
Early GVHD Prediction in Liver Transplantation via Multi-Modal Deep Learning on Imbalanced EHR Data

Yushan Jiang, MS^{1,3}, Shuteng Niu, PhD¹, Dongjin Song, PhD³, Yichen Wang, MB, MS², Jingna Feng, MS¹, Xinyue Hu, MS¹, Liu Yang, MBBS², and Cui Tao, PhD¹

¹Department of Artificial Intelligence and Informatics, Mayo Clinic, Jacksonville, FL 32224, United States,

²Gastroenterology and Hepatology Transplant Center, Mayo Clinic, Jacksonville, FL 32224, United

States, ³School of Computing, University of Connecticut, Storrs, CT 06269, United States

ABSTRACT

Objective: Graft-versus-host disease (GVHD) is a rare but often fatal complication in liver transplantation, with a very high mortality rate. By harnessing multi-modal deep learning methods to integrate heterogeneous and imbalanced electronic health records (EHR), we aim to advance early prediction of GVHD, paving the way for timely intervention and improved patient outcomes.

Materials and Methods: In this study, we analyzed pre-transplant electronic health records (EHR) spanning the period before surgery for 2,100 liver transplantation patients, including 42 cases of graft-versus-host disease (GVHD), from a cohort treated at Mayo Clinic between 1992 and 2025. The dataset comprised four major modalities: patient demographics, laboratory tests, diagnoses, and medications. We developed a multi-modal deep learning framework that dynamically fuses these modalities, handles irregular records with missing values, and addresses extreme class imbalance through AUC-based optimization.

Results: The developed framework outperforms all single-modal and multi-modal machine learning baselines, achieving an AUC of 0.836, an AUPRC of 0.157, a recall of 0.768, and a specificity of 0.803. It also demonstrates the effectiveness of our approach in capturing complementary information from different modalities, leading to improved performance.

Discussion: Our multi-modal deep learning framework substantially improves existing approaches for early GVHD prediction. By effectively addressing the challenges of heterogeneity and extreme class imbalance in real-world EHR, it achieves accurate early prediction.

Conclusion: Our proposed multi-modal deep learning method demonstrates promising results for early prediction of a GVHD in liver transplantation, despite the challenge of extremely imbalanced EHR data.

Key words: multi-modal model, deep learning, GVHD, early risk prediction, electronic health records

INTRODUCTION

Graft-versus-host disease (GVHD) in liver transplantation (LT) is a rare yet fatal immune complication, with a reported incidence of approximately 0.5–2% and mortality rates exceeding 75%^{1–3}. It arises when donor-derived immune cells mount an attack against recipient tissues, triggering systemic inflammation and multi-organ dysfunction. Early detection of GVHD in liver transplantation remains particularly challenging. Its extreme rarity limits large-scale studies and systematic evaluation of risk factors. In addition, patients often present with nonspecific symptoms such as rash and diarrhea, which are easily mis-attributed to infections, drug toxicity, or acute rejection⁴. The absence

of reliable clinical biomarkers or predictive models for pre-transplant risk assessment further hinders early diagnosis. Consequently, GVHD is often recognized only after it has progressed to advanced stages, and such delayed diagnosis contributes to the persistently high mortality, underscoring the urgent need for accurate and timely prediction methods.

Recent advances in electronic health records (EHRs) provide a promising avenue to address this gap. Liver transplant candidates undergo extensive preoperative evaluation, generating rich multi-modal data encompassing demographics, laboratory tests, diagnosis, and medication histories. The collected records enable data-driven clinical discovery beyond traditional clinical assessment⁵. In particular, deep learning methods have demonstrated strong capabilities across diverse clinical prediction tasks on existing EHR benchmarks^{6–8}, including mortality pre-

diction⁹, readmission prediction¹⁰, adverse event detection^{11,12}, and treatment recommendation^{13–15}. By uncovering complex temporal patterns in patient trajectories, deep learning establishes a foundation for predictive models that may enable earlier identification of GVHD, which has been demonstrated in hematopoietic stem cell transplantations^{16–18}.

However, real-world EHRs are highly heterogeneous, irregularly sampled, and plagued by extensive missingness. In addition, the extreme class imbalance of GVHD outcomes in liver transplant patients makes it particularly difficult to learn robust and discriminative patterns for prediction. These challenges hinder most prior methods developed on relatively organized benchmark datasets, leaving early GVHD prediction in liver transplantation largely underexplored within clinical informatics. To the best of our knowledge, the only related work has focused on post-transplant settings and relied on specialized and donor-recipient information¹⁹, which is not readily available. To date, no study has investigated pre-transplant EHR for early GVHD prediction.

To this end, we propose a multi-modal deep learning framework for early GVHD prediction in liver transplantation based solely on pre-transplant EHR data. Our study uses a cohort of 2,100 liver transplant recipients treated at Mayo Clinic between 1992 and 2025, among whom 42 developed GVHD. We integrate four major modalities (*i.e.*, demographics, laboratory tests, diagnoses, and medications) into a unified predictive model. Our main contributions are summarized as follows:

- **Multi-modal framework:** We present the first pre-transplant GVHD prediction framework in liver transplantation, capturing predictive signals across heterogeneous modalities.
- **Robust modeling:** We design modality-specific encoders to handle irregular intervals and substantial missingness, apply cross-modal fusion to extract complementary predictive signals, and incorporate AUC-based optimization to ensure robust learning under extreme class imbalance.
- **Comprehensive validation:** On the Mayo Clinic cohort, our method achieves promising predictive performance (AUC = 0.836, AUPRC = 0.157, recall = 0.768, specificity = 0.803), consistently outperforming both traditional machine learning and state-of-the-art deep learning baselines. Ablation studies further validate the effectiveness of our proposed components, highlighting the benefits of handling irregular intervals and missingness as well as leveraging complementary information across modalities.

