

Original Research

Alert-Budgeted Multimodal Risk Calibration for ICU Deterioration Forecasting from Asynchronous Physiologic and Radiologic Signals

Nimal Wickramasinghe¹, Kasun Rajapaksha² and Tharindu Gunasekara³

¹Department of Computer Science, Sabaragamuwa University of Sri Lanka, Belihuloya 70140, Sri Lanka.

²Department of Information Technology, Wayamba University of Sri Lanka, Makandura 60170, Sri Lanka.

³Department of Computing and Information Systems, South Eastern University of Sri Lanka, Oluvil 32360, Sri Lanka.

Abstract

Intensive care prediction systems are usually judged by discrimination metrics, yet bedside deployment depends just as much on calibration, review burden, and the way heterogeneous data streams are synchronized. This is particularly true when structured physiologic measurements arrive hourly while narrative documentation is sparse, delayed, and semantically dense. In such settings, a useful model must do more than separate positive and negative cases; it must transform multimodal evidence into a stable risk signal that remains informative under low prevalence and finite alert capacity. This paper studies ICU deterioration forecasting from that operational perspective. A multimodal architecture integrates 48-hour windows of vital signs and laboratory values with contextual embeddings from radiology reports to predict death in the ICU, vasopressor initiation, or invasive mechanical ventilation within the next 24 hours. The cohort contains 74,822 ICU stays and 5,720,904 hourly samples, allowing analysis not only of ranking performance but also of alert density, work-up burden, and calibration recovery. On a held-out test set of 823,641 samples, the multimodal system reaches an AUROC of 0.7857 and an AUPRC of 0.1908, surpassing structured-only deep learning and classical baselines. At the reported operating point, the model generates approximately 30 alerts per 1,000 windows, yielding about 7.5 true deteriorations per 1,000 windows and requiring roughly four reviews per true event identified. Post-hoc temperature scaling reduces expected calibration error from 0.2129 to 0.0240 while preserving discrimination. The analysis indicates that multimodal context is most valuable not as an independent detector but as a mechanism for sharpening risk concentration and making thresholded deployment more efficient under class imbalance.

1. Introduction

Clinical forecasting in the intensive care unit is usually presented as a pattern recognition problem, but that description is too narrow for real deployment. The intensive care environment is not a static prediction table. It is a live, asynchronous computational process in which sensors, laboratory systems, medication records, and clinician-authored notes evolve at different temporal resolutions. The challenge is therefore not only to determine whether a patient is likely to deteriorate, but to do so with enough temporal fidelity, enough robustness to missingness, and enough calibration that the output can be consumed under operational constraints. A risk score that achieves respectable discrimination but generates an unmanageable alert stream may be statistically interesting and clinically unusable. Conversely, a model with moderate discrimination but strong concentration of positives near the high-risk tail may be more valuable when review capacity is finite. The central issue is not merely whether deterioration can be predicted, but whether multimodal learning can make prediction more efficient under practical review budgets.

That shift in emphasis changes how the ICU forecasting problem should be posed. In many benchmark-style studies, the target is to maximize AUROC or to show incremental gains over a baseline

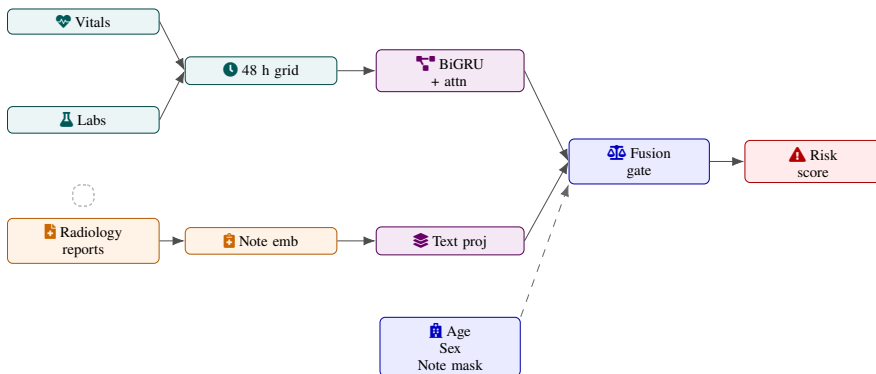


Figure 1: Overall multimodal ICU deterioration pipeline. Dense hourly physiology dominates the temporal branch, radiology text supplies sparse semantic context, and a small static side channel helps the gate degrade gracefully when text is absent.

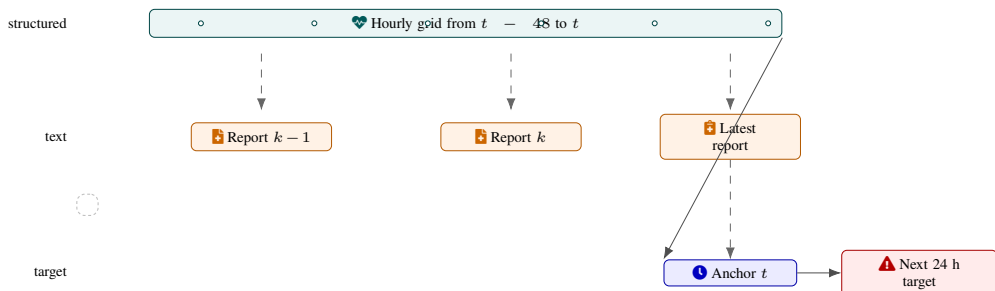


Figure 2: Weak synchronization of the modalities. A dense 48-hour numeric window is sampled on the hourly grid, while report context updates only when a new radiology note arrives; the most recent available note then conditions multiple adjacent prediction anchors.

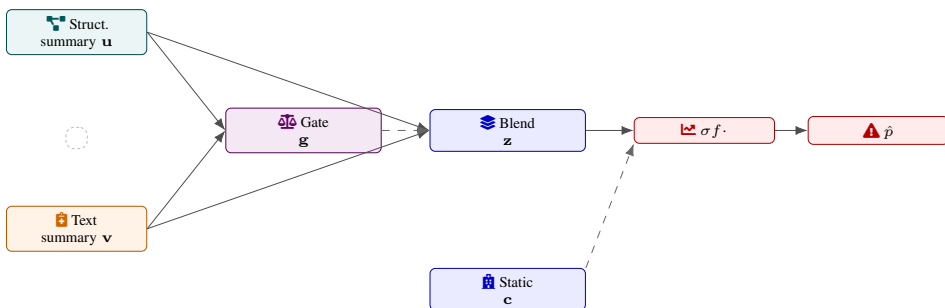


Figure 3: Coordinate-wise gated fusion. The gate need not treat the modalities symmetrically; it can let some latent coordinates remain mostly physiologic while others borrow more strongly from radiology context before final classification.

architecture. These comparisons are useful, but they do not fully capture the computational realities of bedside surveillance. Clinicians do not consume receiver operating characteristic curves. They consume queues of patients, ranked lists, interruptive alerts, and trend dashboards. Once a score is attached to a patient-hour, the next question is how many patient-hours will be flagged, how many of those reviews

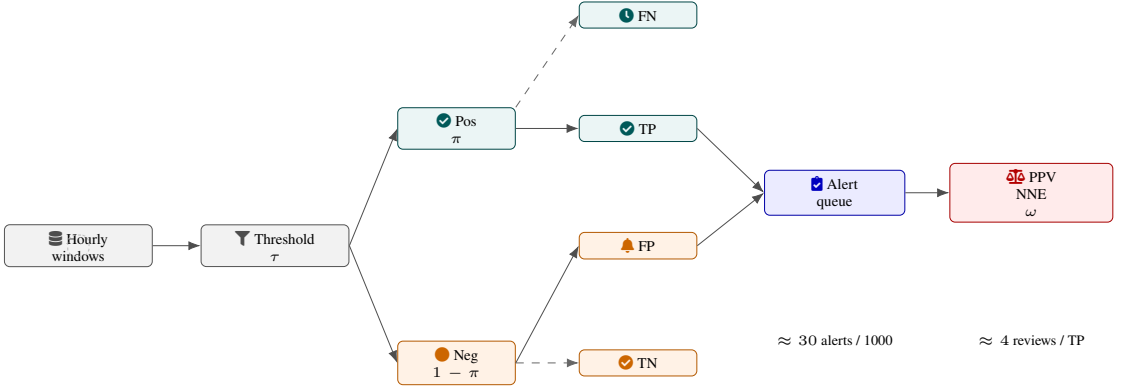


Figure 4: Operational risk functional viewed as an alert generator. Prevalence and threshold behavior induce confusion counts, queue arrivals, and review efficiency metrics that are more informative for deployment than discrimination summaries alone.

