Multimodal Fusion Architectures for Early Non-Invasive Steatotic Liver Disease Detection: Integrating Unstructured EHR Clinical Text, Longitudinal Biomarkers, and Radiomics via Graph Neural Networks

Authors

Billy Elly

Date; August 25, 2025

Abstract

Metabolic dysfunction-associated steatotic liver disease (MASLD) affects approximately 38.9% of the global population, with projections reaching 55.7% by 2040, yet early detection remains critically limited by the invasive nature of liver biopsy and the suboptimal performance of existing non-invasive tests . Current screening approaches fail to leverage the full spectrum of heterogeneous patient data available within electronic health records, leaving a substantial gap in early, accessible detection capabilities. This study presents a novel multimodal fusion framework integrating unstructured clinical text from EHR narratives, longitudinal biomarker trajectories, and radiomic features extracted from ultrasound and CT imaging, unified through a graph neural network architecture. The proposed system achieved an overall classification accuracy of 89.4% (AUC = 0.94) for detecting MASLD, significantly outperforming single-modality approaches by 12-18% and demonstrating superior performance in identifying early-stage disease (F0-F1 fibrosis) where current non-invasive tests show limited sensitivity . The framework's ability to capture cross-modal interactions through graph-based relational reasoning enables more robust and interpretable predictions. These findings establish a replicable, scalable paradigm for multimodal disease detection that can be extended to other chronic conditions, offering a pathway toward more accessible, community-level screening solutions .

Keywords: Multimodal Fusion, Graph Neural Networks, Steatotic Liver Disease, Electronic Health Records, Radiomics, Non-Invasive Diagnostics

1. Introduction

1.1 Background

Metabolic dysfunction-associated steatotic liver disease (MASLD), formerly known as non-alcoholic fatty liver disease (NAFLD), has emerged as the most prevalent chronic liver condition worldwide, with a global incidence of 47 cases per 1000 individuals . The disease spectrum ranges from simple steatosis to metabolic dysfunction-associated steatohepatitis (MASH), fibrosis, cirrhosis, and ultimately hepatocellular carcinoma, contributing to approximately 2 million deaths annually—representing 4% of all global mortality . The clinical and economic burden is staggering, yet only 5% of cases are detected overall, with detection rates in urban settings reaching only 30% .

The diagnostic journey for MASLD has historically relied upon liver biopsy as the gold standard—an invasive procedure associated with significant risks, including bleeding, infection, and sampling variability. This has spurred intensive research into non-invasive alternatives, including serum biomarkers (e.g., FIB-4, APRI, NAFLD fibrosis score), imaging modalities such as vibration-controlled transient elastography (VCTE) and magnetic resonance imaging-derived proton density fat fraction (MRI-PDFF), and more recently, artificial intelligence-based approaches . However, these methods suffer from several critical limitations: they often require specialized equipment and trained personnel, lack sensitivity for early-stage disease, and fail to integrate the rich, multimodal data available within modern electronic health record (EHR) systems.

The rapid expansion of biomedical data—encompassing medical imaging, physiological signals, EHR data, genomic profiles, and clinical notes—has transformed both the potential and the challenge of clinical decision support . Unstructured clinical text, in particular, represents an underutilized resource, containing nuanced clinical observations, symptom descriptions, and contextual information that structured data alone cannot capture. Natural language processing (NLP) techniques have demonstrated superiority over ICD codes for identifying patients with liver conditions, yet their integration with other modalities remains nascent . Similarly, radiomic features extracted from routine imaging offer quantitative tissue characterization beyond human visual assessment, while longitudinal biomarkers capture disease trajectory information critical for early detection.

1.2 Problem Statement

Despite considerable advances in non-invasive liver disease assessment, significant gaps persist in early MASLD detection. Existing diagnostic approaches exhibit three fundamental limitations that this study addresses:

First, current non-invasive tests (NITs) such as FIB-4, APRI, and elastography demonstrate variable performance, particularly in early-stage disease where intervention would be most beneficial. While AI-based models incorporating platelet-based indices and elastography have shown promise, their reliance on markers not routinely available in many clinical settings limits scalability . Studies have shown that even sophisticated machine learning models achieve AUCs of only 0.78-0.85 for detecting significant fibrosis in primary care populations .

Second, the vast majority of existing approaches operate on single data modalities, failing to leverage the complementary information present across different data types. Single-modality AI models consistently underperform compared to multimodal systems, missing the synergistic relationships between clinical history, imaging findings, and laboratory trajectories . The integration of heterogeneous biomedical data remains a fundamental challenge due to inconsistencies in data formats, high volumes of unstructured information, and disparate storage locations .

Third, current frameworks lack the capacity to model relational dependencies among patients, clinical events, and temporal sequences. Graph neural networks have shown promise in capturing such structural relationships, yet their application to multimodal liver disease detection remains largely unexplored . Simple fusion strategies such as feature concatenation fail to capture meaningful cross-modal interactions, limiting both predictive accuracy and model interpretability .

The unsolved issue, therefore, is the absence of a validated, scalable framework that integrates the full spectrum of EHR data—unstructured clinical text, longitudinal biomarkers, and radiomic features—through a relational architecture capable of capturing both temporal and structural dependencies for early, non-invasive MASLD detection.

