Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Do Clinical Notes Improve ICU Mortality Prediction? A Controlled Comparison of Structured and Multimodal EHR Fusion Strategies

Ashish Katyal

aashishkatyal@gmail.com

Birla Institute of Technology and Science, Pilani

Research Article

Keywords: Clinical notes, Electronic health records, ICU mortality prediction, Multimodal learning, Clinical natural language processing, Multimodal data fusion, ClinicalBERT, Intensive care

Posted Date: September 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11124716/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Background

Clinical notes contain information that may complement structured electronic health record (EHR) variables for intensive care unit (ICU) mortality prediction. However, whether incorporating clinical text provides consistent incremental predictive value over structured clinical data remains unclear, particularly when alternative multimodal fusion strategies are evaluated under a common experimental protocol.

Objective

To determine whether clinical notes improve ICU in-hospital mortality prediction beyond structured EHR variables and to assess how different multimodal fusion strategies influence predictive performance.

Methods

A controlled retrospective study was conducted using 20,000 eligible ICU stays from the MIMIC-IV and MIMIC-IV-Note datasets. Structured clinical variables and eligible clinical notes from the first 24 hours of each ICU stay were used to predict in-hospital mortality. Six models were evaluated: a structured-only multilayer perceptron (MLP), a ClinicalBERT-only model, and four multimodal models using concatenation, gated, bilinear, and cross-attention fusion. All models were evaluated on the same patient-level data partitions and class-imbalance treatment and were trained using three random seeds. Performance was assessed using accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUROC), and area under the precision–recall curve (AUPRC).

Results

The structured-only MLP achieved the highest overall performance, with an AUROC of 0.859 ± 0.006, AUPRC of 0.462 ± 0.025, and F1-score of 0.484 ± 0.023. The ClinicalBERT-only model showed substantially lower discrimination (AUROC, 0.732 ± 0.002; AUPRC, 0.259 ± 0.001). Among the multimodal approaches, bilinear fusion achieved the strongest overall performance, with an AUROC of 0.857 ± 0.004, AUPRC of 0.461 ± 0.002, and F1-score of 0.479 ± 0.003, closely approaching the structured-only model. Concatenation achieved the highest mortality recall (0.615 ± 0.034), exceeding the structured-only model (0.543 ± 0.024), but produced more false-positive predictions. Cross-attention yielded lower performance than the other multimodal fusion strategies.

Conclusions

Clinical notes did not provide a consistent improvement in ICU mortality discrimination beyond structured EHR variables under the evaluated conditions. Nevertheless, clinical text contributed complementary information that could increase mortality recall, with its predictive value strongly dependent on the fusion mechanism. These findings highlight the importance of evaluating multimodal clinical AI systems against strong unimodal structured baselines and across multiple performance dimensions rather than assuming that additional clinical modalities necessarily improve prediction.

1. Introduction

Accurate prediction of mortality among critically ill patients is an important task in intensive care medicine because early identification of patients at high risk of death can support clinical assessment, prioritization of care, and allocation of limited ICU resources. Electronic health records (EHRs) provide a large amount of information for developing such prediction models, including demographic characteristics, laboratory measurements, vital signs, diagnoses, procedures, and other routinely recorded clinical variables. The availability of large critical-care datasets has consequently enabled the development of increasingly sophisticated machine-learning and deep-learning approaches for mortality prediction. However, EHR information is inherently heterogeneous. Structured variables provide standardized measurements of patient physiology and clinical status, whereas narrative clinical notes contain contextual information that is difficult to represent through conventional tabular variables.

Clinical notes may therefore provide information that is complementary to structured EHR data. Physician and nursing documentation can contain assessments, clinical reasoning, observations, descriptions of patient status, diagnostic impressions, and other contextual information that may not be explicitly represented in structured fields. Natural language processing (NLP), and particularly pretrained clinical language models, has made it increasingly feasible to transform this unstructured information into representations suitable for predictive modeling. Studies combining clinical text with structured or temporal EHR information have reported improvements in mortality and other clinical outcome prediction tasks, supporting the view that narrative documentation can contain clinically relevant predictive signals. For example, multimodal approaches incorporating clinical text with ICU physiological or structured information have demonstrated improved predictive performance compared with selected unimodal baselines.

Nevertheless, the presence of predictive information in clinical notes does not necessarily imply that adding notes will consistently improve an existing structured-data prediction model. The two modalities differ substantially in dimensionality, sparsity, representation, temporal structure, and information density. Structured ICU variables often contain direct measurements of physiological state, whereas clinical notes may contain redundant, delayed, or highly variable information. Consequently, the incremental contribution of clinical text may depend not only on whether the modality is included, but also on how its representation is combined with structured information. Recent multimodal EHR studies have consequently investigated increasingly sophisticated fusion mechanisms, including hierarchical fusion, attention-based fusion, and other cross-modal architectures.

An important methodological issue is therefore the distinction between **multimodal prediction and incremental predictive value**. A multimodal model may achieve strong predictive performance without demonstrating that the additional modality provides consistent benefit over a strong unimodal baseline. This distinction is particularly important in clinical prediction, where model complexity should be justified by measurable improvement rather than by the availability of additional data sources. Comparisons based on a single performance metric can also obscure this issue. For mortality prediction under class imbalance, for example, a model may improve sensitivity or recall while simultaneously producing substantially more false-positive predictions. Threshold-independent measures such as the area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) therefore provide complementary information to threshold-dependent measures such as precision, recall, and F1 score.

The choice of fusion strategy represents another unresolved issue. Simple concatenation provides a straightforward mechanism for combining modality-specific representations, whereas gated, bilinear, and cross-attention approaches attempt to model interactions between modalities more explicitly. Although increasingly complex fusion mechanisms may capture cross-modal relationships that are not represented by simple concatenation, greater architectural complexity does not necessarily translate into better clinical prediction. Existing studies have frequently focused on proposing a new multimodal architecture and demonstrating its performance against selected baselines. For example, multimodal transformer and fusion-network approaches have reported improvements in mortality prediction when structured EHR variables and clinical text are jointly modeled. However, these findings do not establish whether the observed improvement is attributable to the clinical notes themselves, to the fusion mechanism, or to differences in experimental configuration and baseline strength.

This distinction motivates a controlled comparison in which structured-only, clinical-text-only, and multimodal models are evaluated under the same experimental conditions. Such a design allows the contribution of the clinical-note modality to be examined relative to a strong structured-data baseline while also separating the effect of modality inclusion from the effect of fusion strategy. The issue is especially relevant for ICU mortality prediction, where structured physiological and laboratory measurements collected early during an ICU stay can already provide substantial predictive information. Recent work continues to demonstrate strong performance from multimodal approaches, but also indicates that the relative contribution of structured and textual information is not necessarily uniform across tasks and modalities.

In this study, we investigate whether clinical notes provide incremental predictive value for ICU in-hospital mortality prediction when combined with structured EHR variables from the same early clinical observation period. We use data from MIMIC-IV and MIMIC-IV-Note and restrict the prediction inputs to information available during the first 24 hours of ICU admission. The experimental design compares six models: a structured-data multilayer perceptron (MLP), a ClinicalBERT-based clinical-text model, and four multimodal models using concatenation, gated fusion, bilinear fusion, and cross-attention. All models are evaluated on the same patient-level data partitions and under the same class-imbalance handling

strategy, with performance assessed across threshold-dependent and threshold-independent metrics and repeated across three random seeds.

The study addresses four research questions:

1. **Does adding clinical-note information improve ICU in-hospital mortality prediction compared with structured EHR data alone?**
2. **How does clinical-text-only prediction compare with structured-data-only prediction when both modalities are derived from the first 24 hours of ICU care?**
3. **How does the choice of multimodal fusion mechanism—including concatenation, gated, bilinear, and cross-attention fusion—affect predictive performance?**
4. **Does multimodal fusion provide a consistent advantage across threshold-dependent and threshold-independent performance measures?**

Rather than assuming that additional clinical information necessarily produces better prediction, this study treats the contribution of clinical notes as an empirical question. The controlled comparison is intended to clarify whether the value of narrative clinical information is consistent across modeling strategies and evaluation criteria, and whether increasing architectural sophistication produces measurable benefits over a strong structured-data baseline. The resulting analysis provides a more direct assessment of when multimodal EHR modeling is justified for ICU mortality prediction and highlights the importance of evaluating modality contribution, fusion strategy, and predictive trade-offs jointly.

2. Related Work

2.1 Structured EHR-Based ICU Mortality Prediction

Structured EHR variables have long formed the basis of ICU mortality prediction because they provide relatively standardized measurements of patient physiological and clinical status. Demographic characteristics, vital signs, laboratory measurements, diagnoses, procedures, medications, and severity-related variables can be represented as numerical or categorical features and used with statistical, machine-learning, and deep-learning models. Conventional risk scores and machine-learning approaches demonstrated that a relatively small set of routinely available clinical variables can provide substantial discrimination for ICU mortality. More recent approaches have extended this paradigm by modeling the temporal structure of ICU measurements using recurrent neural networks, attention mechanisms, and Transformer-based architectures.

