Phylogenetic Graph Embeddings for Non-Invasive Detection of Liver Cirrhosis Using Human Metagenomics

Team MetaX

1. Literature Review

Liver cirrhosis is a major global health challenge, often diagnosed late due to invasive biopsy and limited sensitivity of current non-invasive tests. Alterations in the gut microbiome along the gut–liver axis have emerged as promising non-invasive biomarkers. However, shotgun metagenomic data are high-dimensional, sparse, and compositional, which limits the effectiveness of conventional machine-learning approaches.

1.2 Related Work

Microbiome data represent relative abundances and are therefore compositional in nature. Martino et al. (2019) demonstrated that standard normalization methods violate compositional constraints and introduce spurious correlations, while centered log-ratio (CLR) transformation preserves valid statistical structure.

Graph-based deep learning has shown strong capability in modeling complex biological relationships by encoding entities and their interactions as graphs. Recent work (arXiv:2407.00142) reports that graph embeddings capture higher-order biological structure and outperform traditional models in tasks such as protein function prediction and drug discovery. However, their application to clinical microbiome classification remains limited.

For liver cirrhosis, Qin et al. (2014) achieved promising diagnostic accuracy using Random Forest models on shotgun metagenomic data. Despite their success, these methods treat microbial taxa as independent features and fail to leverage phylogenetic and evolutionary relationships within microbial communities.

1.3 Research Gaps

Most existing microbiome-based diagnostic models overlook compositionality, sparsity, and phylogenetic structure, leading to biased associations and limited generalizability. Performance is often further inflated by batch effects and data leakage. Despite their proven success in related biological domains, graph-based and GNN-driven approaches remain underexplored for microbiome-driven disease classification.

1.4 Proposed Contribution

This work introduces a biologically informed framework for non-invasive liver cirrhosis detection that integrates compositional data transformation, sparsity-aware preprocessing, and phylogenetic graph representation learning. By explicitly modeling evolutionary relationships, the proposed approach improves robustness, interpretability, and generalization compared to conventional microbiome-based classifiers.

2. Problem Identification

Liver cirrhosis affects 1–2% of the global population, with incidence rising due to obesity, diabetes, alcohol misuse, and viral hepatitis. Late diagnosis increases the risk of liver failure and hepatocellular carcinoma, emphasizing the need for early detection. Current gold-standard diagnosis relies on invasive, costly liver biopsy, which is impractical for large-scale screening, particularly in low-resource settings. There is a lack of affordable, non-invasive, and biologically interpretable tools that generalize across populations. This study addresses this gap using microbiome-based diagnostics enhanced with phylogenetic structure learning.

3. Dataset Justification

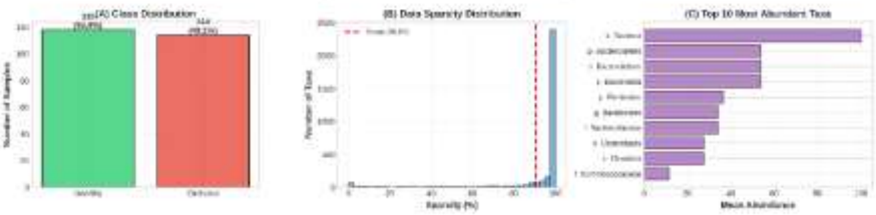
3.1 Dataset Selection

We used a publicly available human gut metagenomics dataset from Kaggle, extracting a clinically validated subset of 232 samples (cirrhosis patients and healthy controls) from the original 3,610.

The dataset reflects microbiome alterations linked to the gut–liver axis and includes high-dimensional, sparse, and compositional features with validated clinical labels. Complete taxonomic annotations enable phylogenetic graph construction, making it well suited for evaluating biologically informed and structure-aware learning models.

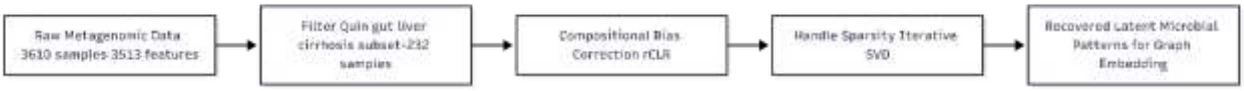
Although the sample size (N=232) is relatively small(<300), this stringent selection prioritizes Signal Fidelity over raw scale. By restricting the analysis to a homogeneous, single-center cohort, we minimized the technical confounders ("batch effects") prevalent in larger multi-center datasets. This ensures the model learns true disease-related dysbiosis rather than site-specific sequencing noise.