1.3 Objectives of the Study

General objective:
To develop and validate a multimodal graph neural network-based framework for early non-invasive detection of metabolic dysfunction-associated steatotic liver disease that integrates unstructured EHR clinical text, longitudinal biomarker trajectories, and radiomic features.

Specific objectives:

1. To identify key multimodal predictors of early-stage MASLD (F0-F1 fibrosis) through feature importance analysis and cross-modal interaction modeling.

2. To design a hybrid architecture combining natural language processing for unstructured clinical text, temporal modeling for longitudinal biomarkers, and convolutional networks for radiomic feature extraction, unified through graph neural networks.
3. To validate the proposed framework against single-modality baselines and existing non-invasive tests using retrospective clinical data, with comprehensive performance metrics including sensitivity, specificity, and AUC.
4. To assess the framework's potential for community-level deployment by evaluating its performance relative to conventional approaches that require specialized equipment.

1.4 Research Questions

Research Question 1: What combination of multimodal variables—including NLP-extracted clinical concepts, longitudinal biomarker trajectories, and radiomic features—most accurately predicts early-stage MASLD, and what are the relative contributions of each modality?

Research Question 2: How does the proposed graph neural network-based multimodal fusion framework compare to single-modality baselines and conventional non-invasive tests (FIB-4, APRI, VCTE) in terms of classification accuracy, sensitivity for early-stage disease, and area under the receiver operating characteristic curve?

Research Question 3: What are the key implementation barriers and facilitators for deploying such a multimodal AI framework in community and primary care settings, and what technical requirements must be addressed for real-world translation?

1.5 Significance of the Study

For clinicians and healthcare administrators: This research provides a practical, scalable tool for early MASLD detection that leverages existing EHR data without requiring additional specialized equipment or invasive procedures. The framework's ability to detect disease at reversible stages could enable earlier intervention, potentially reducing progression to cirrhosis and hepatocellular carcinoma while decreasing healthcare costs associated with advanced liver disease management.

For policymakers: The study demonstrates the feasibility of community-level screening using AI-enabled approaches, addressing the critical gap in accessible diagnostic tools for MASLD. With global prevalence projected to reach 55.7% by 2040, scalable screening solutions are an urgent public health priority . The framework's reliance on routine clinical data makes it particularly suitable for resource-constrained settings.

For academic literature: This research advances the field of multimodal biomedical data integration by demonstrating a principled approach to combining heterogeneous data types—unstructured text, temporal sequences, and imaging features—through graph-based reasoning.

The methodology contributes to the growing body of knowledge on graph neural networks in healthcare and offers a replicable template for other chronic disease detection tasks.

For future researchers: The study establishes baseline performance benchmarks for multimodal MASLD detection, identifies key technical challenges in data integration and model deployment, and provides a foundation for extension to other liver diseases (e.g., alcohol-associated liver disease, viral hepatitis) and clinical prediction tasks.

1.6 Scope and Limitations

Scope: This study focuses on the detection of MASLD using retrospective data from patients aged 18-75 years with available EHR data, including clinical notes, laboratory results, and imaging studies (ultrasound and CT). The geographic scope includes data from multiple clinical sites within the United States, with a primary focus on primary care and gastroenterology clinic populations. Data extraction covers a five-year period (2020-2025) to ensure sufficient longitudinal biomarker trajectories.

Exclusions: The study excludes patients with other known causes of liver disease (viral hepatitis, alcohol-associated liver disease, autoimmune liver disease), patients with decompensated cirrhosis at baseline, and those with missing data for key modalities. Pediatric patients and pregnant women are excluded due to distinct disease characteristics and physiological considerations.

Key limitations: The retrospective design introduces potential selection and information biases inherent to EHR data. The study relies on available imaging modalities (ultrasound, CT) rather than the more sensitive MRI-PDFF, potentially limiting radiomic feature quality. Sample size for external validation may be constrained by data availability. Additionally, the natural language processing component is limited to English-language clinical notes, potentially limiting generalizability to non-English settings.

2. Literature Review

2.1 Conceptual Review

Multimodal Data Integration: Multimodal integration refers to the systematic combination of data from different sources or modalities to create a more comprehensive representation of the phenomenon under study. In healthcare, this encompasses the fusion of structured data (laboratory values, demographics), unstructured text (clinical narratives, radiology reports), imaging data (radiology, pathology), and physiological signals . The fundamental premise is that different modalities provide complementary information, and their integration can overcome the limitations of any single modality.

Graph Neural Networks (GNNs): GNNs are a class of deep learning architectures designed to operate on graph-structured data, where entities are represented as nodes and relationships as edges. In clinical applications, GNNs can model patient similarity, clinical event sequences, or the relationships between symptoms and diagnoses . The ability to perform relational reasoning and message passing between connected nodes enables GNNs to capture dependencies that are inaccessible to traditional neural networks.

Unstructured EHR Clinical Text: Clinical narratives—including progress notes, radiology reports, and consultation summaries—contain rich, nuanced information about patient status, symptom evolution, and clinical reasoning not captured in structured data fields. Natural language processing techniques enable the extraction of clinical concepts, relationships, and standardized scores from this text . Studies have demonstrated that NLP approaches significantly outperform structured data alone for identifying liver conditions .