The strength of structured EHR data is particularly relevant to multimodal prediction because it establishes a strong baseline against which additional modalities should be evaluated. Recent ICU prediction studies continue to demonstrate that structured physiological and temporal variables can provide substantial predictive information, even when other modalities are available. For example, multimodal studies incorporating clinical notes and other data sources have reported that temporal

clinical variables may contribute more strongly to predictive performance than clinical notes in several risk-prediction tasks. Similarly, recent work on ICU mortality prediction has continued to develop models based on structured EHR information, including approaches that transform structured variables into semantic representations before applying pretrained language models.

These findings highlight an important consideration for multimodal modeling: the value of an additional modality should be assessed relative to the predictive strength of structured data rather than against a weak baseline. A multimodal model that performs well in absolute terms does not necessarily demonstrate that clinical notes provide incremental information beyond structured ICU variables. Consequently, controlled comparison with a strong structured-only model is essential when the research question concerns the contribution of clinical text.

2.2 Clinical Notes and Clinical Language Models for Mortality Prediction

Clinical notes provide a complementary source of information because they capture aspects of patient status that are not necessarily represented explicitly in structured EHR fields. Narrative documentation may include symptoms, physical findings, treatment response, clinical assessments, diagnostic impressions, and professional judgment. Earlier approaches to clinical-text prediction relied on manually engineered features, bag-of-words representations, term-frequency measures, or conventional neural language models. The emergence of pretrained language models substantially changed this landscape by enabling contextual representations of clinical narratives.

ClinicalBERT and related clinical language models have been applied to a range of clinical NLP tasks, including outcome prediction. Their ability to encode contextual relationships among clinical terms makes them suitable for extracting information from heterogeneous medical narratives. In ICU settings, clinical notes recorded during the early period of admission have been shown to contain information relevant to subsequent mortality prediction. The thesis literature review similarly identifies evidence that notes available during the first 24 hours can provide predictive information for short-, medium-, and longer-term mortality outcomes.

However, text-only prediction also has important limitations. Clinical documentation varies in length, timing, completeness, and style. Long notes may require truncation or segmentation, while repeated documentation can introduce redundancy. The information contained in a note may also overlap with information already represented by laboratory measurements, vital signs, diagnoses, and other structured variables. Therefore, strong performance from a clinical-text model establishes that notes contain predictive information, but it does not establish that this information provides additional predictive value when structured EHR variables are already available.

This distinction is central to the present study. Rather than treating clinical text as inherently superior because it contains richer semantic information, we evaluate its contribution under a common cohort and observation window and compare it directly with structured-data prediction.

2.3 Multimodal EHR Prediction

The complementary characteristics of structured EHR data and clinical text have motivated a growing body of multimodal prediction research. In principle, structured variables provide quantitative measurements of physiological state, whereas clinical notes provide contextual and semantic information. Combining these modalities can therefore produce a more comprehensive representation of the patient.

Several studies have reported improved predictive performance after integrating structured and unstructured EHR information. A multimodal Transformer, for example, combined clinical notes with structured EHR variables for in-hospital mortality prediction and reported AUROC of 0.877, together with an AUCPR of 0.538 and F1-score of 0.490. The study also investigated ClinicalBERT-based representations and incorporated attribution methods to examine the contribution of model inputs.

Other work has explored multimodal prediction specifically in critical-care settings. Sun et al. proposed a cross-modal contrastive-learning framework that combined structured medical codes with clinical notes for ICU outcome prediction and reported improved mortality and readmission prediction on the eICU-CRD dataset. Their framework used separate representations for structured and textual information and introduced cross-modal learning to capture relationships between the modalities. Wang et al. subsequently proposed a multimodal fusion network using multiple modality-specific encoders and hierarchical fusion for ICU mortality and length-of-stay prediction. Their results further demonstrated the potential of combining heterogeneous EHR modalities.

Although these studies provide evidence that multimodal learning can be effective, their primary objective is generally the development of increasingly sophisticated prediction architectures. As a result, the question of **whether the additional modality itself provides consistent incremental value** can remain difficult to isolate. Differences in cohort construction, observation windows, preprocessing, text encoders, structured representations, class-imbalance handling, and evaluation procedures further complicate direct comparisons across studies. The systematic literature on structured-text EHR fusion has consequently identified standardization, reproducibility, and comparison of fusion strategies as important open issues.

2.4 Fusion Strategies for Structured and Textual EHR Data

The effectiveness of a multimodal model depends not only on the modalities that are included but also on how their representations are combined. Fusion strategies can broadly be implemented through early or feature-level fusion, intermediate representation-level fusion, or late decision-level fusion. Feature concatenation is one of the simplest approaches: modality-specific representations are generated independently and then joined before classification. Its main advantage is architectural simplicity, but concatenation does not explicitly model interactions between modalities.

More complex strategies attempt to learn such interactions. Gated fusion introduces learnable mechanisms that control the contribution of individual modalities, allowing the model to emphasize one source when it is more informative. Bilinear fusion explicitly models multiplicative interactions between representations and can therefore capture relationships that are not represented by simple concatenation. Attention-based approaches extend this idea by allowing information from one modality to influence the representation of another modality selectively.

Recent ICU prediction studies have increasingly adopted attention and cross-modal mechanisms. The cross-modal contrastive-learning framework of Sun et al. explicitly models interactions between structured medical codes and clinical text, while MFNet uses hierarchical multimodal fusion to capture inter-modal relationships across heterogeneous EHR sources. Other multimodal clinical prediction work has similarly investigated early, joint, and late fusion and found that the relative contribution of individual modalities can differ substantially across prediction tasks.

The increasing architectural complexity of fusion mechanisms, however, creates an important empirical question. A more sophisticated fusion mechanism may capture richer cross-modal interactions without necessarily producing better predictive performance. Consequently, comparing multiple fusion strategies under an identical experimental protocol is important for determining whether performance differences arise from the presence of clinical text, the fusion mechanism, or both.

2.5 Evaluation of Multimodal Prediction Under Class Imbalance

Evaluation methodology is particularly important for ICU mortality prediction because mortality is generally a minority outcome. A model can achieve high overall accuracy while performing poorly on the patients of greatest clinical interest. Precision, recall, and F1-score therefore provide important information about minority-class prediction, while AUROC evaluates discrimination across classification thresholds. AUPRC is particularly informative when the positive class is relatively uncommon because it directly characterizes the relationship between precision and recall.

The need for multiple complementary metrics is also relevant to multimodal EHR research. A model may increase mortality recall while simultaneously generating substantially more false-positive predictions. In such cases, a higher recall does not necessarily indicate a uniformly better predictive model. The thesis evaluation framework similarly emphasizes the use of accuracy, precision, recall, F1-score, AUROC, and AUPRC to characterize different aspects of predictive performance.

Recent multimodal studies have also begun to examine modality contribution rather than reporting only the performance of the final multimodal model. For example, a multimodal risk-prediction study combining temporal variables, medical images, and clinical notes used Shapley-value analysis to quantify modality contributions and found that temporal variables were more useful than clinical notes for the prediction tasks considered. This provides an important precedent for treating modality

contribution as an empirical question rather than assuming that additional modalities necessarily produce equivalent predictive benefit.

2.6 Research Gap

The existing literature establishes three important findings. First, structured ICU data provide strong predictive signals for mortality and should therefore constitute a meaningful baseline. Second, clinical notes contain additional contextual information that can be exploited using pretrained clinical language models. Third, multimodal fusion can combine these sources and has produced promising results across ICU and broader clinical prediction tasks.

However, an important methodological gap remains. Much of the multimodal literature is oriented toward developing a new architecture and demonstrating that the resulting model performs well relative to selected baselines. This design does not always isolate the incremental contribution of clinical notes because modality inclusion, encoder choice, fusion mechanism, preprocessing, and other experimental factors may change simultaneously. Furthermore, improvements in one metric may be accompanied by deterioration in another, making conclusions based on a single performance measure potentially incomplete.

A more controlled comparison is therefore needed in which structured-only, clinical-text-only, and multimodal models are evaluated using the **same patient cohort, observation window, data partitions, class-imbalance strategy, and evaluation protocol**. Within the multimodal setting, different fusion mechanisms should also be compared without changing the underlying text encoder or structured-data pathway. Such a design makes it possible to distinguish three related questions: whether clinical notes contain useful predictive information, whether that information provides incremental value beyond structured EHR variables, and whether its usefulness depends on the fusion mechanism.

The present study addresses this gap using MIMIC-IV and MIMIC-IV-Note data and a common first-24-hour ICU observation window. Six models are evaluated: structured-data MLP, ClinicalBERT-only, and four multimodal configurations based on concatenation, gated fusion, bilinear fusion, and cross-attention. ClinicalBERT is kept frozen across the text-based configurations, and all models are evaluated on the same patient-level partitions using the same class-imbalance treatment and three random seeds. This controlled design allows the study to evaluate not simply whether multimodal prediction is effective, but **whether clinical notes improve mortality prediction beyond structured ICU information and under which fusion strategies such improvement, if any, is observed**.

3. Materials and Methods

3.1 Study Design

This study was designed as a controlled empirical comparison of structured, textual, and multimodal electronic health record (EHR) representations for ICU in-hospital mortality prediction. The primary

objective was to determine whether clinical notes provide incremental predictive value beyond structured ICU variables and whether the observed contribution of clinical text depends on the multimodal fusion mechanism.