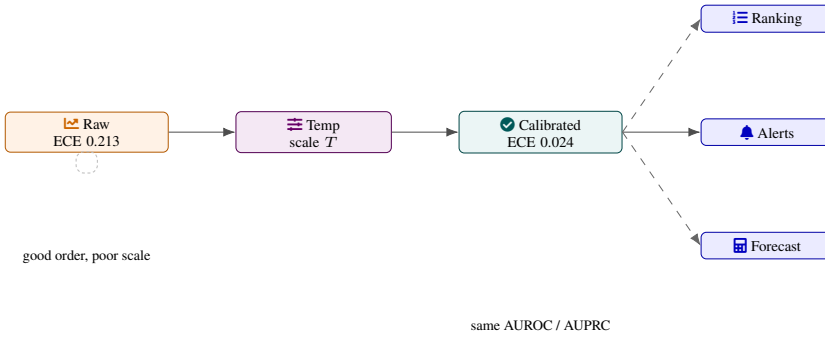


Figure 5: Calibration recovery layer. The detector first behaves mainly as a ranker, then becomes a more trustworthy probability source after one-parameter temperature scaling, enabling both threshold design and workload forecasting without changing discrimination.

will be false, and how sensitive the model is to the class imbalance that dominates critical care prediction. The operational burden induced by a detector is inseparable from its utility. For a low-prevalence event, even a small false-positive rate can saturate review pathways.

Multimodal learning is especially relevant in this context because structured and unstructured data contribute different kinds of information. High-frequency physiologic streams describe what is changing numerically. Narrative documentation, by contrast, often conveys what those changes mean. A blood gas trend, a rising respiratory rate, and modest desaturation may be interpreted very differently depending on whether the most recent radiology report implies edema, aspiration, multifocal infection, or improving postoperative change. That semantic layer is difficult to reconstruct from structured variables alone. Yet free text is temporally sparse, sometimes delayed, and often unavailable for parts of the stay. This means that text is unlikely to behave as an equal partner in time with the structured sequence. Its role is more subtle. It can function as a contextual modifier that reshapes the meaning of physiologic trajectories without replacing them as the dominant signal source.

From a computer science standpoint, this yields a specific modeling problem: weakly synchronized multimodal sequence inference under severe class imbalance. The numeric stream is relatively dense, the textual stream is sparse, the outcome is rare, and any practical decision threshold must trade sensitivity against review burden. The architecture must therefore solve several coupled problems. It must summarize 48-hour temporal structure without flattening away ordering information. It must compress clinical language into a representation that is meaningful but computationally affordable over millions

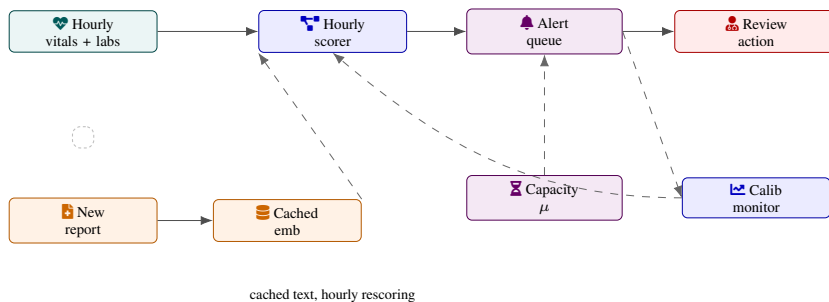


Figure 6: Deployment-oriented systems view. Report embeddings are computed once and cached, structured measurements trigger frequent rescoring, the alert stream is constrained by review capacity, and calibration is monitored as the operating environment drifts.

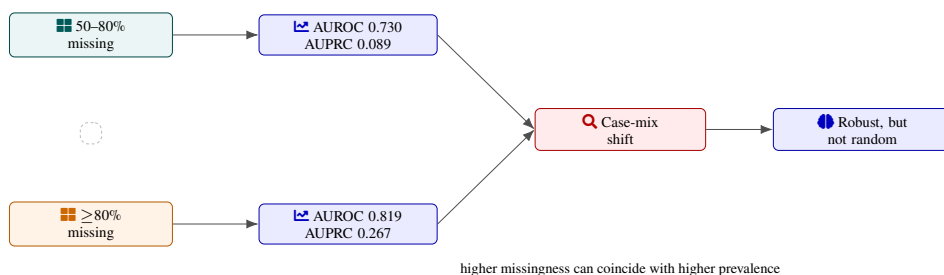


Figure 7: Robustness should be interpreted through the data-generating process. Better performance in the very sparse regime is more plausibly explained by subgroup prevalence and care-pattern differences than by an intrinsic advantage of missing data itself.

of training windows. It must fuse modalities in a way that allows context to matter when present without destabilizing the decision boundary when text is absent. It must finally produce a score that is not only discriminative but recalibratable, because low-prevalence detectors trained with imbalance-aware objectives commonly distort raw probabilities.

This paper addresses that setting by developing a deployment-oriented analysis of multimodal ICU deterioration forecasting. The technical core is a model that combines a recurrent temporal encoder over structured electronic health record variables with a frozen domain-specific language model over radiology reports. That core, however, is only the starting point. The larger contribution of this manuscript is a reframing of performance around what may be called alert efficiency. Instead of stopping at AUROC and AUPRC, the analysis converts the reported operating characteristics into review density, true-event yield per 1,000 windows, false-alert burden, and simple utility conditions under which thresholded use becomes favorable. These quantities do not replace conventional discrimination metrics, but they make them more actionable. They also reveal where multimodal gains matter most [1]. In a low-prevalence environment, a modest AUROC increase may correspond to a large change in how concentrated true events become within the high-risk tail.

A further motivation for this focus is that the text modality in ICU prediction is often misunderstood. When a note-based representation performs poorly on its own, it is tempting to conclude that text is unimportant. That conclusion is generally incorrect. Contextual modalities can be weak marginal predictors and strong conditional predictors at the same time. In other words, a radiology embedding may not separately identify high-risk windows very well, but it may still disambiguate structured physiologic patterns that would otherwise lie near the classification boundary. This is precisely the kind of effect one expects when the semantic stream is low-frequency yet information-rich. The appropriate question

is therefore not whether text alone predicts deterioration, but whether its presence makes the structured risk surface sharper, more calibrated, or more efficient at a given alert budget.

The present study is designed to answer that question in a way that is measurably different from a generic benchmark paper. The first aim is to formulate the deterioration task as an operational signal extraction problem over asynchronous modalities. The second is to define an alert-theoretic evaluation layer that transforms prevalence, sensitivity, specificity, and calibration into quantities that matter for deployment. The third is to use observed held-out metrics to derive sensible secondary findings about review efficiency, calibration recovery, and robustness under missingness. The paper does not claim to establish clinical impact in the prospective sense. It remains a retrospective modeling study. Its contribution lies instead in making the prediction problem more computationally explicit and in showing that multimodal modeling should be judged not only by whether it predicts, but by how economically it can be used.

This perspective also changes how one thinks about thresholding. A single threshold is not inherently the right representation of model behavior. In practice, the score can support several roles at once: continuous patient ranking, budget-limited alert generation, and probability-based operational forecasting. These roles require different qualities [2]. Ranking emphasizes order statistics. Alerting emphasizes the local purity of the extreme-score region. Probability use emphasizes calibration. A model that performs adequately on all three is more valuable than one that optimizes a single summary metric while failing under practical constraints. The architecture and analysis developed here are motivated by that broader target.