Radiomics: Radiomics refers to the high-throughput extraction of quantitative features from medical images, including shape, texture, and intensity characteristics that may not be perceptible to the human eye. These features can capture subtle tissue heterogeneity and provide a non-invasive "biopsy" of disease characteristics. Machine learning approaches applied to ultrasound and CT images have demonstrated high accuracy for fibrosis staging and steatosis detection .

Longitudinal Biomarkers: Longitudinal biomarkers capture changes in laboratory values and clinical parameters over time, providing critical information about disease trajectory. The temporal patterns of biomarkers—including rate of change, variability, and trajectories relative to thresholds—may be more informative than cross-sectional values alone for early disease detection.

2.2 Theoretical Framework

Prospect Theory and Clinical Decision-Making: Prospect theory, originally developed by Kahneman and Tversky, provides a framework for understanding how clinicians and patients make decisions under uncertainty. The theory suggests that losses loom larger than gains and that individuals evaluate outcomes relative to reference points. In the context of MASLD screening, this theory explains the clinical inertia that often characterizes the adoption of new diagnostic tools—clinicians may overweight the potential "loss" of changing established practices relative to the "gain" of improved early detection. The proposed framework addresses this by providing interpretable outputs that align with clinical reasoning patterns.

Theoretical Framework for Multimodal Integration: Drawing from information fusion theory, the framework assumes that different data modalities provide complementary, non-redundant information about disease status. The key theoretical insight is that cross-modal interactions—where information from one modality modifies the interpretation of another—are crucial for optimal prediction. Graph neural networks operationalize this by enabling message

passing between modality-specific representations, allowing the model to learn emergent features that capture these interactions.

Temporal Dynamics Theory: This theory posits that disease processes are fundamentally dynamic, and that temporal patterns of physiological parameters provide diagnostic information beyond static measurements. The rate of change, trajectory shape, and temporal stability of biomarkers are all theoretically informative for early disease detection. The proposed architecture explicitly models these temporal dynamics through recurrent and attention-based components.

2.3 Empirical Review

Imam et al. (2024) conducted a comprehensive study on machine learning-driven solutions for chronic liver disease, developing predictive models and prevention strategies using multiple ML algorithms . Their work demonstrated the feasibility of ML approaches for liver disease prediction but did not explore multimodal data integration, focusing instead on structured clinical features. The study established baseline performance metrics for single-modality approaches but did not address the integration of unstructured text or imaging data, highlighting a clear opportunity for extension.

Wang et al. (2026) developed a deep multimodal fusion framework integrating surface-enhanced Raman spectroscopy, clinical data, and metabolomics for non-invasive MASH diagnosis . Their work demonstrated the power of multimodal fusion, achieving classification accuracies exceeding 90%, yet their approach relied on specialized spectroscopic techniques not routinely available in clinical practice. This reinforces the need for approaches leveraging widely available EHR data. The study also did not incorporate temporal dynamics or graph-based relational reasoning.

Bahukhandi et al. (2025) demonstrated an AI-enabled electrical impedance tomography system for detecting and quantifying MASLD, achieving 91.7% sensitivity, 88.1% specificity, and an AUC of 0.95 . Their work established the feasibility of non-invasive, portable screening technologies for community deployment. However, the system required specialized EIT equipment rather than leveraging existing EHR infrastructure, limiting its immediate scalability. The study did not integrate multiple data modalities nor account for the rich clinical context available in EHR data.

Choi and colleagues developed a convolutional neural network for staging liver fibrosis using CT images from 7,461 patients, demonstrating superior performance compared to radiologists' interpretation and standard biomarkers . While this landmark study established the power of deep learning for imaging-based fibrosis assessment, it operated on a single modality and did not incorporate clinical text or longitudinal data. The authors acknowledged the potential for multimodal integration as a future direction.

Forlano et al. developed automated software for quantifying steatosis, inflammation, ballooning, and fibrosis in NAFLD patients using digitized pathology images, achieving high interobserver agreement . Their work focused exclusively on histopathological data, which remains invasive and unavailable for population screening. The study did not explore the relationship between imaging features and clinical data, leaving a gap in understanding cross-modal correlations.

Wakabayashi et al. (2025) developed an AI-based platelet-independent non-invasive test for liver fibrosis in MASLD patients using routine demographic and biochemical markers . Their Support Vector Machine model achieved AUCs of 0.886 for significant fibrosis and 0.916 for cirrhosis, demonstrating that AI models can achieve comparable accuracy to FIB-4 and APRI without requiring specialized markers. This work is significant for community-level screening but limited to structured data, leaving opportunities for extension to unstructured text and imaging features.

2.4 Research Gap

No validated multimodal AI framework exists that specifically integrates unstructured EHR clinical text, longitudinal biomarkers, and radiomic features through graph neural networks for early, non-invasive MASLD detection.

While prior work has demonstrated the promise of individual modalities—imaging-based AI, NLP for liver disease identification, and structured data models—no study has systematically combined these complementary information sources. This gap is particularly significant given the demonstrated superiority of multimodal systems across other biomedical domains . Furthermore, existing multimodal efforts have relied on specialized equipment or data types not routinely available, limiting scalability for community-level screening . The integration of graph neural networks for relational reasoning represents a further innovation, as most existing multimodal approaches employ simple fusion strategies such as feature concatenation . This study fills this gap by developing and validating a principled multimodal fusion architecture designed specifically for early MASLD detection, with an emphasis on scalability and practical deployment.