Six prediction configurations were evaluated under a common experimental protocol. The configurations comprised: (1) a structured-data-only multilayer perceptron (MLP), (2) a ClinicalBERT-only text model, and four multimodal models using (3) feature concatenation, (4) gated fusion, (5) bilinear fusion, and (6) cross-attention. The same patient cohort, observation window, outcome definition, data partitions, structured-data representation, ClinicalBERT encoder, class-imbalance strategy, and evaluation metrics were used across the experimental configurations. This design was intended to minimize methodological differences between models and provide a direct assessment of modality and fusion effects.

The prediction task was binary classification of in-hospital mortality using information available during the first 24 hours of ICU care. The experimental workflow consisted of cohort construction, temporal restriction of structured and textual information, modality alignment, model-specific representation learning, model training using the training partition, validation-based model monitoring, and final evaluation on an independent test partition.

3.2 Data Source

The study used the Medical Information Mart for Intensive Care IV (MIMIC-IV) together with the linked MIMIC-IV-Note database. MIMIC-IV provides de-identified structured EHR information from hospital and intensive-care encounters, including demographic characteristics, laboratory measurements, vital signs, diagnoses, procedures, medications, and other clinical events. MIMIC-IV-Note provides associated unstructured clinical documentation.

The availability of both structured EHR variables and clinical notes within the linked MIMIC-IV ecosystem makes the dataset suitable for controlled multimodal evaluation. The structured and textual records were aligned using the available patient and ICU-stay identifiers. The thesis implementation used common identifiers to integrate structured data, clinical notes, and outcome labels and excluded records that were invalid or could not be appropriately matched.

3.3 Cohort Construction

The final study cohort consisted of **20,000 eligible ICU stays**. Cohort construction was performed before model development to ensure that all six experimental configurations were evaluated on the same population.

The cohort was divided into training, validation, and test partitions using a **70:15:15** ratio at the patient level. The partitioning variable was subject_id, such that records belonging to the same patient were assigned to only one partition. This patient-level separation was used to reduce the possibility of information leakage between training and evaluation data.

The resulting partitions contained 13,938 ICU stays in the training set, 3,080 in the validation set, and 2,982 in the test set. Across the complete cohort, 17,521 stays corresponded to survivors and 2,479 to patients who died during hospitalization, giving an overall mortality prevalence of **12.40%**.

The mortality distributions were similar across the three partitions: 12.47% in the training set, 12.05% in the validation set, and 12.41% in the test set. This indicates that the class distribution was broadly preserved across the patient-level partitions.

3.4 Observation Window and Outcome Definition

The prediction task was defined as **in-hospital mortality**. To ensure temporal consistency across modalities, the input data were restricted to information available during the **first 24 hours of ICU admission**.

For structured EHR data, variables recorded during this observation period were used to construct the structured representation. Clinical notes eligible for prediction were similarly restricted to the first 24 hours of the ICU stay. Thus, both modalities represented information from the same early clinical period rather than allowing later documentation to contribute to prediction.

The outcome label represented whether the patient died during the hospitalization. The resulting task was formulated as binary classification, with mortality treated as the positive class.

Restricting both modalities to a common temporal window was important for the central research question. It ensured that an apparent advantage of clinical notes could not be attributed simply to their containing information documented substantially later than the structured variables.

3.5 Structured EHR Representation

Structured clinical variables were processed through a dedicated multilayer perceptron (MLP) pathway. The structured representation served as the input to the structured-only baseline and as the structured modality in each multimodal model.

The structured pathway generated a fixed-length representation of the available clinical variables. The MLP was used as the common structured encoder across the multimodal configurations so that differences between fusion strategies would not be confounded by changes in the structured representation.

The structured-only MLP therefore served as the principal baseline against which the incremental contribution of clinical notes was assessed. This was particularly important because a multimodal model should demonstrate improvement over a reasonably strong structured-data model to support a claim of incremental predictive value.

3.6 Clinical-Note Representation

Clinical notes were represented using **Bio_ClinicalBERT/ClinicalBERT**, a pretrained language model adapted to clinical text. ClinicalBERT was used to generate contextualized representations from the eligible clinical notes.

The same ClinicalBERT encoder was used for the text-only and all multimodal configurations. Importantly, the ClinicalBERT parameters were **frozen** during model training. Consequently, the comparison did not involve independent fine-tuning of the language model for different fusion strategies.

Using a common frozen text encoder provided an additional experimental control. Differences between the multimodal configurations could therefore be interpreted primarily in terms of how the structured and textual representations were combined rather than differences in language-model optimization.

The ClinicalBERT-only configuration used the textual representation directly for mortality classification and provided a unimodal baseline for determining the predictive strength of clinical notes independently of structured variables.

3.7 Prediction Models and Fusion Strategies

Six model configurations were evaluated. The first two were unimodal baselines, while the remaining four combined structured and textual representations.

3.7.1 Structured-only model

The structured-only model consisted of an MLP applied to the structured EHR representation. Its purpose was to establish the predictive performance achievable without clinical text and to provide the primary baseline for assessing incremental multimodal value.

3.7.2 ClinicalBERT-only model

The ClinicalBERT-only configuration used contextualized representations generated from clinical notes for mortality prediction. No structured EHR variables were supplied to this model. This configuration established whether information contained in the clinical notes alone could support mortality prediction.

3.7.3 Concatenation fusion

For the concatenation model, the structured and textual representations were combined by direct vector concatenation:

hcat=[hs;ht]\mathbf{h}_{cat}=[\mathbf{h}_{s};\mathbf{h}_{t}]

where hs\mathbf{h}_{s} denotes the structured representation and ht\mathbf{h}_{t} denotes the ClinicalBERT-derived text representation. The resulting joint representation was passed to the classification layer.

Concatenation provides a simple multimodal baseline because it exposes both modality representations to the classifier without explicitly learning multiplicative or attention-based interactions between them.

3.7.4 Gated fusion

Gated fusion introduced a learnable mechanism for controlling the relative contribution of the two modalities. Given structured and textual representations, a gating function was used to learn modality-dependent weighting before classification.

Conceptually, the fused representation can be expressed as:

g = σ(Wg[hs;ht] + bg)\mathbf{g}=\sigma(\mathbf{W}_{g}[\mathbf{h}_{s};\mathbf{h}_{t}]+\mathbf{b}_{g})

where σ\sigma denotes the sigmoid activation and g\mathbf{g} represents the learned gating coefficients. The gate allows the model to modulate the contribution of the modality representations rather than treating them as equally informative.

3.7.5 Bilinear fusion

Bilinear fusion was used to model interactions between structured and textual representations through multiplicative relationships. A general form of the bilinear interaction is:

hbilinear=hsTWht\mathbf{h}_{bilinear}=\mathbf{h}_{s}^{T}\mathbf{W}\mathbf{h}_{t}

where W\mathbf{W} represents learnable interaction parameters.

Unlike simple concatenation, bilinear fusion can represent combinations of structured and textual features through explicit cross-modal interactions. This configuration was therefore included to determine whether a more expressive interaction mechanism could extract additional predictive information from the two modalities.

3.7.6 Cross-attention fusion

The cross-attention configuration used an attention mechanism to model relationships between the structured and textual representations. Rather than simply joining the two representations, cross-attention allows information from one modality to influence the representation of the other through learned attention weights.

This configuration represented the most explicitly interaction-oriented fusion mechanism evaluated in the study. Its inclusion allowed the experiment to determine whether attention-based cross-modal interaction provided an advantage over simpler fusion mechanisms.

The four multimodal configurations therefore represented increasing levels of interaction modeling, from direct concatenation to learned gating, multiplicative bilinear interaction, and cross-attention.

3.8 Model Training

All models were trained using the same training, validation, and test partitions. The training partition was used for model optimization, the validation partition was used for monitoring and model selection, and the test partition was reserved for final performance evaluation.

The multimodal configurations shared the same structured-data pathway and ClinicalBERT representation. The text encoder remained frozen during training, while the downstream prediction and fusion components were trained using the available training data.

To assess the stability of the findings with respect to random initialization and training stochasticity, each model configuration was evaluated using **three random seeds: 42, 52, and 62**. Thus, the six configurations resulted in **18 completed training runs**.

Model performance was summarized across the three seeds using the mean and standard deviation. This approach provided an estimate of both central predictive performance and variation associated with stochastic training.

3.9 Class-Imbalance Handling

Mortality represented the minority outcome in the study cohort, with an overall prevalence of 12.40%. Training therefore used **weighted binary cross-entropy loss** to increase the contribution of mortality cases during optimization.

The positive-class weight was derived from the training partition and was set to **7.0196**. The same class-imbalance strategy was applied across the model configurations to prevent differences in loss weighting from becoming a source of experimental variation.

This design also enabled evaluation beyond overall accuracy. Because a model can achieve high accuracy by favoring the majority survival class, minority-class metrics were explicitly included in the evaluation.

3.10 Evaluation Metrics

Model performance was assessed using both threshold-dependent and threshold-independent metrics.

