Multimodal Deep Learning for Predicting Emergency Department

Diagnostic Discordance Using MIMIC-IV Clinical Text and Time-Series Data

Peter Chika Ozo-ogueji

Department of Math and Statistics, American University

Abstract

Diagnostic errors in the Emergency Department (ED) represent a significant patient safety concern, with missed diagnoses potentially leading to adverse outcomes. This study presents a multimodal deep learning framework for predicting ED diagnostic discordance—cases where the initial working diagnosis differs from the final discharge diagnosis at the ICD family level. Using the MIMIC-IV database, we constructed a cohort of 201,499 ED admissions and developed models integrating clinical text (triage chief complaints and radiology reports) with time-series physiological data (vital signs and laboratory test patterns). Our approach employs a late fusion architecture combining BiLSTM-based text embeddings (256 dimensions) with transformer-encoded time-series features (64 dimensions), connected through a learnable weighted combination. The multimodal fusion model achieved a test AUC of 0.835, significantly outperforming unimodal baselines (text-only: 0.811, time-series-only: 0.750) and the TF-IDF baseline (0.819). Temperature scaling calibration preserved discrimination while improving probability reliability. The learned fusion weight (α = 0.634) indicates that clinical text contributes approximately 63% of the predictive signal, with time-series features providing complementary physiological context. These results demonstrate the value of multimodal integration for clinical decision support and highlight the potential for real-time misdiagnosis risk assessment in emergency care settings.

Keywords: *Emergency Department, Diagnostic Discordance, Multimodal Learning, Clinical NLP, Deep Learning, MIMIC-IV*

1. Introduction

Emergency Departments (EDs) serve as the frontline of healthcare delivery, handling over 130 million visits annually in the United States alone [1]. The high-pressure, time-constrained environment of the ED creates conditions where diagnostic errors can occur, with studies estimating that 5-10% of ED diagnoses contain clinically significant errors [2]. These misdiagnoses can lead to delayed treatment, inappropriate interventions, and in severe cases, preventable mortality. The challenge of accurate ED diagnosis is compounded by the limited information available at the time of initial assessment, the need for rapid decision-making, and the diverse spectrum of presenting complaints.

Recent advances in machine learning and natural language processing have opened new avenues for clinical decision support. Electronic Health Records (EHRs) contain rich multimodal data—including clinical notes, vital signs, laboratory results, and imaging reports —that can be leveraged to identify patients at risk of diagnostic discordance. However, most existing approaches focus on single modalities, potentially missing the complementary information available across different data types.

This study addresses the problem of predicting ED diagnostic discordance using a multimodal fusion approach. We define discordance as cases where the ED working diagnosis misses at least one high-severity condition identified in the final discharge diagnosis at the three-character ICD family level. Our framework integrates clinical text data (triage chief complaints and radiology reports) with time-series physiological measurements (vital signs and laboratory test urgency patterns) through a late fusion architecture with learnable modality weighting.

The primary contributions of this work include: (1) a rigorously constructed cohort from MIMIC-IV with patient-level data splits ensuring no data leakage; (2) a comprehensive text preprocessing pipeline with clinical abbreviation expansion and negation handling; (3) a transformer-based time-series encoder with ablation studies on delta-time and masking strategies; (4) a late fusion architecture achieving state-of-the-art performance for ED misdiagnosis prediction; and (5) calibration analysis ensuring clinically meaningful probability estimates.

2. Related Work

2.1 Diagnostic Error Detection

Diagnostic errors have been recognized as a critical patient safety issue since the landmark IOM report "To Err is Human" [3]. Singh et al. demonstrated that trigger-based approaches using administrative data can identify potential diagnostic errors retrospectively [4]. More recent work has applied machine learning to detect missed diagnoses in specific conditions such as myocardial infarction [5] and stroke [6]. However, these studies typically focus on single conditions rather than general misdiagnosis risk assessment.