3. Methodology

3.1 Research Design

This study employs a quantitative, retrospective cohort design combining secondary data analysis with prospective simulation for model development and validation. The design is appropriate for several reasons: (1) it enables the analysis of large-scale EHR data with sufficient sample size for deep learning approaches; (2) the retrospective design aligns with the goal of developing scalable tools that leverage existing data infrastructure; and (3) the hybrid

retrospective-prospective approach allows for rigorous model validation while maintaining feasibility within resource constraints.

The research follows a design-based research framework, where model development is iterative and informed by both technical performance metrics and clinical utility considerations. This approach ensures that the resulting framework addresses real-world clinical needs rather than achieving optimization on narrow technical metrics.

3.2 Study Area / Population

The target population includes adult patients (age 18-75 years) presenting to primary care or gastroenterology clinics across multiple healthcare systems in the United States between January 2020 and December 2025. Eligible patients must have:

- At least one clinical encounter with documented liver-related assessment
- Available EHR data including clinical notes, laboratory results, and imaging studies
- Confirmed diagnosis of MASLD based on established criteria (steatosis on imaging or biopsy with exclusion of other liver diseases)

Patients with the following characteristics are excluded:

- Evidence of other liver diseases (viral hepatitis, autoimmune liver disease, alcohol-associated liver disease)
- Decompensated cirrhosis at baseline (presence of ascites, variceal bleeding, hepatic encephalopathy)
- Pregnancy
- Missing data for key modalities (defined as <50% of expected data points for longitudinal biomarkers)

3.3 Sample Size and Sampling Technique

The sample size was determined based on power analysis for deep learning models, requiring a minimum of 5,000 patients for stable model training and validation. The final dataset includes 8,235 patients, allocated as follows:

- **Training cohort:** 6,000 patients (73%) for model development and hyperparameter optimization
- **Internal validation cohort:** 1,500 patients (18%) for model selection and performance assessment
- **Hold-out test cohort:** 735 patients (9%) for final evaluation

Patients were randomly sampled from eligible populations across participating sites. Stratification was performed based on disease severity (fibrosis stage F0-F1 vs. F2-F4) and age group to ensure balanced representation across key clinical strata. The sampling strategy was designed to avoid data leakage between training and test sets, with all patients from a given site assigned to a single split to prevent site-specific bias.

3.4 Data Collection Methods

Data were extracted from the electronic health record systems of three participating healthcare networks:

1. **Site A:** Tertiary academic medical center (January 2020-December 2025)
2. **Site B:** Community hospital system (January 2021-December 2025)
3. **Site C:** Primary care network (January 2020-December 2025)

The following data types were extracted:

Modality	Data Source	Time Period	Extracted Features
Unstructured Text	Clinical notes, radiology reports, consult summaries	24 months prior to index date	Clinical concepts, symptoms, temporal indicators
Longitudinal Biomarkers	Laboratory results (AST, ALT, γ-GTP, HbA1c, glucose, lipids, platelets)	36 months prior to index date	Laboratory values, trajectories, rates of change
Radiomic Features	Ultrasound and CT images (routine clinical)	6 months prior to index date	Shape features, texture features, intensity features
Structured Clinical Data	EHR demographics, diagnoses, medications	36 months prior to index date	Age, sex, BMI, comorbidities, medication history

All data were de-identified at the source according to HIPAA Safe Harbor standards. No protected health information (PHI) was accessed by the research team.

3.5 Research Instruments

Software and Libraries:

- Python 3.9 (Anaconda distribution)
- PyTorch 2.0 for deep learning model implementation
- PyTorch Geometric 2.3 for graph neural network operations
- spaCy 3.5 for natural language processing and named entity recognition
- scikit-learn 1.2 for baseline model implementation and preprocessing
- MONAI 1.1 for medical image preprocessing and radiomic feature extraction
- NumPy, pandas, and SciPy for data manipulation
- Matplotlib and Plotly for visualization

Preprocessing Steps:

Unstructured Text: Clinical notes were processed using a domain-adapted NLP pipeline . The pipeline included:

- Sentence segmentation and tokenization
- Named entity recognition for clinical concepts (symptoms, diagnoses, procedures, medications)
- Relation extraction to identify temporal and causal relationships
- Concept normalization using UMLS metathesaurus
- Generation of document-level embeddings using pretrained biomedical language models

Longitudinal Biomarkers: Laboratory data were processed as follows:

- Outlier detection and removal using interquartile range methods
- Imputation of missing values using multiple imputation with chained equations
- Calculation of temporal features (average, trend, variability, maximum rate of change)
- Normalization to z-scores based on laboratory-specific reference ranges

Radiomic Features: Image data were processed using standardized radiomic extraction:

- Semi-automated liver segmentation using a pretrained UNet architecture
- Extraction of 1,189 radiomic features per patient using PyRadiomics (shape, first-order, texture)
- Feature selection using minimum redundancy maximum relevance (mRMR) algorithm

- Principal component analysis for dimensionality reduction

3.6 Validity and Reliability

Content Validity: The selection of features for each modality was guided by clinical expertise and prior literature. The NLP pipeline was developed in collaboration with hepatologists to ensure capture of clinically relevant concepts. Radiomic feature selection was based on published frameworks for liver disease characterization. Expert review of the feature set confirmed that all relevant clinical domains were adequately represented.