The threshold-dependent measures were **accuracy, precision, recall, and F1-score**. Accuracy represents the proportion of correctly classified cases overall. Precision measures the proportion of predicted mortality cases that were actually positive, while recall measures the proportion of mortality cases correctly identified by the model. F1-score is the harmonic mean of precision and recall:

F1 = 2Precision×RecallPrecision+Recall.F1 = 2\frac{Precision \times Recall}{Precision + Recall}.

These measures were used to characterize the trade-off between identifying mortality cases and limiting false-positive predictions.

Two threshold-independent measures were also reported: **area under the receiver operating characteristic curve (AUROC)** and **area under the precision-recall curve (AUPRC)**. AUROC characterizes discrimination across classification thresholds, whereas AUPRC provides a complementary assessment that is particularly informative when the positive class is relatively uncommon.

The combination of these metrics was intentional. A multimodal model could, for example, achieve higher mortality recall while also producing more false-positive predictions. Reporting only one metric could therefore provide an incomplete representation of model behavior.

3.11 Experimental Control and Reproducibility

The principal methodological objective of the experimental design was to isolate the effects of modality inclusion and fusion strategy. Accordingly, the six configurations were evaluated using the same underlying cohort, first-24-hour observation window, patient-level data partitions, outcome definition, class-imbalance strategy, and evaluation protocol.

The text encoder was held constant across text-containing configurations, and ClinicalBERT was frozen rather than independently fine-tuned for each multimodal model. Likewise, the structured representation was generated through a common MLP pathway across the structured-only and multimodal configurations. These controls reduced the possibility that performance differences between fusion strategies were caused by changes in the underlying modality encoders.

The complete experimental design can therefore be viewed as a factorial comparison of modality availability and fusion strategy: the structured-only and ClinicalBERT-only models establish unimodal reference points, while the four multimodal configurations evaluate alternative mechanisms for combining the same two representations.

The primary comparison was between the structured-only model and the multimodal configurations. A multimodal approach was considered to provide evidence of incremental value only when its performance demonstrated an advantage over the structured-only baseline across relevant evaluation measures. Particular attention was given to whether improvements in recall or F1-score were accompanied by improvements in AUROC and AUPRC, thereby distinguishing gains in minority-class detection from broader improvements in discrimination.

All six configurations completed training across all three random seeds, resulting in 18 completed experimental runs without training failures.

4. Results

4.1 Cohort Characteristics

The final study cohort comprised 20,000 eligible ICU stays from the MIMIC-IV and MIMIC-IV-Note datasets. The cohort was divided at the patient level into training, validation, and test partitions. The

training set contained 13,938 ICU stays, the validation set contained 3,080, and the test set contained 2,982. Across the complete cohort, 17,521 stays corresponded to survivors and 2,479 to in-hospital mortality cases, yielding an overall mortality prevalence of 12.40%. The mortality rates were similar across the training, validation, and test partitions (12.47%, 12.05%, and 12.41%, respectively), indicating a comparable outcome distribution across partitions. No patient overlap was present between partitions, and structured and clinical-note data were successfully linked for all 20,000 ICU stays.

The distribution of survival and mortality cases across the three partitions is shown in Fig. 1.

All six prediction configurations were evaluated using the same cohort, observation window, data partitions, and class-imbalance strategy. Each model was trained using three random seeds (42, 52, and 62), resulting in 18 completed training runs. No training failures were observed.

4.2 Overall Predictive Performance

The classification performance of the six models across the three random seeds is presented in Table 3. The structured-only MLP achieved the highest mean F1-score among all models (0.484 ± 0.023), together with an accuracy of 0.856 ± 0.010, precision of 0.437 ± 0.029, and recall of 0.543 ± 0.024.

Among the multimodal configurations, bilinear fusion achieved the highest mean F1-score (0.479 ± 0.003), followed closely by concatenation (0.477 ± 0.006). Gated fusion achieved an F1-score of 0.464 ± 0.010, whereas cross-attention produced the lowest multimodal F1-score (0.443 ± 0.012). Thus, increasing the complexity of the fusion mechanism did not result in a corresponding improvement in classification performance.

The ClinicalBERT-only model produced the lowest overall F1-score (0.328 ± 0.008) and precision (0.232 ± 0.033), although its mean recall was relatively high (0.614 ± 0.175). The high recall combined with low precision indicates that text-only prediction identified a relatively large proportion of mortality cases but also generated a substantial number of false-positive predictions.

A visual comparison of F1-score, AUROC, and AUPRC across all six models is provided in Fig. 2.

4.3 Structured EHR versus Clinical-Text-Only Prediction

The comparison between the two unimodal approaches provides a direct assessment of the predictive strength of structured ICU variables and clinical notes when considered independently.

The structured-only MLP substantially outperformed the ClinicalBERT-only model in overall classification performance. The structured model achieved an F1-score of 0.484 ± 0.023 compared with 0.328 ± 0.008 for ClinicalBERT-only. Accuracy was also higher for the structured model (0.856 ± 0.010 versus 0.687 ± 0.094), as was precision (0.437 ± 0.029 versus 0.232 ± 0.033).

In contrast, ClinicalBERT-only achieved a higher mean mortality recall than the structured model (0.614 ± 0.175 versus 0.543 ± 0.024). However, the substantially lower precision indicates that this increase in sensitivity was accompanied by a greater false-positive burden.

The difference between the two modalities was also evident in threshold-independent discrimination. The structured-only model achieved a mean AUROC of 0.859 ± 0.006 and AUPRC of 0.462 ± 0.025, whereas ClinicalBERT-only achieved AUROC of 0.732 ± 0.002 and AUPRC of 0.259 ± 0.001.

These findings indicate that clinical notes contained predictive information regarding mortality, but the clinical-text representation alone did not provide discrimination comparable to the structured ICU variables under the evaluated experimental conditions.

4.4 Comparison of Multimodal Fusion Strategies

The four multimodal approaches produced different performance profiles despite using the same structured and clinical-text representations.

Bilinear fusion achieved the strongest overall multimodal performance, with accuracy of 0.846 ± 0.009, precision of 0.414 ± 0.016, recall of 0.570 ± 0.038, and F1-score of 0.479 ± 0.003. Concatenation achieved the highest mortality recall among all multimodal models (0.615 ± 0.034), but its precision was lower (0.391 ± 0.020), resulting in an F1-score of 0.477 ± 0.006.

Gated fusion produced an F1-score of 0.464 ± 0.010, while cross-attention achieved 0.443 ± 0.012. Cross-attention also exhibited the largest variability in recall across seeds (0.574 ± 0.077), suggesting less stable threshold-dependent performance under the evaluated configuration.

The relative performance of the four multimodal strategies demonstrates that the addition of a more complex fusion mechanism did not necessarily improve prediction. Bilinear fusion provided the closest overall multimodal performance to the structured-only baseline, whereas cross-attention provided no measurable advantage over the simpler fusion approaches.

4.5 Threshold-Independent Discrimination

Threshold-independent performance further supported the observation that structured EHR information provided the strongest overall discrimination.

The structured-only MLP achieved the highest AUROC (0.859 ± 0.006) and AUPRC (0.462 ± 0.025). Bilinear fusion produced nearly identical discrimination, with AUROC of 0.857 ± 0.004 and AUPRC of 0.461 ± 0.002. The differences between structured-only and bilinear fusion were therefore only 0.002 for AUROC and 0.001 for AUPRC.

Concatenation achieved an AUROC of 0.856 ± 0.007 and AUPRC of 0.448 ± 0.006, while gated fusion achieved 0.854 ± 0.006 and 0.454 ± 0.015, respectively. Cross-attention produced lower discrimination, with AUROC of 0.836 ± 0.008 and AUPRC of 0.412 ± 0.017.

ClinicalBERT-only again produced the lowest threshold-independent performance, with AUROC of 0.732 ± 0.002 and AUPRC of 0.259 ± 0.001.

Because mortality represented only 12.40% of the cohort, precision–recall analysis provides an important complementary perspective to ROC analysis. The approximate no-skill AUPRC was 0.124, and all six models exceeded this baseline. Nevertheless, the structured-only and bilinear models remained the strongest approaches according to AUPRC.

To further visualize discrimination across operating thresholds, mean ROC and precision–recall curves across the three random seeds are presented in Fig. 3.

4.6 Error and False-Positive Analysis

The threshold-dependent results reveal an important distinction between mortality sensitivity and overall classification performance.

The concatenation model achieved the highest mortality recall (0.615 ± 0.034), indicating that it identified the largest proportion of mortality cases among the evaluated models. However, this increased sensitivity was accompanied by 1,070 false-positive predictions, compared with 780 for the structured-only model.

The ClinicalBERT-only model generated an even larger false-positive burden, with 2,374 false-positive predictions. This finding is consistent with its relatively high recall but substantially lower precision. The structured-only model therefore provided a more balanced trade-off between mortality detection and false-positive predictions.

Bilinear fusion produced the closest multimodal approximation to this balance. Its F1-score (0.479 ± 0.003), AUROC (0.857 ± 0.004), and AUPRC (0.461 ± 0.002) remained close to those of the structured-only model, while its recall was higher (0.570 ± 0.038 versus 0.543 ± 0.024).

A compact visualization of the confusion matrices provides a complementary view of these differences.

