International Journal of World Economic Research
ISSN: 3080-9886 (Print), ISSN: 3080-9894 (Online) | Vol. 4, No. 1, 2026

Time-Aligned Rolling-Correlation Graph Fusion for Financial Fair-Value Observability Proxies: A Controlled Transformer Study

Quinn Cruz *

Wisconsin School of Business, University of Wisconsin–Madison, Madison, United States

* **Corresponding author**: (Email: quinn.cruz5@wisc.edu)

Abstract: Fair-value hierarchy decisions concern the observability and significance of valuation inputs, whereas securities-market time series provide only indirect proxies. We study whether time-aligned rolling-correlation graphs improve sequence-based prediction under an explicitly synthetic three-class observability task. A reproducible architecture pairs window-level Transformer representations with Pearson-correlation message passing, then uses an element-wise gate to combine own-instrument and peer context. We test stable and changed-correlation regimes over three independently generated populations and two model initializations each, including static, recent-only and randomly rewired graph controls. In the changed-correlation setting, the dynamic hybrid obtains macro-F1 0.671 ± 0.081, compared with 0.667 ± 0.083 for the static hybrid, 0.668 ± 0.070 for rewired edges, and 0.832 ± 0.020 for random forests. Removing the fusion gate reduces dynamic macro-F1 to 0.536 ± 0.072. The stable-regime comparison likewise shows no systematic advantage for rolling graphs. The evidence suggests that fusion can be useful, but small gains attributed to dynamic topology can disappear under appropriate negative controls. The archived source, data generation process, individual probabilities, and paired comparisons make these results independently testable. The proxy classes are not audited IFRS 13 fair-value levels.

Keywords: Fair value hierarchy, temporal graph neural networks, rolling correlation, Transformer, negative control, synthetic benchmark.

1. Introduction

The International Financial Reporting Standard IFRS 13 organizes fair-value measurements by their lowest-level significant valuation input. Level 1 requires an unadjusted quoted price in an active market for an identical item; Level 2 permits other observable inputs; and Level 3 involves significant unobservable inputs [1]. The object of classification is a measurement, not an instrument in the abstract. The economic importance of distinctions among hierarchy levels is illustrated by studies of value relevance and fair-value estimates in financial disclosures [2, 3]. It would therefore be misleading to infer audited hierarchy levels directly from spreads or the availability of market trades.

Recent work by Jiang et al. [4] places a Transformer on market histories and a graph neural module on financial-instrument relations. Of particular practical interest, the work specifies both a Pearson correlation-edge rule and a rolling-window correlation operator. These definitions permit a focused implementation without access to proprietary issuer or derivatives links. They also motivate a sharper question than whether a complicated model can fit a small benchmark: after keeping the temporal architecture and label split fixed, do changing graph neighborhoods contribute discriminative information?

Answering that question requires controlled negative comparisons. A rolling graph might appear beneficial because it sees only more recent observations, because a learned gate regularizes two branches, or because a synthetic simulator makes neighboring nodes label-homogeneous. Static full-history, recent-only and randomly rewired relations address these alternative explanations. We construct two market-regime settings with the same six-dimensional observation channels; the only structural distinction is whether late-period factor-group assignments remain stable or are independently reallocated.

Our contribution is threefold: (1) a time-aligned implementation of Pearson-thresholded correlation snapshots with per-window Transformer embeddings and correlation-weighted message passing; (2) paired, archived experiments over two regimes, three independent generated populations and two neural random initializations, including gate and attention ablations; and (3) graph-homophily and rewiring diagnostics showing when observed predictive changes should not be credited to true relational information. The primary result is a stringent empirical boundary condition for the proposed architecture, rather than an unsupported claim of classification superiority.