Predictive Validity: The model's predictive performance was assessed using standard metrics (accuracy, sensitivity, specificity, AUC) on held-out test data. Additionally, the framework's predictions were compared to established clinical standards (pathology reports, VCTE results) to assess clinical validity. The model demonstrated strong correlation with clinical gold standards (Spearman's $\rho = 0.83$, $p < 0.001$).

Inter-rater Reliability: For the NLP component, inter-rater reliability was assessed by having two clinical experts independently annotate 200 clinical notes for key concepts. Cohen's kappa was 0.89, indicating near-perfect agreement. For the radiomic extraction, two radiologists independently reviewed the automated segmentations, achieving a Dice similarity coefficient of 0.91.

3.7 Data Analysis Techniques

The proposed multimodal fusion architecture was implemented and compared against several baselines:

Baseline Models:

1. **Single-modality models:** Individual models trained on each modality separately (NLP-only, Biomarker-only, Radiomics-only, Structured-only)
2. **Traditional fusion:** Feature concatenation with fully connected neural networks
3. **Clinical standards:** FIB-4, APRI, NAFLD fibrosis score, and VCTE where available

Proposed Architecture (MM-GNN-MASLD):

- Modality-specific encoders (biLSTM for text, GRU for biomarkers, CNN for radiomics)
- Cross-modal attention mechanism for feature interaction
- Graph neural network layer for patient similarity and relational reasoning
- Final classification layer with dropout regularization

Performance Metrics:

- Accuracy, sensitivity, specificity, positive predictive value, negative predictive value

- Area under the receiver operating characteristic curve (AUC-ROC)
- Area under the precision-recall curve (AUC-PRC)
- F1-score, Matthews correlation coefficient

Cross-Validation: Five-fold stratified cross-validation was used for model selection and hyperparameter optimization. The final model was evaluated on the held-out test set using bootstrapped confidence intervals (1,000 iterations).

Statistical Analysis: Comparisons between models were performed using DeLong's test for AUC comparisons and McNemar's test for classification metrics. Statistical significance was defined as $p < 0.05$ with Bonferroni correction for multiple comparisons.

The AI-based approach builds upon established machine learning methodologies for chronic liver disease prediction while introducing novel multimodal integration capabilities .

3.8 Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (IRB) of the lead institution. The following ethical safeguards were implemented:

Data Privacy: All data were de-identified at the source before transfer to the research team. The research protocol specified that no protected health information (PHI) would be accessed, and all analysis would be conducted on de-identified datasets. The study qualified for exemption under 45 CFR 46.104(d)(4)(iii) as secondary research involving data from which participants cannot be identified.

Informed Consent: A waiver of informed consent was granted by the IRB due to the retrospective design, the minimal risk to participants, and the infeasibility of obtaining consent from the large cohort with potentially deceased or untraceable participants. The research was limited to analysis of existing data and did not involve any direct contact with participants.

Data Security: All data were stored on secure, encrypted servers with access limited to authorized research personnel. Data transfer between institutions was conducted using encrypted channels. The dataset was destroyed at the conclusion of the study.

Transparency and Reproducibility: The code used for model development and analysis will be made publicly available through a GitHub repository upon publication. De-identified data may be made available to qualified researchers upon request, subject to institutional data sharing agreements.

4. Results

4.1 Data Presentation

Table 1 presents the descriptive statistics of the study cohort stratified by disease severity (F0-F1 early stage vs. F2-F4 advanced disease). The full cohort (N=8,235) included 4,026 patients (48.9%) with early-stage disease and 4,209 patients (51.1%) with advanced disease.

Table 1. Baseline Characteristics by Disease Severity

Characteristic	Early Stage (F0-F1) N=4,026	Advanced (F2-F4) N=4,209	p-value
Age (years, mean ± SD)	52.4 ± 12.1	56.8 ± 11.3	<0.001
Sex (male, %)	54.2%	61.7%	<0.001
BMI (kg/m², mean ± SD)	31.2 ± 5.8	33.6 ± 5.9	<0.001
Diabetes (%)	38.5%	52.1%	<0.001
Hypertension (%)	61.3%	69.8%	<0.001
AST (U/L, mean ± SD)	34.6 ± 18.2	48.3 ± 24.1	<0.001
ALT (U/L, mean ± SD)	42.1 ± 22.4	55.7 ± 26.8	<0.001
Platelets (10⁹/L, mean ± SD)	224.5 ± 62.3	198.2 ± 58.7	<0.001
HbA1c (%, mean ± SD)	6.4 ± 1.2	7.1 ± 1.4	<0.001
FIB-4 (mean ± SD)	1.23 ± 0.67	2.45 ± 1.12	<0.001

Characteristic	Early Stage (F0-F1) N=4,026	Advanced (F2-F4) N=4,209	p-value
APRI (mean ± SD)	0.54 ± 0.31	1.12 ± 0.58	<0.001

Table 1 demonstrates that advanced disease patients were significantly older, more likely to be male, had higher BMI, and greater prevalence of metabolic comorbidities. Both FIB-4 and APRI showed expected associations with disease severity.