2.2 Clinical Natural Language Processing

Clinical NLP has advanced significantly with the introduction of domain-specific language models. ClinicalBERT [7] and BioBERT [8] demonstrated the value of pretraining on medical corpora. More recently, transformer architectures have been applied to various clinical prediction tasks. For ED applications, prior work has used chief complaint text for triage acuity prediction [9] and early warning systems [10]. Our work extends this by combining text with physiological time-series data.

2.3 Multimodal Clinical Learning

Multimodal fusion in healthcare has shown promise for tasks such as mortality prediction [11] and length of stay estimation [12]. Khadanga et al. demonstrated that combining clinical notes with structured data improves ICU outcome prediction [13]. Late fusion approaches, which combine modality-specific representations at the decision level, have proven effective when modalities have different characteristics or missing data patterns [14]. Our work applies late fusion with learnable weights to the ED misdiagnosis prediction problem.

3. Methodology

3.1 Problem Formulation

We formulate ED diagnostic discordance prediction as a binary classification task. For each admission i, we aim to estimate the probability of diagnostic discordance given multimodal input features available at ED discharge time. The prediction target is defined as:

$$\hat{y}_i = f\_\theta(x_i(\leq T\_ED)) \approx y_i$$

(1)

where $x_i(\leq T\_ED)$ represents all features timestamped on or before ED discharge time T_ED for admission i. The binary label $y_i \in \{0,1\}$ is defined as:

$$y_i = \mathbb{1}\{\nexists f \in F\_DC \cap T : f \notin F\_ED\}$$

(2)

where F_ED denotes the set of ICD family codes assigned in the ED, F_DC denotes the discharge diagnosis family codes, and T represents the set of high-severity (life-threatening) diagnosis families. Thus, $y_i = 1$ indicates discordance where the ED missed at least one critical diagnosis.

3.2 Data and Cohort Construction

We utilized the MIMIC-IV database (version 3.1) and MIMIC-IV-ED, comprising de-identified EHR data from Beth Israel Deaconess Medical Center. The cohort was constructed with the following inclusion criteria: (1) valid ED timestamps where intime ≤ outtime; (2) hospital admission via the ED; (3) both ED and discharge diagnoses available; and (4) one-to-one mapping between ED stays and hospital admissions. The final cohort consists of 201,499 ED admissions from 106,985 unique patients, with a discordance rate of 27.8% for high-severity missed diagnoses.

Data splits were performed at the patient level (70% train, 15% validation, 15% test) to prevent information leakage from patients with multiple admissions. This resulted in 141,300 training admissions, 30,273 validation admissions, and 29,923 test admissions.

3.3 Clinical Text Processing

Clinical text data comprised triage chief complaints (425,087 records covering 99.8% of stays) and radiology reports (215,507 records after temporal filtering). The combined text representation for each admission is:

$$X\_text = [TRIAGE: t_1...t_n] \oplus [SEP] \oplus [RADIOLOGY: r_1...r_m]$$

(3)

where ⊕ denotes concatenation, and t_i, r_j are token sequences from triage and radiology notes respectively. The preprocessing pipeline included clinical abbreviation expansion and negation handling with special tokens.

3.4 Time-Series Feature Engineering

Physiological time-series data $X\_ts \in \mathbb{R}$^(T×D) includes vital signs (heart rate, respiratory rate, oxygen saturation, mean arterial pressure) and laboratory test urgency patterns, where T is the sequence length and D = 5 features. Mean arterial pressure (MAP) was computed as:

$$MAP = DBP + ⅓(SBP - DBP)$$

(4)

Time-series windows were constructed using an 8-hour window with 2-hour steps, selected through grid search. For irregularly sampled data, we compute delta-time features δ_t representing time since last observation for each variable.