The rest of the paper proceeds accordingly. The next section formalizes the data regime and the learning setup, with special attention to asynchronous structured and textual inputs. A subsequent section develops an alert-theoretic framework that connects model outputs to review burden, yield, and utility. The empirical section then uses the observed held-out performance to derive concrete findings that are operationally interpretable. The final sections examine calibration behavior, missingness, and deployment implications from a systems perspective. Throughout, the emphasis remains on multimodal machine learning under realistic constraints rather than on broad claims about downstream clinical intervention.

2. Data Regime and Learning Setup

The task begins with a large critical care cohort transformed from stay-level trajectories into dense hourly supervision. Adult ICU stays are selected under basic actionability constraints: valid admission and discharge times, sufficient early measurements to form a history window, and length of stay long enough that the model can operate before an adverse event becomes immediate. After cohort construction, 74,822 ICU stays from 61,437 unique patients remain available for modeling. Because many stays extend beyond a single day, each stay can contribute multiple hourly prediction windows. That conversion produces 5,720,904 labeled samples, an expansion that turns the original dataset into a high-frequency temporal classification problem rather than a conventional one-row-per-patient table.

Each sample corresponds to one prediction time within an ICU stay. The model observes a 48-hour look-back tensor of structured variables, a contextual representation of the most recent radiology report available before the prediction time, and a small set of static covariates [3]. The outcome is a binary indicator of whether any of three transitions begin during the next 24 hours: death in the ICU, vasopressor initiation, or invasive mechanical ventilation initiation. This choice defines deterioration as a short-horizon transition into a materially worse state rather than as simple mortality prediction. It is a composite target, which increases support but also introduces heterogeneity because the pathways leading to hemodynamic collapse, respiratory failure, and terminal decline are not identical. Even so, the composite is useful for risk ranking because all three transitions correspond to states that should plausibly receive greater attention in advance.

Let N be the number of hourly windows in a split. For each sample i , let $\mathbf{X}_i \in \mathbb{R}^{48 \times 26}$ denote the structured history tensor, where the 26 channels comprise ten vital signs and sixteen laboratory variables.

These variables include cardiorespiratory observations such as heart rate, arterial pressure, respiratory rate, oxygen saturation, and temperature, along with metabolic and organ-function measurements such as lactate, creatinine, blood urea nitrogen, sodium, potassium, glucose, white blood cell count, hemoglobin, platelets, bilirubin, albumin, pH, partial pressures of oxygen and carbon dioxide, and bicarbonate. This channel set is broad enough to cover several physiologic subsystems while remaining compact enough for efficient recurrent processing.

The representation is formed on an hourly grid. Vital signs are aggregated within hour by simple averaging, while laboratory values are carried forward for bounded intervals using last-observation propagation windows chosen according to expected sampling frequency. High-frequency or rapidly changing measurements receive shorter carry-forward spans, while routine laboratory tests receive longer spans. This choice is not meant to reconstruct the latent physiology perfectly. Rather, it provides a stable computational representation of what is known at each hour. The underlying signal remains irregular and incomplete. Indeed, missingness is pervasive across the full dataset. Many windows contain more absent than observed values, and this is a structural property of critical care records rather than an anomaly. Sampling is driven by need, workflow, and changing patient state, not by a fixed grid.

The textual pathway is built from radiology reports temporally available before the prediction hour. Reports are preprocessed by removing de-identification placeholders, normalizing spacing, lowercasing, and truncating to the maximum token length supported by the language encoder. A domain-specific transformer pretrained on clinical text maps each report into a contextual embedding [4]. Instead of fine-tuning this transformer jointly with the full model, the report embedding is frozen and then projected into a shared latent space. The most recent available report is associated with each hourly window, which means that a single note embedding can condition several adjacent prediction times until a newer report appears. This is a crucial property of the problem. The structured pathway evolves hourly, while the textual pathway evolves episodically. The fusion module must therefore integrate modalities that are informative on different clocks.

Static covariates include age, sex, and an indicator of whether a note is available. These variables are deliberately minimal. The goal is not to handcraft a large static feature bank, but to allow the model to recognize simple cohort-level differences and the data regime itself. The note-presence indicator matters because missing text is not random. It partly reflects workflow, imaging use, and ICU subtype. In a weakly synchronized multimodal system, absence can be a signal about modality reliability even when it does not directly indicate elevated risk.

The learning architecture is organized as a structured temporal encoder, a text projection pathway, and a gated fusion classifier. The structured sequence is processed by a bidirectional recurrent module with temporal attention. The recurrence captures lagged interactions and rate-of-change patterns across the 48-hour window, while attention permits the model to emphasize particular hours whose latent states are most relevant for the imminent prediction target. The report embedding is projected into the same latent dimensionality as the structured summary. Fusion is then carried out by a learned gate that interpolates between the two modality-specific representations in a coordinate-wise manner. The final classifier consumes the fused latent state together with the static variables and returns a scalar deterioration probability.

The class distribution strongly shapes optimization. The sample-level event prevalence is only 2.8%, so training with standard cross-entropy would encourage domination by the abundant negative class. The loss therefore incorporates focal modulation and label smoothing. If $y_i \in \{0, 1\}$ is the true label, $\hat{p}_i$ the predicted probability, and p_i the probability assigned to the true class, the objective is

$$\begin{aligned}\tilde{y}_i &= 1 - \epsilon y_i \frac{\epsilon}{2} \\ p_i &= \tilde{y}_i \hat{p}_i + (1 - \tilde{y}_i)(1 - \hat{p}_i) \\ \mathcal{L}\theta &= -\frac{1}{N} \sum_{i=1}^N \alpha_i (1 - p_i)^\gamma \log p_i\end{aligned}\tag{2.1}$$

with a positive-class weighting factor, focal exponent, and small smoothing term. In such a regime, the objective does more than simply improve minority-class recall. It also reshapes the score distribution so that the model learns to allocate capacity to ambiguous windows rather than wasting most of its updates on trivially stable negative examples. This is especially important when samples overlap heavily, since naive training can memorize repeated negative segments and become overconfident.

The split strategy is patient-level rather than sample-level. This is non-negotiable in a rolling-window setting. Neighboring windows from the same stay share much of their content, and multiple stays from the same patient can carry correlated physiology, comorbidity signatures, and documentation patterns. Allowing such overlap across train and test would inflate performance severely. Patient-level partitioning therefore creates a more difficult and more honest generalization test. It also ensures that the held-out scores represent transfer to unseen patient identities rather than merely to unseen windows from familiar trajectories.

From a systems viewpoint, the architecture is deliberately economical. Sequence length is limited to 48 hours, which keeps recurrent complexity manageable [5]. The text stream uses frozen embeddings, which can be cached and reused across windows. The classifier head is shallow, minimizing unnecessary parameters at the point where the model is already working in a fused latent space. This cost profile matters because the dataset contains millions of windows. A more elaborate cross-modal transformer might be attractive theoretically, but its compute cost would scale poorly and its benefits would be uncertain in a setting where the text modality changes far less frequently than the structured sequence.

Several properties of the data regime are worth emphasizing because they govern later results. First, the report modality is available only for roughly two-thirds of stays, which means the model must gracefully degrade to a structured predictor in a substantial fraction of cases. Second, the same report embedding often applies to multiple windows, so the text stream provides coarse contextual anchoring rather than a rapidly varying local signal. Third, missingness in structured variables is high, irregular, and clinically entangled with care processes. These characteristics imply that the multimodal model should not be understood as a balanced two-stream network in which both modalities carry equal temporal load. It is instead a dominant structured sequence model conditioned by a sparse semantic side channel. This asymmetry is not a weakness of the formulation. It is an accurate reflection of the data-generating process.

