Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 9, No. 2, 2026

Graph representations and propagation architectures for predicting three-level liver stiffness reference categories in NHANES

Linjie Tang, Zhitao Yu

College of Electrical Engineering, Southwest Minzu University, Chengdu 610225, China

Abstract: Objective To compare feature graphs and patient similarity graphs across representative graph neural network architectures for predicting liver stiffness reference categories, with attention to probability scoring and identification o f the high category. Methods We used 24 clinical predictors to estimate three liver stiffness measurement (LSM) categories, low (<8 kPa), intermediate (8 to <12 kPa), and high (≥12 kPa). The development cohort comprised 6,812 NHANES participants from 2017–March 2020. An exploratory cross-period cohort comprised 4,557 participants from August 2021–August 2023. Five propagation architectures were evaluated with both graph representations. Prior-informed graph attention (PIGA), two multi-hop propagation models, and three tabular baselines brought the comparison to 16 models. Five-fold cross-validation grouped participants by survey stratum and primary sampling unit (PSU). The primary metric was unweighted multiclass negative log-likelihood (NLL). Paired intervals were estimated by stratified PSU resampling of fixed out-of-fold predictions. Results Graph representation rankings depended on architecture. Patient graphs reduced NLL by 0.01233 for GAT (95% interval 0.00857–0.01622) and 0.00875 for GATv2 (0.00510–0.01244). The GIN feature graph reduced NLL by 0.02081 (0.01329–0.02856). These three differences retained their directions across periods. Feature graphs had higher out-of-fold high-category recall point estimates in all five architecture pairs. PIGA had the highest out-of-fold quadratic weighted kappa (0.3031) and high-category recall (25.07%). Logistic regression had the lowest out-of-fold NLL (0.3143), followed among graph models by PIGA (0.3173). The highest intermediate-category recall was 2.46%. Conclusion Feature and patient graph rankings varied with propagation architecture and evaluation objective. Joint assessment of probability scores and category-specific identification provides a basis for selecting graph models for this NHANES LSM reference-category task.

Keywords: Graph neural networks, tabular learning, liver stiffness, class imbalance, complex surveys, NHANES.

1. Introduction

Vibration-controlled transient elastography (VCTE) provides a liver stiffness measurement (LSM) that supports assessment of liver fibrosis [1–3]. Routine laboratory and anthropometric measurements can be used to predict LSM reference categories. The National Health and Nutrition Examination Survey (NHANES) provides VCTE alongside demographic, examination, and laboratory data, allowing models to be compared using a common set of inputs.

Graph representations of clinical tabular data can aggregate information among variables within a participant or among participants with similar measurements. Clinical graph learning and tabular graph methods have used a range of node definitions and relationships [4–7]. Feature graphs and patient graphs differ in the entities represented by nodes, their neighborhoods, and the way predictions are obtained. Their relative performance therefore warrants comparison within the same clinical task across several propagation architectures.

Model selection also depends on the evaluation objective. Negative log-likelihood (NLL) scores predicted probabilities, the area under the receiver operating characteristic curve (AUROC) measures ranking, and category-specific recall measures identification under a specified classification rule. In an imbalanced outcome, overall accuracy is strongly influenced by the most common category. Probability ranking and minority-category identification can consequently yield different model rankings. Including tabular baselines and category-specific metrics helps establish which aspects of prediction are associated with a graph representation [8].

We compared feature graphs and patient graphs across five propagation architectures using the same 24 predictors, three LSM reference categories, and PSU-grouped cross-validation. PIGA, multi-hop propagation models, and three tabular models provided additional comparators. We examined whether NLL rankings between graph representations varied by architecture and whether high-category identification followed the same rankings. Survey-weighted and exploratory cross-period analyses assessed performance under different evaluation weights and survey periods. This design connects graph representation, propagation architecture, and evaluation objective in model selection for clinical tabular prediction.

2. Related Work

2.1. Prediction of Elastography Phenotypes

Studies with paired liver biopsies have examined the relationship between VCTE measurements and fibrosis stages [3]. Using NHANES, Han et al. predicted continuous LSM with XGBoost and examined models incorporating different insulin resistance indices. Noureddin et al. described elastography-defined steatosis and fibrosis, with machine learning used to classify controlled attenuation parameter (CAP)-defined steatosis [9, 10]. These studies support prediction of elastography phenotypes from routine measurements. The present study estimates probabilities of three LSM reference categories and compares graph representations using a common predictor set.

2.2. Nodes and Relationships in Clinical Graph Learning

Clinical graph models can represent medical events within an encounter or individual participants. The graph convolutional Transformer developed by Choi et al. represents medical events within a visit and uses data-derived priors to guide structure learning [4]. Lu and Uddin constructed weighted patient networks from medical claims and aggregated information across patients to predict chronic diseases [5]. In NHANES, Chen et al. represented anatomically adjacent periodontal sites as graph nodes for prediction using mortality follow-up [6]. In our task, functional groups represent relationships among variables within a participant, whereas training-fold nearest neighbors represent similarity among participants. Both graph representations use the same routine measurements.

2.3. Tabular Graph Construction and Model Benchmarking

Table 2 Graph represents table columns as nodes and learns a shared weighted feature graph for sample representation [7]. In comparisons across tabular tasks, Grinsztajn et al. found gradient-boosted trees to be strong competitors to deep learning models [8]. We apply each of five propagation architectures to feature and patient graphs, with logistic regression, XGBoost, and a multilayer perceptron (MLP) as tabular baselines, to examine relative performance and probability prediction quality.

3. Data and Methods

3.1. Data Sources, Cohorts, And Predictors

We used the NHANES 2017–March 2020 pre-pandemic combined files and the August 2021–August 2023 files [1, 2, 11, 12]. Demographic, VCTE, body measurement, standard biochemistry, glycated hemoglobin, total cholesterol, high-density lipoprotein cholesterol, and complete blood count data were linked by participant identifier (SEQN). Eligibility required age ≥20 years, a complete VCTE examination (LUAXSTAT=1), and a finite, positive median LSM. NHANES defines a complete examination by fasting for at least 3 h, at least 10 valid measurements, and an interquartile range-to-median ratio <30% [1, 2]. Participants missing any predictor were excluded.

In the development period, 7,396 participants met the age and imaging criteria, and 6,812 had complete predictors. These participants represented 24 masked strata and 49 stratum–PSU combinations. The corresponding cross-period numbers were 5,277 eligible participants and 4,557 complete cases, representing 15 masked strata and 30 stratum–PSU combinations. Cohort selection is summarized in Figure 1A.

All models used the 24 predictors in Table 1. Aspartate aminotransferase (AST) and alanine aminotransferase (ALT) corresponded to LBXSASSI and LBXSATSI and were used to calculate the AST/ALT ratio (AAR) [13, 14]. AST, ALT, GGT, ALP, TBIL, GLU, HbA1c, TC, HDL, CR, UREA, UA, K, NA, CL, PHOS, PLT, and FE were transformed as ln(1+max(x,0)). The remaining predictors retained their original scales. Predictors were then standardized using the mean and standard deviation of the relevant training partition. LSM and CAP were excluded from the inputs.