3.5 Model Architectures

Text Encoder (BiLSTM with Attention): The bidirectional LSTM processes token embeddings E ∈ ℝ^(L×d) where L is sequence length and d = 128. The LSTM hidden states are computed as:

$$f_t = \sigma(Wf \cdot x_t + Uf \cdot h_{t-1} + bf) \ \ [forget\ gate]$$
$$i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i) \ \ [input\ gate]$$
$$\tilde{c}_t = tanh(Wc \cdot x_t + Uc \cdot h_{t-1} + bc) \ \ [cell\ candidate]$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \ \ [cell\ state]$$
$$o_t = \sigma(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o) \ \ [output\ gate]$$
$$h_t = o_t \odot tanh(c_t) \ \ [hidden\ state]$$

(5)

where ⊙ denotes element-wise multiplication, σ is the sigmoid function, and W, U, b are learnable parameters. The bidirectional representation concatenates forward and backward hidden states:

$$h_t = [\overrightarrow{h}_t ; \overleftarrow{h}_t] = BiLSTM(x_t)$$

(6)

Attention pooling computes a weighted sum of hidden states to produce the final 256-dimensional text embedding:

$$\alpha_t = softmax(w^T h_t)$$
$$h\_text = \Sigma_t\ \alpha_t \cdot h_t$$

(7)

Time-Series Encoder (Transformer): The transformer encoder processes vital sign sequences using scaled dot-product self-attention:

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d_k})V$$

(8)

where Q = XW_Q, K = XW_K, V = XW_V are query, key, and value projections, and $\sqrt{d_k}$ is the scaling factor preventing gradient vanishing. A learnable [CLS] token is prepended to capture sequence-level representation h_ts ∈ ℝ^64.

Late Fusion Architecture: Each modality encoder produces a logit through separate classification heads:

$$z\_text = MLP\_text(h\_text) \in \mathbb{R}$$
$$z\_ts = MLP\_ts(h\_ts) \in \mathbb{R}$$

(9)

The fused prediction combines logits through a learnable weight α:

$$z\_fusion = \sigma(\alpha)\cdot z\_text + (1 - \sigma(\alpha))\cdot z\_ts$$

(10)

where σ(α) ensures the weight remains in [0,1]. The final probability is $\hat{y} = \sigma(z\_fusion)$.

3.6 Training Procedure

All models were trained using weighted binary cross-entropy loss to address class imbalance:

$$L = -1/N\ \Sigma_i\ [w_1\cdot y_i\cdot \log(\hat{y}_i) + w_0\cdot(1-y_i)\cdot \log(1-\hat{y}_i)]$$

(11)

where w_1 = N_neg/N_pos = 2.47 compensates for class imbalance. The AdamW optimizer was used with learning rate 10^{-3} and weight decay 10^{-4}. Early stopping with patience of 5 epochs prevented overfitting.

Temperature scaling was applied post-hoc for probability calibration:

$$P\_calibrated = \sigma(z / T)$$

(12)

where T > 0 is optimized on the validation set to minimize negative log-likelihood. The optimal temperature T = 1.106 was found via L-BFGS optimization.

4. Results

4.1 Baseline Comparisons

We established a strong baseline using TF-IDF vectorization (10,000 features, 1-2 grams) with logistic regression. The early fusion variant (combined text + modality indicators) achieved a test AUC of 0.819, demonstrating that the prediction task is feasible with traditional methods.

4.2 Deep Learning Model Performance

Table 1 presents the comprehensive model comparison. The multimodal late fusion model achieved the highest AUC of 0.835, outperforming all unimodal approaches. Evaluation metrics are defined as:

$$Precision = TP / (TP + FP)$$
$$Recall = TP / (TP + FN)$$
$$F_1 = 2\cdot Precision\cdot Recall / (Precision + Recall)$$
$$AUC = P(f(x^+) > f(x^-))$$

(13)

Table 1: Model Performance Comparison on Test Set

Model	AUC	F1	Precision	Recall	Specificity
TF-IDF Baseline	0.819	0.611	0.512	0.756	0.732
BiLSTM (Text)	0.814	0.489	0.703	0.375	0.939

Model	AUC	F1	Precision	Recall	Specificity
Text-Only Fusion	0.811	0.608	0.531	0.710	0.746
Time-Only Fusion	0.750	0.557	0.455	0.716	0.654
Multimodal Fusion	**0.835**	**0.637**	**0.559**	**0.741**	**0.764**

4.3 Modality Contribution Analysis

The learned fusion weight α converged to 0.634 after applying the sigmoid transformation, indicating that clinical text contributes approximately 63% of the predictive signal while time-series features provide 37%. This aligns with clinical intuition—the chief complaint and radiology findings capture diagnostic reasoning directly, while vital signs provide complementary physiological context.

