Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.38289

Personalized Clinical Risk Stratification Based on Multimodal Medical Data Fusion

Yutong Li*, Yifan Qi

School of Computing, City University of Hong Kong, Hong Kong, China
**Corresponding Author. Email: liyutong02571@outlook.com*

Abstract. A representation of the patient capturing physiological data, contextual data, and radiological information is required for personalized clinical risk stratification. In this paper, we propose a multi-modal approach that combines structured information extracted from electronic health records, unstructured information from clinical notes, and chest X-rays collected from the MIMIC-IV and MIMIC-CXR-JPG databases within the first 24 hours in intensive care. We encode structured information using TabTransformer, unstructured text with BioClinicalBERT, and radiology images with DenseNet121. These modality-specific representations are then mapped to a shared latent space and fused by a patient-specific gated attention mechanism to predict in-hospital mortality and classify into low, intermediate, and high-risk groups. The fused representation gave an AUROC of 0.912 ± 0.006, an AUPRC of 0.478 ± 0.014, and a Brier score of 0.074 ± 0.003, outperforming concatenation of the features, which scored an AUROC of 0.889 ± 0.007. When structured, text, and radiological information were dropped, the model's discrimination performance declined to 0.853 ± 0.009, 0.881 ± 0.008, and 0.894 ± 0.007, respectively. The mortality rates among the risk-stratified groups ranged from 1.9 ± 0.4% to 28.7 ± 1.6%, indicating clear clinical separation.

Keywords: clinical risk stratification, multimodal medical data, data fusion, deep learning, personalized prediction

1. Introduction

Clinical risk stratification informs care escalation, monitoring frequency, and resource distribution within the critical-care setting; however, traditional risk stratification methods often rely on a reduced set of structured variables to represent the patient state. This method ignores the information present in the narrative and imaging domains despite the use of both by clinicians when assessing severity. Recent reviews on multimodal machine learning in healthcare point out that the complementary use of heterogeneous sources is beneficial in clinical representation if properly managed [1]; image and clinical multimodal approaches have shown the importance of representation alignment, modality interaction, and fusion in determining performance and usability [2].

The existence of linked databases of critical care provides the opportunity to study this question in a controlled fashion. The MIMIC-IV database contains detailed demographics, lab values, vital

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.38289

signs, medications, diagnoses, and clinical documentation [3]. Multimodal studies on chest radiographs and clinical data have shown that the combination of different sources may improve predictive power and has exposed technical challenges inherent in the process, such as asynchronous acquisition of heterogeneous sources and the lack of perfect matching between them [4]. Finally, clinical notes contain relevant prognostic information that is not captured in structured variables; models based on early notes in the ICU have produced significant discrimination in multiple mortality horizons [5].

One aspect that remains to be addressed is the patient-specific nature of the fusion process itself. For example, a chest radiograph will play a greater role in representing a patient with acute pulmonary disease, whereas for other patients, lab values or clinical documentation will be more predictive. This is not possible with fixed fusion weights. A gated multimodal model, which integrates the structured data, clinical notes, and chest X-rays from the first 24 hours after ICU admission, is thus proposed for this study. Separate encoders preserve the modality-specific structure, a shared latent space allows for comparable weight allocation, and adaptive gates govern the relative contribution of each modality in mortality risk assessment. Continuous probabilities are converted into low-, intermediate-, and high-risk categories through thresholding.

2. Literature review

2.1. Personalized clinical risk stratification

Risk stratification has moved from scoring at the bedside using fixed rules to representation learning, which allows the modeling of nonlinear relationships between patient features. The problem with strictly structured data is not the lack of predictive features but rather the reduction of patient state into measures that do not capture diagnostic reasoning or visibly apparent burden of disease. Multimodal image and text research illustrates that multiple information sources for biomedicine contain different levels of semantic content and that specific alignment is necessary prior to prediction tasks [6]. Modern risk prediction research based on EHRs reveals similar findings about differences in the importance of physiological data, images, and clinical notes per outcome, which supports models that maintain information prior to fusion [7].

2.2. Multimodal medical data fusion

In multimodal fusion for medical applications, one must consider different aspects such as dimensionality differences, statistical differences, and clinical significance. Disentangled fusion algorithms have already demonstrated that shared and modality-specific data should be disentangled in the case when signals from electronic health record and chest radiographs are not consistent with each other [8]. Confidence-based multimodal systems also confirm that fusion performance can be improved when the system estimates the confidence of the representations of each modality rather than considering their evidences to be of equal importance [9]. The systematic review on the structured and unstructured EHR fusion provides additional evidence about the importance of the adaptive fusion and reproducible comparative evaluation [10].