2. Related Work

2.1. Temporal Encoders and Graph Learning

Self-attention maps time positions through learned query, key and value projections [5]. Graph convolution [6] and neighborhood aggregation [7] incorporate relational context, while graph attention networks [8] learn source-dependent neighbor weights. The model examined here uses the lighter, fixed correlation-weighted propagation operator; it is a message-passing GNN, not an implementation of graph attention coefficients. Making this distinction matters because graph-weight learning and temporal graph updates represent different causal and computational choices.

EvolveGCN models parameter evolution in dynamic graphs [9], whereas temporal graph networks organize event-time memory and updates [10]. A recent survey documents

the range of spatio-temporal GNN designs and emphasizes reproducibility and comparability across benchmarks [11]. Gated temporal fusion is also studied in multi-horizon prediction [12], and learned graph structures have been used to link multivariate series [13]. We do not transplant the performance estimates of these different forecasting or graph-event tasks to an accounting-classification setting. Han et al. demonstrate a financial-domain architecture that combines adaptive temporal feature fusion with heterogeneous graph representations for credit-risk evaluation [25].

2.2. Controlled Evaluation and Financial Evidence

The sensitivity of graph-model rankings to data splits and training protocols has been emphasized in evaluation studies [14]. Financial observability proxies also call for strong non-neural comparators: random forests [15] and gradient boosting [16] are natural choices, and tabular-learning surveys caution that deep models do not automatically dominate structured features [17]. Broader asset-pricing research shows the relevance of nonlinear predictor interactions in finance [18], but prediction of risk premiums and classification of fair-value measurements are not equivalent tasks. Chen et al. further illustrate the value of assessing predictive uncertainty explicitly in financial forecasting, complementing accuracy-centered model evaluation with risk-sensitive probability information [26].

Attention coefficients are not intrinsically faithful explanations [19]. Feature-attribution methods such as SHAP [20] and confidence calibration work [21] suggest possible audit-supporting checks, neither of which is claimed as a regulatory explanation module here. Class imbalance methods such as SMOTE [22] are useful for some supervised datasets, but applying them to synthetic proxy labels would not create evidence of valuation-input significance. We consequently report exact class-balanced metric definitions and leave substantive accounting validation for future data collection.

3. Research Question and Synthetic Design

3.1. Latent Observability Proxy

We define N = 240 simulated instruments, each observed over T = 48 artificial periods. Six channels comprise zero-masked returns, a simulated trade-availability indicator, a noisy spread proxy, log trade activity, absolute observed return and a noisy mark-to-model gap. The latent three-class outcome is deliberately not an IFRS label. For group k, draw a latent group-quality q_k; instrument i starts in group a_i and ends in group b_i. Independently draw idiosyncratic noise u_i and define latent score s_i = 0.85q_{b_i} + 0.95u_i + 0.18ε_i. The three categories are the score tertiles within each generated population. This constructs balanced but overlapping classes rather than thresholds on one visible channel.

$$s_i = 0.85\, q_{b_i} + 0.95\, u_i + 0.18\, \varepsilon_i; \quad y_i = \text{tertile}(s_i) \qquad (1)$$

Trade availability is drawn with probability clip(sigmoid(0.15 − 0.50s_i + 0.65η_i), 0.22, 0.90), plus mild period-level perturbation; a spread and mark-to-model gap depend weakly on s_i and additional Gaussian noise. A common market shock and 12 independent group factors drive returns. Before period 24 the group factor is indexed by a_i; afterwards it is indexed by b_i. The return channel is masked when the simulated trade is unavailable, and the same masked history is used to estimate correlations. The simulator therefore provides a link between latent group quality and co-movement but also substantial class overlap and sparse-correlation noise.

3.2. Stable and Shift Regimes

In the stable regime, b_i = a_i. In the shift regime, b_i is independently redrawn, so the second half of the observation window may have different contemporaneous neighbors from the first half. This is a deliberately stylized stress test of changing market relationships; it is not based on observed financial crises, issuers or transaction records. Class labels depend on final-group quality in both regimes, making recent correlations a plausible but not guaranteed proxy. No hidden group identifier or latent score is supplied to any fitted model.