Table 1. Clinical predictors, groups, and units

Group	Predictor	Abbreviation	Unit or coding
Liver biochemistry	Aspartate aminotransferase	AST	U/L
	Alanine aminotransferase	ALT	U/L
	γ-Glutamyl transferase	GGT	U/L
	Alkaline phosphatase	ALP	U/L
	Total bilirubin	TBIL	μmol/L
	Albumin	ALB	g/L
	Total protein	TP	g/L
Glucose and lipid metabolism	Standard biochemistry glucose	GLU	mmol/L
	Glycated hemoglobin	HbA1c	%
	Total cholesterol	TC	mg/dL
	High-density lipoprotein cholesterol	HDL	mg/dL
Renal and related markers	Creatinine	CR	μmol/L
	Urea nitrogen	UREA	mmol/L
	Uric acid	UA	μmol/L
Electrolytes and phosphorus	Potassium	K	mmol/L
	Sodium	NA	mmol/L
	Chloride	CL	mmol/L
	Phosphorus	PHOS	mmol/L
Blood-related markers	Platelet count	PLT	10^9/L
	Serum iron	FE	μmol/L
Demographics and anthropometrics	Age	AGE	Years
	Sex	SEX	1=male, 2=female
	Body mass index	BMI	kg/m²
Derived predictor	AST/ALT ratio	AAR	Dimensionless

Note. TC and HDL were retained in mg/dL. AAR was calculated from the original AST and ALT measurements.

3.2. Prediction Task

The prediction unit was the participant. Median LSM from

the same survey examination was encoded into three ordered reference categories, low (<8 kPa), intermediate (8≤LSM<12 kPa), and high (≥12 kPa). The 8 and 12 kPa thresholds were drawn from LSM stratification used in clinical pathways for steatotic liver disease and were used here to define contemporaneous elastography reference categories [15]. Models estimated conditional probabilities $p_k(x) = P(Y = k|x)$ for k=0, 1, 2, summing to 1. All models used a multiclass objective and produced three probabilities. NLL was calculated from the probability assigned to the observed category, and the largest probability determined the predicted category. Category distributions are reported in Table 2.

Table 2. LSM reference categories among complete-case participants

Reference category	Development cohort			Exploratory cross-period evaluation		
	n	Unweighted %	Weighted %	n	Unweighted %	Weighted %
Low (<8 kPa)	6,118	89.81	90.86	3,984	87.43	88.16
Intermediate (8 to <12 kPa)	460	6.75	5.77	359	7.88	7.36
High (≥12 kPa)	234	3.44	3.37	214	4.70	4.48
Total	6,812	100.00	100.00	4,557	100.00	100.00

Note. Weighted percentages were calculated among participants with complete predictors in each period. Percentages were rounded independently.

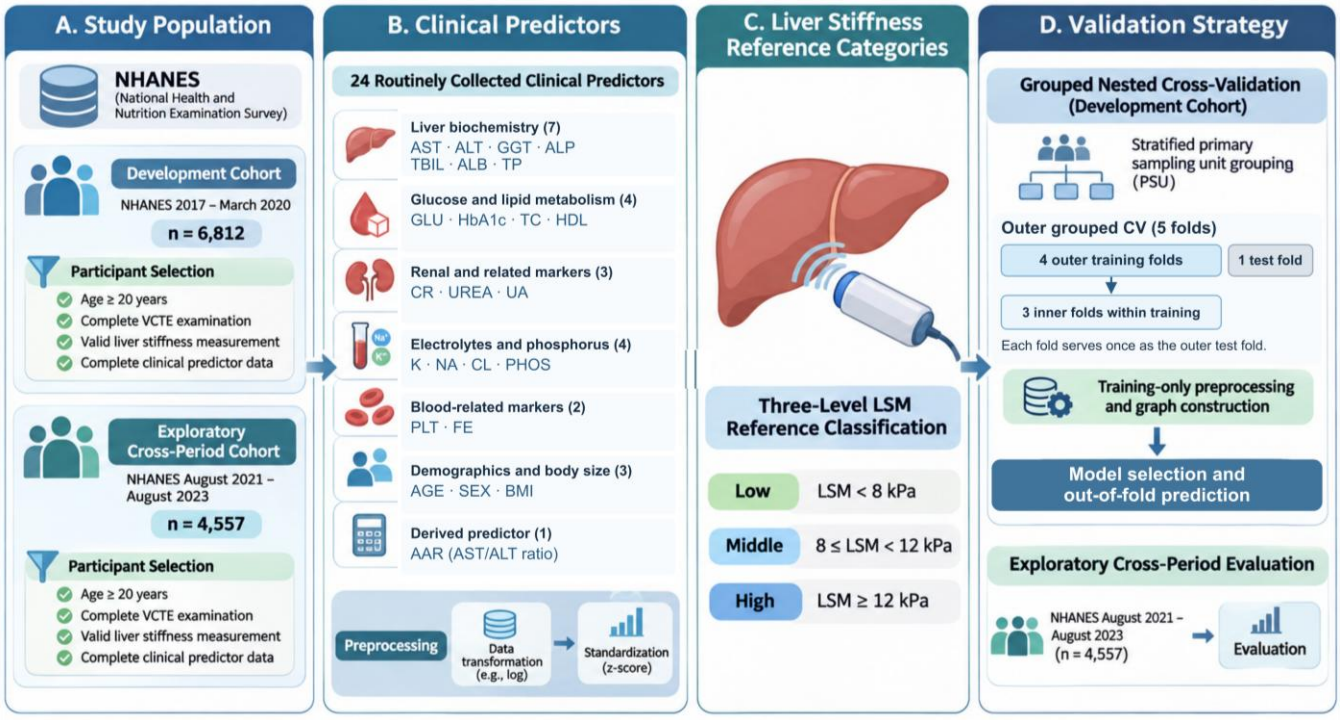

Figure 1. Study cohorts, clinical predictors, and validation design.

A, The age and imaging criteria identified 7,396 and 5,277 eligible participants in the two periods, of whom 6,812 and 4,557 had complete predictors. B, The 24 predictors and preprocessing. C, VCTE-derived LSM reference categories. D, Development-cohort cross-validation grouped participants by survey stratum and PSU, with three inner folds restricted to the outer training partition. The later period was used for exploratory evaluation.

3.3. Graph Construction and Prediction

Feature graphs. Each participant was represented by a graph of 24 variable nodes. For the five paired architectures, the initial representation of variable j combined a value projection and a learned variable identity embedding, $h_j^{(0)} = W_v z_j + b_v + e_j$, with dimension 64. Encoding parameters were shared across participants. Liver biochemistry was divided into AST/ALT, GGT/ALP/TBIL, and ALB/TP groups. Additional groups represented glucose and lipid metabolism, renal and related markers, electrolytes and phosphorus, and blood-related markers. All pairs of nodes within a group were connected bidirectionally. AAR was connected to AST and ALT. Age, sex, and BMI had no group edges. The final node representations were mean-pooled to predict category probabilities. PIGA combined clinical groups and absolute training-fold Spearman correlations into a prior matrix that fixed first-layer attention. Its subsequent two layers learned attention weights, and its classifier combined graph features with a parallel tabular MLP branch.