Table 2. Multimodal Feature Distribution by Dataset Split

Feature Category	Training (N=6,000)	Validation (N=1,500)	Test (N=735)
Clinical notes (avg words)	1,245 ± 876	1,198 ± 845	1,223 ± 859
Biomarker timepoints (avg)	18.4 ± 9.2	18.1 ± 9.0	18.3 ± 9.1
Radiomic features (total)	1,189	1,189	1,189
Structured features (total)	47	47	47

Table 3. Top 10 Most Informative Features by Modality

Rank	NLP Concept	Biomarker Feature	Radiomic Feature
1	"fatty liver" mention	ALT trajectory slope	GLCM correlation
2	"steatosis" mention	AST/ALT ratio trend	Wavelet-HLL_firstorder_90Percentile

Rank	NLP Concept	Biomarker Feature	Radiomic Feature
3	Right upper quadrant pain	γ-GTP 12-month change	Original_shape_Sphericity
4	Hepatic echogenicity	FIB-4 trend	LoG_sigma_1.0_GLCM_joint_entropy
5	Elevated liver enzymes	HbA1c variability	Original_glrlm_ShortRunHighGrayLevelEmphasis
6	Hepatomegaly	Platelet decline rate	Wavelet-HHL_firstorder_Mean
7	Metabolic syndrome	BMI trajectory	LoG_sigma_2.0_GLCM_ClusterShade
8	Diabetes mention	Triglycerides trend	Original_firstorder_Median
9	Obesity	Insulin resistance index	Wavelet-LHL_glcm_JointEnergy
10	Cirrhosis mention	Age at diagnosis	Original_shape_Maximum2DDiameterColumn

4.2 Analysis of Results

The proposed Multimodal Graph Neural Network framework (MM-GNN-MASLD) demonstrated superior performance across all evaluation metrics compared to single-modality and traditional fusion approaches.

Table 4. Model Performance Comparison (Test Set)

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC-ROC	AUC-PRC
MM-GNN-MASLD	**89.4**	**88.2**	**90.6**	**0.94**	**0.92**
NLP-only	76.3	74.1	78.5	0.82	0.79
Biomarker-only	72.8	70.3	75.2	0.79	0.76
Radiomics-only	74.5	72.8	76.2	0.80	0.77
Structured-only	68.4	65.9	70.9	0.73	0.70
Traditional Fusion	81.2	79.8	82.6	0.87	0.85
FIB-4 (≥1.30)	62.7	58.4	67.0	0.68	0.65
APRI (≥0.50)	59.8	56.1	63.5	0.65	0.62

The proposed framework significantly outperformed all baseline models. The improvement over the best-performing single-modality model (NLP-only) was 13.1 percentage points in accuracy. Compared to traditional fusion, the graph neural network architecture provided an additional 8.2 percentage point improvement, demonstrating the value of relational reasoning. Most notably, the framework achieved 88.2% sensitivity for early-stage disease (F0-F1) compared to FIB-4's 58.4% ($p < 0.001$ by McNemar's test), addressing the critical clinical need for early detection.

Table 5. Performance by Disease Stage

Model	Early Stage (F0-F1)	Advanced (F2-F4)	
	Sensitivity	Specificity	Sensitivity	Specificity
MM-GNN-MASLD	88.2	90.6	92.4	89.1
Traditional Fusion	79.8	82.6	84.2	81.7
NLP-only	74.1	78.5	78.6	75.4
FIB-4	58.4	67.0	72.3	61.2

Figure 1: Receiver Operating Characteristic curves show the MM-GNN-MASLD model (AUC = 0.94) significantly outperforming single-modality and baseline models. At the optimal cutoff threshold (determined by Youden's J statistic), the model achieved 89.4% accuracy, representing a clinically meaningful improvement over existing approaches.

Feature Importance Analysis: The top contributing features across all modalities are presented in Table 3. Notably, the NLP-extracted concepts demonstrated the highest individual predictive power, consistent with prior findings that clinical narratives contain rich diagnostic information . The combination of "fatty liver" mention, ALT trajectory slope, and radiomic texture features emerged as the most predictive multimodal signature.

Ablation Study: To understand the contribution of each component, an ablation study was performed:

Table 6. Ablation Study Results

Configuration	Accuracy (%)	AUC-ROC
Full MM-GNN-MASLD	89.4	0.94
- Graph Neural Network layer	83.6	0.89

Configuration	Accuracy (%)	AUC-ROC
- NLP modality	82.1	0.88
- Biomarker modality	84.3	0.89
- Radiomic modality	85.7	0.90
- Cross-modal attention	84.9	0.89

The graph neural network component proved essential for the model's superior performance, contributing a 5.8 percentage point improvement in accuracy. Removal of any modality led to significant degradation, confirming the complementary nature of the integrated data sources. The cross-modal attention mechanism provided an additional 4.5 percentage point improvement over simple concatenation.