Collectively, these results demonstrate that multi-modal heterogeneous EHR modeling with imbalance-aware optimization enables earlier GVHD identification, supporting

timely intervention and ultimately improving patient outcomes in liver transplantation.

MATERIALS AND METHODS

Primary Cohort for Multi-modal GVHD Prediction

We curated a cohort of liver transplantation (LT) recipients treated at the Mayo Clinic between 1992 and 2025 to enable multi-modal prediction of GVHD²⁰. All patients who underwent LT during this period were initially eligible. To ensure that all utilized data reflected adult clinical profiles, we excluded patients who were younger than 18 years at any point within the three-year pre-transplant observation window. This criterion is equivalent to excluding individuals younger than 21 years at the time of transplantation. After applying the age filter, the cohort comprised 42 GVHD cases and 8,350 controls. To balance computational feasibility while preserving all positive cases, we retained all 42 GVHD cases and randomly selected 2,058 controls, yielding a final analytic cohort of 2,100 adult LT recipients, of whom 42 (2.0%) developed GVHD post-transplant. GVHD outcomes were identified from structured diagnostic codes and subsequently validated through manual chart review by transplant hepatologists to ensure clinical accuracy. The primary prediction target was the occurrence of GVHD following LT, formulated as a binary classification task characterized by severe class imbalance.

All available pre-transplant records were extracted and processed into structured formats as model inputs for each patient. Four major data modalities were used: demographic information, laboratory test results, diagnosis data, and medication administration records.

Demographics. This modality includes 4 features: gender, race, ethnicity, and age. The median age at transplantation was 59 years (range: 22–76), with a mean of 56.7 years. Overall, 65.1% of patients were male and 34.9% female. Most patients identified as White (87.2%), followed by other race (4.4%), Asian (3.6%), Black (3.5%), and unknown (1.4%). In terms of ethnicity, 84.0% were non-Hispanic, 14.0% Hispanic, and 1.9% were unknown or declined to answer.

Diagnosis. This modality includes 20 symptom-related diagnoses extracted from structured records up to the three years preceding transplantation. Diagnoses were aggregated every three months as a union of visits, yielding a temporal sequence of diagnosis representations. The included diagnoses span hepatic, respiratory, metabolic, infectious, and treatment-related conditions relevant to transplant risk.

Drug. This modality captures the daily administration of 280 commonly used inpatient drugs during the 24 observations preceding liver transplantation, aggregated at a 1-day interval. For each day, the feature represents the

count of administered doses for each drug. The drugs span a wide range of classes, including antibiotics and antimicrobials (*e.g.*, meropenem, vancomycin), electrolyte management agents (*e.g.*, sodium bicarbonate, potassium chloride), analgesics and sedatives (*e.g.*, fentanyl, morphine, midazolam), endocrine/metabolic therapies (*e.g.*, various forms of insulin), immunosuppressants and corticosteroids (*e.g.*, tacrolimus, prednisone), and liver-specific agents (*e.g.*, lactulose, rifaximin). The remaining drugs encompass a heterogeneous group of pharmacologic categories.

Lab Test. This modality includes 74 laboratory test variables, extracted using a fixed 1-hour context window across the 24 most recent observations prior to transplantation. Each observation captures the last available test result within the corresponding window. The selected tests span diverse clinical categories, including liver function (*e.g.*, AST, ALT, bilirubin), renal function (*e.g.*, BUN, creatinine), coagulation (*e.g.*, INR, APTT, fibrinogen), hematology (*e.g.*, hemoglobin, platelets, WBC), and electrolytes/metabolites (*e.g.*, sodium, potassium, bicarbonate, calcium). The overall missing ratio was 72.9%, reflecting the inherent sparsity of laboratory data due to on-demand and irregular testing patterns in real-world clinical practice. Note that all three temporal modalities were constructed from records spanning one or multiple clinical visits. As such, the time steps are not necessarily consecutive and may reflect irregular intervals between observations.

The study protocol was approved by the Mayo Clinic Institutional Review Board (IRB), and all data access and analyses were conducted in accordance with institutional guidelines, HIPAA regulations, and applicable ethical standards.

Table 1. Summary statistics across all modalities.

Modality	#Features	#Steps	Interval
Demographics	4	–	–
Lab Test	74	24	1-hour
Diagnosis	20	12	3-month
Drug	280	24	1-day

Multi-modal Deep Learning Framework for GVHD Prediction in Liver Transplantation

Notation. First, we introduce the notations for the model inputs of each modality. We denote the whole EHR input as: $\mathbf{X} = \{\mathbf{X}^{\text{demo}}, \mathbf{X}^{\text{lab}}, \mathbf{X}^{\text{dx}}, \mathbf{X}^{\text{drug}}\}$, where $\mathbf{X}^{\text{demo}} \in \mathbb{R}^{F_{\text{de}}}$, $\mathbf{X}^{\text{lab}} \in \mathbb{R}^{T_{\text{lab}} \times F_{\text{lab}}}$, $\mathbf{X}^{\text{dx}} \in \mathbb{R}^{T_{\text{dx}} \times F_{\text{dx}}}$, $\mathbf{X}^{\text{drug}} \in \mathbb{R}^{T_{\text{dr}} \times F_{\text{dr}}}$, $T_{(\cdot)}$ denotes the number of time steps and $F_{(\cdot)}$ denotes the feature dimension.