Patient graphs. Each participant was a node represented by 24 standardized predictors. We selected 20 nearest neighbors per training node by Euclidean distance, excluded the node itself, and used the symmetric union of neighbor relationships. Validation and test nodes received messages from their training neighbors. Edges among held-out participants were excluded, and held-out participants sent no messages to training nodes. Edges were determined from predictors. Graphs were constructed separately within each inner and outer training partition. Cross-period evaluation used the full development cohort as the reference graph.

Each architecture pair shared propagation depth, hidden dimension, optimizer, and data partitions. Node encoding, prediction readout, and training batches were adapted to the graph representation. The paired comparisons therefore evaluated the complete feature-graph and patient-graph models under each architecture. Figure 2 contrasts their nodes, relationships, and outputs.

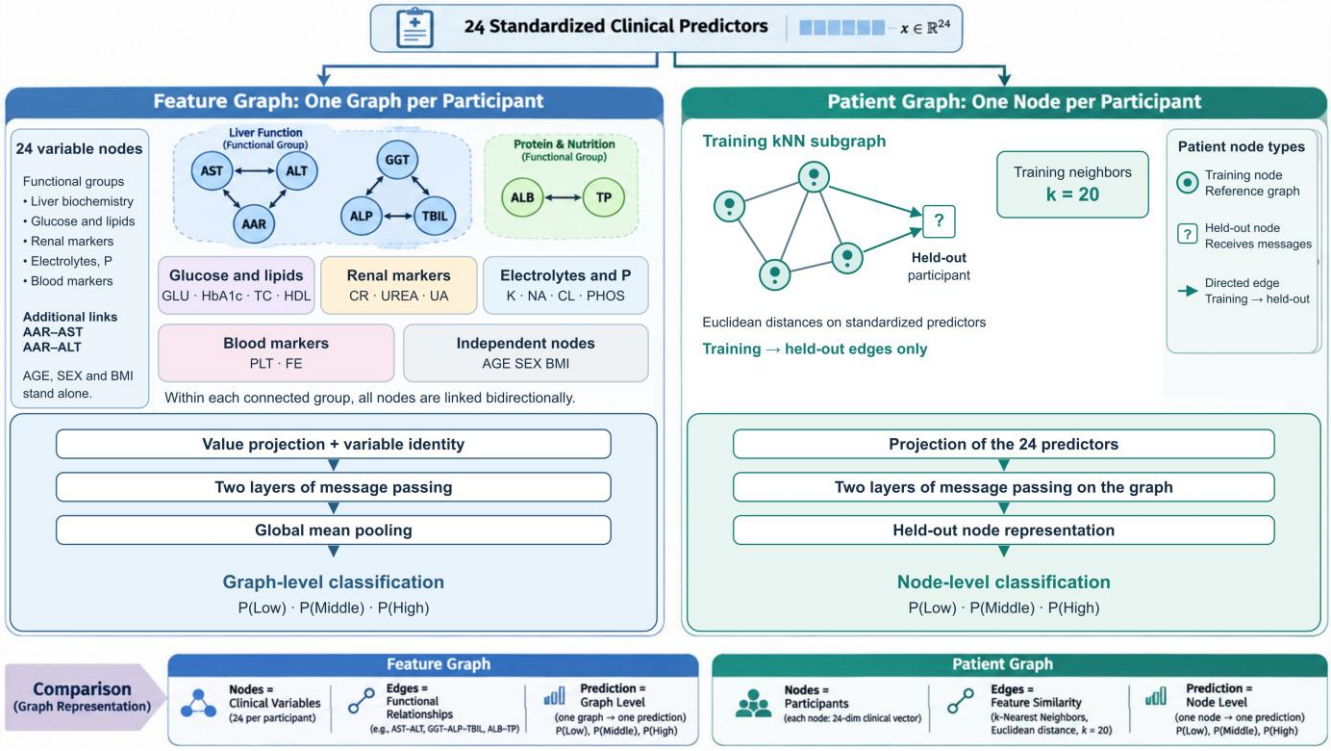

Figure 2. Feature and patient graph construction.

Feature graphs represent clinical variables as nodes and pool node features for graph-level classification. Patient graphs represent participants as nodes, construct 20-nearest-neighbor relationships using standardized training predictors, and classify held-out nodes using incoming training-node messages. The two-layer propagation diagrams summarize the shared setting of the five paired architectures.

3.4. Propagation Architectures and Tabular Comparators

Five architectures were applied to both graph representations, yielding five paired comparisons. PIGA was evaluated on feature graphs, and APPNP and GPR-GNN on patient graphs, for a total of 13 graph models (Table 3). The model set and evaluation are summarized in Figure 3. Tabular comparators were L2-regularized multinomial logistic regression, XGBoost [16], and an MLP with two hidden layers. All used the same 24 predictors.

GCN [17], GraphSAGE [18], GIN [19], GAT [20], and GATv2 used two neighborhood aggregation layers with residual connections and layer normalization. GCN aggregated using normalized adjacency weights. GraphSAGE combined neighbor means with the central node representation. GIN used an MLP to update the sum of neighbor representations and a weighted central-node representation. GAT and GATv2 used four attention heads. In GATv2 [21], the ranking of neighbor attention weights can depend on the query node.

PIGA adopted prior-guided attention and interlayer Kullback–Leibler (KL) constraints from the graph convolutional Transformer [4]. In the clinical grouping matrix G, off-diagonal within-group entries and the AAR–AST and AAR–ALT entries were 1, with other entries 0. Off-diagonal entries of C were absolute Spearman correlations among training-sample variables, and its diagonal entries were 1. The mixture $S=0.6G+0.4C$ was row-normalized to obtain P, retaining diagonal entries. First-layer attention was fixed at $A^{(1)}=P$. Layers 2 and 3 learned attention weights from node features. The training objective combined cross-entropy with $0.05[KL(P\|A^{(2)})+KL(A^{(2)}\|A^{(3)})]$, averaging each KL divergence across attention-matrix rows. Value, variable identity, and variable type embeddings were concatenated into 192-dimensional node representations. Mean-pooled graph features were concatenated with the 64-dimensional output of an MLP receiving all predictors, then classified. PIGA was evaluated as an additional comparator alongside the five architecture pairs.

Feature-graph GCN used conventional normalization with self-loops. Patient-graph GCN and GPR-GNN used edge weights frozen from the training graph. Training-node degrees were defined as training-graph in-degree plus 1, and validation or test node degrees as $k+1=21$. An edge weight was the reciprocal square root of the product of its endpoint degrees. These two patient-graph models used no additional self-loops. Training-node features remained fixed when validation or test nodes were added. The two GCN representations thus used their respective normalization settings.

APPNP and GPR-GNN first obtained three class scores from an MLP with one 64-unit hidden layer, then propagated those scores on the patient graph. APPNP used 10 iterations, retained the initial scores with restart weight 0.1 at each iteration, and included self-loops and propagation dropout [22]. GPR-GNN learned a linear combination of scores propagated over hops 0–10 [23]. Initial coefficients were 0.1×0.9^{h} for $h=0, \ldots, 10$. The remaining mass beyond hop 10 was excluded from the final initial coefficient.