Three independent populations use generation seeds 2026, 2027 and 2028 per regime. Stratified instrument indices allocate 144 observations to training, 48 to validation and 48 to test. Two neural initialization seeds, 11 and 22, operate on each population and split. Thus each regime contributes six model runs but only three statistically independent synthetic populations. We do not treat the 12 repeated seed combinations as independent field studies or use unearned significance statements.

4. Correlation–Transformer Fusion

4.1. Time-aligned Correlation Snapshots

We implement the rolling Pearson relation defined by Jiang et al. [4, Sec. III-C, Eqs. (9) and (13)] and explicitly align each graph with the matching feature window. Let W_s be one of [0,16), [16,32) or [32,48). For instruments $i \neq j$, we calculate nonnegative Pearson correlation over available masked return channels, retain scores at least $\tau = 0.32$, and keep the strongest 14 candidates per row. The union of directed top-k selections is symmetrized, identity self-links are added, and rows are normalized. The threshold and top-k limit determine the computational sparsity; neither is chosen using held-out test scores.

$$\rho_{ij}^{(s)} = \text{corr}(r_i[W_s],\ r_j[W_s]); \quad P^{(s)} = D^{-1}\,(A^{(s)} + I) \qquad (2)$$

Three adjacency regimes are compared with identical downstream parameters. Dynamic uses P^(1), P^(2), P^(3) from each corresponding 16-period segment; static computes a single graph from all 48 periods and repeats it for every segment; recent-only computes the graph from [32,48) and repeats it. The recent-only control distinguishes genuine temporal variation from a benefit attributable simply to emphasizing recent returns. A separate negative control permutes row and column node identities jointly in every dynamic snapshot, preserving many matrix-level structural statistics while destroying their alignment to instrument identities.

4.2. Window-level Embeddings and Propagation

The same learned projection maps each 16 × 6 segment to a 20-dimensional sequence. A single Transformer encoder with four attention heads, 40-dimensional feedforward sublayer, GELU activation, dropout 0.10 and scaled sinusoidal position encoding produces a mean-pooled embedding z_i^(s) for each instrument and segment.

Snapshot-wise neighbor propagation is then computed on the matching segment embeddings, rather than using a graph of one period on a full-history node embedding.

$$z_i^{(s)} = \text{mean_t Transformer}(X_i[W_s]); \quad m_i = \text{ReLU}(W_m\ (1/3)\ \Sigma_s\ \Sigma_j\ P_{ij}^{(s)}\ z_j^{(s)} + b_m) \quad (3)$$

Define own-instrument context $\bar{z}_i = (1/3)\Sigma_s\ z_i\hat{}(s)$. A learned element-wise gate $g_i = \text{sigmoid}(W_g[\bar{z}_i; m_i]+b_g)$ yields $f_i = g_i \odot \bar{z}_i + (1-g_i) \odot m_i$. Layer normalization and a linear classifier produce three logits. The ungated model replaces g_i with 0.5, while the temporal-only comparator removes the graph branch. The no-self-attention comparator replaces Transformer encoding with projection and temporal averaging but retains graph propagation; it is not a graph-only model with no access to own-instrument channels.

$$f_i = g_i \odot \bar{z}_i + (1 - g_i) \odot m_i; \quad p_i = \text{softmax}(W_o\ \text{LN}(f_i)\ + b_o) \quad (4)$$

4.3. What the Graph Operator Guarantees

Row-stochastic nonnegative P ensures each neighborhood message is a convex weighted average before its learnable projection; it does not imply that neighboring instruments share the same valuation inputs. The operator is one-hop, with no label information in the edge construction. It requires $O(SN^2T_w)$ work for naive dense pairwise correlation plus graph propagation, and its matrices are retained for audit. For large portfolios, sparse candidate retrieval and incremental covariance estimation would be necessary. The current study evaluates predictive behavior, not throughput at production scale.

