Multimodal Deep Learning for Predicting Emergency Department Diagnostic Discordance Using MIMIC-IV Clinical Text, Time-Series, and Chest X-Ray Imaging Data

Peter Chika Ozo-ogueji

Department of Mathematics 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 multimodal-aligned cohort of 50,967 ED admissions integrating three complementary data sources: clinical text (triage chief complaints and radiology reports), time-series physiological data (vital signs and laboratory test patterns), and chest X-ray imaging features.

Our approach employs a late fusion architecture combining BiLSTM-based text embeddings (256 dimensions) with transformer-encoded time-series features (64 dimensions) and Vision Transformer (ViT)-extracted chest X-ray embeddings (768 dimensions) through learnable softmax-normalized fusion weights. The multimodal fusion model achieved a test AUC of 0.807, matching the best bimodal configuration (Text+Time) while providing improved recall (70.1%) and F1-score (0.528). Comprehensive ablation studies across all seven model configurations (three unimodal, three bimodal, one multimodal) revealed that text contributes 61.4% of the predictive signal, with time-series features providing 21.5% and imaging 17.1%. Isotonic regression calibration achieved an Expected Calibration Error (ECE) of 1.89%, substantially improving probability reliability for clinical deployment. These results demonstrate the value of multimodal integration for clinical decision support and highlight both the potential and limitations of incorporating imaging data for real-time misdiagnosis risk assessment in emergency care settings.

Keywords: Emergency Department, Diagnostic Discordance, Multimodal Learning, Multimodal Fusion, Clinical NLP, Deep Learning, Medical Imaging, 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 multimodal machine learning have demonstrated significant potential for healthcare applications [3,4]. Electronic Health Records (EHRs) contain rich multimodal data—including clinical notes, vital signs, laboratory results, and imaging studies—that can be leveraged to identify patients at risk of diagnostic discordance. However, most existing approaches focus on single modalities or bimodal combinations, potentially missing the complementary information available across different data types. Recent work on multimodal fusion in healthcare, such as the MedPatch framework [5] and reviews of multimodal machine learning approaches [6], have highlighted the importance of integrating diverse clinical data sources for improved predictive performance.

This study addresses the problem of predicting ED diagnostic discordance using a multimodal fusion approach. We extend prior bimodal work by incorporating chest X-ray imaging features alongside clinical text and time-series physiological data. 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: (1) clinical text data (triage chief complaints and radiology reports) processed through BiLSTM encoders; (2) time-series physiological measurements (vital signs and laboratory test urgency patterns) processed through transformer encoders; and (3) chest X-ray imaging features extracted via Vision Transformer (ViT) and DINOv2 self-supervised learning through a late fusion architecture with learnable modality weighting.

The primary contributions of this work include: (1) a rigorously constructed multimodal-aligned cohort from MIMIC-IV with patient-level data splits ensuring no data leakage across 50,967 admissions; (2) comprehensive ablation studies across all seven model configurations (three unimodal, three bimodal, one multimodal) enabling systematic evaluation of modality contributions; (3) a late fusion architecture with softmax-normalized learnable weights achieving interpretable modality importance quantification; (4) calibration analysis using multiple techniques (temperature scaling, isotonic regression, Platt scaling) achieving ECE < 2%; and (5) analysis of the limiting factors when incorporating imaging data into ED prediction models.

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" [7]. Singh et al. demonstrated that trigger-based approaches using administrative data can identify potential diagnostic errors retrospectively [8]. More recent work has applied machine learning to detect missed diagnoses in specific conditions such as myocardial infarction [9] and stroke [10]. 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 [11] and BioBERT [12] 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 [13] and early warning systems [14]. Our work extends this by combining text with both physiological time-series data and medical imaging.

2.3 Multimodal Clinical Learning

Multimodal fusion in healthcare has shown promise for tasks such as mortality prediction [15] and length of stay estimation [16]. Khadanga et al. demonstrated that combining clinical notes with structured data improves ICU outcome prediction [17]. Recent reviews have comprehensively examined multimodal machine learning approaches in healthcare, emphasizing the importance of integrating imaging data, text, time series, and tabular data for improved clinical decision-making [6]. Late fusion approaches, which combine modality-specific representations at the decision level, have proven effective when modalities have different characteristics or missing data patterns [18].

2.4 Medical Imaging in Clinical Prediction

Vision Transformers (ViT) and self-supervised learning approaches like DINOv2 have revolutionized medical image analysis [19]. Recent work has demonstrated the value of combining chest X-ray features with clinical data for ICU predictions [5,20]. The MedPatch framework specifically addresses confidence-guided multi-stage fusion for combining clinical time-series, chest X-ray images, and clinical notes [5]. Our work applies similar multimodal integration to the ED misdiagnosis prediction problem, extending beyond ICU-focused applications.

