandan (2025) [18].

This observation motivates two evaluation tasks beyond standard metrics. The first is shortcut sensitivity analysis, which measures how much predictions change when missingness patterns are perturbed while preserving observed values. For example, one can artificially mask a subset of observed values and see whether risk increases systematically, which would suggest that the model treats missingness as risk. Another perturbation is to modify event counts or time-since-last-observation features while holding values fixed. Such tests do not represent real data, but they can reveal brittleness and overreliance on observation features. The goal is not to claim causal effects, but to understand model behavior.

The second task is policy-shift robustness. If the measurement policy changes, the distribution of missingness features changes. One can simulate mild policy shifts by subsampling measurement events, reducing measurement frequency uniformly or in channel-specific ways, and evaluating whether ranking and calibration degrade. Event subsampling can be done with a channel-specific retention probability. While this does not replicate clinician-driven changes, it provides a controlled stress test that probes whether the model relies on fine-grained measurement frequency [19]. If performance collapses under mild subsampling, the model is likely brittle to policy changes.

Distribution shift also arises in text. Note availability can change with documentation policies, template updates, and note-type inclusion criteria. A model that benefits from text may degrade if notes become less frequent or if note content changes. Modality dropout during training can improve robustness by ensuring the model remains competent without notes, but evaluation should still stratify by note

availability. Additionally, one should test the model on time-separated cohorts to capture temporal drift in documentation practices.

Calibration evaluation must be integrated into these shift assessments. A model can maintain AUROC under shift while calibration degrades. Therefore, one should compute reliability curves and calibration errors in each stratum. Tail calibration is especially important because alerts operate in the high-score region. If the model is overconfident in the top decile, it can trigger many false alerts [20]. Tail calibration can be evaluated by computing empirical event rates among the highest predicted risk quantiles and comparing them to mean predicted probabilities.

Finally, evaluation should reflect the intended deployment policy. If the system will trigger alerts with a refractory period and hysteresis, metrics should be computed on the induced alert events rather than on raw hourly samples. This connects model performance to operational burden, such as alerts per patient-day, and to clinical utility, such as median lead time before events. A model that performs well under sample-level metrics but triggers repeated alerts may be unacceptable in practice. Conversely, a model with modest sample-level metrics might provide good event-level detection when combined with a well-designed policy.

7. Systems Implementation, Streaming Inference, and Auditable Causality

A forecasting model is not only an algorithm but also a system. In ICU settings, the system must ingest streaming structured data, handle late-arriving values, integrate asynchronous notes, and produce risk outputs with low latency. It must also be auditable: for any output, one should be able to reconstruct what information was available and what computations were performed. This requirement is both a safety constraint and a debugging necessity [21].

Streaming inference favors event-driven encoders. A continuous-time recurrent encoder can update state at each measurement event and propagate state forward in time as needed. In a discretized decision system, one can query the hidden state at hourly decision times by integrating the ODE between events. This avoids recomputing over a fixed window at each hour. The computational cost becomes proportional to the number of events, which can be efficient when measurements are sparse. For high-frequency vitals, one may still need to aggregate or downsample events to avoid overwhelming the encoder, but the event-driven structure remains useful for representing irregular labs and interventions.

Late-arriving data introduces a causality challenge. Suppose a lab value measured at time t is recorded in the system at time $t + \Delta$. A real-time system at time $t + \epsilon$ cannot use that value for predictions at time $t + \epsilon$ if it is not available, even though its chart time is earlier. Therefore, the system must distinguish between event time and availability time, using availability time to determine what can be conditioned on. The same is true for notes [22]. This implies that the data pipeline must preserve availability times and not overwrite them during preprocessing. In addition, the system should log the exact set of events used for each prediction, enabling retrospective verification.

Auditable causality can be formalized by defining a prediction trace. For each decision time, the trace contains identifiers of all events and notes that were visible, along with their timestamps and any preprocessing transformations applied. The model consumes this trace and produces a logit and probability. If calibration is applied post-hoc, the trace also records the calibration parameters used. This makes it possible to reproduce outputs exactly and to investigate discrepancies when outcomes differ. From a computer-science perspective, this is a provenance problem and can be addressed with structured logging and deterministic preprocessing.