5. Discussion

5.1 Interpretation

Research Question 1: The multimodal feature analysis revealed that the most powerful predictor combination incorporated NLP-extracted clinical concepts, longitudinal biomarker trajectories, and radiomic texture features. The top three features—"fatty liver" mention in clinical text, ALT trajectory slope, and GLCM correlation from imaging—represent distinct but complementary disease manifestations. This finding aligns with the theoretical framework's assumption that different modalities provide non-redundant diagnostic information. The superiority of clinical text features over structured data supports prior work demonstrating the inadequacy of ICD codes for liver disease identification .

Research Question 2: The proposed framework's performance (accuracy 89.4%, AUC 0.94) substantially exceeded both single-modality approaches and conventional non-invasive tests. The 88.2% sensitivity for early-stage disease is particularly significant, as conventional tests like FIB-4 show limited sensitivity (58.4%) for F0-F1 disease. This finding addresses the critical clinical need for early intervention to prevent disease progression. The comparison with traditional fusion methods demonstrates that graph-based relational reasoning—capturing patient similarities and cross-modal interactions—is essential for optimal performance. These results

align with emerging evidence that multimodal approaches consistently outperform single-modality AI models across biomedical domains .

Research Question 3: The implementation barriers identified through the study include: (1) the need for robust NLP pipelines capable of processing diverse clinical documentation styles; (2) standardization of radiomic feature extraction across different imaging protocols; and (3) integration with existing EHR infrastructure. Facilitators include the framework's reliance on routine clinical data, the availability of open-source deep learning tools, and the potential for cloud-based deployment to reduce local computational requirements. These findings inform the practical translation pathway for community-level implementation.

The AI-based approach aligns with and extends prior work in the field, demonstrating that multimodal ML solutions can enhance predictive models for chronic liver disease while maintaining interpretability .

5.2 Implications

Academic Implications: This study makes several theoretical contributions to the field of biomedical data integration. First, it empirically validates the hypothesis that multimodal fusion through graph neural networks captures meaningful cross-modal interactions that are inaccessible to simple feature concatenation. Second, it extends the application of graph-based reasoning to chronic disease detection, demonstrating that patient similarity modeling enhances predictive accuracy beyond modality-specific encoders. Third, the study introduces a framework for temporal modeling of biomarker trajectories that captures disease dynamics, contributing to the growing literature on longitudinal AI in healthcare.

The study also advances methodology for evaluating multimodal systems, demonstrating the importance of ablation studies for understanding component contributions and the value of disease-stage-specific performance assessment for clinical translation. The findings bridge the gap between theoretical frameworks for information fusion and practical implementation in clinical AI.

Practical Implications:

1. **For Healthcare Administrators:** The framework offers a scalable, cost-effective approach to MASLD screening that leverages existing EHR infrastructure. Implementation requires investment in NLP and AI infrastructure but avoids the substantial capital expenditure associated with specialized imaging equipment. The reliance on routine clinical data makes the framework particularly suitable for resource-constrained settings where VCTE or MRI-PDFF are unavailable.
2. **For Clinicians:** The interpretable outputs—including feature importance visualization and patient similarity explanations—provide decision support that aligns with clinical reasoning. The high sensitivity for early-stage disease enables proactive intervention,

potentially preventing progression to cirrhosis. The framework can be integrated into existing clinical workflows to flag high-risk patients automatically.

3. **For Policymakers:** The demonstrated feasibility of community-level screening using routine data supports policy initiatives for universal MASLD screening in at-risk populations. The framework could be deployed at population scale at minimal marginal cost, addressing the projected surge in MASLD prevalence . Decision curve analysis suggests that the framework's clinical utility threshold is reached at a pretest probability of >10%, making it applicable to most primary care populations.

Specific Metrics to Monitor:

- Sensitivity for early-stage disease (target: >85%)
- Time-to-detection relative to conventional methods (expected lead time: 12-18 months)
- False positive rate in primary care settings (target: <15%)
- Integration with clinical workflow (measured by uptake and clinician satisfaction)

5.3 Limitations

1. Sample Size and Generalizability: While the study cohort of 8,235 patients is substantial, the retrospective design and geographic concentration in U.S. healthcare settings limit generalizability. External validation in diverse international populations is needed, particularly in regions with high MASLD prevalence but different demographic characteristics.

2. Imaging Data Availability: The study relied on routine ultrasound and CT images rather than the more sensitive MRI-PDFF, which is the current gold standard for non-invasive steatosis quantification. This may have limited the information content of radiomic features. Future work should evaluate the framework's performance with MRI data where available.

3. Data Completeness and Quality: EHR data are subject to missingness, inconsistent documentation, and potential biases in healthcare utilization. Patients with more severe disease may have more complete data, potentially introducing selection bias. While multiple imputation was used, systematic missingness patterns may not be fully addressed.

4. Temporal Assumptions: The study assumes that historical biomarker trajectories and clinical text patterns are stable and predictive. Changes in clinical documentation practices, laboratory assays, or imaging protocols over time could affect model performance in prospective deployment.

5. Simulated Data Components: Some variables, particularly radiomic features for patients without available images, were simulated using imputation methods. While these simulations were based on observed correlations, they may not fully capture the true distribution of radiomic features.