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:

$$\hat{y}_i = f\_\theta(x\_text, x\_time, x\_image) \approx y_i$$

where the binary label $y_i \in \{0,1\}$ indicates whether the ED missed at least one critical diagnosis. The label is defined as $y_i = 1$ when there exists a high-severity ICD family code in the discharge diagnosis that was not present in the ED working diagnosis.

3.2 Data and Cohort Construction

We utilized the MIMIC-IV database (version 3.1), MIMIC-IV-ED, and MIMIC-CXR for chest X-ray imaging data. The cohort construction followed a hierarchical alignment process:

Text and Time-Series Alignment: The initial bimodal cohort consisted of 189,159 ED admissions with both text and time-series data available, derived from 201,499 total admissions after applying inclusion criteria.

Multimodal Alignment: Matching with chest X-ray imaging data via hospital admission ID (hadm_id) reduced the cohort to 50,967 admissions (26.9% retention rate). This significant reduction reflects the clinical reality that not all ED patients receive chest X-rays during their visit.

Data splits were performed at the patient level (70% train, 15% validation, 15% test) to prevent information leakage. The final multimodal cohort consists of 35,640 training, 7,606 validation, and 7,721 test admissions with a positive class rate of 19.8% (class imbalance ratio 4.06:1).

3.3 Modality-Specific Processing

Clinical Text Processing: Clinical text data comprised triage chief complaints and radiology reports. The preprocessing pipeline included clinical abbreviation expansion and negation handling. A BiLSTM encoder with attention pooling produced 256-dimensional text embeddings for each admission.

Time-Series Feature Engineering: Physiological time-series data includes vital signs (heart rate, respiratory rate, oxygen saturation, mean arterial pressure) and laboratory test urgency patterns. Time-series windows were constructed using an 8-hour window with 2-hour steps. A transformer encoder with learnable [CLS] token produced 64-dimensional time-series features.

Chest X-Ray Image Processing: Chest X-ray images were processed using a Vision Transformer (ViT) backbone with DINOv2 self-supervised pretraining. Pooled image features (768 dimensions) were extracted, along with biomarker-specific features (16 dimensions) for radiological findings. The combined image representation totals 784 dimensions, though our fusion model uses the primary 768-dimensional pooled features.

3.4 Late Fusion Architecture

Each modality encoder feeds into a dedicated classification head consisting of two fully-connected layers with batch normalization, ReLU activation, and dropout (p=0.3):

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

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

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

The fused prediction combines logits through learnable weights with softmax normalization:

$$w = softmax([w\_T, w\_Ti, w\_I] / \tau)$$

$$z\_fusion = w\_T \cdot z\_text + w\_Ti \cdot z\_time + w\_I \cdot z\_image$$

where τ is a learnable temperature parameter. The softmax normalization ensures weights sum to 1 and provides interpretable modality contributions. Final probability is $\hat{y} = \sigma(z\_fusion)$. During training, modality dropout (p=0.3) is applied to improve robustness to missing modalities.

3.5 Training Procedure

All models were trained using weighted binary cross-entropy loss with pos_weight = 4.06 to address 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. Batch size was 128 with training for up to 20 epochs.

For probability calibration, we evaluated three post-hoc methods: (1) Temperature scaling optimizing T on the validation set via L-BFGS; (2) Isotonic regression fitting a non-parametric monotonic function; (3) Platt scaling fitting a logistic regression model.

4. Results

4.1 Comprehensive Model Comparison

Table 1 presents the performance comparison across all seven model configurations. The multimodal fusion model achieved the highest F1-score (0.528) and tied with Text+Time for the highest AUC (0.807).

Table 1: Model Performance Comparison on Test Set (n=7,721)

Model	AUC	F1	Precision	Recall	Specificity	Accuracy	Brier
Text-Only	0.798	0.522	0.437	0.648	0.794	0.765	0.169
Time-Only	0.670	0.395	0.265	0.775	0.470	0.530	0.256
Image-Only	0.635	0.375	0.257	0.692	0.507	0.544	0.242
Text+Time	0.807*	0.522	0.431	0.664	0.784	0.760	0.165
Text+Image	0.798	0.524	0.434	0.663	0.787	0.762	0.168
Time+Image	0.698	0.416	0.286	0.762	0.531	0.577	0.241
Multimodal	0.807*	0.528*	0.424	0.701*	0.765	0.752	0.172

**Best performance in column. Multimodal achieves best F1 and recall while matching best AUC.*

4.2 Modality Contribution Analysis

The learned fusion weights after softmax normalization reveal the relative importance of each modality:

Text: 61.4% | Time-Series: 21.5% | Image: 17.1%

Text dominates the prediction, consistent with clinical intuition—chief complaints and radiology reports contain direct diagnostic reasoning. Time-series features provide complementary physiological context, while imaging contributes the smallest but still meaningful signal. The relatively lower contribution of imaging may reflect: (1) the multimodal subset represents patients who received chest X-rays, potentially a specific clinical subpopulation; (2) imaging information may already be partially captured in radiology reports; (3) chest X-ray findings may have less direct relevance to diagnostic discordance compared to clinical reasoning.

