TGCA: A Text-Guided Cross-Attention Multimodal Fusion Frameworkfor Prognosis Prediction

Jiawei Wang[1], Qieshi Zhang[1], Jun Cheng[1], Yinzhong Ma[1], Yabin Ji[2]

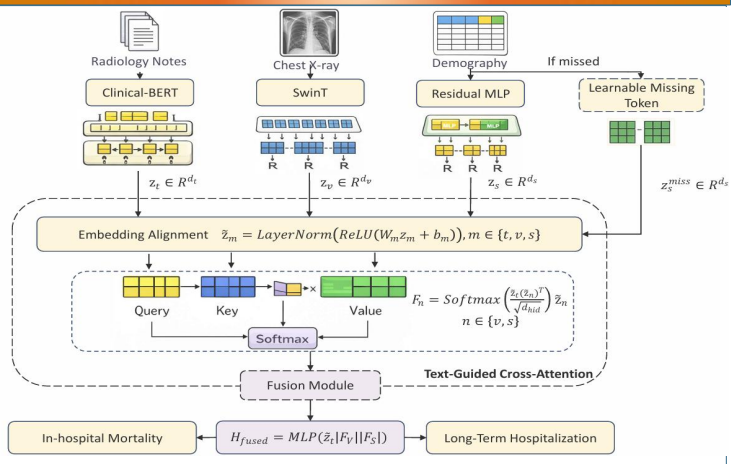
[1]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

[2]Southern Medical University, Guangzhou, China

Abstract

Accurate prognosis prediction is a cornerstone of modern clinical decision-making, where integrating massive heterogeneous clinical data provides a vital foundation for forecasting patient outcomes. However, existing models often rely on single-modality data or naive feature fusion, struggling to bridge the nonlinear gap between modalities and comprehensively capture the complex physiological states of patients. To address this, we propose a Text-Guided Cross-attention Mechanism (TGCA), which uses expert descriptions from medical reports to guide the model's focus toward pathological image regions and key tabular physiological indicators, effectively uncovering potential cross-modal information. The framework adopts a decoupled multi-channel architecture, utilizing ClinicalBERT to extract text semantics, Swin Transformer (SwinT) to capture image features, and Residual Multilayer Perceptron (ResMLP) to process physiological data before fusion. Simultaneously, to address modality missingness, we innovatively introduce a learnable missing labeling strategy, effectively reducing the bias caused by missing data. Comparative experiments show the model significantly outperforms other methods on the Medical Information Mart for Intensive Care (MIMIC-IV) integrated dataset. Furthermore, ablation experiments demonstrate the effectiveness of our fusion and modality missingness strategies.

Methodology



The proposed architecture integrates radiology notes, chest X-rays, and demographic information for clinical outcome prediction. Clinical-BERT, Swin Transformer, and a residual MLP encode textual, visual, and structured modalities, respectively. To address incomplete records, missing demographic inputs are replaced with learnable tokens. Aligned embeddings are then processed through text-guided cross-attention, where note-derived representations query visual and structured features. The resulting attended features are concatenated and passed into a fusion MLP to robustly predict in-hospital mortality and long-term hospitalization risk jointly.

Conclusion

This study addresses multimodal clinical prognostic analysis and modality missingness by proposing a text-guided cross-attention predictive framework. Through integrating imaging, medical reports, and physiological indicators with learnable missing data markers, the model achieves leading-edge Mortality and PLOS prediction performance, improving accuracy and robustness. It tackles heterogeneous medical data challenges and supports critical care decision-making with significant clinical and academic value.

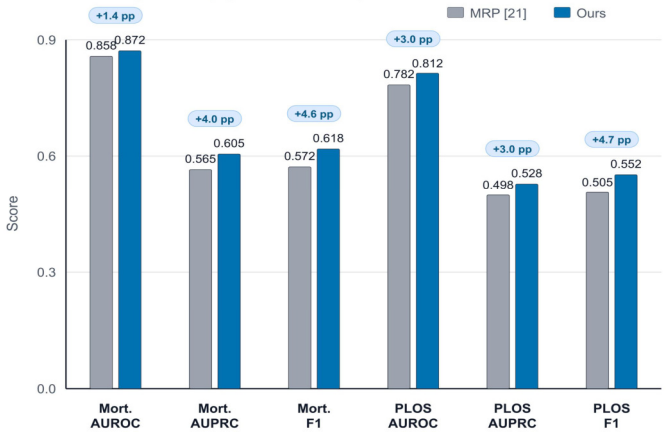
Background & Motivation

- Although accurate prognosis relies on integrating heterogeneous multimodal data, existing models struggle to bridge the nonlinear semantic gap and handle missing modalities, thereby failing to accurately characterize the dynamic physiological states of critically ill patients.
- A decoupled multi-channel encoding framework is proposed, integrating multiple methods for heterogeneous data to achieve joint representation of cross-modal data.
- A text-guided cross-attention mechanism is designed, utilizing clinical text as a semantic query to dynamically integrate visual and tabular features, effectively capturing the potential correlation between expert observations and objective patient data.
- A learnable missing data labeling strategy is introduced to enhance the model's robustness to incomplete data and improve prediction accuracy in real-world clinical scenarios.

Main Results

A Main performance vs strongest baseline

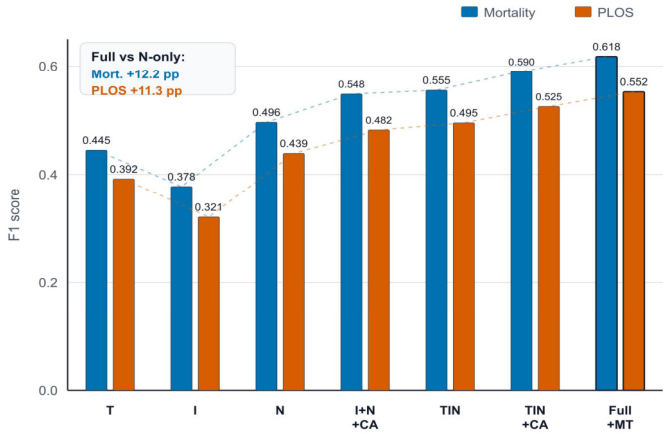
Table 1: Ours against MRP [21] on the three most reported metrics



Compared with the strongest baseline MRP, our model achieves consistent improvements across AUROC, AUPRC, and F1 for both Mortality and PLOS, with the most notable gains observed in F1.

B Ablation: F1 gains from components

Table 2: compact view of component contribution on Mortality and PLOS



T: Tabular, I: Image, N: Notes. CA: Cross-Attention,MT: Learnable Missing Tokens.

The ablation results show that multimodal fusion, text-guided cross-attention, and learnable missing tokens progressively improve predictive performance, with Full+MT achieving the highest F1 scores for both Mortality and PLOS.