4.4 Calibration Analysis

Calibration quality was assessed using the Expected Calibration Error (ECE):

$$ECE = \Sigma_m (|B_m|/n) \cdot |acc(B_m) - conf(B_m)| \quad (14)$$

where B_m represents samples in bin m, $acc(B_m)$ is accuracy within the bin, and $conf(B_m)$ is average predicted confidence. Temperature scaling with T = 1.106 preserved discrimination (AUC unchanged at 0.835) while improving probability reliability, essential for clinical decision support.

5. Discussion

Our multimodal fusion approach demonstrates that integrating clinical text with physiological time-series data improves ED diagnostic discordance prediction compared to unimodal approaches. The 0.835 AUC achieved is clinically meaningful and comparable to other well-established clinical prediction tasks such as ICU mortality prediction (AUC 0.82-0.87) and clinical deterioration detection (AUC 0.83-0.90).

An unexpected finding was that the simple TF-IDF baseline performed competitively with deep learning text encoders. This suggests that for this task, vocabulary-based features capture much of the diagnostic information present in clinical text. However, the multimodal fusion still outperformed all baselines, indicating that time-series integration provides value beyond text-only approaches.

Several limitations should be acknowledged. First, the study uses data from a single academic medical center, potentially limiting generalizability. Second, the label definition based on ICD code discordance is a proxy for actual diagnostic error—some discordances may reflect diagnostic evolution rather than error. Third, approximately 6% of samples were dropped during multimodal alignment due to missing time-series data.

6. Conclusion

This study presents a multimodal deep learning framework for predicting ED diagnostic discordance that achieves state-of-the-art performance by combining clinical text and time-series physiological data. The late fusion architecture with learnable modality weighting provides interpretable insights into the relative contribution of each data source. Future work should explore external validation across multiple institutions, integration of additional modalities, and prospective evaluation of the model's impact on clinical workflows and patient outcomes.

References

[1] Centers for Disease Control and Prevention. National Hospital Ambulatory Medical Care Survey: 2021 Emergency Department Summary Tables. 2023.

[2] Newman-Toker DE, et al. Serious misdiagnosis-related harms in malpractice claims: The 'Big Three'. Diagnosis. 2019;6(3):227-240.

[3] Kohn LT, Corrigan JM, Donaldson MS. To Err is Human: Building a Safer Health System. National Academies Press. 2000.

[4] Singh H, et al. Types and origins of diagnostic errors in primary care settings. JAMA Internal Medicine. 2013;173(6):418-425.

[5] Than M, et al. Machine learning to predict the likelihood of acute myocardial infarction. Circulation. 2019;140(11):899-909.

[6] Heo J, et al. Machine learning-based model for prediction of outcomes in acute stroke. Stroke. 2019;50(5):1263-1265.

[7] Alsentzer E, et al. Publicly available clinical BERT embeddings. NAACL Clinical NLP Workshop. 2019.

[8] Lee J, et al. BioBERT: a pre-trained biomedical language representation model. Bioinformatics. 2020;36(4):1234-1240.

[9] Fernandes M, et al. Clinical decision support for chief complaint classification. J Med Internet Res. 2020;22(8):e17734.

[10] Kwon JM, et al. An algorithm based on deep learning for predicting in-hospital cardiac arrest. JAMIA. 2018;25(11):1442-1450.

[11] Harutyunyan H, et al. Multitask learning and benchmarking with clinical time series data. Scientific Data. 2019;6:96.

[12] Rajkomar A, et al. Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine. 2018;1:18.

[13] Khadanga S, et al. Using clinical notes with time series data for ICU management. EMNLP Clinical NLP Workshop. 2019.

[14] Baltrusaitis T, et al. Multimodal machine learning: A survey and taxonomy. IEEE TPAMI. 2019;41(2):423-443.