Table 3. Graph representations, model sizes, and propagation mechanisms

Architecture	Representation	Hidden dimension and depth	Propagation mechanism
GCN [17]	Feature and patient	64 dimensions, 2 layers	Normalized adjacency aggregation
GAT [20]	Feature and patient	64 dimensions, 2 layers	Four-head neighborhood attention
GraphSAGE [18]	Feature and patient	64 dimensions, 2 layers	Neighbor means and central-node updates
GIN [19]	Feature and patient	64 dimensions, 2 layers	Neighbor sum, learned central coefficient, and MLP update
GATv2 [21]	Feature and patient	64 dimensions, 2 layers	Four-head attention with query-dependent neighbor ranking
PIGA	Feature	192-dimensional nodes, 3 layers, 64-dimensional tabular branch	Fixed first-layer prior, learned dense attention, and interlayer KL constraints
APPNP [22]	Patient	64-unit prediction hidden layer, 10 propagation steps	Fixed personalized PageRank score propagation
GPR-GNN [23]	Patient	64-unit prediction hidden layer, hops 0–10	Learned multi-hop score weights

Note. Five paired architectures yielded 10 graph models. PIGA, APPNP, and GPR-GNN provided three additional graph models.

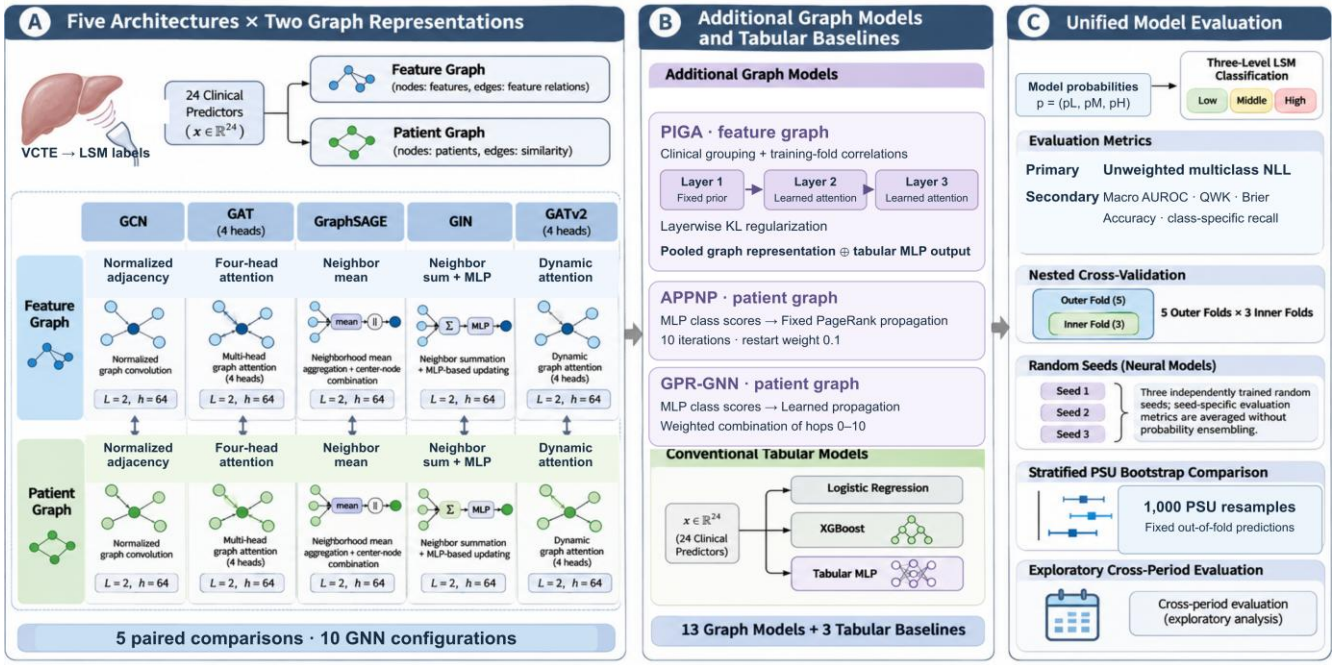

Figure 3. Architecture pairs and model evaluation.

A, Five architectures were evaluated with both graph representations. B, PIGA, APPNP, GPR-GNN, and three tabular baselines. PIGA concatenated its graph and tabular branches before classification. C, The primary metric was unweighted multiclass NLL. Neural-model metrics were calculated by seed and averaged. Paired NLL intervals used stratified PSU resampling of fixed out-of-fold predictions.

3.5. Validation and Model Selection

The development cohort underwent five-fold stratified group cross-validation using StratifiedGroupKFold. The combination of survey period, masked stratum, and PSU formed an indivisible group. Within each outer training partition, three-fold grouped validation selected the training duration for neural models, regularization strength for logistic regression, and tree depth and boosting rounds for XGBoost. Neural models were refitted using three random seeds per outer fold. For each seed, predictions from the five held-out folds were pooled to calculate out-of-fold (OOF) metrics. Metrics were then averaged across the three seeds. Probability predictions were retained separately by seed. Logistic regression and XGBoost each yielded one set of OOF predictions.

Inner neural training was capped at 100 epochs and stopped after 15 epochs without improvement in validation NLL. The median of the three best inner-fold epochs, rounded upward, determined the refitting duration in each outer training partition. Logistic regression selected C from 0.1, 1, and 10 by inner OOF NLL. XGBoost selected depth from 3, 4, and 5, with learning rate 0.05 and row and column subsampling fractions of 0.8. Inner training was capped at 500 boosting rounds, with early stopping after 25 rounds without improvement. The median best round count for the selected depth was rounded upward.

Cross-period models were refitted on the full development cohort. Neural training duration was the upward-rounded median of the selected durations across the five outer folds. Logistic regression and XGBoost hyperparameters were chosen by mean inner OOF NLL across the five outer training partitions. The 2021–2023 period had been examined in earlier analyses and was used here for exploratory cross-period evaluation.

3.6. Metrics, Survey Weights, And Paired Comparisons

The primary metric was unweighted multiclass NLL,

$-n^{-1}\sum_{i=1}^{n}\log p_{y_i}(x_i)$, the mean negative log-probability assigned to the observed category. Lower values indicate better probability scores. Secondary metrics were macro-average one-vs-rest AUROC, quadratic weighted kappa (QWK), multiclass Brier score, accuracy, and recall for each category. The Brier score summed squared differences between the three predicted probabilities and the one-hot outcome, then averaged over participants. At cumulative LSM thresholds of ≥8 and ≥12 kPa, probabilities for the relevant categories were summed to calculate binary AUROC and average precision (AP).

Survey-weighted evaluation used WTMECPRP examination weights for development and WTPH2YR two-year phlebotomy weights for the cross-period cohort. These metrics describe weighted performance among participants with complete predictors. Training and model selection used unweighted losses, and survey weights received no additional adjustment for missing imaging or predictors [11–14]. NLL and Brier score jointly reflect discrimination and calibration and were used to assess overall probability prediction quality.