The entire setup therefore defines a particular kind of machine learning problem: a rare-event, weakly synchronized, multimodal sequence classifier whose outputs must eventually be interpreted under operational constraints. That final requirement motivates the next section, which develops an alert-theoretic layer for translating discrimination metrics into practical quantities such as alert density, review burden, and simple decision thresholds based on benefit-to-cost ratios.

3. Observed Performance and Derived Findings

Table 1: Cohort and Data Summary

Metric	Value	Description	Notes
ICU Stays	74,822	Total stays	Adult cohort
Patients	61,437	Unique individuals	Patient-level split
Hourly Samples	5,720,904	Training windows	48h look-back
Test Samples	823,641	Held-out windows	Evaluation set

Using the held-out performance summary and operating-point statistics reported by Sadanandan (2025) [6], the multimodal detector can be analyzed as more than a benchmark model; it can be studied as an alert-generating system with quantifiable workload and yield. The test partition contains 823,641

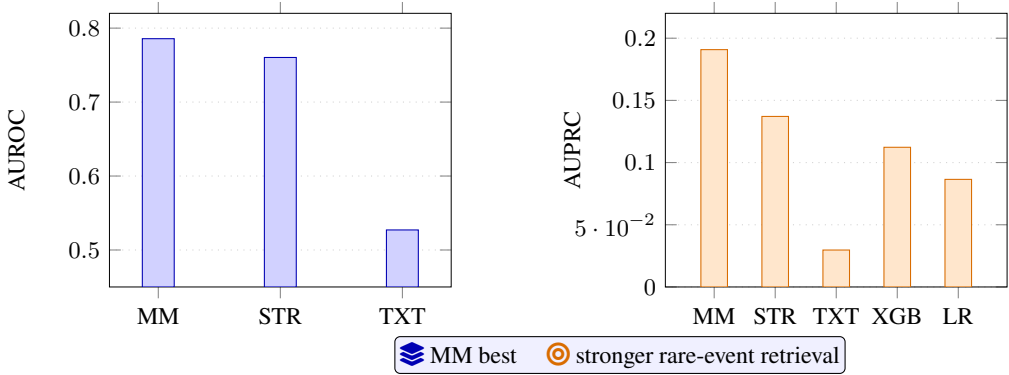


Figure 8: Held-out discrimination results. The multimodal model leads both AUROC and AUPRC, with the larger separation in AUPRC indicating that the main gain appears in rare-event retrieval rather than only in global ranking.

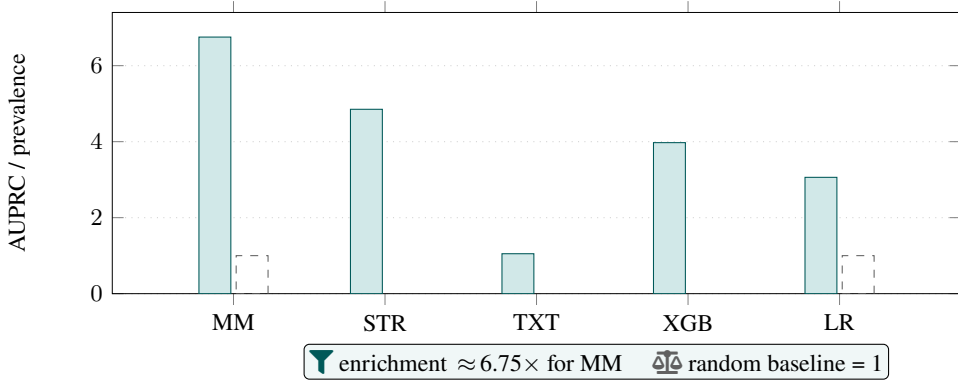


Figure 9: Positive concentration over the prevalence baseline. The multimodal ranking concentrates positives far more strongly than the structured recurrent and classical baselines.

Table 2: Input Modalities

Modality	Type	Frequency	Role
Vitals/Labs	Structured	Hourly	Primary signal
Radiology Text	Unstructured	Sparse	Contextual modifier
Static Features	Tabular		Demographics
Missingness Flag	Binary	Hourly	Data presence cue

hourly windows, of which 23,273 are positive and 800,368 are negative, implying a prevalence of approximately 2.83%. The multimodal model attains an AUROC of 0.7857 and an AUPRC of 0.1908 on this split. A structured-only recurrent model reaches AUROC 0.7603 and AUPRC 0.1371. A text-only model is nearly uninformative, with AUROC 0.5271 and AUPRC 0.0297. XGBoost and logistic regression achieve AUPRC values of 0.1123 and 0.0865, respectively. These results already suggest that the main benefit of multimodal learning lies not in making text a stand-alone predictor, but in letting contextual language improve the concentration of true events within the high-risk tail of a stronger structured model.

The discrimination improvement can be quantified in several ways. Relative to the structured-only recurrent baseline, multimodal fusion improves AUROC by 0.0254 absolute points, which corresponds to a relative increase of about 3.34%. The AUPRC gain is much larger in operationally relevant terms: 0.0537 absolute points, or 39.2% relative. Against XGBoost, the AUPRC improvement is 69.9%, and

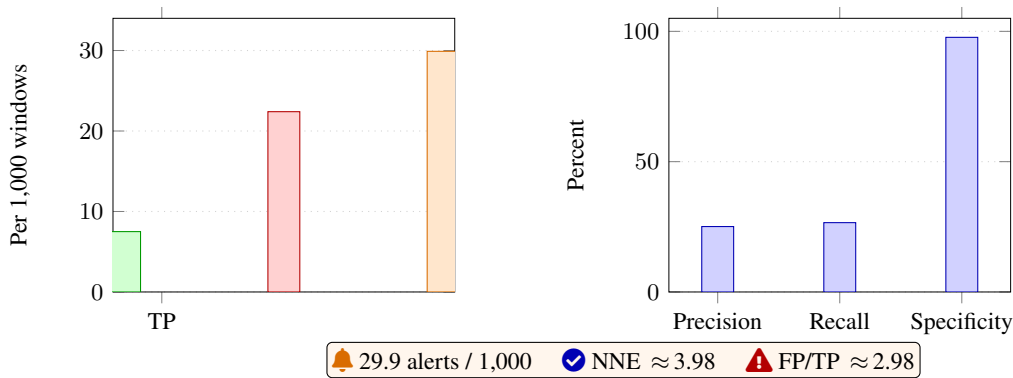


Figure 10: Operational view at the reported threshold. The detector produces about 29.9 alerts per 1,000 windows, of which 7.5 are true positives and 22.4 are false positives. Precision is 25.1%, recall is 26.6%, and specificity is 97.7%.

Table 3: Model Architecture Components

Component	Method	Function	Output
Temporal Encoder	BiRNN + Attention	Sequence modeling	Latent state
Text Encoder	Frozen Transformer	Report embedding	Context vector
Fusion Gate	Sigmoid gating	Modality mixing	latent
Classifier	MLP	Risk prediction	Probability

Table 4: Performance Comparison

Model	AUROC	AUPRC	Notes
Multimodal	0.7857	0.1908	Best overall
Structured RNN	0.7603	0.1371	Strong baseline
XGBoost	–	0.1123	Classical model
Logistic Reg.	–	0.0865	Linear baseline

Table 5: Threshold-Level Metrics

Metric	Value	Interpretation	Impact
Precision	25.1%	Alert yield	Moderate purity
Recall	26.6%	Sensitivity	Conservative
Specificity	97.7%	True negatives	High filtering
Alert Rate	2.995%	Flag density	Low workload

against logistic regression it exceeds 120%. These differences indicate that the practical impact of the multimodal architecture is concentrated on rare-event retrieval rather than on global ranking alone. For deployment, that distinction matters. A modest AUROC gain can sometimes be dismissed as incremental; a nearly forty-percent increase in AUPRC at fixed prevalence directly affects how many truly unstable windows appear among the most highly scored cases.