DOI: 10.54254/2755-2721/2026.38289

3. Methodology

3.1. Dataset construction and outcome definition

The overall framework of patient-specific multimodal fusion for personalized clinical risk stratification is illustrated in Figure 1. Patient-level records were formed by joining MIMIC-IV and MIMIC-CXR-JPG based on patient identification and hospital admission information. The criteria for inclusion were adults of 18 years of age or above, where an ICU admission record had structured observations, one clinical note, and one frontal chest radiograph within 24 hours. In case of multiple radiographs, the one closest to the end of the observation period without going past the prediction window was selected. The outcome variable was mortality of the patient within 24 hours in the hospital post-initial observation period. The structured variables included demographics, comorbidities, vitals, complete blood count, electrolyte level, kidney markers, glucose, lactate, and treatment of major respiratory or vasoactive problems. Physiological outliers were treated as missing. Patient-level data partitioning was done before preprocessing, such that multiple admissions by the same patient could not span across training, validation, and testing sets.

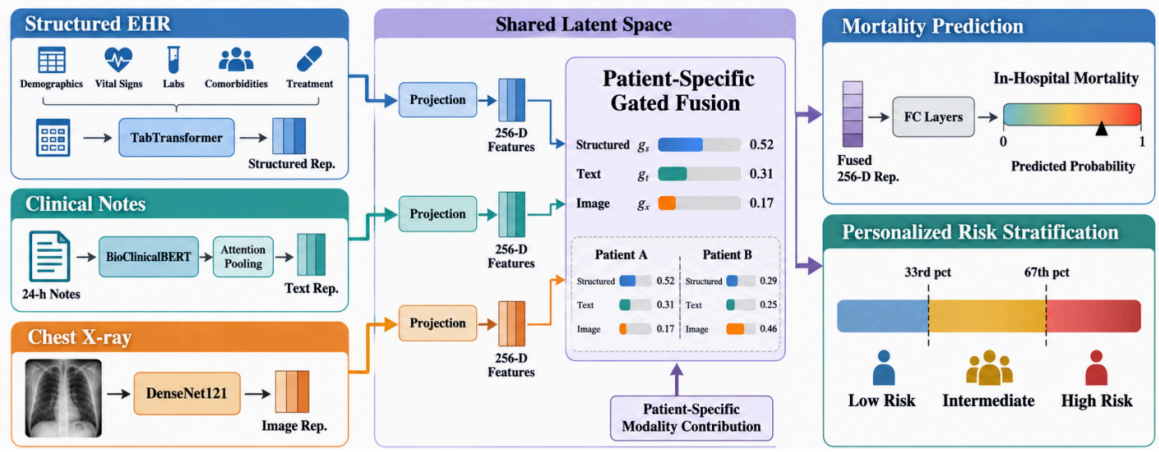

Figure 1. Patient-specific multimodal fusion framework for personalized clinical risk stratification

3.2. Modality-specific data processing

Continuous structured variables underwent winsorization at the 0.5th and 99.5th percentiles of the training cohort and standardization using training statistics. Median imputation was done along with missingness variables whereas categorical variables were turned into trainable embeddings. Clinical notes were arranged chronologically with removal of administrative boilerplate but keeping negations and temporal markers intact. Long clinical notes were split into overlapping 384-token sequences with a 64-token overlap. Each sequence was then embedded using BioClinicalBERT and the [CLS] vector from each sequence was pooled using attention pooling to create a single 768-dimensional embedding. The selected CXR image was scaled down to 224 × 224 pixels and underwent intensity normalization. DenseNet121 with medical pretraining weights was employed for feature extraction.

3.3. Multimodal feature encoding

The structured features were encoded through a four-layered TabTransformer with eight heads and a hidden size of 128. The BioClinicalBERT model yielded the 768-dimensional narrative vector, while

DOI: 10.54254/2755-2721/2026.38289

DenseNet121 global average pooling returned a 1,024-dimensional radiographic vector. Three separate projection heads were used to project the aforementioned encodings into a latent space of 256 dimensions, which was achieved via a fully connected layer, followed by GELU, layer normalization, and 0.20 dropout. The modality-specific projections retained the statistical dependencies of the source while ensuring dimensional uniformity. The use of modality dropout of 0.10 during training helped mitigate over-reliance on a particular modality.