For each architecture pair, the contrast was feature-graph NLL minus patient-graph NLL. Comparisons with logistic regression used the logistic regression model fitted in the corresponding experiment. We drew 1,000 bootstrap resamples of PSUs with replacement within each development-cohort stratum, retaining the original number of PSUs per stratum. Paired differences were calculated on the same resampled participants using fixed OOF predictions. Percentiles 2.5 and 97.5 defined the 95% intervals. Within each resample, neural-model NLL was averaged across seeds before calculating contrasts, matching the primary aggregation. These intervals quantify sampling-unit variation conditional on fixed predictions. Refitting variation was excluded, and intervals were unadjusted for multiple comparisons.

3.7. Training and Implementation

The five paired architectures used GELU activations, residual connections, layer normalization, and dropout 0.2. PIGA used 64-dimensional value, identity, and type embeddings, with dropout 0.1. The tabular MLP had two 64-unit hidden layers, GELU activations, dropout 0.2, and layer normalization in the first hidden layer. Prediction MLPs in APPNP and GPR-GNN used dropout 0.2. All neural models used AdamW with learning rate 0.001 and weight decay 0.0001 to minimize unweighted cross-entropy. PIGA additionally used the KL regularization term described above. Patient-graph models updated parameters using the full training graph at each step. Feature-graph models and the tabular MLP used batches of 256.

The outer split seed was 42. Inner split seeds were 10000 plus the zero-based outer fold index. Inner neural fits used seed 917, and outer and cross-period fits used seeds 0, 1, and 2. The PSU bootstrap seed was 20261116. Models were implemented using PyTorch, PyTorch Geometric, scikit-learn, and XGBoost.

4. Results

4.1. Graph Representation Rankings Depend on Propagation Architecture

In OOF evaluation, the GAT and GATv2 patient graphs had lower NLL by 0.01233 [0.00857, 0.01622] and 0.00875 [0.00510, 0.01244], respectively. The GIN feature graph had lower NLL by 0.02081 [0.01329, 0.02856]. All three contrasts retained their directions in the cross-period evaluation. The OOF intervals crossed zero for GCN and GraphSAGE, and the GCN point estimate changed direction across periods (Figure 4). Feature graphs had higher OOF high-category recall point estimates in all five architecture pairs (Figure 6).

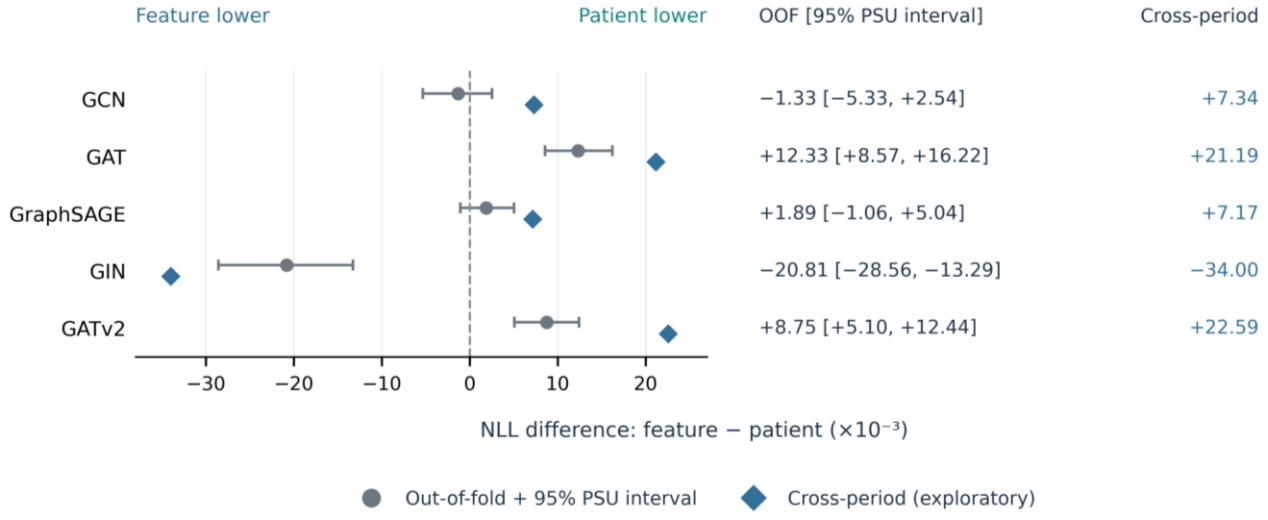

Figure 4. Paired comparisons of graph representations.

Differences are feature-graph minus patient-graph unweighted NLL, with negative values favoring feature graphs. Circles and error bars show OOF estimates and conditional 95% PSU intervals in 6,812 development participants. Diamonds show exploratory cross-period estimates in 4,557 participants. Values are expressed in units of 10^{-3}. Neural-model metrics were averaged across three seeds. Intervals used 1,000 stratified PSU resamples of fixed predictions, excluded refitting variation, and were unadjusted for multiple comparisons. Cross-period intervals were not estimated.

4.2. Probability Scores and High-Category Identification Yield Different Rankings

Logistic regression had the lowest OOF NLL (0.3143), and PIGA had the lowest NLL among graph models (0.3173). The paired PIGA-minus-logistic-regression contrast was 0.00302, with a 95% interval of [−0.00008, 0.00579]. PIGA had the highest OOF QWK (0.3031) and high-category recall (25.07%) point estimates, compared with 0.2768 and 21.37%

for logistic regression. XGBoost had the highest macro AUROC point estimate (0.8319) (Table 4).

The GCN feature graph had the lowest NLL among the 10 models in the five architecture pairs (0.3208). Its NLL exceeded logistic regression by 0.00656 [0.00209, 0.01092]. For 12 of the 13 graph models, the lower bound of the paired NLL interval relative to logistic regression was above zero. Figure 5 shows all NLL contrasts and an OOF detail panel. QWK, AUROC, and recall rankings were compared descriptively using point estimates.