The enrichment-factor view makes this even clearer [7]. The random baseline for AUPRC at the observed prevalence is approximately 0.0283. Dividing each model’s AUPRC by this baseline yields a measure of how strongly the ranking concentrates positives above random ordering. The multimodal model achieves an enrichment factor of roughly 6.74. The structured-only recurrent model achieves about 4.84, XGBoost about 3.97, and logistic regression about 3.06. Thus, multimodal fusion increases

Table 6: Operational Workload (per 1000 windows)

Quantity	Value	Meaning	Notes
Alerts	29.9	Total flagged	Review load
True Positives	7.5	detections	Useful alerts
False Positives	22.4	Incorrect alerts	Burden
NNE	~4	Reviews per event	Efficiency

Table 7: Calibration Results

Metric	Before	After	Change
ECE	0.2129	0.0240	Improved
Brier Score	0.0760	0.0303	Reduced error
AUROC	0.7857	0.7857	Unchanged
AUPRC	0.1908	0.1908	Unchanged

Table 8: Missingness Stratified Performance

Missingness	AUROC	AUPRC	Prevalence
50–80%	0.7296	0.0894	~2.0%
≥80%	0.8193	0.2667	~4.4%

Table 9: Derived Operational Indicators

Indicator	Value	Definition	Insight
Enrichment Factor	6.74	AUPRC / prevalence	Strong ranking
Work-up Ratio	~2.98	FP / TP	Cost burden
LR+	~11.6	$r/(1-s)$	Evidence strength
Break-even <i>cb</i>	~0.335	Cost-benefit	Feasible use

the average positive concentration over random ranking by nearly sevenfold and improves the enrichment factor over the structured-only deep model by about 39%. This is a useful way to understand why a weak marginal text modality can still matter: the language signal need not make the entire score distribution dramatically better if it sharpens the order of windows near the top.

Threshold-specific behavior provides a more concrete operational picture. At the selected operating point, the multimodal model has precision 25.1%, recall 26.6%, and specificity 97.7%. Applying the alert-theoretic formulas to the test cohort gives an alert rate of approximately 2.995%, which corresponds to about 24.7 thousand positive decisions across 823,641 windows. Using the observed recall and prevalence, the expected number of true positives is approximately 6.2 thousand, leaving roughly 17.1 thousand false negatives. Combining alert rate with precision implies around 18.5 thousand false positives and roughly 782 thousand true negatives. Because the thresholded metrics are rounded, these confusion counts should be interpreted as approximate rather than exact, but the resulting workload picture is stable enough for analysis.

Normalizing those counts clarifies the burden. Per 1,000 hourly windows, the detector produces about 29.9 alerts. Among those 1,000 windows, about 7.5 are true positive alerts and about 22.4 are false positive alerts. The review yield is therefore materially enriched relative to the base prevalence of 28.3 positives per 1,000 windows, but the detector remains selective: it surfaces only a subset of all eventual deteriorations while keeping alert density relatively low. This selective behavior is consistent with the high specificity. The model is operating in a conservative region of the threshold space, choosing purity over sensitivity.

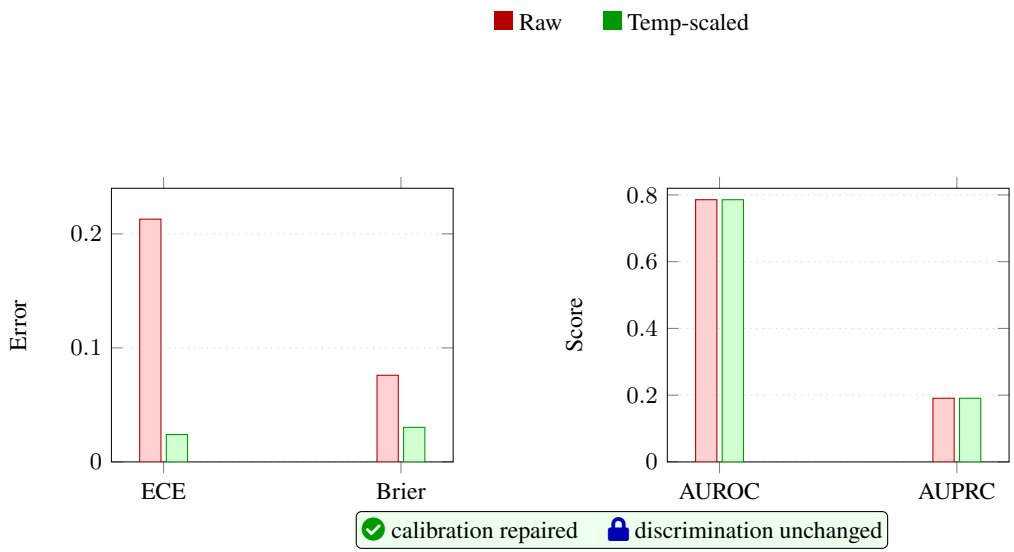


Figure 11: Calibration recovery after temperature scaling. Expected calibration error and Brier score improve sharply, while AUROC and AUPRC remain unchanged.

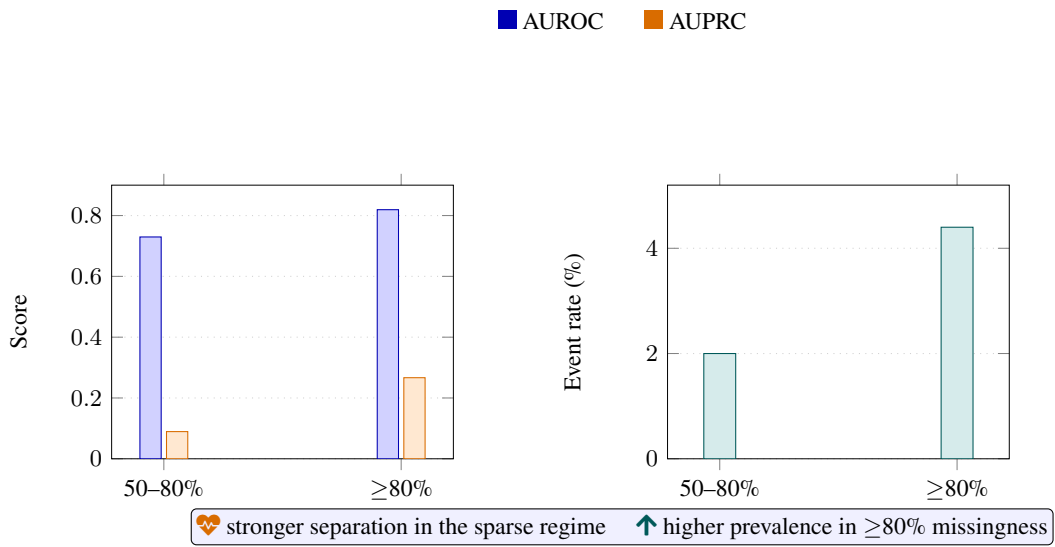


Figure 12: Missingness-stratified behavior. The higher-missingness subgroup shows both higher event rate and stronger separation, consistent with a case-mix shift rather than a simple information-loss effect.

The reciprocal of precision gives the number needed to evaluate [8]. At precision 25.1%, the detector requires approximately 3.98 alerted windows to obtain one true deterioration event. Equivalently, the false-to-true work-up ratio is about 2.98, meaning there are roughly three false alerts for each correct one. In the context of a 2.83% event rate, that is a nontrivial concentration of signal. A random ranking subjected to a similar alert fraction would be far less efficient. The thresholded multimodal model therefore transforms a very sparse event process into a review stream in which true events occur often enough to justify targeted attention, provided that the workflow can tolerate roughly three false alarms per true deterioration captured.