Each temporal modality is accompanied by a global time index $g^{(\cdot)} \in [0, 1]^{T_{(\cdot)}}$, which represents $T_{(\cdot)}$ irregular steps across one or multiple clinical visits within the full look-back

period for each patient. Due to the on-demand nature of laboratory testing, a binary mask $\mathbf{M} \in \mathbb{R}^{T_{\text{lab}} \times F_{\text{lab}}}$ is also provided, where $M_{i,j} = 1$ indicates that the j -th lab test at time step i is observed, and 0 otherwise.

Modality Encoding. Next, we introduce the encoding component for each modality, as shown in Figure 1. Each encoder maps the raw input $\mathbf{X}^{(\cdot)}$ into a compact latent representation. For temporal modalities, the encoder further accounts for irregular time steps and missing patterns.

Demographics Encoder. For static demographic features $\mathbf{X}^{\text{demo}}$, we apply a two-layer feed-forward network to embed these attributes into a latent representation with dimensionality d_{demo} :

$$\mathbf{H}^{\text{demo}} = \text{FFN}(\mathbf{X}^{\text{demo}}) \in \mathbb{R}^{d_{\text{demo}}}$$

which provides static contextual information that anchors the integration with dynamic temporal signals.

Irregular Diagnosis Encoder. Given the sparse and irregular nature of clinical encounters, each diagnosis sequence is modeled using a Gated Recurrent Unit (GRU)²¹ to iteratively capture dependencies among diagnostic events. To account for non-uniform intervals within and across visits, the global time index $g^{\text{dx}} \in [0, 1]^{N_{\text{dx}}}$ is concatenated with the input feature vector at each step, thereby informing the recurrent process:

$$\mathbf{H}^{\text{dx}} = \text{GRU}(\mathbf{X}^{\text{dx}} \| g^{\text{dx}}) \in \mathbb{R}^{d_{\text{dx}}}$$

where $\|$ denotes the concatenation operator across feature dimension, d_{dx} denotes the dimensionality of obtained representations.

Irregular Missing-aware Lab Test Encoder. We first construct the lab-specific temporal embedding $\mathbf{P}$ as a trigonometric function of the global time index g^{lab} , modulated by learnable frequency components, which amortizes the effect of missing values in individual irregular lab test records.

$$\mathbf{P} = \mathbf{A} \Psi_{\sin}^{\top}(g^{\text{lab}}) + \mathbf{B} \Psi_{\cos}^{\top}(g^{\text{lab}}) \in \mathbb{R}^{T_{\text{lab}} \times F_{\text{lab}}},$$

where $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{T_{\text{lab}} \times K}$ are learnable coefficient matrices, and $\Psi_{\sin}, \Psi_{\cos} \in \mathbb{R}^{K \times F_{\text{lab}}}$ are Fourier basis functions applied to global time index g^{lab} , with $K \ll T_{\text{lab}}$ components.

To enrich the representations, we perform a masked dimension extension: the observed lab value at each time step is fused with the temporal embedding and lifted into a higher-dimensional space via a generic two-layer feed-forward network, while the missing values are encoded solely by the temporal embedding:

$$h_{i,j}^{\text{lab},0} = \begin{cases} \text{FFN}(\mathbf{x}_{i,j}^{\text{lab}}, \mathbf{p}_{i,j}), & M_{i,j} = 1 \\ \text{FFN}(\mathbf{p}_{i,j}), & M_{i,j} = 0 \end{cases}$$

This design enables missing lab entries to be imputed through learnable time-informed embeddings, while simultaneously augmenting observed values with temporal context. The fused representations $\mathbf{H}^{\text{lab},0}$ augmented with

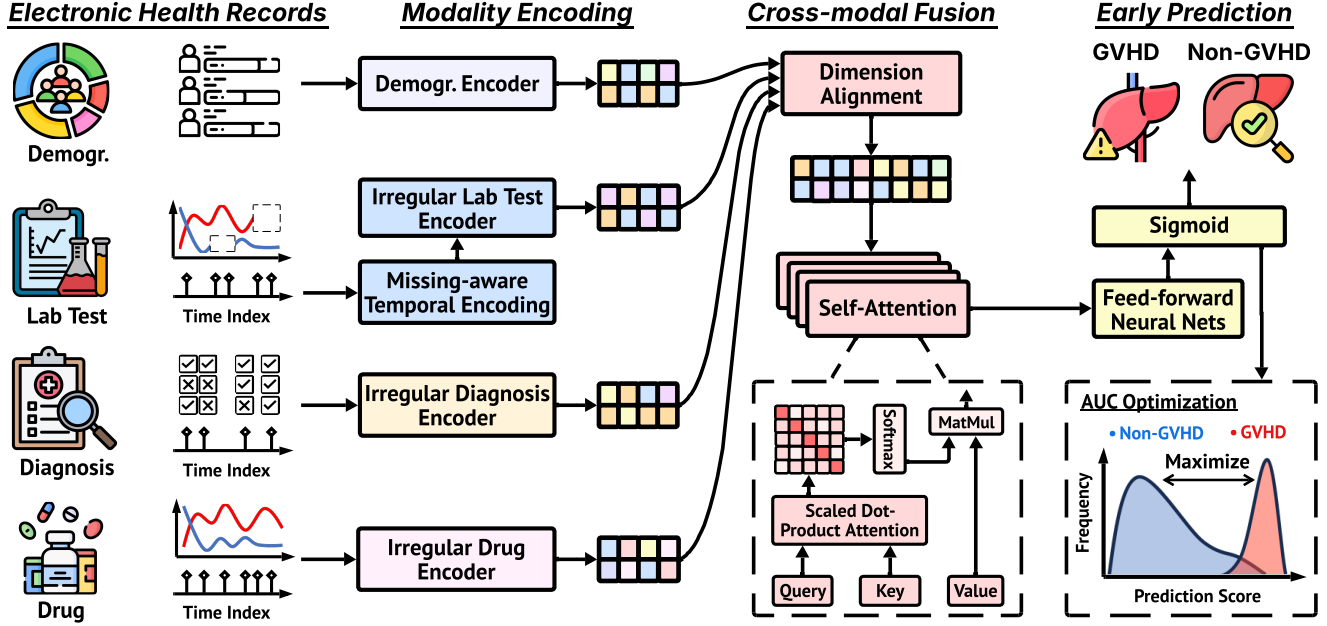


Figure 1. Early prediction of Graft-versus-host disease (GVHD) in liver transplantation using multi-modal EHRs.