4.7 Stability Across Random Seeds

The three-seed evaluation also revealed differences in the stability of model performance.

Bilinear fusion demonstrated the lowest standard deviation in F1-score among the six models (0.003), with a mean F1-score of 0.479. Its AUROC and AUPRC were also relatively stable across seeds (0.857 ± 0.004 and 0.461 ± 0.002, respectively). Concatenation showed somewhat greater F1 variability (0.006), followed by gated fusion (0.010) and cross-attention (0.012).

The structured-only model achieved the highest mean F1-score but exhibited greater variability across seeds (0.484 ± 0.023). ClinicalBERT-only showed particularly large variation in recall (0.614 ± 0.175),

indicating that its relatively high average recall was less stable across training runs.

These results suggest that bilinear fusion provided the most stable multimodal performance under the experimental conditions, although its average predictive performance remained slightly below the structured-only baseline.

4.8 Summary of Findings with Respect to the Research Questions

The results provide a consistent answer to the four research questions.

RQ1: Does adding clinical-note information improve ICU in-hospital mortality prediction compared with structured EHR data alone?

No consistent overall improvement was observed. The structured-only model achieved the highest F1-score, AUROC, and AUPRC. Bilinear fusion approached the structured baseline but did not surpass it on the principal metrics.

RQ2: How does clinical-text-only prediction compare with structured-only prediction?

Structured EHR variables provided substantially stronger overall predictive performance than ClinicalBERT-only representations. ClinicalBERT-only nevertheless achieved relatively high mortality recall, indicating that clinical notes contained predictive information but generated a substantially higher false-positive burden.

RQ3: How do different multimodal fusion mechanisms affect performance?

Fusion strategy influenced performance. Bilinear fusion was the strongest multimodal approach across the principal metrics, whereas concatenation achieved the highest mortality recall. Cross-attention did not provide an advantage over the simpler fusion mechanisms and produced the lowest multimodal F1-score, AUROC, and AUPRC.

RQ4: Does multimodal fusion provide a consistent advantage across threshold-dependent and threshold-independent metrics?

No. Multimodal models did not consistently outperform the structured-only baseline across F1-score, AUROC, and AUPRC. The principal exception was mortality recall, where concatenation achieved a 7.21-percentage-point improvement over structured-only prediction. This gain, however, was accompanied by a larger false-positive burden.

Overall, the results indicate that **clinical notes were predictive of ICU mortality but did not provide a consistent incremental advantage over structured ICU variables under the evaluated experimental conditions**. The predictive contribution of the clinical-text modality was conditional on the fusion mechanism and on the performance criterion considered.

5. Discussion

5.1 Principal Findings

This study examined whether clinical notes provide incremental predictive value beyond structured electronic health record (EHR) variables for ICU in-hospital mortality prediction and whether this contribution depends on the mechanism used to combine the two modalities. Using a common cohort of 20,000 ICU stays, a shared 24-hour observation window, patient-level data partitions, a frozen ClinicalBERT representation, and four multimodal fusion strategies, the results provide three principal findings.

First, the structured-only model achieved the strongest overall predictive performance. It produced the highest mean F1-score, AUROC, and AUPRC, indicating that structured ICU variables available during the first 24 hours contained substantial predictive information for the mortality outcome. Second, clinical notes were predictive but did not consistently provide an incremental advantage when added to structured EHR information. Bilinear fusion approached the structured-only baseline closely, but did not surpass it on the principal discrimination measures. Third, the effect of multimodal integration depended on the fusion mechanism and evaluation criterion. Concatenation produced the highest mortality recall but also generated substantially more false-positive predictions, whereas bilinear fusion provided the strongest and most stable overall multimodal performance.

These findings challenge the assumption that incorporating an additional EHR modality will necessarily improve predictive performance. Instead, they suggest that the predictive value of clinical notes should be evaluated relative to a strong structured-data baseline and across complementary performance measures.

5.2 Do Clinical Notes Provide Incremental Predictive Value?

The central question of this study was whether clinical notes improve mortality prediction when structured ICU variables are already available. The results provide limited evidence for a consistent incremental benefit.

The structured-only model achieved an AUROC of 0.859 ± 0.006 and AUPRC of 0.462 ± 0.025. The strongest multimodal alternative, bilinear fusion, achieved AUROC of 0.857 ± 0.004 and AUPRC of 0.461 ± 0.002. Thus, the addition of clinical notes did not increase threshold-independent discrimination under the evaluated conditions. The corresponding F1-score was also slightly lower for bilinear fusion than for the structured-only model.

This result should not be interpreted as evidence that clinical notes contain no useful information. The ClinicalBERT-only model achieved a mortality recall of 0.614 ± 0.175, demonstrating that the notes contained signals associated with mortality. Moreover, concatenation increased recall to 0.615 ± 0.034

compared with 0.543 ± 0.024 for structured-only prediction. The absence of an improvement in overall discrimination therefore appears to reflect the limited incremental contribution of the available textual information relative to the strong predictive signal already present in structured variables, rather than the complete absence of information in clinical notes.

An important implication is that multimodal prediction and incremental predictive value are not equivalent. A multimodal model may use information from both modalities and still fail to outperform a strong unimodal baseline. Consequently, claims that clinical notes improve EHR prediction should be based on direct comparisons with an appropriately matched structured-only model rather than on the performance of a multimodal model alone.

5.3 Why Did Structured EHR Variables Remain Dominant?

The stronger performance of the structured-only model may reflect the nature of the prediction task and the observation window. The study used information available during the first 24 hours of ICU admission. During this period, structured ICU variables can directly capture physiological measurements, laboratory results, interventions, and other clinically recorded variables that are closely associated with acute severity.

Clinical notes represent a different form of information. They can contain assessment, context, diagnostic reasoning, and narrative descriptions that may not be explicitly represented in structured fields. However, their predictive contribution depends on whether this information is sufficiently distinct from the structured variables and whether it is represented consistently enough for the model to exploit.

The results are consistent with a scenario in which a substantial proportion of the clinically relevant signal contained in the eligible notes overlaps with information already encoded in structured ICU variables. Under such circumstances, adding the text modality increases representational complexity without necessarily increasing discriminative information.

The findings therefore emphasize the importance of evaluating **incremental information content**, rather than assuming that additional modality diversity automatically translates into better prediction.

5.4 The Role of the Fusion Mechanism

The comparison of four multimodal strategies demonstrates that the method used to combine modalities materially affects performance.

Bilinear fusion produced the strongest overall multimodal results and remained close to the structured-only baseline across F1-score, AUROC, and AUPRC. Its relatively small performance variation across random seeds also indicates stable behavior under the evaluated experimental configuration. In contrast, concatenation achieved a similar F1-score but emphasized sensitivity more strongly, producing the highest mortality recall among the multimodal models.

Gated fusion produced intermediate performance, while cross-attention performed least favorably among the multimodal configurations on the principal metrics. These results indicate that greater architectural complexity does not necessarily produce greater predictive value. In particular, cross-attention did not provide a measurable advantage over simpler fusion mechanisms in this setting.

One possible interpretation is that the modalities did not require highly complex cross-modal interactions to capture their useful complementary information. When structured variables already provide strong predictive representations, a more complex fusion mechanism may introduce additional parameters and optimization requirements without adding sufficiently informative interactions. However, the present experiment does not isolate the precise mechanism responsible for these differences, and explanations concerning optimization or representation redundancy should therefore be regarded as hypotheses rather than demonstrated causal mechanisms.

The practical implication is that multimodal architecture selection should be empirical. Simpler fusion mechanisms should not be regarded as inferior merely because they model fewer complex interactions.

5.5 Sensitivity Versus False-Positive Burden

The threshold-dependent results reveal an important distinction that would be obscured by reporting a single aggregate metric.

Concatenation achieved the highest mortality recall, indicating that it identified a larger proportion of patients who experienced the outcome. However, this increase in sensitivity was accompanied by a substantially greater false-positive burden. The concatenation model produced 1,070 false-positive predictions compared with 780 for the structured-only model. ClinicalBERT-only generated 2,374 false positives, consistent with its low precision despite relatively high recall.

This finding illustrates why accuracy or recall alone is insufficient for evaluating multimodal mortality prediction, particularly when the outcome is relatively uncommon. A model that identifies more mortality cases by substantially increasing false alarms may not provide a better overall prediction system.

The results also demonstrate why AUPRC is particularly informative in this setting. With mortality prevalence of 12.40%, the no-skill precision–recall baseline is approximately 0.124. The structured-only and bilinear models achieved the highest AUPRC values, indicating that their favorable performance was not limited to a particular classification threshold.

Accordingly, the appropriate choice between multimodal configurations may depend on the intended application. If maximizing sensitivity is the primary objective, concatenation may be attractive. If a more balanced and stable predictive profile is desired, bilinear fusion provides a stronger multimodal alternative. These conclusions are empirical rather than clinical recommendations, and prospective evaluation would be required before translating such trade-offs into clinical decision thresholds.

5.6 Implications for Multimodal EHR Modeling

The findings have several methodological implications for research on multimodal clinical prediction.

First, a strong unimodal structured baseline is essential. Without it, a multimodal model that achieves good predictive performance may be incorrectly interpreted as demonstrating the value of clinical text. In the present study, the structured-only model performed at least as well as every multimodal configuration on the principal overall metrics.