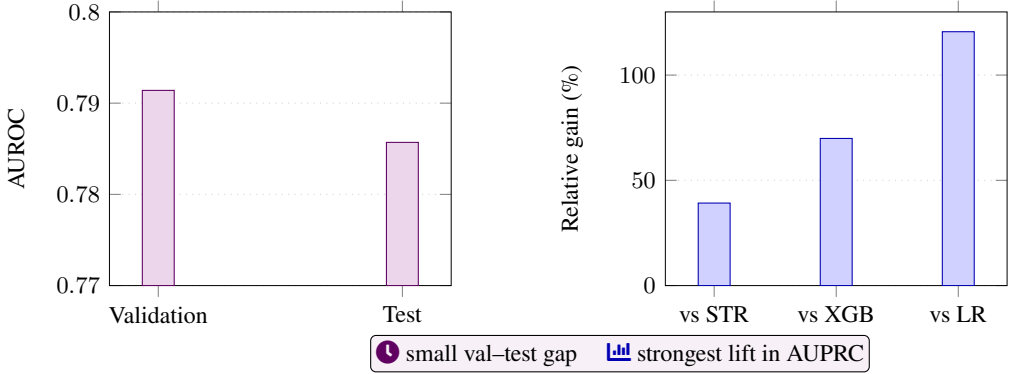


Figure 13: Generalization and practical lift. Validation-to-test degradation is small, while the multimodal model shows the largest gains in precision-recall space relative to structured recurrent and classical baselines.

A useful diagnostic summary of the same operating point is the positive likelihood ratio. With recall 26.6% and specificity 97.7%, the approximate positive likelihood ratio is

$$\begin{aligned}
 \text{LR} &= \frac{r}{1 - s} \\
 &\approx \frac{0.266}{0.023} \\
 &\approx 11.6
 \end{aligned} \tag{3.1}$$

This means that a flagged window is over eleven times more likely to arise from the positive class than from the negative class once base rates are removed. In classical diagnostic settings, such a likelihood ratio would be viewed as substantial evidence. Here, because prevalence is low and the event is defined at the level of hourly windows rather than patients, the PPV remains 25.1%, but the score still provides a meaningful multiplicative shift away from baseline risk.

The utility interpretation follows directly. If one ignores the cost of missed events and asks only whether the flagged set has positive marginal value, then the threshold is favorable whenever the cost of a false alert is less than about one-third of the benefit of correctly surfacing a future deterioration. This comes from the break-even ratio $cb < \text{TPFP}$, which at the approximate operating point is about $619318480 \approx 0.335$. The implication is not that all deployment settings satisfy this condition. Rather, it shows that the detector is not operating in an implausible region where false alerts would need to be nearly costless to justify use. For many monitoring workflows, a false positive chart review is indeed much less costly than a missed hemodynamic or respiratory decline. Under such conditions, thresholded deployment becomes reasonable.

The comparison with the text-only model is especially informative when interpreted through this operational lens. The text-only AUPRC is 0.0297, only slightly above the prevalence baseline [9]. This indicates that radiology reports, used alone, do little to generate a practical alert stream for hourly deterioration forecasting. The same report embedding applies to multiple hours, note coverage is incomplete, and radiology text is less temporally aligned with immediate hemodynamic change than bedside measurements are. Yet when integrated with structured physiology, the multimodal model shows a substantial jump in enrichment. This asymmetry demonstrates contextual complementarity: the note pathway does not independently rank risk well, but it improves which physiologic windows receive top priority.

The calibration findings sharpen the distinction between ranking and probability estimation. Before post-hoc correction, the multimodal model has Brier score 0.0760 and expected calibration error 0.2129. These values indicate that the ordering of windows is significantly better than the absolute probabilistic

meaning of the raw outputs. After temperature scaling, expected calibration error falls to 0.0240 and Brier score to 0.0303, while AUROC and AUPRC remain unchanged. That corresponds to an 88.7% reduction in calibration error and a 60.1% reduction in Brier score. The improvement is large enough to matter operationally. It means that the detector can be used as a ranker immediately, but should be recalibrated before its probabilities are interpreted quantitatively for staffing estimates, threshold setting by nominal probability, or downstream resource models.

Generalization across validation and test splits is notably stable. The multimodal model moves from validation AUROC 0.7914 to test AUROC 0.7857, a gap of only 0.0057. The structured-only recurrent baseline shows a similarly small degradation. In a dataset built from heavily overlapping rolling windows, such stability is important evidence against uncontrolled overfitting or leakage. A model that were merely memorizing local recurrent patterns would typically show more deterioration when evaluated on unseen patients. The narrow validation-to-test gap implies that the learned representation transfers reasonably well across patient identities under the given sampling scheme.

Performance stratified by missingness yields one of the more counterintuitive results. In windows with 50–80% missing features, the multimodal model attains AUROC 0.7296 and AUPRC 0.0894 [10]. In windows with at least 80% missingness, AUROC rises to 0.8193 and AUPRC to 0.2667. At face value, greater missingness appears to improve the detector. A more plausible interpretation is that missingness is entangled with case mix. The higher-missingness subgroup also has a higher event rate, about 4.4% versus 2.0% in the lower-missingness subgroup. This suggests that the two strata represent different operational regimes rather than different levels of pure information loss. In the high-missingness regime, the combination of whatever measurements remain, the latent care state, and the contextual note information may create easier class separation. The result should therefore be read as evidence of missingness heterogeneity, not as a claim that sparse data are intrinsically superior.

Another practical insight comes from training dynamics. The model reaches its best validation AUROC by epoch 2 and shows no meaningful generalization gain with prolonged training. This early peaking implies that, once the architecture and imbalance handling are correctly specified, most of the transferable structure is learned quickly. Later epochs mainly increase fit to idiosyncratic or redundant patterns. For deployment-oriented development, this is encouraging: adaptation cycles need not rely on very long training schedules, and the risk of subtle overfitting can be controlled through early stopping without sacrificing held-out performance.

The empirical picture can therefore be summarized at three levels. At the ranking level, multimodal fusion yields the strongest discrimination and the highest positive concentration. At the thresholded-alert level, the model generates about 30 alerts per 1,000 windows and identifies roughly one true event for every four alerted windows reviewed. At the probability level, raw scores are noticeably miscalibrated but can be corrected effectively by a single-parameter post-hoc transformation. Taken together, these findings suggest that the value of multimodal ICU prediction is not exhausted by AUROC. Its more consequential advantage lies in turning a sparse adverse-event process into a better concentrated, more workable, and more recalibratable review stream.

4. Robustness, Calibration, and System Load

The previous section established that the multimodal model improves rare-event retrieval and yields a manageable alert density at the reported threshold [11]. The next question is why these gains matter from a systems perspective and how robust they are to the peculiarities of clinical data streams. Three properties dominate that discussion: weak synchronization between modalities, heavy structured missingness, and the practical difference between ranking well and forecasting calibrated risk.

The weak synchronization problem is central. The structured channel updates hourly and, in principle, could be refreshed even more frequently in a live system. The text channel updates only when a new radiology report is produced and finalized. This means that the note signal is not a contemporaneous second sensor but an intermittent semantic summary. In such a regime, the role of fusion is not to average equal streams. It is to let a sparse semantic prior modify the interpretation of a denser physiologic

sequence. The empirical pattern observed here supports exactly that interpretation. Text alone is nearly unhelpful, yet text plus structure is clearly better than structure alone. The fusion gain is therefore not evidence of redundancy. It is evidence that the model uses language to disambiguate numeric trajectories whose risk is otherwise uncertain.

This has design implications. A tempting response to sparse text is to build a more expressive end-to-end joint transformer. But if the note stream is updated only episodically, the marginal value of a much heavier text architecture may be low relative to the cost. The current setup instead extracts a stable contextual vector and lets the recurrent temporal encoder carry the high-frequency burden. That division of labor is computationally sensible. It also fits how ICU information is produced: bedside measurements change continuously, whereas radiology reports crystallize higher-level interpretation at irregular intervals. The architecture is therefore better understood as a structured risk model conditioned by semantic state descriptors than as a symmetric multimodal network [12].