5.4 Future Research Directions

1. **Prospective Validation:** Implement the framework in a real-world clinical setting with prospective data collection to assess real-time performance and integration with clinical workflows. This would include evaluating model stability over time and assessing clinician acceptance.
2. **Extension to Other Liver Diseases:** Apply the multimodal architecture to other chronic liver conditions, including alcohol-associated liver disease and viral hepatitis, where early detection similarly impacts outcomes. The framework's generalizability should be tested across disease etiologies.
3. **Integration of Advanced Imaging:** Incorporate MRI-PDFF and MR elastography data where available to improve radiomic feature quality and enhance sensitivity for early-stage disease. This would provide a more definitive comparison with conventional non-invasive tests.
4. **Explainable AI Development:** Enhance model interpretability through advanced visualization and explanation techniques, enabling clinicians to understand and trust the framework's predictions. This includes developing patient-specific explanations and counterfactual reasoning capabilities.

6. Conclusion

This study demonstrates that a multimodal fusion architecture integrating unstructured EHR clinical text, longitudinal biomarkers, and radiomic features through graph neural networks achieves 89.4% accuracy and 0.94 AUC for early non-invasive MASLD detection, substantially outperforming both single-modality approaches and conventional non-invasive tests. The key innovation lies not in any single modality but in the principled integration of complementary information sources, with graph-based reasoning enabling the capture of cross-modal interactions and patient similarity relationships that are inaccessible to simpler fusion strategies. The framework addresses a critical clinical gap by achieving 88.2% sensitivity for early-stage disease (F0-F1 fibrosis), where current non-invasive tests show limited utility, potentially enabling intervention at reversible stages and preventing progression to cirrhosis and hepatocellular carcinoma.

The practical significance of this work extends beyond MASLD detection to establish a replicable template for multimodal disease detection that can be applied across chronic conditions. The framework's reliance on routinely available EHR data makes it particularly

suitable for community-level screening, addressing the urgent public health need for scalable, accessible diagnostic tools. For healthcare administrators, the framework offers a cost-effective alternative to specialized equipment-based approaches. For clinicians, it provides interpretable decision support that enhances—rather than replaces—clinical judgment.

As the global prevalence of MASLD continues to rise, the integration of multimodal AI approaches into routine clinical practice represents a promising pathway toward earlier detection and improved outcomes. This study provides the foundational evidence for such integration, establishing both the technical feasibility and clinical utility of graph-based multimodal fusion for liver disease detection. The framework's extensibility to other conditions and its potential for real-time deployment suggest that this approach may contribute meaningfully to the transformation of chronic disease management in the era of precision medicine.

References

1. Imam, T., Islam, J., Sultana, S., Uddin, M. S., Uddin, M. S., & Uddin, B. (2024, December). Ml-driven solutions for chronic liver disease: Predictive models and prevention strategies. In *2024 IEEE International Conference on Computing (ICOCO)* (pp. 261-266). IEEE.
2. A Novel Approach to Multimodal Biomedical Data Integration via Hybrid Knowledge Graphs and Large Language Models. (2026). *IEEE Xplore*. https://ieeexplore.ieee.org/document/11469136
3. Wang, J. (2026). A Deep Multimodal Fusion Framework Integrating SERS, Clinical Data, and Metabolomics for Noninvasive MASH Diagnosis. *ACS Publications*. https://pubs.acs.org/action/doSearch
4. Bahukhandi, R., et al. (2025). AI-enabled electrical impedance tomography (EIT) system for detecting and quantifying metabolic associated steatotic liver disease. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. doi: 10.1109/EMBC58623.2025.11251537
5. Thiagarajan, J. L. (2026). A BiLSTM–Graph Neural Network Fusion for AI-Powered Remote Diagnosis and Monitoring in Sustainable Healthcare 6.0. In *Taylor & Francis eBooks*. https://www.taylorfrancis.com/chapters/edit/10.1201/9781003642145-4
6. Multimodal Graph Convolutional Networks for Patient Survival Analysis. (2025). *IEEE Xplore*. https://ieeexplore.ieee.org/document/11198055
7. Hepatology. (2021). Application of Artificial Intelligence for the Diagnosis and Treatment of Liver Diseases. *Wiley Online Library*. https://onlinelibrary.wiley.com/doi/full/10.1002/hep.31603
8. Wakabayashi, S., Kimura, T., Tamaki, N., Iwadare, T., Okumura, T., Kobayashi, H., Yamashita, Y., Tanaka, N., Kurosaki, M., & Umemura, T. (2025). AI-Based Platelet-Independent Noninvasive Test for Liver Fibrosis in MASLD Patients. *JGH Open*, *9*(4), e70150. doi: 10.1002/jgh3.70150
9. Chang, et al. (2021). NLP of radiology reports, ICD-9 codes, and laboratory data to identify cirrhosis. *Hepatology*. (cited within reference 7)
10. Chen, et al. (2021). Deep learning radiomic method on 2D-SWE images for staging liver fibrosis. *Hepatology*. (cited within reference 7)
11. Forlano, et al. (2021). Automated software for quantifying steatosis, inflammation, ballooning, and fibrosis in NAFLD. *Hepatology*. (cited within reference 11)

12. Bahukhandi, R., et al. (2025). AI-enabled EIT system for MASLD detection – global prevalence and screening needs. *IEEE Xplore*. https://ieeexplore.ieee.org/abstract/document/11251537