Second, modality contribution should be evaluated using more than one metric. The clinical notes contributed to mortality sensitivity in some configurations, but this did not translate into improved AUROC, AUPRC, or F1-score. Reporting only one performance measure could therefore lead to a substantially different interpretation of the same models.

Third, fusion mechanisms should be compared under controlled conditions. Because the structured pathway, text representation, cohort, observation window, and evaluation protocol were held constant in this study, differences among multimodal configurations can be interpreted more directly as differences associated with the fusion strategy rather than changes in the underlying data or encoder.

Finally, the results suggest that the value of multimodal EHR modeling may depend on the clinical task, data representation, observation window, and information overlap between modalities. Clinical notes may provide greater incremental value when they contain information that is unavailable or poorly represented in structured fields. Establishing when this condition occurs is an important direction for future multimodal clinical prediction research.

5.7 Relation to Existing Multimodal Prediction Research

Previous studies have demonstrated that combining structured EHR information with clinical text can improve clinical outcome prediction. Multimodal architectures have incorporated clinical notes alongside structured variables using approaches ranging from direct feature concatenation to more sophisticated attention-based and hierarchical fusion mechanisms. These studies establish the potential value of multimodal EHR representations but do not imply that the addition of clinical text will consistently improve performance across datasets and tasks.

The present study complements this literature by focusing specifically on **controlled modality comparison and incremental predictive value**. Rather than evaluating only whether a multimodal architecture achieves high predictive performance, the experiment compares structured-only, text-only, and multimodal configurations under a common experimental protocol. This distinction is important because the apparent benefit of multimodal learning can depend on the strength of the unimodal baselines, the fusion strategy, and the evaluation metric.

The results therefore support a more cautious interpretation of multimodal EHR prediction: clinical text can be informative without necessarily providing measurable improvement over structured information, and multimodal fusion should be justified empirically rather than assumed to be beneficial.

5.8 Clinical and Methodological Implications

From a methodological perspective, the results suggest that multimodal EHR systems should be evaluated according to the specific information contribution of each modality. A model that requires clinical text processing introduces additional computational and engineering requirements, including text preprocessing, language-model inference, storage, and model integration. If the resulting system does not improve discrimination over a strong structured baseline, the justification for this additional complexity becomes less clear.

At the same time, the higher recall observed with concatenation indicates that clinical notes may remain useful when the objective places greater emphasis on identifying mortality cases. This suggests that modality contribution may be application-dependent rather than uniformly positive or negative.

For clinical AI research, the principal implication is therefore not that structured data should replace clinical notes. Rather, **the incremental value of clinical notes should be demonstrated for the specific task and evaluation objective under consideration**. Such evaluation can help determine whether the additional modality provides information that is sufficiently complementary to justify its computational and methodological cost.

5.9 Limitations

Several limitations should be considered when interpreting these findings.

First, the analysis was conducted using a retrospective dataset derived from MIMIC-IV and MIMIC-IV-Note. Although the cohort was relatively large, the study does not establish how the models would perform on patients from other institutions, healthcare systems, or documentation practices.

Second, the observation window was restricted to the first 24 hours of ICU admission. Clinical notes generated later during hospitalization may contain additional information that could alter the incremental contribution of the text modality. The present findings therefore apply specifically to the evaluated early-ICU prediction setting.

Third, ClinicalBERT was used as a frozen text representation across the text-containing models. Fine-tuning the language model or using a more recent clinical language model could potentially change the relative contribution of clinical notes.

Fourth, the study evaluated four fusion mechanisms under a common experimental configuration. The results should therefore not be interpreted as evidence that bilinear fusion is universally superior to gated fusion, concatenation, or cross-attention. Different encoder dimensions, training strategies, preprocessing procedures, or datasets could produce different rankings.

Fifth, the study focused on in-hospital mortality as a single outcome. The incremental value of clinical notes may differ for other outcomes, including readmission, length of stay, deterioration, or specific clinical events.

Finally, the evaluation was retrospective and did not assess prospective clinical utility, calibration in clinical practice, decision-curve performance, or the effect of model predictions on clinical decision-making. Further external and prospective validation would be necessary before considering clinical deployment.

5.10 Overall Interpretation

Taken together, the findings indicate that clinical notes contain predictive information for ICU mortality but did not provide a consistent incremental advantage over structured ICU variables within the evaluated 24-hour prediction setting. The structured-only model achieved the strongest overall performance, while bilinear fusion provided the closest multimodal alternative and the most stable multimodal results. Concatenation increased mortality recall but at the cost of a substantially larger false-positive burden, demonstrating that modality integration can change the operating characteristics of a model without improving its overall discrimination.

The broader implication is that multimodal EHR prediction should not be framed simply as a question of whether more modalities can be incorporated. The more relevant question is whether the additional modality contributes **new predictive information that remains useful after accounting for the information already present in other EHR representations**. Under the conditions evaluated in this study, clinical notes provided evidence of predictive value but not of consistent incremental discrimination. This finding supports controlled, modality-aware evaluation as an important methodological requirement for multimodal clinical AI.

6. Limitations

This study has several limitations that should be considered when interpreting the comparative performance of structured, clinical-text, and multimodal models.

First, the evaluation was conducted using a single retrospective dataset, MIMIC-IV and its linked MIMIC-IV-Note component. Although the study included 20,000 ICU stays and used patient-level partitioning, the data originate from a single healthcare system. Differences in patient populations, clinical workflows, documentation practices, measurement patterns, and institutional coding practices may affect the relative contribution of structured variables and clinical notes. Consequently, the observed performance and the relative ranking of fusion strategies should not be assumed to generalize directly to other hospitals or healthcare systems. External validation using independent clinical datasets is required.

Second, the study evaluated only one prediction outcome, namely in-hospital mortality, using information available during the first 24 hours of ICU admission. This design provides a controlled early-prediction setting but does not establish whether the contribution of clinical notes would be different for other outcomes or for longer observation periods. Notes generated later during an ICU stay may contain additional diagnostic, prognostic, or clinical-management information that is not available within the defined observation window.

Third, the clinical-text representation was generated using a frozen Bio_ClinicalBERT/ClinicalBERT encoder. The language model was not fine-tuned for the mortality-prediction task. Although using a common frozen encoder strengthened the experimental control by keeping the text representation consistent across the text-containing models, it also limits the conclusions that can be drawn about the maximum predictive value obtainable from clinical notes. Fine-tuned clinical language models, alternative pretrained models, or task-specific text representations may produce different results.

Fourth, the study compared four fusion mechanisms under a common experimental configuration: concatenation, gated fusion, bilinear fusion, and cross-attention. The results therefore characterize these strategies under the specific encoders, representations, observation window, and training procedure used in this study. They should not be interpreted as evidence that one fusion mechanism is universally superior to the others. In particular, the finding that bilinear fusion performed best among the evaluated multimodal models does not establish its superiority across datasets, clinical tasks, or model configurations.

Fifth, the study focused primarily on predictive discrimination and classification performance. Although accuracy, precision, recall, F1-score, AUROC, and AUPRC were evaluated across three random seeds, other dimensions of clinical model evaluation were outside the scope of the experiment. Calibration, decision-curve analysis, clinical utility, interpretability, and the consequences of different operating thresholds were not systematically evaluated. These factors are important when determining whether a model's predictions could provide meaningful support in clinical practice.

Finally, the study was conducted as a retrospective research evaluation and did not assess prospective clinical deployment, real-time integration with hospital information systems, or the effect of model predictions on clinical decision-making. The observed differences between structured and multimodal models should therefore be interpreted as evidence concerning predictive performance under the experimental conditions, rather than evidence of clinical effectiveness.

Despite these limitations, the controlled design provides a useful basis for assessing the incremental contribution of clinical notes relative to structured ICU information. Future studies should extend the comparison to independent institutions and datasets, additional clinical outcomes, alternative and fine-tuned clinical language models, temporally richer note representations, and prospective evaluations of calibration and clinical utility.

7. Conclusion

This study investigated whether clinical notes provide incremental predictive value beyond structured EHR variables for ICU in-hospital mortality prediction and examined how this contribution is affected by the choice of multimodal fusion strategy. Using a controlled experimental design based on 20,000 ICU stays, a common first-24-hour observation window, patient-level data partitions, and three random seeds, six prediction configurations were systematically compared.

The results show that structured ICU variables provided the strongest overall predictive signal under the evaluated conditions. The structured-only MLP achieved the highest overall F1-score, AUROC, and AUPRC, while the multimodal models did not consistently exceed this baseline. Bilinear fusion was the strongest multimodal approach and closely approached the structured-only model across the principal performance measures. Clinical notes nevertheless demonstrated predictive value, particularly with respect to mortality recall. Concatenation achieved the highest mortality recall among the evaluated models, but this improvement was accompanied by a substantially greater false-positive burden.

These findings indicate that the addition of clinical notes should not be assumed to provide an automatic improvement in ICU mortality prediction. Instead, their contribution appears to be conditional on the fusion mechanism and the performance criterion being considered. In particular, improved sensitivity does not necessarily translate into improved overall discrimination or classification performance.