Modality fusion in a system must handle missing notes gracefully. If note embeddings are computed by a large transformer, computing them repeatedly for each decision time is inefficient. An event-driven note embedding cache computes the embedding once at note arrival and stores it [23]. The text context state is then updated. This yields low incremental cost. However, transformer encoders can be sensitive to template changes and de-identification artifacts. A robust system therefore monitors embedding statistics, such as mean and variance over time windows, to detect shifts that might indicate changes in note content distribution.

The calibration component is also a systems artifact. Temperature scaling requires a scalar parameter, which can be fit on a calibration set and then deployed. Over time, the optimal temperature may drift. A system can therefore include a calibration maintenance job that periodically refits the temperature on recent labeled data, using a moving window. Because labels may be delayed, the calibration window may lag by days or weeks. Nonetheless, periodic recalibration can improve probability reliability under shift [24]. The system must then version calibration parameters and ensure that the correct version is applied for each prediction, which again is a provenance requirement.

Robustness monitoring should include both data-level and model-level signals. Data-level signals include missingness distributions, measurement rates per channel, note arrival rates, and basic summary statistics of values. Model-level signals include score distributions, fraction of scores above operational thresholds, and calibration residuals when labels become available. Drift detection can be implemented by comparing these signals to training baselines using statistical tests or control charts. When drift is detected, the system can trigger a recalibration procedure or raise an alert for human review.

Finally, integration into clinical workflows requires mapping probabilities to actions. A probability stream at hourly resolution is rarely actionable without a policy. Policies such as hysteresis thresholds and refractory periods reduce alert fatigue. Because these policies depend on probability semantics, calibration is directly tied to operational burden. Therefore, system design should treat the policy as part of the overall model, with evaluation and monitoring that reflect the policy's induced alerts [25].

8. Discussion: Separating Physiologic Signal from Observation-Policy Signal

A central tension in ICU forecasting is that the observation process carries information. Clinicians measure more when they are concerned, so measurement frequency can predict deterioration. Conditioning on measurement frequency can improve predictive performance, but it can also make the model brittle under policy changes. This tension is not resolved by simply excluding missingness features, because missingness can also reflect real physiologic stability patterns and because excluding it can reduce performance in the current workflow. The goal is therefore to build models that can use observation-process information while being robust to reasonable changes.

The event-stream and intensity modeling perspective provides one path. By explicitly modeling measurement intensity, one can quantify how much the hidden state encodes propensity and how much that propensity contributes to risk predictions. One can then evaluate models under subsampled measurement scenarios to see whether performance is stable. If performance degrades severely, one can adjust training with augmentation or regularization to reduce reliance on propensity. This is a computer-science approach to robustness that treats measurement policy as a domain variable.

Another path is to create multi-task objectives that include auxiliary tasks tied to physiologic state rather than to workflow [26]. For example, predicting future values of certain physiologic variables can encourage the encoder to represent dynamics. If the encoder can forecast future vitals, it likely encodes physiologic structure beyond observation patterns. Similarly, reconstructing masked measurements can encourage robustness to missingness. These auxiliary tasks can be implemented with self-supervised losses, which is attractive because labeled deterioration events are sparse. However, self-supervised tasks can also learn workflow artifacts if not designed carefully. For instance, reconstructing missing labs may encourage the model to infer lab ordering patterns. Therefore, auxiliary tasks should be chosen to emphasize within-patient dynamics rather than across-patient workflow regularities.

Text adds another layer to this separation problem. Notes can encode clinician concern and planned interventions, which are part of the observation-policy signal. They can also encode contextual facts not present in structured fields, such as nuanced respiratory effort, mental status changes, or imaging interpretation [27]. Distinguishing these roles is difficult. A practical approach is to treat text as context modulation rather than as a primary predictor, using fusion operators that limit the direct influence of text on logits unless supported by structured evidence. Decayed text context and modulation mechanisms can help.

Evaluation should reflect these distinctions. If a model's performance gain comes largely from observation-policy signal, it may fail under policy shift. Stratified evaluation by missingness and note availability can reveal whether gains are concentrated in certain regimes. Additionally, one can evaluate on subsets where observation patterns are controlled, such as patients with similar measurement rates, to see whether the model still discriminates. While such controlled subsets may be smaller, they provide insight into what the model is using.

Finally, probability semantics must be interpreted cautiously. A calibrated probability is calibrated with respect to the deployed environment, which includes the observation policy. If the observation policy changes, the probability semantics change [28]. Therefore, calibration maintenance is necessary, and one should avoid overinterpreting the probability as an intrinsic physiologic risk independent of workflow. From a decision-support standpoint, probabilities are still useful if they are reliable predictors of outcomes under the current care process, but their interpretation must remain anchored to that process.