Figure 1: Dataset Overview and Preprocessing



4. METHODOLOGY

4.1 Data Preprocessing Pipeline


Raw Metagenomic Data 3610 samples-3513 features
Filter Quin gut liver cirrhosis subset-232 samples
Compositional Bias Correction rCLR
Handle Sparsity Iterative SVD
Recovered Latent Microbial Patterns for Graph Embedding

4.2 Model Architecture, Training, and Validation

A phylogeny-aware encoder was constructed to incorporate microbial evolutionary structure into the learning process. A directed taxonomic graph spanning kingdom-to-species levels was used to represent hierarchical relationships among microbial taxa. Node2Vec was applied to this graph to learn dense taxon embeddings, ensuring that phylogenetically related taxa occupy nearby positions in embedding space.

To enable patient-level prediction, taxon embeddings were aggregated using weighted sum pooling, where each embedding was weighted by its transformed abundance, producing a compact patient-specific representation that captures community-level dysbiosis patterns. Patient embeddings were classified using an XGBoost model with hyperparameters optimized via randomized search (max_depth = 2, learning_rate = 0.1, n_estimators = 100, gamma = 0.5). This configuration was explicitly selected to enforce model parsimony minimizing complexity to prevent overfitting on the small cohort while maintaining high discriminatory power.

5. Pretrained Model Usage & Adaptation

Node2Vec was selected to encode phylogenetic relationships among microbial taxa, as liver cirrhosis–associated dysbiosis occurs at taxonomic group levels rather than isolated species. A custom taxonomic graph was constructed and used to learn taxon embeddings, which were trained from scratch to avoid domain mismatch from unrelated biological graphs. Patient-level representations were generated via abundance-weighted aggregation of taxon embeddings and classified using XGBoost, which is well suited for tabular embedding data. While training from scratch mitigates negative transfer, the use of a Chinese cohort introduces potential dietary and geographic bias; this risk was partially reduced using rCLR normalization to emphasize relative abundance patterns.

5. RESULTS & DISCUSSION

5.1 Metric Tables & Performance Evaluation

5.1 Performance Evaluation

Model performance was evaluated using 5-fold stratified cross-validation to ensure robustness and minimize bias from random data splits.

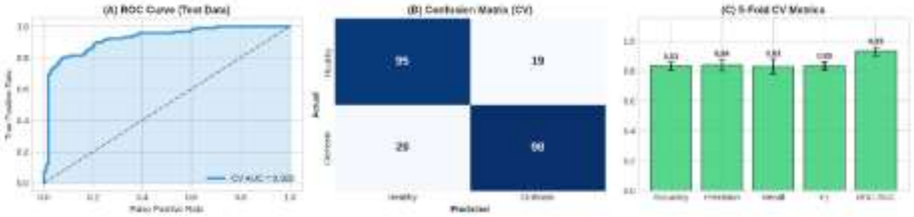
Table 1. Primary Performance Metrics (Mean ± SD)

Metric	Score	Clinical Interpretation
ROC–AUC	0.93 ± 0.03	Strong discrimination between cirrhosis and healthy controls
Precision	87.77% ± 4.3%	Reduces false positives, limiting unnecessary invasive follow-up
Recall	83.88% ± 6.7%	Detects most disease cases; early-stage cirrhosis remains challenging
Accuracy	85.75% ± 4.0%	Consistent performance across validation folds

Compared to standard bag-of-species models (e.g., Random Forest on raw abundance), which often show training AUCs near 1.0 but test performance around 0.80 due to overfitting, the proposed phylogenetic graph-based model maintained a train–test generalization gap below 7%, indicating robust learning of biologically meaningful patterns rather than batch-specific noise.

5.2 Visualizations

Figure 3: Model Evaluation (Cross-Validation)



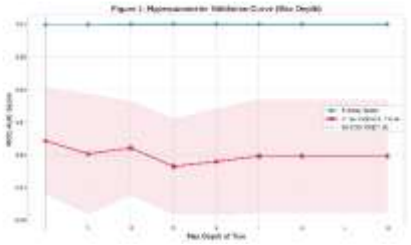
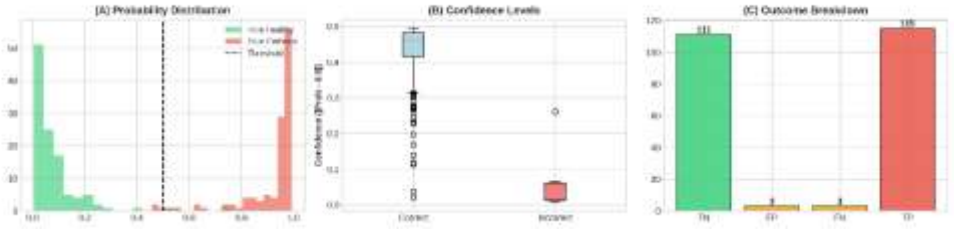