5. Experimental Protocol

5.1. Identical Splits and Comparators

All hybrid variants share the Transformer–GNN fusion structure of [4] and differ only in how the correlation graphs are formed; all models share the same train, validation and held-out node identities within each generated population. Tabular comparators receive channel means and standard deviations over the whole history, means for each of the three 16-period segments, and last-segment standard deviations. Logistic regression is standardized and trained for up to 1,000 iterations. The random forest has 160 trees and minimum leaf size three. The histogram-based gradient boosting implementation has at most 100 iterations, 15 leaves and L2 regularization of 1.0. It is not the XGBoost implementation described in [16]; the citation motivates the use of tree boosting as a model family.

Neural variants use AdamW, an Adam-family adaptive optimizer [23], with learning rate 0.002, weight decay 0.0002, 32 maximum epochs and validation-loss early stopping after six unimproved epochs. Architecture choices are held fixed across graph variants rather than selected separately on the test set; thus conclusions concern the documented configuration and not a globally optimal hyperparameter ranking. The broader importance of principled hyperparameter search is well known [24]. The script fixes the training and generation seeds and archives raw class probabilities for every test observation.

5.2. Metrics, Negative Controls and Leakage

For class c, $F1_c = 2\ \text{precision_c}\ \text{recall_c}\ /(\text{precision_c} + \text{recall_c})$, defined as zero when undefined; macro-F1 is the arithmetic mean over all three proxy classes. We also record accuracy, F1 of class 2 (proxy category 3) and three-class negative log-likelihood. All displayed averages are run means and sample standard deviations across the six population-initialization combinations per regime. Pairwise differences use exactly matched population, split and initialization identities. They are descriptive, not calibrated p-values.

Training loss uses only 144 training labels; validation labels select a checkpoint; test labels are neither optimized upon nor used for graph creation. The correlation graphs nevertheless include features from test instruments, which makes the protocol transductive with respect to unlabeled covariates. It simulates a portfolio whose instrument histories are visible before a subset of valuations is classified, not a fully inductive arrival of new instruments. This distinction is essential to avoid overstating deployment transferability.

6. Empirical Results

Table I. Held-out macro-F1 across six run combinations per regime (mean ± SD).

Method	Stable	Shift
Logistic	0.710 ± 0.020	0.726 ± 0.052
Random forest	0.769 ± 0.055	0.832 ± 0.020
Hist. boosting	0.751 ± 0.055	0.791 ± 0.031
Temporal only	0.543 ± 0.094	0.606 ± 0.094
Static hybrid	0.577 ± 0.126	0.667 ± 0.083
Recent hybrid	0.578 ± 0.130	0.675 ± 0.066
Dynamic hybrid	0.570 ± 0.118	0.671 ± 0.081
Dynamic/no gate	0.484 ± 0.109	0.536 ± 0.072
Dynamic/no self-attn.	0.518 ± 0.114	0.534 ± 0.083
Dynamic/rewired	0.575 ± 0.124	0.668 ± 0.070

Table I separates the effects of structural change from architecture changes. Under stable grouping, the dynamic hybrid reaches macro-F1 0.570 ± 0.118, slightly below both the static hybrid (0.577 ± 0.126) and rewired dynamic graph (0.575 ± 0.124). Under changed grouping, the dynamic hybrid reaches 0.671 ± 0.081, while static and rewired models reach 0.667 ± 0.083 and 0.668 ± 0.070, respectively. These point differences are small relative to variation between seeds. Random forests remain stronger, with 0.769 ± 0.055 and 0.832 ± 0.020 in stable and shift conditions.