The principal methodological implication is that multimodal clinical prediction should be evaluated against strong unimodal baselines using complementary threshold-dependent and threshold-independent metrics. Such controlled comparisons can distinguish genuine incremental information from improvements that arise primarily from changes in model operating characteristics.

Future work should evaluate these findings across independent institutions and clinical outcomes, investigate fine-tuned and alternative clinical language models, incorporate temporally richer representations of clinical notes, and assess calibration and clinical utility in prospective settings. Such studies will help determine when clinical text provides sufficiently complementary information to justify its integration into multimodal clinical prediction systems.

References

1. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S et al (2023) MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 10:1. 10.1038/s41597-022-01899-x
2. Johnson A, Pollard T, Horng S, Celi LA, Mark R (2023) MIMIC-IV-Note: Deidentified free-text clinical notes. 10.13026/1n74-ne17. Version 2.2. PhysioNet
3. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M et al (2016) MIMIC-III, a freely accessible critical care database. Sci Data 3:160035. 10.1038/sdata.2016.35
4. Huang K, Altosaar J, Ranganath R, ClinicalBERT (2019) Modeling clinical notes and predicting hospital readmission. arXiv. 10.48550/arXiv.1904.05342
5. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T et al (2019) Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop; pp. 72–78. 10.18653/v1/W19-1909
6. Calvert J, Mao Q, Hoffman JL, Jay M, Desautels T, Mohamadlou H et al (2016) Using electronic health record collected clinical variables to predict medical intensive care unit mortality. Ann Med

Surg (Lond) 11:52–57. 10.1016/j.amsu.2016.09.002

7. Niu K, Lu Y, Peng X, Zeng J (2022) Fusion of sequential visits and medical ontology for mortality prediction. J Biomed Inf 127:104012. 10.1016/j.jbi.2022.104012
8. Bardak B, Tan M (2021) Improving clinical outcome predictions using convolution over medical entities with multimodal learning. Artif Intell Med 117:102112. 10.1016/j.artmed.2021.102112
9. Lyu W et al (2022) A multimodal transformer: Fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction. In: AMIA Annual Symposium Proceedings
10. Sun M, Yang X, Niu J, Gu Y, Wang C, Zhang W et al (2024) A cross-modal clinical prediction system for intensive care unit patient outcome. Knowl Based Syst 283:111160. 10.1016/j.knosys.2023.111160
11. Wang C, Yang X, Sun M, Gu Y, Niu J, Zhang W et al (2024) Multimodal fusion network for ICU patient outcome prediction. Neural Netw 180:106672. 10.1016/j.neunet.2024.106672
12. Teles AS, de Moura IR, Silva F, Roberts A, Stahl D (2025) EHR-based prediction modelling meets multimodal deep learning: A systematic review of structured and textual data fusion methods. Inf Fusion 118:102981. 10.1016/j.inffus.2025.102981
13. Wang Y, Li W (2025) Integrating multimodal EHR data for mortality prediction in ICU sepsis patients. Stat Med. 10.1002/sim.70060
14. Chen R, Cai Y, Zhang S, Huo Z, Song M, Li W et al (2026) Contextualized biomedical language processing enhances ICU survival prediction. iScience 29(1):114442. 10.1016/j.isci.2025.114442
15. Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 10(3):e0118432. 10.1371/journal.pone.0118432
16. Davis J, Goadrich M (2006) The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning; pp. 233–240. 10.1145/1143844.1143874
17. Devlin J, Chang MW, Lee K, Toutanova K, BERT (2019) Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; pp. 4171–4186. 10.18653/v1/N19-1423
18. Johnson AEW, Stone DJ, Celi LA, Pollard TJ (2018) The MIMIC Code Repository: enabling reproducibility in critical care research. J Am Med Inf Assoc 25(1):32–39. 10.1093/jamia/ocx084

Figures

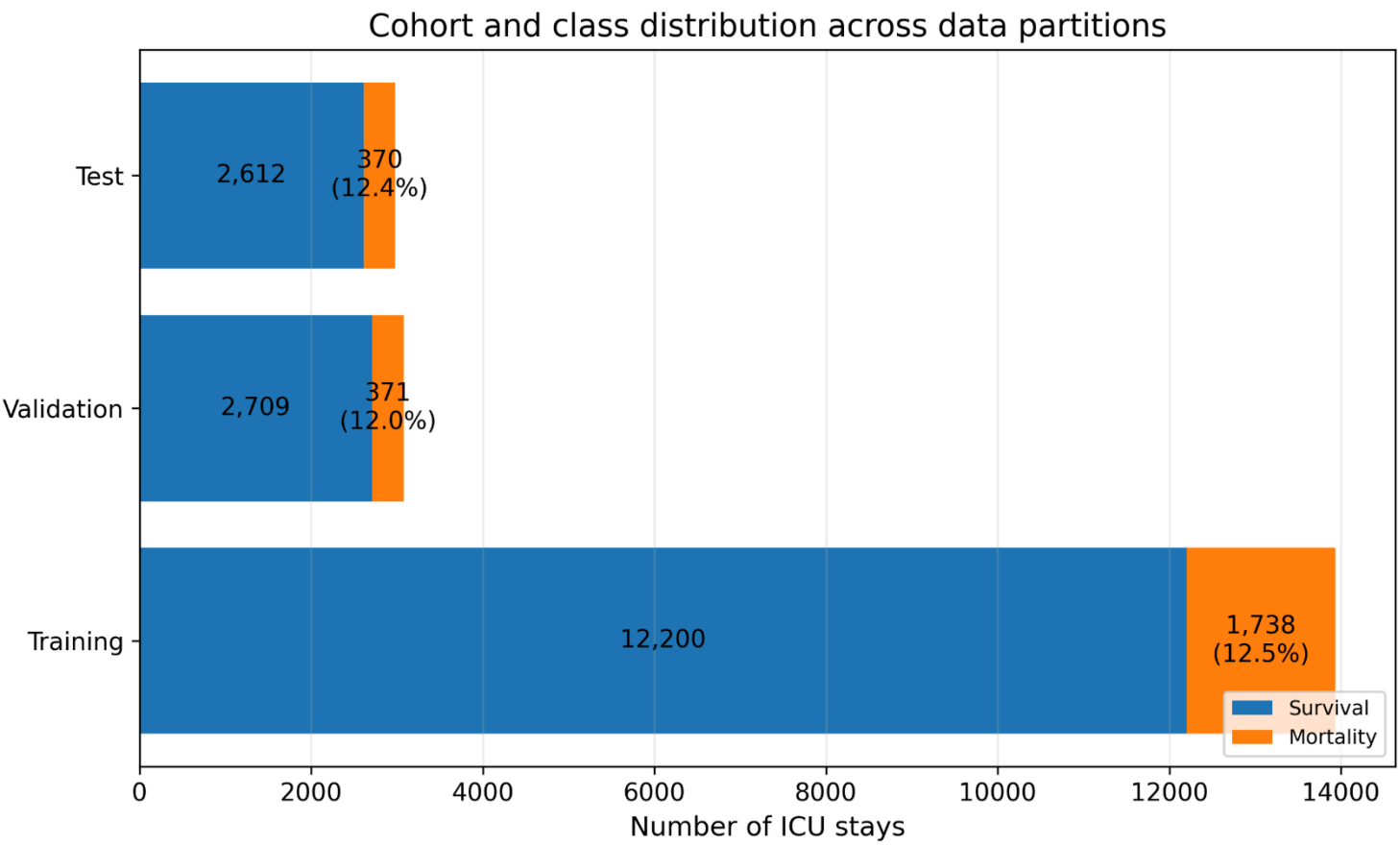


Figure 1

Distribution of survival and mortality cases across the training, validation, and test partitions.

The figure shows the number of survival and in-hospital mortality cases in each patient-level data partition.