4.3 Calibration Analysis

Table 2 presents calibration performance across different post-hoc calibration methods.

Table 2: Calibration Method Comparison

Method	AUC	ECE (%)	Brier Score
Uncalibrated	0.807	18.07	0.172
Temperature Scaling (T=1.088)	0.807	18.57	0.172
Isotonic Regression	0.806	1.89*	0.125*
Platt Scaling	0.807	2.17	0.125

**Best calibration performance. Isotonic regression achieves ECE < 2%.*

Isotonic regression achieved the best calibration with ECE of 1.89% and Brier score of 0.125, substantially improving probability reliability while maintaining discrimination. This is essential for clinical deployment where well-calibrated probabilities enable appropriate threshold selection for different clinical scenarios.

5. Discussion

Our multimodal fusion approach demonstrates that integrating clinical text, physiological time-series data, and chest X-ray imaging features provides the best balance of discrimination and recall for ED diagnostic discordance prediction. The multimodal model achieves the highest F1-score (0.528) and recall (70.1%) among all configurations, identifying more true misdiagnosis cases—a critical consideration for patient safety.

An important finding is that the multimodal model matches but does not substantially exceed the Text+Time bimodal configuration in AUC (both 0.807). This may be attributed to several factors consistent with recent multimodal learning literature [6,18]:

Modality Overlap: Radiology reports already encode much of the diagnostic information present in chest X-ray images, creating redundancy between text and imaging modalities.

Population Selection Bias: The requirement for chest X-ray availability reduces the cohort from 189K to 51K samples (27% retention). Patients receiving chest X-rays may represent a specific clinical subpopulation, potentially limiting generalizability.

Task Specificity: Chest X-ray findings may have limited direct relevance to diagnostic discordance at the ICD family level, particularly for conditions not primarily diagnosed through chest imaging.

The learned fusion weights (Text: 61.4%, Time: 21.5%, Image: 17.1%) provide interpretable insights consistent with clinical practice. Text contributes most of the predictive signal because clinical notes capture diagnostic reasoning directly. Time-series features provide complementary physiological context that may not be explicitly documented. Imaging contributes meaningfully but less than other modalities, suggesting that its unique contribution beyond text is limited for this task.

Our calibration analysis revealed that traditional temperature scaling, effective for many classification tasks, provided minimal improvement for this problem. Isotonic regression and Platt scaling achieved substantial ECE reduction (<2.2%), consistent with recent recommendations for clinical prediction models [21]. Well-calibrated probabilities enable appropriate threshold selection: lower thresholds for high-sensitivity screening (catching more potential misdiagnoses) versus higher thresholds for high-specificity flagging (reducing alert fatigue).

Several limitations should be acknowledged. First, the study uses data from a single academic medical center (Beth Israel Deaconess), 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, the 73% sample loss during multimodal alignment may introduce selection bias. Fourth, the ViT/DINOv2 image features were pre-extracted rather than end-to-end trained, which may limit adaptation to the specific prediction task.

6. Conclusion

This study presents a multimodal deep learning framework for predicting ED diagnostic discordance that integrates clinical text, time-series physiological data, and chest X-ray imaging features. Our comprehensive evaluation across all seven model configurations reveals that multimodal fusion achieves the best F1-score and recall while matching the best AUC, demonstrating value for patient safety applications where identifying potential misdiagnoses is paramount.

The learned fusion weights provide interpretable insights into modality importance: text dominates (61.4%), followed by time-series (21.5%) and imaging (17.1%). While imaging's contribution is relatively modest, the multimodal model's improved recall suggests it captures complementary information not present in other modalities. Isotonic regression calibration achieves ECE < 2%, ensuring clinically meaningful probability estimates.

Future work should explore: (1) external validation across multiple institutions; (2) end-to-end training of image encoders for task-specific adaptation; (3) attention mechanisms for interpretable cross-modal interactions; (4) integration of additional modalities such as structured EHR data; and (5) prospective evaluation of the model's impact on clinical workflows and patient outcomes. Our results highlight both the promise and current limitations of multimodal fusion for ED clinical decision support.

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] Acosta JN, et al. Multimodal biomedical AI. Nature Medicine. 2022;28:1773-1784.

[4] Huang SC, et al. Fusion of medical imaging and electronic health records using deep learning. Nature Medicine. 2020;26(8):1181-1189.

[5] AlSaad R, et al. MedPatch: Confidence-Guided Multi-Stage Fusion for Multimodal Clinical Data. arXiv:2508.09182. 2025.

[6] Tayebi Arasteh S, et al. Review of multimodal machine learning approaches in healthcare. Information Fusion. 2024;102:102690.

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

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

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

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

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

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

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

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

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

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

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

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

[19] Chen RJ, et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. CVPR. 2022.

[20] Shamout FE, et al. An artificial intelligence system for predicting the deterioration of COVID-19 patients. NPJ Digital Medicine. 2021;4:80.

[21] Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. ICML. 2005.