g^{lab} are then encoded by a GRU to capture the temporal evolution of lab test trajectories.

$$\mathbf{H}^{\text{lab}} = \text{GRU}(\mathbf{H}^{\text{lab},0} \| g^{\text{lab}}) \in \mathbb{R}^{d_{\text{lab}}}$$

Irregular Drug Encoder. Given the high-dimensional and irregular daily drug records, we concatenate the input with global time index to enrich temporal awareness. We apply two-dimensional convolutional neural networks whose kernels span the entire feature dimension, while sliding along the temporal axis. This design captures local temporal dependencies together with global co-usage patterns across medications. Global average pooling is subsequently applied to obtain a compact representation:

$$\mathbf{H}^{\text{drug}} = \text{Conv2D}(\mathbf{X}^{\text{drug}} \| g^{\text{drug}}) \in \mathbb{R}^{d_{\text{drug}}}$$

Empirically, this design outperforms GRU-based alternatives on the drug modality, as convolutional filters better exploit the high-dimensional co-occurrence structure while alleviating sparsity.

Cross-modal Fusion for GVHD Prediction. With modality-specific representations $\mathbf{H}^{\text{demo}} \in \mathbb{R}^{d_{\text{demo}}}$, $\mathbf{H}^{\text{dx}} \in \mathbb{R}^{d_{\text{dx}}}$, $\mathbf{H}^{\text{lab}} \in \mathbb{R}^{d_{\text{lab}}}$, and $\mathbf{H}^{\text{drug}} \in \mathbb{R}^{d_{\text{drug}}}$, we first perform an optional per-modality dimension alignment via MLP to obtain a common embedding size d :

$$\tilde{\mathbf{H}}^m = \Phi_m(\mathbf{H}^m) \in \mathbb{R}^d, \quad m \in \{\text{demo}, \text{lab}, \text{dx}, \text{drug}\}$$

After alignment, these representations are stacked into a unified collection of modality vectors.

$$\mathbf{Z} = [\tilde{\mathbf{H}}^{\text{demo}} \| \tilde{\mathbf{H}}^{\text{lab}} \| \tilde{\mathbf{H}}^{\text{dx}} \| \tilde{\mathbf{H}}^{\text{drug}}] \in \mathbb{R}^{4 \times d}$$

To contextualize salient cross-modal patterns for GVHD prediction, we project $\mathbf{Z}$ into queries, keys, and values, and apply a multi-head scaled dot-product attention with residual connections and normalization. This design captures cross-modal interactions²², uncovering predictive signals for GVHD from the interplay of lab trajectories, medications, and diagnostic context.

$$\mathbf{Z}' = \text{LayerNorm}(\mathbf{Z} + \text{MultiHeadAttention}(\mathbf{Z})).$$

Finally, the fused representation $\mathbf{Z}'$ is flattened, fed into a two-layer feed-forward neural networks and activated by a sigmoid function:

$$\hat{y} = \text{Sigmoid}(\text{FFN}(\mathbf{Z}'))$$

where $\hat{y} \in [0, 1]$ denotes the prediction score of GVHD.

Learning Objective. Given the extreme imbalance of GVHD outcomes, we adopt a direct AUC optimization framework instead of conventional cross-entropy. Specifically, we employ the Pairwise AUC Margin Loss^{23,24}, which encourages higher prediction scores for positive (GVHD) patients compared to negative ones. Formally, for a positive instance i and negative instance j , the objective is:

$$\mathcal{L}_{\text{AUC}} = \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{N}} \log(1 + \exp(-(\hat{y}_i - \hat{y}_j)))$$

where $\mathcal{P}$ and $\mathcal{N}$ denote positive and negative sets, respectively, $\exp(\cdot)$ is the exponential function e^x , and $\log(\cdot)$ is the natural logarithm.

To improve efficiency under severe imbalance, we use a dual sampler that adaptively pairs each positive case with

Table 2. Performance Evaluation of the proposed method and baselines.

Method ↓	Graft-versus-host disease (GVHD)			
	AUC	AUPRC	Recall	Specificity
Logistic Reg.	0.563 ± 0.003	0.053 ± 0.008	0.358 ± 0.060	0.808 ± 0.056
SVC	0.723 ± 0.010	0.055 ± 0.003	0.719 ± 0.029	0.627 ± 0.022
Random Forest	0.633 ± 0.006	0.048 ± 0.008	0.374 ± 0.051	0.789 ± 0.039
AdaBoost	0.617 ± 0.030	0.039 ± 0.009	0.615 ± 0.149	0.605 ± 0.090
Gradient Boost	0.513 ± 0.018	0.031 ± 0.003	0.162 ± 0.128	0.814 ± 0.084
XGBoost	0.660 ± 0.013	0.054 ± 0.006	0.657 ± 0.020	0.646 ± 0.004
RAIM	0.710 ± 0.022	0.135 ± 0.014	0.644 ± 0.066	0.688 ± 0.053
F-LSTM	0.738 ± 0.012	0.107 ± 0.015	0.676 ± 0.067	0.703 ± 0.060
F-CNN	0.764 ± 0.012	0.147 ± 0.008	0.737 ± 0.015	0.651 ± 0.011
F-Transformer	0.718 ± 0.009	0.072 ± 0.009	0.732 ± 0.087	0.606 ± 0.073
Dipole	0.725 ± 0.006	0.120 ± 0.021	0.715 ± 0.040	0.641 ± 0.057
Ours	0.836 ± 0.007	0.157 ± 0.014	0.768 ± 0.061	0.803 ± 0.053