3.4. Gated multimodal fusion model

The fusion module was implemented in PyTorch 2.2. The projected vectors formed a patient context from which a gating network generated three non-negative modality weights summing to one. These weights were recalculated for every patient, allowing physiological instability, narrative severity, or radiographic abnormality to dominate according to the joint representation. Weighted vectors were concatenated so modality identity remained explicit after gating.

For patient i and modality m, the personalized gate and mortality probability were defined as shown in Equation (1):

$$g_{i,m} = \frac{exp\{\boldsymbol{a}_m^{\top} tanh[\boldsymbol{W}_g(\boldsymbol{z}_{i,s} \| \boldsymbol{z}_{i,t} \| \boldsymbol{z}_{i,x}) + \boldsymbol{b}_g]\}}{\sum_{r \in \{s,t,x\}} exp\{\boldsymbol{a}_r^{\top} tanh[\boldsymbol{W}_g(\boldsymbol{z}_{i,s} \| \boldsymbol{z}_{i,t} \| \boldsymbol{z}_{i,x}) + \boldsymbol{b}_g]\}} \tag{1}$$

Here, $\boldsymbol{z}$ denotes projected structured, textual, and imaging representations, g is the patient-specific modality weight. The classifier used 384- and 128-unit fully connected layers. Positive-class-weighted binary cross-entropy addressed imbalance. AdamW used a 1×10^{-4} learning rate, 1×10^{-5} weight decay, batch size 32, gradient clipping at 1.0, and cosine decay. Training stopped after ten epochs without validation-AUROC improvement. In order to avoid the gate collapsing towards a single dominant modality source, the weak entropy regularization term was used early on in training and annealed away, while allowing for strong patient-specific weighting when the modality had considerably more information to offer.

3.5. Personalized risk stratification

The continuously computed mortality probability was used as the individualized probability estimate. The 33rd and 67th percentiles of the risks in the training cohort defined the fixed low-, medium-, and high-risk categories, which remained static for validation and testing. The gate values for the individual patient were included with each estimate, thus allowing patients in the same risk group to have different contribution patterns.

4. Experimental design and validation

4.1. Comparative and ablation experiments

The participants were split into 70% for training, 15% for validation, and 15% for testing via mortality stratified patient-level sampling. The comparison was made between structured logistic regression, TabTransformer, BioClinicalBERT, DenseNet121, and concatenated projection of three embeddings. The splits for each model were identical, and selection of the models was based on training and validation. The ablation study included removing the structured data, clinical text, and

DOI: 10.54254/2755-2721/2026.38289

CXR data independently. In addition, another comparison used unweighted concatenation without patient-specific gating but with the same encoders' outputs. The analysis was performed using five different random seeds. The test metrics were provided as mean ± standard deviation, and paired AUROC difference was computed using patient-level bootstrap of 1,000 patients with Holm correction.

4.2. Performance and clinical validation

The discrimination metrics included AUROC and AUPRC while the metrics for calibration included Brier Score and Expected Calibration Error. Cutoff values derived from training fixed the thresholds used to predict mortality in the 3 risk groups. Subgroup analysis was done based on sex and age groups (less than 65 and above 65), and respiratory/cardiovascular/sepsis-based diagnoses.

5. Results

5.1. Multimodal prediction performance and ablation analysis

Among 18,462 patients, 2,176 died in hospital. Gated fusion achieved AUROC 0.912 ± 0.006, AUPRC 0.478 ± 0.014, F1-score 0.521 ± 0.016, and Brier score 0.074 ± 0.003. Direct concatenation reached AUROC 0.889 ± 0.007*. Removing structured, textual, or imaging data reduced AUROC to 0.853 ± 0.009*, 0.881 ± 0.008*, and 0.894 ± 0.007*, respectively. Expected calibration error was 0.024 ± 0.004* versus 0.041 ± 0.006 for direct concatenation (see Table 1).