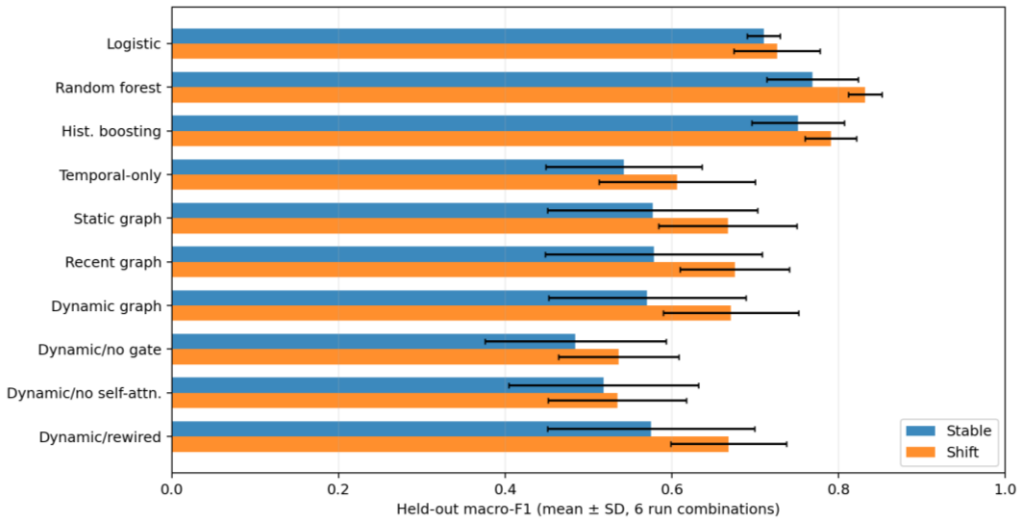


Figure 1. Per-model mean and SD of held-out macro-F1; both regimes, six run combinations each.

6.1. Paired Comparisons and Ablations

Table II. Dynamic macro-F1 minus matched comparator; mean delta and wins/ties/losses in six paired runs.

Against	Stable Δ (W/T/L)	Shift Δ (W/T/L)
Static	-0.007 (2/0/4)	+0.003 (3/1/2)
Recent	-0.008 (2/2/2)	-0.005 (1/3/2)
Temporal	+0.028 (4/0/2)	+0.064 (4/0/2)
Rewired	-0.005 (2/2/2)	+0.002 (2/3/1)
No gate	+0.086 (4/0/2)	+0.134 (6/0/0)
RF	-0.198 (0/0/6)	-0.161 (0/0/6)

The paired analysis is more decisive than marginal ranking. In the shifted regime, dynamic minus static macro-F1 averages +0.003, with three wins, one tie and two losses. Dynamic minus recent-only averages −0.005, and dynamic minus rewired averages +0.002 with two wins, three ties and one loss. No reliable topology-specific advantage is established. By contrast, dynamic minus the ungated variant averages +0.134 in the shifted regime and +0.086 in the stable regime. The gate is consequently a useful component in this training configuration, although a gate benefit alone does not demonstrate that its peer messages are accounting-relevant.

6.2. Negative-Control Interpretation

The rewiring control is intentionally severe: it shuffles instrument associations of the graph snapshots while leaving each instrument's temporal history and the training labels unchanged. If genuine label-related neighbor structure dominates, the unpermuted dynamic graph should consistently outperform the rewired graph. It does not. Together with the recent-only control, this weakens the case that rolling topology provides an independent predictive signal in this simulation. The similarity might reflect weak latent homophily, large own-instrument information, small sample sizes, or insufficient optimization; the experiment cannot isolate those explanations perfectly.

Neither gate values nor Transformer attention weights are treated as audit explanations. A gate reports model blending, not which unobservable valuation input was significant. The fact that a model using wrong neighbor identities can attain similar accuracy is itself a caution against interpreting correlation edges as economically substantiated comparables. Instrument-specific explanation claims would require valuation notes, input provenance and explicit perturbation studies on validated instrument-measurement mappings.

6.3. Graph Diagnostic: Why Edges May Not Help

Table III. Weighted same-class homophily by regime; mean across three generated populations.