a dynamically selected subset of negatives, thereby avoiding exhaustive pairwise comparisons while preserving informative gradients. Optimization is carried out with the Proximal Epoch Stochastic Gradient (PESG) optimizer²³, which integrates stochastic gradient descent with dual variable updates for provable convergence and stable AUC maximization. Importantly, this AUC-oriented objective emphasizes correct ranking of high-risk GVHD patients against the majority of non-GVHD cases, aligning directly with the clinical need for early identification of rare but life-threatening complications. This prioritization ensures the model remains sensitive to early warning signals, even when positive cases are extremely scarce.

Baselines, Implementation, and Experiment Setup

We first introduce the details for replication of multi-modal baseline methods. For traditional baselines, we construct multi-modal patient-level features by aggregating across all available modalities. Demographics are directly included as static covariates. For diagnoses, we apply a simple binary indicator, marking whether a patient has any diagnosis codes across the observation window. Laboratory results are first filled using mean imputation across the cohort. From the imputed values, we compute mean, minimum, and maximum for each lab test across time, and concatenate these as fixed-length features. For drugs, we collapse drug exposures by summing over time, yielding total counts per drug. Finally, the selected modality-specific features are concatenated into a single vector. This design reduces heterogeneous temporal EHR data into standardized, patient-level inputs that can be directly consumed by traditional machine learning models (e.g., logistic regression, random forest, boosting methods). We implemented deep learning baselines including RAIM²⁵, Dipole²⁶, F-LSTM/CNN¹¹, and a Transformer-based vari-

ant (F-Transformer). For these baselines, we iterate over three temporal modalities as the major modality while aggregating the others as covariates (following the same procedure used in the traditional baselines) for cross-modal integration and interaction. We reported the best performance across the three configurations, using the AUC score as the primary evaluation criterion. All deep learning baselines are also optimized using Pairwise AUC Margin Loss to alleviate class imbalance.

Next, we introduce the experiment settings for evaluation. All input features are standardized using a standard scaler. The number of learnable frequency components is set to $K = 12$. All modality encoders use a hidden dimension of 32 ($d_{\text{demo}} = d_{\text{lab}} = d_{\text{dx}} = d_{\text{drug}} = 32$), and the feed-forward layers have a hidden size of 128. Training is performed with a batch size of 64 and a dual sampler ratio of 0.5 to balance positive and negative pairs, using the PESG optimizer with a learning rate of 0.001 for AUC maximization. For other deep learning baselines, we use a hidden size of 128 and optimize with Adam²⁷ using the same batch size and learning rate. We adopt stratified 5-fold cross-validation that preserves the class imbalance, repeated over 3 random seeds. We report the mean and standard deviation of AUC, AUPRC, Recall, and Specificity.

RESULTS

Results of Baseline Comparisons, Ablation Study, and Model Analysis

Table 1 summarizes the characteristics of the four electronic health record modalities included in this study. Demographic data consisted of 4 static features. Laboratory data consisted of 74 features measured across 24

Table 3. The ablation study of input modalities for early GVHD prediction.

Modalities				Graft-versus-host disease (GVHD)			
Demo.	Dx.	Lab.	Drug	AUC	AUPRC	Recall	Specificity
✓	✓	✗	✗	0.708 ± 0.032	0.081 ± 0.038	0.611 ± 0.079	0.703 ± 0.044
✓	✗	✓	✗	0.692 ± 0.008	0.061 ± 0.015	0.680 ± 0.056	0.642 ± 0.066
✓	✗	✗	✓	0.762 ± 0.009	0.118 ± 0.013	0.585 ± 0.095	0.817 ± 0.051
✓	✓	✓	✗	0.754 ± 0.011	0.060 ± 0.007	<u>0.729 ± 0.036</u>	0.659 ± 0.057
✓	✗	✓	✓	0.771 ± 0.008	0.128 ± 0.030	0.681 ± 0.062	0.756 ± 0.041
✓	✓	✗	✓	<u>0.822 ± 0.017</u>	<u>0.147 ± 0.012</u>	0.671 ± 0.038	<u>0.813 ± 0.064</u>
✓	✓	✓	✓	0.836 ± 0.007	0.157 ± 0.014	0.768 ± 0.061	0.803 ± 0.053

one-hour intervals. Diagnosis data comprised 20 features aggregated into 12 steps, each representing a 3-month interval. Medication (Drug) data included 280 features recorded over 24 steps with a daily resolution. The granularity and temporal ranges were determined in consultation with clinical experts and validated through empirical experiments for each modality.

Table 2 presents the performance of the proposed framework compared with traditional machine learning methods and deep learning baselines for early GVHD prediction. Among the traditional machine learning methods, Support Vector Classifier (SVC) achieved the highest AUC (0.723 ± 0.010) and recall (0.719 ± 0.029), while Gradient Boosting obtained the highest specificity (0.814 ± 0.084). Overall, deep learning baselines generally outperformed traditional methods, particularly in terms of AUC and AUPRC. Within the deep learning group, F-CNN yield the stronger performance in terms of AUC (0.764 ± 0.012) and recall (0.737 ± 0.015), whereas TANet reached the highest AUPRC (0.148 ± 0.016). Notably, the proposed framework demonstrated the most consistent improvements across evaluation metrics, achieving the best overall AUC (0.836 ± 0.007), the highest AUPRC (0.157 ± 0.014), a leading recall (0.768 ± 0.061), and the second-best specificity (0.803 ± 0.053).