Table 1. Multimodal prediction and ablation results

Model / Setting	AUROC	AUPRC	F1-score	Balanced Accuracy	Brier Score
Logistic regression	0.801 ± 0.011	0.301 ± 0.015	0.411 ± 0.019	0.746 ± 0.016	0.101 ± 0.005
TabTransformer	0.858 ± 0.009	0.377 ± 0.014	0.463 ± 0.018	0.781 ± 0.015	0.091 ± 0.004
BioClinicalBERT	0.824 ± 0.010	0.338 ± 0.016	0.438 ± 0.020	0.763 ± 0.017	0.097 ± 0.005
DenseNet121	0.769 ± 0.012	0.264 ± 0.014	0.391 ± 0.018	0.721 ± 0.018	0.108 ± 0.006
Direct concatenation	0.889 ± 0.007	0.431 ± 0.013	0.496 ± 0.017	0.816 ± 0.013	0.081 ± 0.004
Gated multimodal fusion	0.912 ± 0.006*	0.478 ± 0.014*	0.521 ± 0.016*	0.836 ± 0.012*	0.074 ± 0.003*
Without structured data	0.853 ± 0.009*	0.361 ± 0.015	0.456 ± 0.018	0.777 ± 0.015	0.093 ± 0.005
Without clinical text	0.881 ± 0.008*	0.416 ± 0.014	0.487 ± 0.017	0.808 ± 0.014	0.084 ± 0.004
Without chest X-ray	0.894 ± 0.007*	0.443 ± 0.013	0.503 ± 0.016	0.821 ± 0.013	0.080 ± 0.004

**P < 0.05 after Holm-adjusted bootstrap comparison with the relevant reference model.*

5.2. Personalized risk stratification and subgroup validation

Mortality increased from 1.9 ± 0.4% in low risk to 7.6 ± 0.8% in intermediate risk and 28.7 ± 1.6% in high risk*. High-risk gate weights were 0.44 ± 0.12 for structured data, 0.34 ± 0.11 for text, and 0.22 ± 0.09 for imaging; imaging weight rose to 0.37 ± 0.12 in major respiratory diagnoses. AUROC was 0.910 ± 0.009 in women, 0.914 ± 0.008 in men, 0.917 ± 0.008 below age 65, and 0.906 ± 0.009 at older ages. Respiratory, cardiovascular, and sepsis subgroups achieved 0.918 ± 0.010, 0.904 ± 0.011, and 0.909 ± 0.010.

DOI: 10.54254/2755-2721/2026.38289

6. Conclusion

The multimodal architecture combines structured clinical data, early clinical text narratives, and chest X-rays into a modality-aware representation for stratifying in-hospital mortality. The TabTransformer, BioClinicalBERT, and DenseNet121 encoders capture information modality-wise, whereas gating fusion weights the evidential impact on a per-patient basis before making predictions. Our model provided an improvement in terms of discrimination, precision-recall, and calibration compared to both unimodal and concatenated baselines, and the ablation study demonstrated complementarity of the information from all three modalities. The ability to separate mortality based on fixed risk strata and subgroup stability confirms that our multimodal approach is clinically meaningful.

Author contribution

Yutong Li and Yifan Qi contributed equally to this paper.

References

[1] Krones, F., Marikkar, U., Parsons, G., Szmul, A., & Mahdi, A. (2025). Review of multimodal machine learning approaches in healthcare. *Information Fusion, 114*, 102690.
[2] Warner, E., Lee, J., Hsu, W., Syeda-Mahmood, T., Kahn, C. E., Jr., Gevaert, O., & Rao, A. (2024). Multimodal machine learning in image-based and clinical biomedicine: Survey and prospects. *International Journal of Computer Vision, 132*, 3753–3769.
[3] Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L. W. H., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data, 10*, 1.
[4] Hayat, N., Geras, K. J., & Shamout, F. E. (2022). MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. *Proceedings of Machine Learning Research, 182*, 479–503.
[5] Mahbub, M., Srinivasan, S., Danciu, I., Peluso, A., Begoli, E., Tamang, S., & Peterson, G. D. (2022). Unstructured clinical notes within the 24 hours since admission predict short, mid & long-term mortality in adult ICU patients. *PLOS ONE, 17*(1), e0262182.
[6] Sun, Z., Lin, M., Zhu, Q., Xie, Q., Wang, F., Lu, Z., & Peng, Y. (2023). A scoping review on multimodal deep learning in biomedical images and texts. *Journal of Biomedical Informatics, 146*, 104482.
[7] Wang, Y., Yin, C., & Zhang, P. (2024). Multimodal risk prediction with physiological signals, medical images and clinical notes. *Heliyon, 10*(5), e26772.
[8] Yao, W., Yin, K., Cheung, W. K., Liu, J., & Qin, J. (2024). DrFuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(15), 16416–16424.
[9] Jorf, B. A., & Shamout, F. E. (2025). MedPatch: Confidence-guided multi-stage fusion for multimodal clinical data. *Proceedings of Machine Learning Research, 298*.
[10] Teles, A. S., de Moura, I. R., 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. *Information Fusion, 118*, 102981.