9. Conclusion

This paper presented a technically grounded framework for multimodal ICU deterioration forecasting that treats irregular sampling and informative missingness as primary modeling objects rather than preprocessing nuisances. By representing structured measurements as marked event streams and introducing explicit observation-intensity modeling, the framework makes the measurement policy visible to the learning system and provides tools to analyze robustness under policy shift. Modality-stable fusion operators and causal masking constraints were developed to integrate asynchronous text context while preserving definability under modality absence and minimizing leakage risk. Learning objectives were discussed from the standpoint of identifiability and probability semantics, motivating a modular separation between discrimination learning under imbalance and post-hoc calibration recovery.

A central message is that performance improvements in clinical forecasting must be interpreted through the lens of the observation process. Missingness patterns can be predictive yet confounded, and text can provide complementary semantic context while also encoding workflow-driven signals. Therefore, evaluation should be stratified by missingness regimes and modality availability, and it should include shift-oriented stress tests that probe dependence on measurement frequency. Deployment should be treated as a systems problem requiring auditable causality, streaming inference, calibration maintenance, and monitoring for drift in both data and scores. Under these principles, multimodal forecasting models can be engineered to deliver reliable probabilistic outputs that remain meaningful under realistic operational constraints and evolving clinical workflows [29].

References

- [1] M. T. Huber, D. Y. Ling, A. S. Rozen, S. Y. Terauchi, P. Sharma, J. Fleischer-Black, L. A. Schoenherr, R. N. Hutchinson, C. Lindvall, C. A. Jones, R. T. Guerry, and A. Berlin, "Top ten tips palliative care clinicians should know about leveraging the electronic health record for data collection and quality improvement.," *Journal of palliative medicine*, vol. 26, pp. 849–855, 12 2022.
- [2] J. Chen, X. Yin, and J. Ning, "A fine-grained and secure health data sharing scheme based on blockchain," *Transactions on Emerging Telecommunications Technologies*, vol. 33, 4 2022.
- [3] T. Gizaw, M. Bogale, and T. Alemayehu, "Evaluation of the electronic health record system in maternal and child health centers of marie stopes international ethiopia," *Gates Open Research*, vol. 3, pp. 1655–, 11 2019.
- [4] R. Karmakar and A. Basu, *Implementation of a Reversible Watermarking Technique for Medical Images*, pp. 533–571. IGI Global, 6 2022.
- [5] L. Choi, C. Beck, E. McNeer, H. L. Weeks, M. L. Williams, N. T. James, X. Niu, B. Abou-Khalil, K. A. Birdwell, D. M. Roden, C. M. Stein, C. A. Bejan, J. C. Denny, and S. L. V. Driest, "Development of a system for postmarketing population pharmacokinetic and pharmacodynamic studies using real-world data from electronic health records.," *Clinical pharmacology and therapeutics*, vol. 107, pp. 934–943, 2 2020.