Table 4. Predictive performance of the 16 models

Model	Out-of-fold evaluation			Exploratory cross-period evaluation		
	NLL (95% interval) ↓	AUROC ↑	QWK ↑	NLL ↓	AUROC ↑	QWK ↑
Logistic regression	0.3143 [0.292, 0.336]	0.8284	0.2768	0.3849	0.7937	0.2936
XGBoost	0.3166 [0.293, 0.338]	0.8319	0.1870	0.3882	0.7905	0.1702
Tabular MLP	0.3166 [0.294, 0.338]	0.8274	0.2148	0.3944	0.7913	0.2511
PIGA feature	0.3173 [0.294, 0.340]	0.8261	0.3031	0.3941	0.7932	0.2769
GCN feature	0.3208 [0.296, 0.344]	0.8270	0.1904	0.4027	0.7860	0.2284
GAT feature	0.3351 [0.308, 0.359]	0.8068	0.1579	0.4291	0.7658	0.1691
GraphSAGE feature	0.3241 [0.299, 0.346]	0.8197	0.2015	0.4115	0.7814	0.2221
GIN feature	0.3247 [0.301, 0.346]	0.8205	0.2018	0.3989	0.7837	0.2495
GATv2 feature	0.3333 [0.306, 0.357]	0.8096	0.1743	0.4278	0.7709	0.1794
GCN patient	0.3222 [0.299, 0.344]	0.8190	0.1711	0.3954	0.7787	0.1898
GAT patient	0.3228 [0.298, 0.346]	0.8232	0.1465	0.4079	0.7762	0.1542
GraphSAGE patient	0.3222 [0.298, 0.345]	0.8236	0.1415	0.4043	0.7802	0.1771
GIN patient	0.3455 [0.317, 0.373]	0.7961	0.0203	0.4329	0.7670	0.0388
GATv2 patient	0.3246 [0.300, 0.348]	0.8196	0.1138	0.4052	0.7802	0.1322
APPNP patient	0.4055 [0.388, 0.422]	0.7026	0.0000	0.4587	0.6913	0.0000
GPR-GNN patient	0.3700 [0.350, 0.389]	0.7441	0.0000	0.4328	0.7243	0.0000

Note. AUROC is the macro-average one-vs-rest AUROC. NLL intervals are conditional on fixed OOF predictions. Neural-model metrics are averages across seeds, with probability predictions retained separately. Bold values indicate the best point estimate in each column.

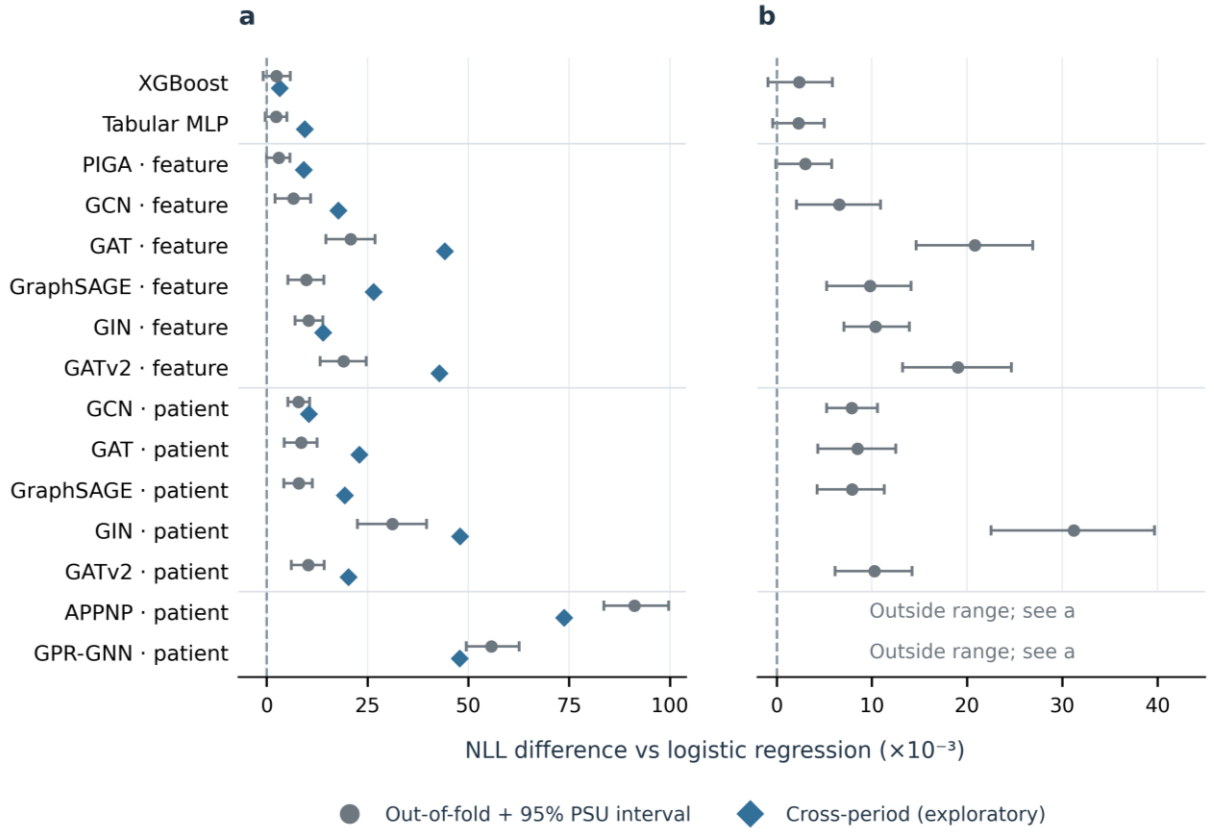

Figure 5. Model NLL contrasts relative to logistic regression.

Each contrast subtracts logistic regression NLL in the corresponding analysis. Positive values indicate higher NLL, in units of 10^{-3}. Panel a shows all 15 comparator models, with OOF circles and conditional intervals and exploratory cross-period diamonds. Panel b enlarges the OOF range from −2 to 45. APPNP and GPR-GNN lie beyond that range and are shown in panel a. Panels share model order and have different horizontal ranges. The development and cross-period cohorts

contained 6,812 and 4,557 participants. Neural-model metrics were averaged across three seeds. Logistic regression and XGBoost each had one prediction set. OOF intervals used 1,000 stratified PSU resamples of fixed predictions, excluded refitting variation, and were unadjusted for multiple comparisons. Cross-period intervals were not estimated.

4.3. Probability Ranking and Three-Category Classification

OOF accuracy ranged from 89.77% to 90.21%, close to the low-category proportion of 89.81%. The highest intermediate-category recall was 2.46%. APPNP and GPR-GNN assigned every participant to the low category in all three OOF seed runs, giving zero recall for the intermediate and high categories. Their macro AUROCs were 0.7026 and 0.7441, respectively, reflecting variation in the ranking of predicted probabilities. Both models also classified every cross-period participant as low (Figure 6). Logistic regression had the highest cross-period high-category recall, at 22.43%.

For the binary outcome LSM≥12 kPa, OOF AUROC was 0.8825 for logistic regression and 0.8819 for PIGA, with AP of 0.3401 and 0.3432. Their high-category recalls under three-category maximum-probability classification were 21.37% and 25.07%. Probability ranking for a cumulative binary outcome and three-category assignment thus described different aspects of performance.