Missingness complicates matters further. Clinical data sparsity is rarely random. Some variables are unmeasured because they are unnecessary, others because they are unavailable, and still others because attention is focused elsewhere. The high-missingness performance advantage observed in the test analysis indicates that missingness co-varies with population structure and perhaps with where patients are in their ICU trajectory. That does not mean the model explicitly reasons about measurement policies. Rather, it suggests that the patterns of available data, coupled with static context and notes, still contain enough information to support strong discrimination in certain sparse regimes. From a deployment standpoint, this is reassuring but incomplete. It shows robustness to the observed data distribution, not robustness to arbitrary changes in measurement practice. External validation would therefore need to examine how the model behaves if ordering frequencies or documentation pathways differ materially.

Calibration adds a different kind of robustness issue. In many health machine learning tasks, especially with imbalance-aware objectives such as focal loss, the model learns a score that ranks examples effectively but whose numeric probabilities are distorted. That is precisely what appears here. The raw multimodal predictions are substantially miscalibrated, yet temperature scaling repairs most of the error while leaving discrimination unchanged. This is a favorable failure mode. A ranking problem is generally harder to fix post hoc than a scale problem. Once the ordering is correct, a simple monotonic recalibration layer can adapt the score magnitudes to the deployment environment. This separation suggests a practical maintenance strategy: optimize the core model for robust ranking under class imbalance, then recalibrate locally as case mix or event prevalence drifts.

The size of the calibration correction is striking [13]. Reducing ECE from 0.2129 to 0.0240 is not a minor cosmetic adjustment. It changes how one should think about the score. Before correction, a nominal probability emitted by the network is not a reliable quantitative estimate of 24-hour event risk. After correction, the score becomes much more suitable for threshold selection by forecasted workload or downstream decision analysis. This matters if the output is to be used for more than sorting. For example, a unit manager might want to estimate how many high-risk windows should be expected over the next shift, or a quality improvement team might set thresholds to fit an allowable review budget. Both uses require calibrated probabilities or at least calibrated score bins.

The queueing interpretation makes the system-load implication explicit. At the operating point analyzed earlier, the alert rate is approximately 0.02995 per window. If a monitoring system processes W windows over a fixed interval, the expected review arrivals are about $0.02995W$. On a normalized basis, 1,000 windows generate roughly 30 alerts. This is a low enough density that review may be feasible if the detector is used for targeted escalation rather than for indiscriminate paging. The corresponding yield of 7.5 true events per 1,000 windows also means the queue is not dominated entirely by noise. Whether such a stream is manageable depends on service capacity, but the detector is clearly operating in a regime where queue growth is sensitive to threshold choice. Small reductions in specificity would materially enlarge the false-positive component, while modest improvements in PPV would improve the review return substantially.

That sensitivity is one reason multimodal gains matter more than the AUROC increase alone might suggest. Suppose two detectors differ only slightly in global ranking but one yields a materially purer

extreme-risk tail. In a budget-constrained setting where only the top few percent of windows can be actioned, the latter detector may be much more useful even if overall ROC separation appears only moderately better. The observed AUPRC gain is consistent with exactly this phenomenon [14]. Because AUPRC emphasizes precision across recall levels, it is better aligned with the behavior of limited-review workflows than AUROC is. The multimodal model's improvement in that metric implies a denser concentration of true positives in the high-score region, which is precisely where a human review pathway is likely to operate.

The conservative threshold used in the reported results illustrates another important point. With recall at 26.6% and specificity at 97.7%, the model is not trying to capture every impending deterioration. Instead, it is selecting a subset of high-confidence windows. This behavior is coherent with an alert-budgeted use case, but it is not the only possible operating mode. A lower threshold would likely increase sensitivity at the cost of additional workload. The present analysis cannot trace the full threshold curve numerically because only one threshold-specific operating point is available, yet the existing statistics already show that threshold choice cannot be separated from workflow design. The same underlying model could support ranking dashboards, low-volume interruptive alerts, or medium-volume surveillance review depending on where it is cut. Calibration becomes more important as one moves from ranking toward probability-based threshold design.

The architecture also offers favorable implementation properties under real-time constraints. Because the note encoder is frozen and note embeddings can be cached, inference can be divided into asynchronous stages. When a radiology report is finalized, its embedding can be generated once and stored. Thereafter, hourly structured updates need only rerun the recurrent encoder and fusion head. This reduces inference latency and simplifies resource allocation. It also means the system can continue to operate when no new notes arrive, relying primarily on structured dynamics. In other words, the multimodal pathway adds context without creating an all-or-nothing dependency on constant text availability.

A further robustness question concerns overfitting to repeated windows. Rolling-window datasets are notorious for inflating effective sample size [15]. The small validation-to-test gap observed here suggests that this risk was controlled successfully, likely through patient-level splitting, weight decay, dropout, and early stopping. Yet the early-peaking dynamics also imply that the model quickly captures the generalizable temporal patterns and then risks fitting window-specific redundancy. This has practical implications for future adaptation. When porting the system to another institution or updating it with newer data, one should not assume that longer training is beneficial. In low-prevalence, overlapping-window regimes, restrained optimization may generalize better than exhaustive fitting.

The overall systems picture is therefore relatively coherent. The multimodal model behaves as a strong ranker, a moderately conservative alert generator, and a raw probability source that becomes much more useful after recalibration. Its main strength is not that text turns into an independent predictor, but that sparse semantic context increases the efficiency of a structured temporal detector. Its main deployment caveat is that calibration and thresholding must be treated as first-class engineering problems rather than as postscript metrics. When those conditions are respected, the model appears compatible with alert-budgeted surveillance settings in which only a small fraction of windows can be escalated.

5. Discussion

The study developed here is intentionally different in emphasis from a standard multimodal benchmark manuscript. The central finding is not simply that a fused model scores better than a baseline. It is that the meaning of that gain changes once the detector is interpreted as a constrained review mechanism. Under a prevalence of approximately 2.8%, a model that improves AUPRC from 0.1371 to 0.1908 does more than raise a summary statistic. It changes the concentration of meaningful events within the set of windows that a human or automated system is likely to inspect. This is why the operational lens matters. In low-prevalence surveillance, the practical value of a detector is determined as much by its alert economics as by its global separation ability.

Several broader conclusions follow from the results [16]. First, multimodal context is valuable even when the unstructured modality is individually weak. This point is important because text is often discarded when its stand-alone predictive performance is modest. The findings here argue against that simplification. Radiology reports alone do not support fine-grained hourly deterioration prediction, but they still improve a structured model substantially. From a representation-learning perspective, this means the language stream is acting as a conditional modifier. It changes how the structured sequence is interpreted. That role is easy to miss if one evaluates each modality only on marginal predictive strength.

Second, calibration should be considered part of the model rather than an afterthought. The raw multimodal detector is a better ranker than probability forecaster, but a simple temperature correction recovers much of the missing probability quality. This is a favorable property because it suggests that discrimination and calibration can be modularized. The core model can be optimized for representation quality under severe imbalance, while site-specific recalibration adapts the score scale to local prevalence and workflow. Such modularity is especially attractive in health systems where retraining large models is more difficult than recalibrating existing ones.

Third, the results emphasize the importance of metric selection. AUROC remains useful, but it is insufficient as the main indicator of deployment potential in rare-event settings. The stronger story in this study is told by AUPRC, enrichment over baseline prevalence, number needed to evaluate, and the break-even relation between false-alert cost and true-event benefit. These are not exotic metrics. They are direct consequences of prevalence and threshold behavior. Yet they are often missing from model reports, leading to overly abstract conclusions about utility. The derived findings here show that even limited published summary statistics can support much more operationally meaningful interpretations [17].

Fourth, the missingness analysis highlights the degree to which model performance depends on the data-generating process rather than just on the architecture. The better discrimination observed in higher-missingness windows likely reflects case-mix structure and care patterns, not some inherent advantage of sparse data. This observation should caution against simplistic conclusions about robustness. A model can be robust to the empirical distribution on which it was trained while still being sensitive to institutional shifts in ordering behavior or documentation frequency. For multimodal health AI, robustness is therefore inseparable from workflow realism.