Table 3 reports the ablation results evaluating the contribution of different modalities for early GVHD prediction. Using demographics combined with diagnoses achieved an AUC of 0.708 ± 0.032 , while demographics with laboratory tests reached an AUC of 0.692 ± 0.008 . Incorporating drugs along with demographics resulted in stronger performance (AUC 0.762 ± 0.009 , AUPRC 0.118 ± 0.013), with the highest specificity observed in this setting (0.817 ± 0.051). Overall, adding multiple modalities consistently enhanced performance. For instance, combining demographics, diagnoses, and drugs yielded an AUC of 0.822 ± 0.017 and an AUPRC of 0.147 ± 0.012 , both higher than any two-modality configuration. The best overall performance was observed when all four modalities were integrated, with results surpassing all ablation settings.

The left subplot of Figure 2 illustrates the contribution of

different model components to early GVHD prediction. Removing the global time index resulted in a decrease of AUC to 0.805 and AUPRC to 0.146, with recall and specificity at 0.747 and 0.727, respectively. Excluding the missing-aware design further reduced AUC to 0.800 and AUPRC to 0.125, accompanied by recall of 0.739 and specificity of 0.712. Without the cross-modal fusion module, the model achieved an AUC of 0.801 and an AUPRC of 0.141, with recall of 0.710 and specificity of 0.731. When all components were included, our model achieved the strongest performance.

The right subplot compares different optimization strategies for handling class imbalance in early GVHD prediction. Using Binary Cross Entropy (BCE) alone resulted in an AUC of 0.772, an AUPRC of 0.107, a recall of 0.715, and a specificity of 0.710. Incorporating progressive sampling with BCE improved performance slightly, yielding an AUC of 0.781, an AUPRC of 0.122, recall of 0.713, and specificity of 0.723. The AUC optimization consistently achieved the best performance across all evaluation metrics, improving by a clear margin compared with the other variants.

DISCUSSION

In this study, we developed a multi-modal deep learning framework to predict GVHD prior to liver transplantation surgery, using the EHRs of a real-world cohort from Mayo Clinic. Based on the reported results, we highlighted three key observations: (1) **Complementary predictive value of modalities.** Different temporal modalities provide distinct but complementary signals for early GVHD prediction, where effective integration strategies are crucial. (2) **Importance of handling data irregularity and missingness.** Robust methods that account for irregular and missing records improve predictive performance. (3) **Addressing severe class imbalance.** The extreme imbalance between GVHD and non-GVHD cases hinders model performance. Our integration of direct AUC optimization proved effective in mitigating this issue, leading to more reliable

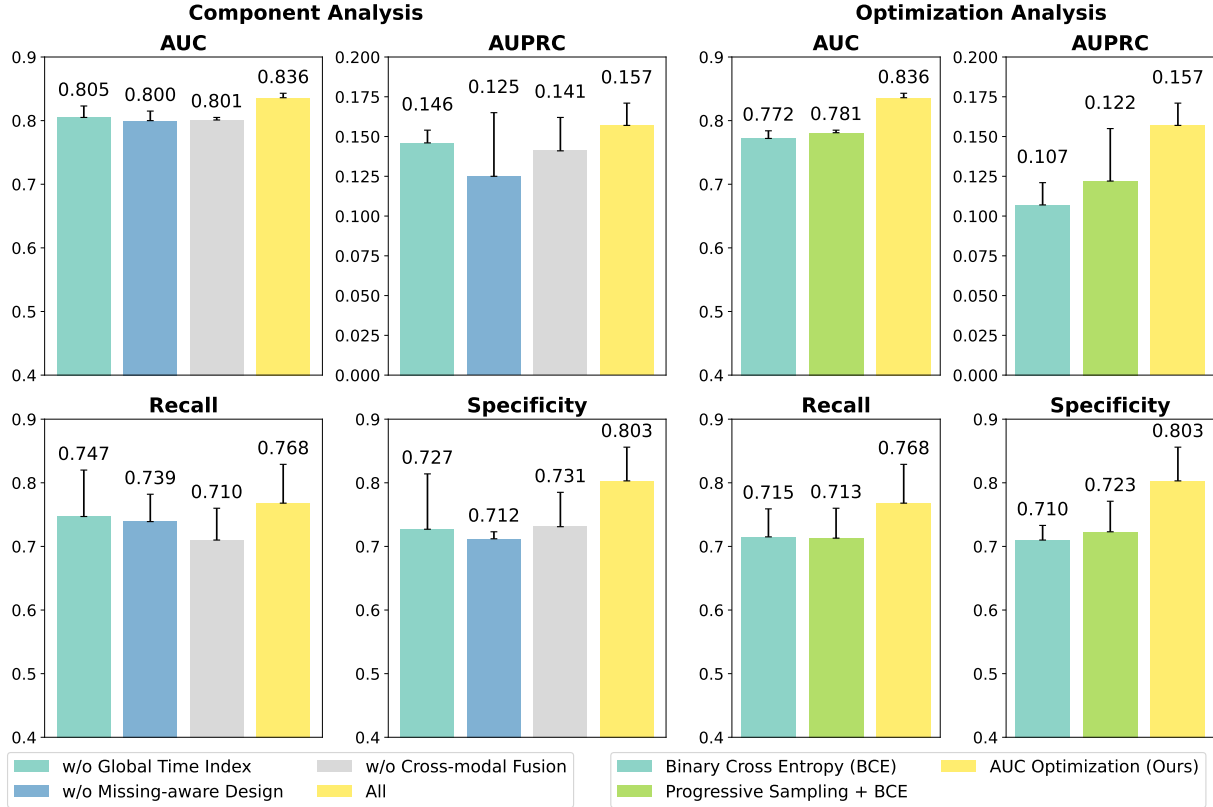


Figure 2. Left: Analysis of model components, Right: Comparison of optimization methods handling class imbalance.

early detection.