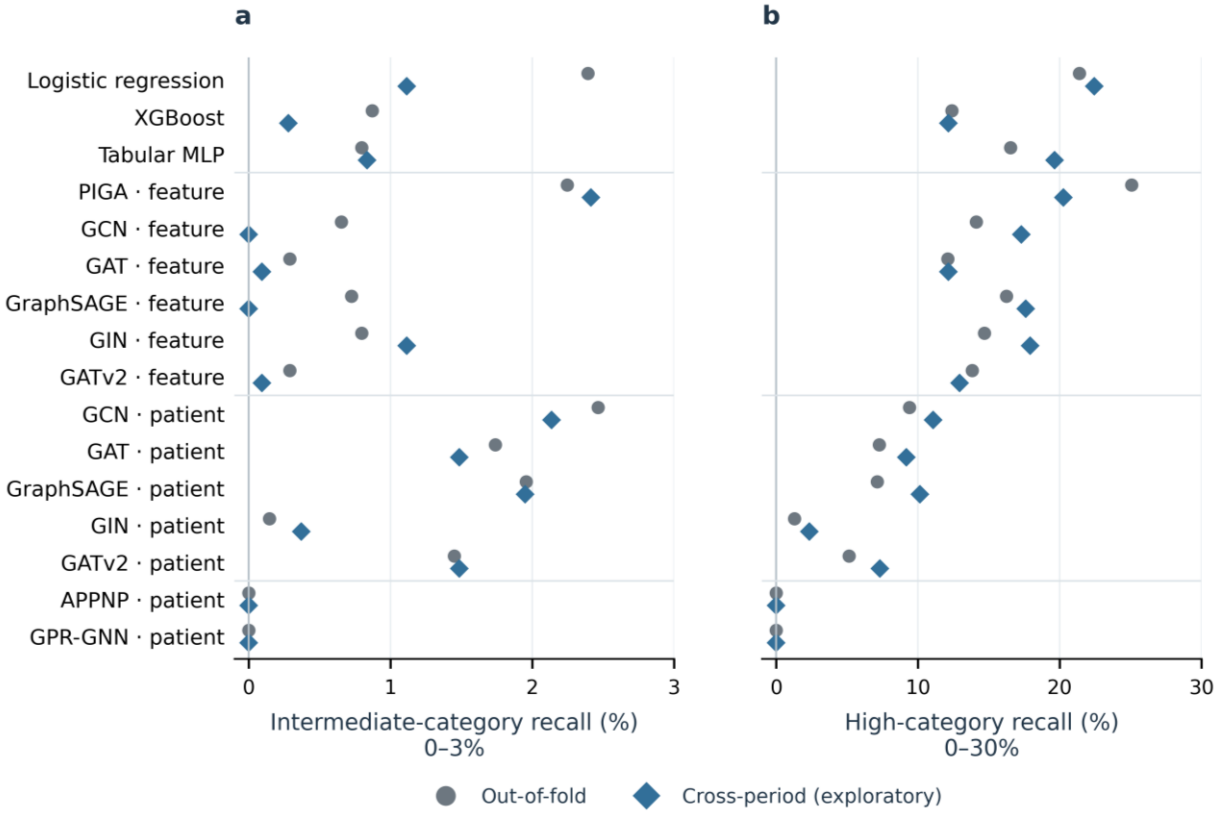

Figure 6. Intermediate- and high-category recall across periods.

Panel a shows intermediate-category recall on a 0–3% axis. Panel b shows high-category recall on a 0–30% axis. Circles denote OOF estimates and diamonds denote exploratory cross-period estimates, compared descriptively. Model order is shared between panels. The intermediate-category counts were 460 and 359, and the high-category counts were 234 and 214, in the development and cross-period cohorts. Recall used the maximum-probability category. Neural-model recall was averaged across three seeds, with probability predictions retained separately. Logistic regression and XGBoost each had one prediction set.

4.4. Survey-weighted and Cross-Period Performance

XGBoost had lower weighted OOF NLL than logistic regression (0.2901 versus 0.2907) and the highest weighted macro AUROC (0.8447). PIGA had the highest weighted QWK (0.2413). In cross-period evaluation, logistic regression had the lowest weighted NLL (0.3632) and highest weighted QWK (0.3031), whereas PIGA had the highest weighted macro AUROC (0.8059). Weighted NLL and macro AUROC for all models are reported in Table 5.

Relative to OOF evaluation, cross-period unweighted NLL increased by 0.0532–0.0945 across models. Brier scores increased, and macro AUROC and accuracy decreased. Logistic regression had the lowest unweighted cross-period NLL (0.3849). Every graph model had a positive NLL contrast relative to logistic regression in both periods (Figure 5). Changes in QWK and minority-category recall varied by model (Table 4 and Figure 6).

Table 5. Survey-weighted NLL and macro AUROC

Model	Out-of-fold evaluation		Exploratory cross-period evaluation	
	Weighted NLL ↓	Weighted AUROC ↑	Weighted NLL ↓	Weighted AUROC ↑
Logistic regression	0.2907	0.8369	0.3632	0.8053
XGBoost	0.2901	0.8447	0.3654	0.8053
Tabular MLP	0.2908	0.8394	0.3727	0.8024
PIGA feature	0.2917	0.8365	0.3716	0.8059
GCN feature	0.2948	0.8377	0.3810	0.7998
GAT feature	0.3079	0.8236	0.4099	0.7756
GraphSAGE feature	0.2997	0.8278	0.3906	0.7941
GIN feature	0.3003	0.8284	0.3764	0.7975
GATv2 feature	0.3071	0.8237	0.4086	0.7808
GCN patient	0.2963	0.8278	0.3742	0.7892
GAT patient	0.2980	0.8345	0.3874	0.7847
GraphSAGE patient	0.2971	0.8342	0.3832	0.7895
GIN patient	0.3166	0.8111	0.4113	0.7782
GATv2 patient	0.2987	0.8307	0.3848	0.7894
APPNP patient	0.3810	0.7209	0.4407	0.6990
GPR-GNN patient	0.3431	0.7618	0.4131	0.7329

Note. WTMECPRP was used for development and WTPH2YR for cross-period evaluation. Weights were applied to evaluation metrics. Training losses were unweighted. Values are descriptive weighted point estimates, with the best in each column shown in bold.

5. Discussion

5.1. Graph Representation, Propagation, And Evaluation Objective

Under a common predictor set, reference outcome, and grouped validation design, the relative performance of feature and patient graphs varied by propagation architecture. Patient graphs had lower NLL for GAT and GATv2, whereas feature graphs had lower NLL for GIN. Feature graphs had higher OOF high-category recall point estimates in all five architecture pairs. Together, these observations show that graph selection depends on both the propagation architecture and the emphasis placed on probability scoring or identification of a particular category.

Previous clinical graph studies have modeled within-visit event relationships or between-patient similarity [4, 5]. Table2Graph maps columns into a shared feature graph [7]. Our comparison places within-participant variable relationships and between-participant similarity in the same clinical task. Holding predictors and reference labels constant makes the resulting rankings directly interpretable as comparisons of graph-model combinations within this task. The node definition, propagation architecture, and evaluation objective jointly inform the choice of representation.

5.2. Probability Prediction and High-Category Identification

Logistic regression had the lowest unweighted NLL, whereas PIGA had the highest OOF QWK and high-category recall point estimates. Strong performance of conventional models on tabular tasks has been observed previously [8]. The present results further show how probability scores and category identification can favor different models in the same cohort. NLL scores the probability of the observed category, AUROC assesses ranking, and QWK and recall depend on the category assigned by the maximum-probability rule.