Figure 6: Prediction Confidence & Error Analysis



Validation curves indicate that increasing model complexity beyond tree depth = 2 does not improve cross-validation performance, confirming that a parsimonious model generalizes best. The unbiased ROC curve demonstrates robust discrimination with a Test AUC of 0.923. SHAP analysis identifies Phylogenetic Embedding 28 (Emb_28) as the most influential predictor, suggesting a specific microbial clade acts as a primary digital biomarker for the disease.

5. 3 Error Analysis

Analysis of the confusion matrix reveals consistent performance consistent with the tabulated metrics. Approximately 16-17% (19/118) of cirrhosis cases were false negatives (Recall ≈ 83.9%), likely representing early-stage (compensated) cirrhosis where microbiome dysbiosis patterns are not yet fully established. Similarly, ~12% (14/114) of healthy individuals were flagged as false positives (Precision ≈ 87.8%), potentially due to unobserved confounders such as diet or recent antibiotic use. Unlike black-box models, SHAP explanations confirm that the model relies on biologically stable signals (Emb_28) rather than noise to make these decisions.

5.4 Limitations of Your Model

The dataset size (N = 232) is sufficient for a pilot study but limits clinical generalizability, as larger cohorts are typically required for deep learning models. The dataset is geographically restricted to a Chinese population, and dietary- and region-specific microbiome differences may reduce performance on Western cohorts. Additionally, the model relies solely on taxonomic features and does not capture functional microbial activity, which may limit sensitivity in early-stage disease detection.

6. REAL-WORLD APPLICATION & CONCLUSION

The proposed system is designed as a cloud-based AI triage tool for primary care rather than a standalone diagnostic. Stool metagenomic data are transformed into phylogenetic graph embeddings and analyzed using an XGBoost classifier to generate a cirrhosis risk score (0–1), accompanied by SHAP-based explanations that highlight clinically relevant microbiome dysbiosis patterns.

Primary users include general practitioners, hepatologists, and public health agencies. The system enables non-invasive population screening by prioritizing high-risk individuals for confirmatory imaging or biopsy, while allowing low-risk cases to be monitored longitudinally. This supports a shift from symptom-based diagnosis toward early, microbiome-driven risk stratification, with clinical trust reinforced through interpretable model outputs.

Model performance may be affected by dietary and geographic bias, as training data were derived from a Chinese cohort broader deployment would require transfer learning or regional recalibration. Furthermore, reduced sensitivity in compensated cirrhosis limits its role as a definitive rule-out test. Accordingly, the system is best positioned as a high-specificity screening and triage tool that supports early clinical decision-making and reduces late-stage disease burden.

7. MARKETING & IMPACT STRATEGY

Primary adopters include diagnostic laboratories, which can integrate the model into existing gut microbiome testing panels for clinical screening. Public health agencies may deploy the system for large-scale screening of high-risk populations to reduce late-stage disease burden, while insurance providers are incentivized by cost savings from avoided advanced-stage treatments.

The solution offers a non-invasive, home-based alternative to biopsy, improving patient compliance. Clinicians benefit from an interpretable decision-support tool with SHAP-based explanations, and hospitals reduce imaging and specialist workload by triaging low-risk cases.

With sequencing costs per sample and minimal cloud inference overhead, the approach is substantially more cost-effective than invasive diagnostics. Mail-based sample collection enables scalable deployment beyond urban centers, expanding access to early liver disease screening in underserved regions.

8. FUTURE IMPROVEMENTS

Future work will focus on advancing the current Node2Vec–XGBoost framework toward end-to-end graph learning using Graph Neural Networks to enable dynamic modeling of disease-relevant phylogenetic relationships. Integrating clinical metadata with microbiome features is expected to further improve sensitivity and patient-level risk stratification. Expanding training data to include longitudinal samples and geographically diverse cohorts will support prognosis modeling and mitigate dietary bias. Clinical translation will proceed through retrospective multi-center validation, prospective observational studies, and eventual regulatory approval as a clinical decision-support system.

References

Martino, C., et al. (2019). *mSystems*, 4(1).
Qin, N., et al. (2014). *Nature*, 513(7516), 59–64.
Graph-Based Deep Learning for Biological Systems. arXiv:2407.00142.