Our first key finding is that multi-modal EHR data capture diverse clinical signals essential for early GVHD prediction. As shown in Table 3, incorporating additional modalities consistently improves predictive performance across all combinations. Among the individual temporal sources, drug exposure data stood out with a relatively high AUC (0.762 ± 0.009) and the highest specificity (0.817 ± 0.051), suggesting that medication patterns prior to transplant may contain strong predictive cues for downstream complications. Diagnosis history also contributed positively, especially when combined with drug data, yielding the second-best overall performance (AUC: 0.822 ± 0.017 ; AUPRC: 0.147 ± 0.012 ; Specificity: 0.813 ± 0.064). This synergy likely reflects the clinical context in which diagnosis and medication co-evolve, providing a proxy for evolving patient status. Adding laboratory measurements further strengthened predictive accuracy, as they capture dynamic indicators of immune activation, organ dysfunction, and other early clinical signs of GVHD, culminating in the best results when all four modalities are used.

As reported in Table 2, our method outperformed both traditional machine learning and deep learning baselines. Traditional models, which rely on summarizing statistics concatenated across modalities, are limited in their ability

to capture temporal dependencies. Similarly, several deep learning baselines focus on one primary temporal stream while incorporating other modalities as static covariates for cross-modal interactions. In contrast, our framework performs end-to-end learning across all temporal modalities, enabling joint modeling of fine-grained temporal patterns and inter-modality dependencies. This design enables richer cross-modal interactions, such as associating changes in diagnosis with concurrent medication adjustments, thereby uncovering subtle yet clinically meaningful patterns. By using self-attention across modalities, our framework learns to align temporally heterogeneous signals in a dynamic manner. As shown in Figure 2, removing the cross-modal fusion mechanism resulted in a notable drop in AUC (from 0.836 to 0.801) and AUPRC (from 0.157 to 0.141), underscoring the importance of effective integration in capturing synergistic signals across heterogeneous clinical modalities.

Our second key observation underscores the importance of robust mechanisms for handling data irregularity and missingness in multi-modal EHRs. As shown in the component analysis of Figure 2, removing either the global time index or the missing-aware design resulted in consistent performance degradation across all evaluation metrics. The global time index serves as a unified temporal reference that aligns asynchronous events across modal-

ities, while the missing-aware design captures patterns of informative absence, which may reflect underlying clinical intent or patient stability. Notably, removing the missing-aware component led to a substantial drop in AUPRC, from 0.157 to 0.125, reflecting degraded detection of rare GVHD cases under imbalance.

Our third key observation emphasizes the importance of directly addressing the extreme class imbalance in GVHD prediction. As shown in the optimization analysis of Figure 2, conventional training using BCE loss yields suboptimal performance, particularly in AUPRC (0.107). Even when combined with progressive sampling that emphasizes rare positive instances in the early training stage, the improvement is marginal. In contrast, our approach using direct AUC optimization achieves a substantial boost in both AUC (from 0.772 to 0.836) and AUPRC (from 0.107 to 0.157), indicating better ranking and detection of rare positive cases. Additionally, specificity improves from 0.710 to 0.803 without sacrificing recall. The reported performance gains underscore the value of AUC optimization in learning more discriminative decision boundaries for early GVHD prediction.

CONCLUSION REMARK AND FUTURE WORK

In this study, we present a multi-modal deep learning framework for early prediction of graft-versus-host disease (GVHD) prior to liver transplantation, using heterogeneous pre-transplant EHR data. Using a real-world cohort from Mayo Clinic, we integrated demographics, laboratory tests, diagnoses, and medications into a unified predictive model. Through modality-specific encoding, cross-modal fusion, and direct AUC optimization, our method achieved strong performance under extreme class imbalance. Extensive evaluations underscore the advantages of our proposed method over baselines and the effectiveness of its components, showing the potential of early GVHD detection in liver transplantation.

This work has two primary limitations: (1) potential cohort bias, as the dataset was derived from a single institution, and (2) limited model interpretability. In future work, we plan to expand the cohort by incorporating more diverse patient populations from the Mayo Clinic Platform²⁸ and to enhance explainability through case-based reasoning^{14,29,30} and modality gating mechanisms^{31,32} that quantify feature contributions across data modalities.

FUNDING

This project was funded by the National Institute of Health, grant number R01AG083039. It was performed during the first author's summer research internship at Mayo Clinic AI & I department. Dongjin Song gratefully acknowledges the support of the National Science Foundation under Grant

No. 2338878, as well as the generous research gifts from NEC Labs America and Morgan Stanley.