PIGA's OOF high-category recall was 25.07%. The highest intermediate-category recall across models was 2.46%, showing a different pattern of identification between the two minority categories. APPNP and GPR-GNN produced participant-specific probabilities while assigning all participants to the low category. Reporting probability metrics alongside category-specific classification results makes these behaviors visible and relates model performance to the intended prediction objective.

5.3. Survey Weights, Cross-Period Performance, And Study Scope

Weighting changed the lowest-NLL development model from logistic regression to XGBoost, demonstrating that the relative importance assigned to participants can affect model rankings. Across periods, the highest weighted macro AUROC and lowest weighted NLL also belonged to different models, further distinguishing probability ranking from probability scoring.

Cross-period NLL and Brier scores were higher, and macro AUROC was lower. GAT, GATv2, and GIN retained the direction of the NLL contrast between graph representations, whereas GCN changed direction. Category proportions and fitting schemes differed between periods. Development performance used OOF predictions, while cross-period performance used models refitted on the full development cohort.

The findings apply to NHANES adults with complete predictors and contemporaneous LSM reference categories. Histological confirmation was outside the outcome definition [15]. The cross-sectional data support assessment of contemporaneous prediction, and the previously examined later period provides exploratory evidence of transfer across survey periods. Feature and patient models differed in node encoding, readout, and normalization, and neural models used fixed depths, dimensions, and training hyperparameters. The contributions of individual prior components and edge designs were not isolated. Complete-case selection can affect the composition of the analysis cohorts. Weighted estimates describe participants with complete data, and the PSU intervals quantify sampling-unit variation conditional on fixed predictions.

6. Conclusion

In the NHANES three-level LSM reference-category task, graph representation rankings depended on architecture and evaluation objective. Patient graphs had lower NLL for GAT and GATv2, whereas feature graphs had lower NLL for GIN. Feature graphs had higher OOF high-category recall point estimates across all five architecture pairs. Logistic regression had the lowest unweighted NLL in both periods, and PIGA had the highest OOF QWK and high-category recall point estimates. Joint comparison of graph representations, propagation architectures, and evaluation objectives supports model selection for this clinical tabular prediction task.

Declarations

Ethics approval

This study analyzed publicly available, anonymized NHANES data from the National Center for Health Statistics (NCHS). The original NHANES surveys were approved by the NCHS Research Ethics Review Board[24].

Competing interests

Linjie Tang and Zhitao Yu declare no competing interests related to this study.

Author contributions

Linjie Tang was responsible for Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing – original draft, and Writing – review & editing. Zhitao Yu was responsible for Validation.

Data availability

The 2017–March 2020 pre-pandemic combined files and August 2021–August 2023 files are available from the NCHS NHANES website[11,12]. The accompanying materials include complete performance summaries and a dictionary mapping predictors to their original fields.

Code availability

Model implementations, experiment configurations, statistical analysis scripts, and figure-generation code have been archived. They are available from the corresponding author during review and will be made public after acceptance.

Acknowledgment

This work was supported by the Innovative Research Project of Southwest Minzu University.

References

[1] National Center for Health Statistics. (2022). *P_LUX* [Data file]. CDC. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_LUX.htm

[2] National Center for Health Statistics. (2024). *LUX_L* [Data file]. CDC. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/LUX_L.htm

[3] Eddowes, P. J., Sasso, M., Allison, M., et al. (2019). [Article title missing]. Gastroenterology, 156(6), 1717–1730. https://doi.org/10.1053/j.gastro.2019.01.042

[4] Choi, E., Xu, Z., Li, Y., Dusenberry, M. W., Flores, G., Xue, E., & Dai, A. M. (2020). [Article title missing]. Proceedings of the AAAI Conference on Artificial Intelligence, 34(1), 606–613. https://doi.org/10.1609/aaai.v34i01.5400

[5] Lu, H., & Uddin, S. (2021). [Article title missing]. Scientific Reports, 11, 22607. https://doi.org/10.1038/s41598-021-01964-2

[6] Chen, T. P., Yu, H. C., Lin, W. Y., & Chang, Y. C. (2026). [Article title missing]. Journal of Dental Sciences, 21(2), 1148–1155. https://doi.org/10.1016/j.jds.2025.12.027

[7] Zhou, K., Liu, Z., Chen, R., Li, L., Choi, S. H., & Hu, X. (2022). [Article title missing]. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI), 2420–2426. https://doi.org/10.24963/ijcai.2022/336

[8] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). [Article title missing]. Advances in Neural Information Processing Systems, 35. https://proceedings.neurips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract-Datasets_and_Benchmarks.html

[9] Han, K., Tan, K., Shen, J., et al. (2022). [Article title missing]. Frontiers in Public Health, 10, 1008794. https://doi.org/10.3389/fpubh.2022.1008794

[10] Noureddin, M., Ntanios, F., Malhotra, D., et al. (2022). [Article title missing]. Hepatology Communications, 6(7), 1537–1548. https://doi.org/10.1002/hep4.1935

[11] National Center for Health Statistics. (n.d.). NHANES 2017-2020 overview brief. CDC. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overviewBrief.aspx?Cycle=2017-2020

[12] National Center for Health Statistics. (2024). NHANES 2021-2023 overview brief. CDC. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overviewbrief.aspx?Cycle=2021-2023

[13] National Center for Health Statistics. (n.d.). *P_BIOPRO* [Data file]. CDC. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/P_BIOPRO.htm

[14] National Center for Health Statistics. (n.d.). *BIOPRO_L* [Data file]. CDC. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/BIOPRO_L.htm

[15] Rinella, M. E., Neuschwander-Tetri, B. A., Siddiqui, M. S., et al. (2023). [Article title missing]. Hepatology, 77(5), 1797–1835. https://doi.org/10.1097/HEP.0000000000000323

[16] Chen, T., & Guestrin, C. (2016). [Article title missing]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

[17] Kipf, T. N., & Welling, M. (2017). [Article title missing]. *International Conference on Learning Representations*. arXiv. https://arxiv.org/abs/1609.02907

[18] Hamilton, W. L., Ying, Z., & Leskovec, J. (2017). [Article title missing]. Advances in Neural Information Processing Systems. https://proceedings.neurips.cc/paper_files/paper/2017/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html

[19] Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). [Article title missing]. *International Conference on Learning Representations*. shturl.cc/cGZsotWEwkly1yADIKCX1Bi6XkIG8Z

[20] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). [Article title missing]. *International Conference on Learning Representations*. arXiv. https://arxiv.org/abs/1710.10903

[21] Brody, S., Alon, U., & Yahav, E. (2022). [Article title missing]. *International Conference on Learning Representations*. arXiv. https://arxiv.org/abs/2105.14491

[22] Klicpera, J., Bojchevski, A., & Günnemann, S. (2019). [Article title missing]. *International Conference on Learning Representations*. arXiv. https://arxiv.org/abs/1810.05997

[23] Chien, E., Peng, J., Li, P., & Milenkovic, O. (2021). [Article title missing]. *International Conference on Learning Representations*. arXiv. https://arxiv.org/abs/2006.07988

[24] National Center for Health Statistics. (n.d.). NHANES Institutional Review Board information. CDC. https://www.cdc.gov/nchs/nhanes/about/erb.html