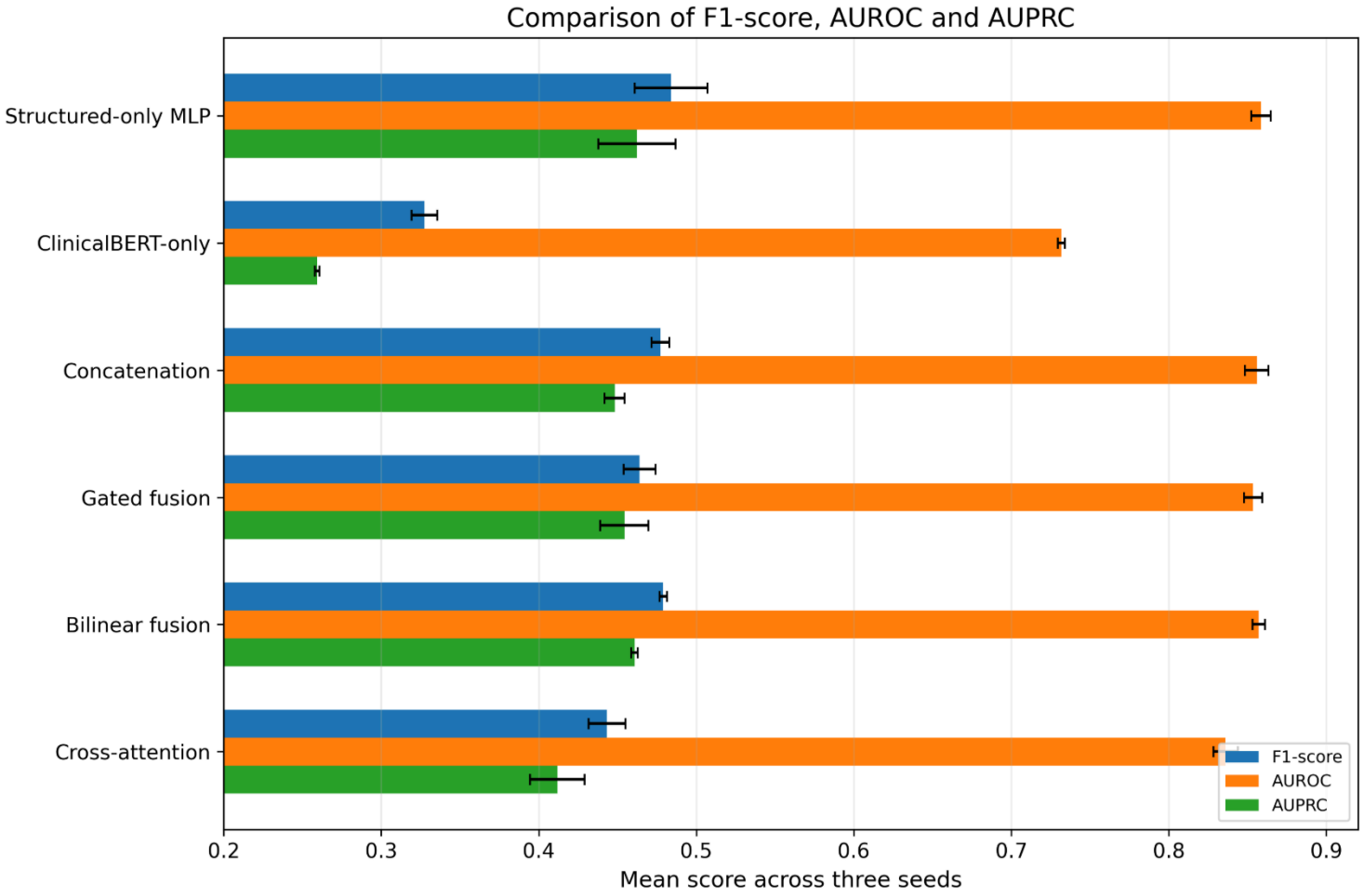


Figure 2

Comparison of F1-score, AUROC, and AUPRC across the six prediction models.
Values represent mean performance across three random seeds. The figure compares the structured-only MLP, ClinicalBERT-only model, and four multimodal fusion strategies.

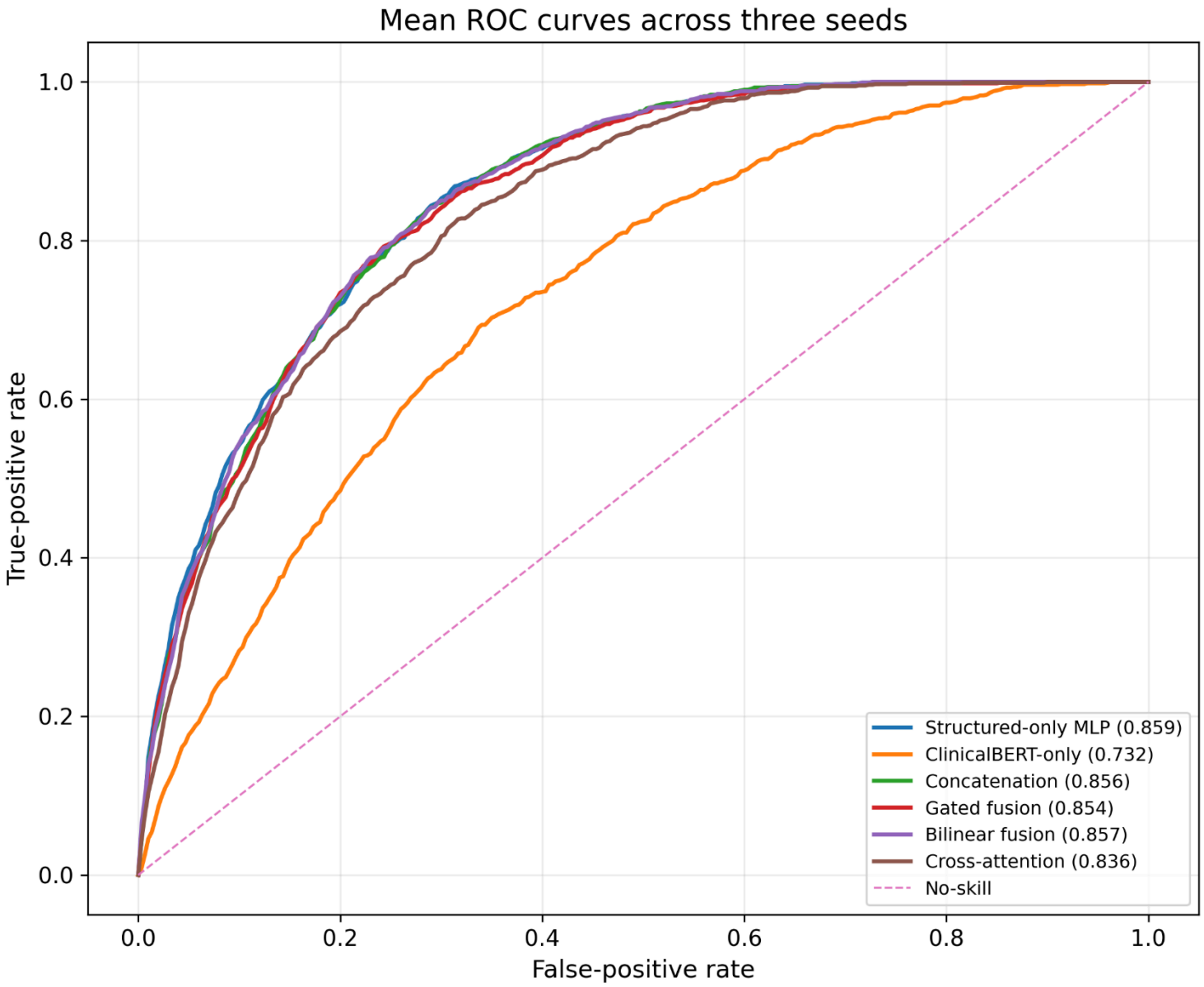


Figure 3

Mean receiver operating characteristic and precision–recall curves across three random seeds. *(A) Mean ROC curves showing discrimination across classification thresholds. (B) Mean precision–recall curves, with the no-skill AUPRC corresponding approximately to the cohort mortality prevalence of 0.124.*

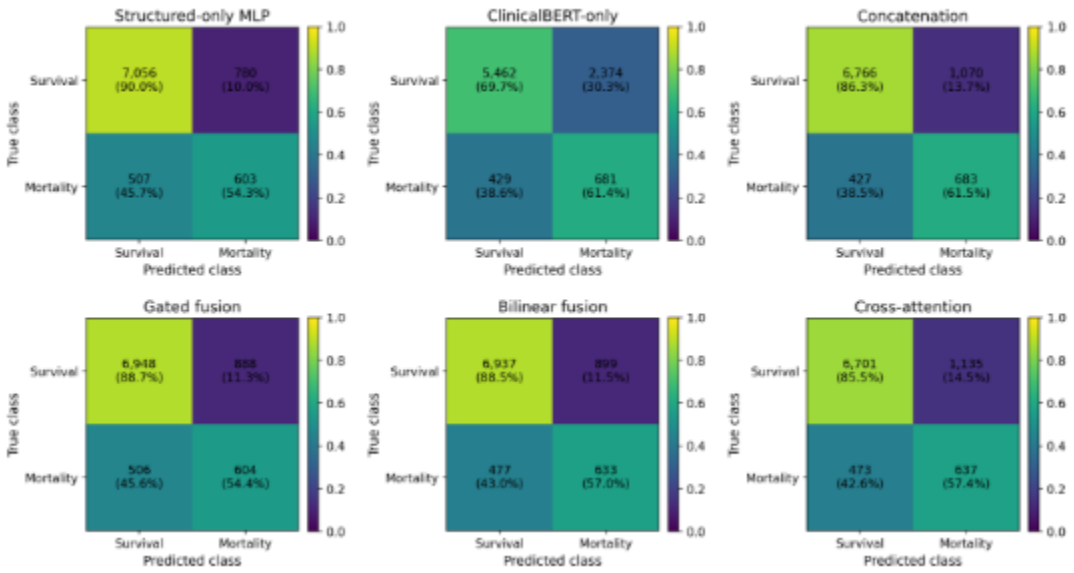


Figure 4

Confusion matrices for the structured-only, ClinicalBERT-only, concatenation, and bilinear fusion models.

Confusion matrices are shown to illustrate differences in mortality detection, false-negative errors, and false-positive burden. The selected models represent the structured baseline, clinical-text-only baseline, highest-recall multimodal model, and strongest overall multimodal model, respectively.