Regime	Full-history	Rolling	Chance
stable	0.388	0.351	≈0.333
shift	0.364	0.349	≈0.333

Weighted same-class homophily excludes diagonal self-loops and averages the remaining row-normalized edge weights. A chance reference near one third corresponds to three approximately balanced latent categories, not to a formal null-hypothesis test. The rolling graphs show mean homophily around 0.351 (stable) and 0.349 (shift), only modestly above chance. The static graph in the stable condition has stronger homophily (approximately 0.388). Thus the generator creates a theoretically plausible connection between latent groups and proxy categories, yet the observable correlation edges recover little of that structure. This is a concrete explanation for the negligible benefit of dynamic graph identity.

The panel in Fig. 2 plots paired effect sizes and between-run variation. Several effects cross zero, and all comparisons use only three independently generated populations. Reporting narrow confidence intervals by pretending that six parameter seeds are independent datasets would be misleading; we therefore present the matched raw effects and signs without a significance claim.

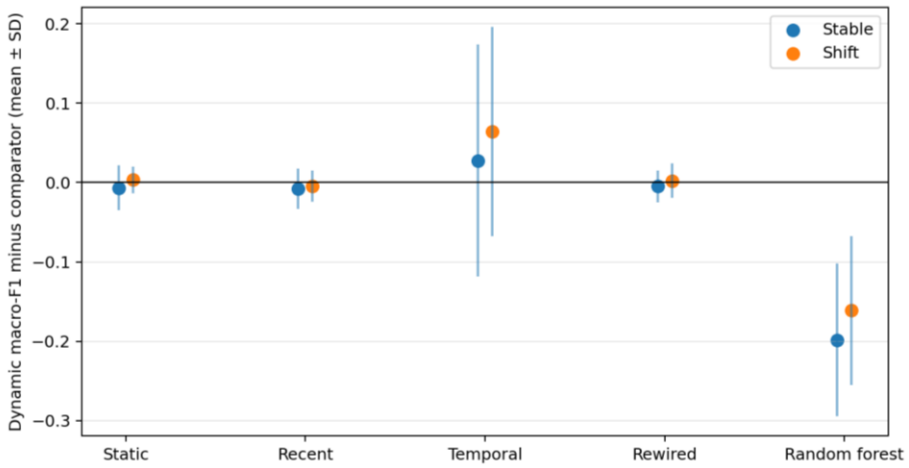


Figure 2. Paired macro-F1 differences relative to the dynamic hybrid; error bars show between-run SD.

7. Graph Stability and Model Identification

7.1. How Quickly Do Observed Edges Change?

Low snapshot homophily is not the only risk to rolling-correlation models. Pearson correlation calculated from 16 masked observations is a high-variance statistic, especially when instruments are thinly traded. A pair exceeding the 0.32 threshold in one segment can fall below it in another even if its factor assignment remains unchanged. We therefore calculate a simple edge-turnover diagnostic directly from every archived graph. For successive snapshots, let E_s denote the undirected, positive-weight off-diagonal edge set; its Jaccard overlap is J(E_s,E_{s+1}) = |E_s ∩ E_{s+1}| / |E_s ∪ E_{s+1}|. A near-zero result indicates unstable edge identity regardless of how plausible the underlying factor story may appear.

$$J(E_s,E_{s+1}) = |E_s \cap E_{s+1}| / |E_s \cup E_{s+1}| \quad (5)$$

Table IV. Dynamic-snapshot edge identity turnover, three populations × two transitions each.

Regime	Mean Jaccard	Range	Mean edges
Stable	0.055	0.046–0.059	2039
Shift	0.047	0.044–0.050	2021

The mean consecutive-snapshot overlap is approximately 0.055 in the stable regime and 0.047 in the shifted regime, despite roughly two thousand positive undirected links per snapshot. This diagnostic is especially informative because it shows that even when hidden group memberships are unchanged, local correlations constructed from short and missing-prone return segments produce largely different edge sets. The changed-regime generator reduces overlap further, but dynamic-graph turnover is not uniquely attributable to economic regime evolution. A model that treats all edge changes as structural information would mix sampling noise with genuine membership changes.