- [6] A. J. Holmgren, M. Kuznetsova, D. C. Classen, and D. W. Bates, "Assessing hospital electronic health record vendor performance across publicly reported quality measures.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 28, pp. 2101–2107, 8 2021.
- [7] K. S. Kaswan, L. Gaur, J. S. Dhatteval, and R. Kumar, *AI-Based Natural Language Processing for the Generation of Meaningful Information Electronic Health Record (EHR) Data*, pp. 41–86. CRC Press, 8 2021.
- [8] G. T. Nguyen and B. R. Yehia, "Documentation of sexual partner gender is low in electronic health records: Observations, predictors, and recommendations to improve population health management in primary care," *Population health management*, vol. 18, pp. 217–222, 10 2014.
- [9] D. L. Willett, V. Kannan, L. Chu, J. R. Buchanan, F. Velasco, J. D. Clark, J. S. Fish, A. R. Ortuzar, J. E. Youngblood, D. Bhat, and M. A. Basit, "Snomed ct concept hierarchies for sharing definitions of clinical conditions using electronic health record data.," *Applied clinical informatics*, vol. 9, pp. 667–682, 8 2018.
- [10] J. F. Vos, A. Boonstra, A. Kooistra, M. A. Seelen, and M. van Offenbeek, "The influence of electronic health record use on collaboration among medical specialties," *BMC health services research*, vol. 20, pp. 1–11, 7 2020.
- [11] D. Levi, M. V. Vazquez, M. V. Giussi, S. Esteban, and A. Baum, "Medinfo - representation of the transgender population in electronic health records: Implementation strategy in the public health care system of buenos aires city.," *Studies in health technology and informatics*, vol. 264, pp. 1131–1134, 8 2019.
- [12] 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.
- [13] M. Udoh, H. Okeke, L. Edet, and N. Osuchukwu, "Healthcare providers' knowledge and perceptions of electronic health records and perceived effect on health service delivery in tertiary health facilities in uyo, nigeria," *World Journal of Public Health*, 9 2023.
- [14] S. M. Shortreed, A. J. Cook, R. Y. Coley, J. F. Bobb, and J. C. Nelson, "Challenges and opportunities for using big health care data to advance medical science and public health.," *American journal of epidemiology*, vol. 188, pp. 851–861, 3 2019.
- [15] T. Poongodi, D. Sumathi, P. Suresh, and B. Balusamy, *Deep Learning Techniques for Electronic Health Record (EHR) Analysis*, pp. 73–103. Germany: Springer Singapore, 7 2020.
- [16] S. Müller, "Is there a civic duty to support medical ai development by sharing electronic health records?," *BMC medical ethics*, vol. 23, pp. 134–134, 12 2022.
- [17] S. Kumar and M. S. Koti, "Efficient authentication for securing electronic health records using algebraic structure," in *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, vol. 3, pp. 366–370, IEEE, 12 2021.
- [18] 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.
- [19] A. Knuppel, A. Boyd, J. Macleod, N. Chaturvedi, and D. M. Williams, "The long covid evidence gap: comparing self-reporting and clinical coding of long covid using longitudinal study data linked to healthcare records," 2 2023.
- [20] M. Gottumukkala, "Design, development, and evaluation of an automated solution for electronic information exchange between acute and long-term postacute care facilities: Design science research (preprint)," 11 2022.
- [21] K. K. A., J. S. P., K. S. S., D. H. C., and P. P. L., "Big data and electronic health record in pharmacy: A systematic review," *Journal of Advanced Scientific Research*, vol. 12, pp. 39–45, 10 2021.
- [22] N. A. W and S. F. Nuryati, "Evaluasi implementasi sistem electronic health record (ehr)di rumah sakit akademik universitas gadjah madaberdasarkan metode analisis pieces," *Jurnal Manajemen Informasi Kesehatan Indonesia*, vol. 3, 4 2015.
- [23] W.-J. Tuan, N. Sehgal, and A. Zgierska, "Using electronic health record's data to assess daily dose of opioids prescribed for outpatients with chronic non-cancer pain.," *Family medicine and community health*, vol. 9, pp. e001277–, 11 2021.
- [24] S. Upadhyay and H.-F. Hu, "A qualitative analysis of the impact of electronic health records (ehr) on healthcare quality and safety: Clinicians' lived experiences.," *Health services insights*, vol. 15, pp. 11786329211070722–11786329211070722, 3 2022.

- [25] A. Wright, J. Feblowitz, L. Samal, A. B. McCoy, and D. F. Sittig, “The medicare electronic health record incentive program: provider performance on core and menu measures.,” *Health services research*, vol. 49, pp. 325–346, 12 2013.
- [26] N. S. Almaghrabi and B. A. Bugis, “Patient confidentiality of electronic health records: A recent review of the saudi literature,” *Dr. Sulaiman Al Habib Medical Journal*, vol. 4, pp. 126–135, 7 2022.
- [27] N. G. Weiskopf, D. A. Dorr, C. Jackson, H. P. Lehmann, and C. A. Thompson, “Healthcare utilization is a collider: an introduction to collider bias in ehr data reuse.,” *Journal of the American Medical Informatics Association : JAMIA*, vol. 30, pp. 971–977, 2 2023.
- [28] D. B. Srinivas, D. K. M, R. H. P, and L. H, “Securing sharable electronic health records on cloud storage,” in *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1054–1059, IEEE, 2 2023.
- [29] S. Khairat, L. C. Zalla, A. Gartland, and C. Seashore, “Association between proficiency and efficiency in electronic health records among pediatricians at a major academic health system,” *Frontiers in digital health*, vol. 3, pp. 689646–689646, 9 2021.