The work also suggests a way to think about text in critical care systems more generally. Free text should not always be modeled as another dense stream. In many hospital settings, it is better treated as semantic state annotation: lower-frequency, richer-context information that updates intermittently and conditions the interpretation of denser numeric data. That view has consequences for architecture. It favors computationally light fusion mechanisms, cached embeddings, and explicit handling of note absence over heavy symmetric architectures that presume synchronized modalities. The present results fit that design philosophy well. The text channel matters, but it matters in a targeted, contextual way.

There are, however, clear limitations that bound the claims. The model is derived from a single institution and from one note family only. The outcome is composite, and the analysis remains retrospective. The alert-theoretic findings are therefore best viewed as deployment-oriented implications of observed test behavior, not as guarantees of real-world effect. A prospective implementation would need to study user interaction, review latency, alert fatigue, and the possibility that clinicians already anticipate many of the flagged events. None of those questions can be settled by retrospective forecasting performance alone.

Future work could develop several extensions while remaining within the same computer science framing [18]. Event-specific heads for mortality, vasopressor initiation, and ventilation would clarify whether multimodal context benefits all components equally. More expressive missingness modeling could distinguish between stale observation carry-forward and informative nonmeasurement. Integrating nursing notes and physician assessments would test whether higher-frequency text improves not just contextual disambiguation but also the temporal precision of the detector. Finally, explicit top- k or budget-aware training objectives might align the learned ranking even more closely with review-constrained deployment.

Even without those extensions, the present analysis contributes a useful reframing. It shows that the success of multimodal ICU prediction should not be judged only by whether a fused model outruns a baseline on AUROC. The more informative question is whether semantic context increases the density of genuinely high-risk windows within the portion of the score distribution that a real system can act on. In this study, the answer appears to be yes. The multimodal detector generates a sparse enough alert stream to be operationally plausible, concentrates true events more strongly than structured-only and classical alternatives, and becomes substantially more usable after straightforward recalibration. Those are meaningful properties for any real-time clinical monitoring system built under finite attention.

6. Conclusion

This paper examined ICU deterioration forecasting through an alert-budgeted multimodal lens rather than as a conventional benchmark exercise. A recurrent encoder over structured physiologic histories was combined with contextual embeddings from radiology reports, and the resulting detector was evaluated not only by discrimination but also by review yield, workload, and calibration recovery. On 823,641 held-out hourly windows, the multimodal system achieved AUROC 0.7857 and AUPRC 0.1908, outperforming structured-only deep learning and classical baselines. At the reported operating point, the model produced roughly 30 alerts per 1,000 windows, with about 7.5 true positives and 22.4 false positives per 1,000 windows, corresponding to approximately four alerted reviews per true deterioration captured.

The empirical pattern indicates that sparse clinical text is most effective as contextual information rather than as an independent detector. Its contribution is expressed primarily through stronger positive concentration and better operational efficiency of the high-risk tail. The study also shows that raw probabilities from the imbalance-trained network are substantially improvable through post-hoc temperature scaling, which sharply reduces calibration error without affecting ranking. Taken together, these findings support a specific computational view of the problem: multimodal ICU forecasting is a weakly synchronized, low-prevalence sequence learning task whose usefulness depends on how efficiently it converts heterogeneous evidence into a manageable review stream. Under that view, the multimodal architecture studied here appears technically credible, operationally interpretable, and meaningfully distinct from structured-only prediction [19].

References

- [1] H. A. W, A. Kandula, and X. Wang, *ICBCT - A Differential-Privacy-Based Blockchain Architecture to Secure and Store Electronic Health Records*. ACM, 3 2021.
- [2] N. Hampiholi, "Through the lens of principled data practice a groundbreaking exploration into ethical healthcare platforms," *International Journal of Engineering Applied Sciences and Technology*, vol. 8, pp. 26–30, 9 2023.
- [3] R. G. Mishuris, J. Yoder, D. Wilson, and D. M. Mann, "Integrating data from an online diabetes prevention program into an electronic health record and clinical workflow, a design phase usability study," *BMC medical informatics and decision making*, vol. 16, pp. 88–88, 7 2016.
- [4] G. Ser, A. Robertson, and A. Sheikh, "A qualitative exploration of workarounds related to the implementation of national electronic health records in early adopter mental health hospitals," *PloS one*, vol. 9, pp. 1–9, 1 2014.
- [5] R. Patel, J. Oram, N. Hebden, Z. Payne, M. Morse, and C. Gadd, "Derivation and validation of the management and supervision tool (mast) risk of crisis (roc) algorithm using electronic health record (ehr) data," *BJPsych Open*, vol. 8, pp. S65–S65, 6 2022.
- [6] 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.
- [7] F. Rezaeibagha and Y. Mu, "Distributed clinical data sharing via dynamic access-control policy transformation," *International journal of medical informatics*, vol. 89, pp. 25–31, 2 2016.
- [8] D. Maxwell-Downing, "Clinical issues—january 2013," *AORN Journal*, vol. 97, pp. 140–148, 12 2012.

- [9] S. Biswal, C. Xiao, L. M. Glass, E. Milkovits, and J. Sun, "Aaai - doctor2vec: Dynamic doctor representation learning for clinical trial recruitment," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 557–564, 4 2020.
- [10] N. X. T. Le, M. T. Wiley, A. Loza, V. Hristidis, and R. El-Kareh, "Prediction of medical concepts in electronic health records: Similar patient analysis.," *JMIR medical informatics*, vol. 8, pp. e16008–, 7 2020.
- [11] H. Angier, S. Likumahuwa, S. C. Finnegan, T. Vakarcs, C. C. Nelson, A. Bazemore, M. Carrozza, and J. E. DeVoe, "Using geographic information systems (gis) to identify communities in need of health insurance outreach: An ochin practice-based research network (pbrn) report," *Journal of the American Board of Family Medicine : JABFM*, vol. 27, pp. 804–810, 11 2014.
- [12] S. Singh, M. Rakhra, A. Malik, and D. Singh, "Blockchain-based ehr system for indian healthcare industry using aadhar," in *2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, pp. 997–1001, IEEE, 1 2023.
- [13] S. Tang, P. Davarmanesh, Y. Song, D. Koutra, M. W. Sjoding, and J. Wiens, "Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data," *Journal of the American Medical Informatics Association : JAMIA*, vol. 27, pp. 1921–1934, 10 2020.
- [14] D. R. Masys, G. P. Jarvik, N. F. Abernethy, N. R. Anderson, G. J. Papanicolaou, D. N. Paltoo, M. A. Hoffman, I. S. Kohane, and H. P. Levy, "Technical desiderata for the integration of genomic data into electronic health records," *Journal of biomedical informatics*, vol. 45, pp. 419–422, 12 2011.
- [15] H. Hemingway, F. W. Asselbergs, J. Danesh, R. Dobson, N. Maniadakis, A. P. Maggioni, G. J. M. W. van Thiel, M. Cronin, G. Brobert, P. Vardas, S. D. Anker, D. E. Grobbee, and S. Denaxas, "Big data from electronic health records for early and late translational cardiovascular research: challenges and potential," *European heart journal*, vol. 39, pp. 1481–1495, 8 2017.
- [16] A. F. Muzaffar, S. Abdul-Massih, J. M. Stevenson, and S. Alvarez-Arango, "Use of the electronic health record for monitoring adverse drug reactions.," *Current allergy and asthma reports*, vol. 23, pp. 417–426, 5 2023.
- [17] K. Bosak and N. Abusalim, "Coordinating resources for a hybrid effectiveness and implementation trial across an integrated digital network," 8 2022.
- [18] N. A. Walton, D. K. Johnson, T. N. Person, J. C. Reynolds, and M. S. Williams, "Pilot implementation of clinical genomic data into the native electronic health record: Challenges of scalability," *ACI Open*, vol. 04, pp. e162–e166, 12 2020.
- [19] 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.