REFERENCES

- Guo ZY, He XS, Wu LW, Zhu XF, Ju WQ, Wang DP, et al. Graft-versus-host disease after liver transplantation: a report of two cases and review of literature. *World Journal of Gastroenterology*: WJG. 2008;14(6):974.
- Wood A, Egtesad B, Lindenmeyer CC. Graft-versus-host disease after liver transplantation. *Clinical Liver Disease*. 2020;15(2):81-4.
- Akbulut S, Yilmaz M, Yilmaz S. Graft-versus-host disease after liver transplantation: a comprehensive literature review. *World journal of gastroenterology*: WJG. 2012;18(37):5240.
- Murali AR, Chandra S, Stewart Z, Blazar BR, Farooq U, Ince MN, et al. Graft versus host disease after liver transplantation in adults: a case series, review of literature, and an approach to management. *Transplantation*. 2016;100(12):2661-70.
- Wang J, Luo J, Ye M, Wang X, Zhong Y, Chang A, et al. Recent advances in predictive modeling with electronic health records. In: *IJCAI: proceedings of the conference*. vol. 2024; 2024. p. 8272.
- Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Scientific data*. 2016;3(1):1-9.
- Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV. *PhysioNet* Available online at: <https://physionet.org/content/mimiciv/10/> (accessed August 23, 2021). 2020:49-55.
- Pollard TJ, Johnson AE, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific data*. 2018;5(1):1-13.
- Ma L, Zhang C, Wang Y, Ruan W, Wang J, Tang W, et al. Concare: Personalized clinical feature embedding via capturing the healthcare context. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 34; 2020. p. 833-40.
- Wang X, Luo J, Wang J, Yin Z, Cui S, Zhong Y, et al. Hierarchical pretraining on multimodal electronic health records. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Conference on Empirical Methods in Natural Language Processing. vol. 2023; 2023. p. 2839.
- Tang S, Davarmanesh P, Song Y, Koutra D, Sjoding MW, Wiens J. Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*. 2020;27(12):1921-34.
- Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*. 2016;29.
- Wang L, Yu W, He X, Cheng W, Ren MR, Wang W, et al. Adversarial cooperative imitation learning for dynamic treatment regimes. In: *Proceedings of The Web Conference 2020*; 2020. p. 1785-95.
- Jiang Y, Yu W, Song D, Cheng W, Chen H. Interpretable skill learning for dynamic treatment regimes through imitation. In: *2023 57th Annual Conference on Information Sciences and Systems (CISS)*. IEEE; 2023. p. 1-6.

15. Jiang Y, Yu W, Song D, Wang L, Cheng W, Chen H. FedSkill: Privacy preserved interpretable skill learning via imitation. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2023. p. 1010-9.
16. Tang S, Chappell GT, Mazzoli A, Tewari M, Choi SW, Wiens J. Predicting acute graft-versus-host disease using machine learning and longitudinal vital sign data from electronic health records. *JCO clinical cancer informatics*. 2020;4:128-35.
17. Jo T, Arai Y, Kanda J, Kondo T, Ikegame K, Uchida N, et al. A convolutional neural network-based model that predicts acute graft-versus-host disease after allogeneic hematopoietic stem cell transplantation. *Communications Medicine*. 2023;3(1):67.
18. Song L, Wu X, Xu M, Xue L, Yu X, Cheng Z, et al. Development of an ensemble prediction model for acute graft-versus-host disease in allogeneic transplantation based on machine learning. *BMC Medical Informatics and Decision Making*. 2025;25(1):234.
19. Cooper JP, Perkins JD, Warner PR, Shingina A, Biggins SW, Abkowitz JL, et al. Acute Graft-Versus-Host Disease After Orthotopic Liver Transplantation: Predicting This Rare Complication Using Machine Learning. *Liver Transplantation*. 2022;28(3):407-21.
20. Abdelhameed A, Tao C, Yang L. Deep learning prediction modeling of major adverse cardiovascular events following liver transplantation. *Journal of Hepatology*. 2023;78:S463.
21. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:14123555*. 2014.
22. Jiang Y, Ning K, Pan Z, Shen X, Ni J, Yu W, et al. Multi-modal time series analysis: A tutorial and survey. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2; 2025. p. 6043-53.
23. Yuan Z, Yan Y, Sonka M, Yang T. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 3040-9.
24. Yuan Z, Zhu D, Qiu ZH, Li G, Wang X, Yang T. LibAUC: A Deep Learning Library for X-risk Optimization. In: 29th SIGKDD Conference on Knowledge Discovery and Data Mining; 2023. .
25. Xu Y, Biswal S, Deshpande SR, Maher KO, Sun J. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In: Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining; 2018. p. 2565-73.
26. Ma F, Chitta R, Zhou J, You Q, Sun T, Gao J. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining; 2017. p. 1903-11.
27. Kingma DP. Adam: A method for stochastic optimization. *arXiv preprint arXiv:14126980*. 2014.
28. Yu Y, Hu X, Rajaganapathy S, Feng J, Abdelhameed A, Li X, et al. Launching Insights: A Pilot Study on Leveraging Real-World Observational Data from the Mayo Clinic Platform to Advance Clinical Research. *Mayo Clinic Platform*; 2025. *ArXiv preprint arXiv:2504.16090*. Available from: <https://arxiv.org/abs/2504.16090>.
29. Jiang Y, Yu W, Lee G, Song D, Shin K, Cheng W, et al. TimeXL: Explainable multi-modal time series prediction with llm-in-the-loop. In: The Thirty-Ninth Annual Conference on Neural Information Processing Systems; 2025. .
30. Ni J, Chen Z, Cheng W, Zong B, Song D, Liu Y, et al. Interpreting convolutional sequence model by learning local prototypes with adaptation regularization. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management; 2021. p. 1366-75.
31. Jiang M, Ji S. Cross-modality gated attention fusion for multimodal sentiment analysis. *arXiv preprint arXiv:220811893*. 2022.
32. Yun S, Choi I, Peng J, Wu Y, Bao J, Zhang Q, et al. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. *Advances in Neural Information Processing Systems*. 2024;37:98782-805.