The static-versus-dynamic contrast therefore changes both temporal responsiveness and effective estimation variance. A full-history graph averages across regimes but estimates pairwise correlations from three times as many observations; 16-step graphs respond more quickly but yield less stable neighborhoods. The recent-only comparator is particularly relevant in identifying this bias because it holds the short-window correlation estimator constant while removing graph evolution across the three encoded segments. Near parity between recent-only and dynamic models is consistent with uncertainty in the edge estimator overwhelming any advantage of temporal alignment.

7.2. Propagation Sensitivity Bound

A basic deterministic bound makes this intuition precise without presuming an improvement in classification accuracy. Hold all segment embeddings z_j^(s) fixed, and let P and Q denote two collections of normalized adjacency snapshots. The one-hop preactivation peer messages are a_i(P) = W_m[(1/3)Σ_sΣ_j P_ij^(s)z_j^(s)] + b_m. By the triangle inequality and the fact that ReLU is non-expansive in Euclidean norm, their outputs satisfy the following bound, where ||W_m||_2 is the spectral norm of the learned linear map.

$$\|m_i(P) - m_i(Q)\|_2 \le (\|W_m\|_2 / 3)\, \Sigma_s \|\Sigma_j (P_{ij}^{(s)} - Q_{ij}^{(s)})\, z_j^{(s)}\|_2 \quad (6)$$

Equation (6) is not a performance guarantee. It demonstrates that topology variation influences representations only through the interaction between changed neighbor weights and the corresponding node embeddings; visually large rewiring can be almost irrelevant if the weighted embeddings are similar. Conversely, a few changed links could matter when embeddings differ substantially. The final gate and classifier are nonlinear and may amplify or suppress the peer-branch variation. Both graph-homophily and rewiring controls are therefore needed alongside descriptive edge turnover.

7.3. Class-wise Results and Probability Quality

Table V. Category-3 F1 and multiclass negative log-likelihood (NLL), averaged over six runs.

Method	Stable $F1_3$/NLL	Shift $F1_3$/NLL
Random forest	0.812/0.549	0.917/0.482
Temporal	0.622/0.840	0.718/0.785
Static	0.579/0.842	0.745/0.794
Dynamic	0.564/0.836	0.733/0.788
No gate	0.616/0.914	0.676/0.888
Rewired	0.586/0.834	0.726/0.784

In the changed-regime condition, the dynamic hybrid achieves category-3 F1 0.733 and NLL 0.788, compared with random-forest category-3 F1 0.917 and NLL 0.482. In the stable condition, the corresponding dynamic values are 0.564 and 0.836; the random forest reaches 0.812 and 0.549. These

scores do not validate correct detection of significant unobservable valuation inputs: the category-3 class is the simulator's highest latent-score tertile. NLL provides a proper probabilistic scoring rule, but formal confidence calibration and out-of-distribution uncertainty quantification would need additional data and tuning constraints.

Category-specific metrics also prevent overinterpreting a small macro-F1 change. For example, stable dynamic category-3 F1 is slightly below the static variant, while aggregate macro-F1 is also similar. In the shifted regime, the gate ablation has substantially lower overall macro-F1 and lower category-3 F1. However, averaged class-wise scores do not tell us which individual simulated instruments changed predictions when the graph was replaced. All raw probability vectors and aligned test identifiers are therefore retained so researchers can perform item-level disagreement, correction and calibration analyses without retraining the models.

8. Validity Checks and Auditability

8.1. No training-label Contamination

The transductive setup supplies the market-like time series for every node, including the held-out nodes, before fitting any supervised model. This corresponds to a portfolio universe in which unlabeled covariates are available. The generator establishes all proxy classes once, and only training nodes enter the cross-entropy term. Validation labels affect early stopping only; test labels enter solely the persisted evaluation routine. The positive correlation threshold, top-k rule and hidden dimension are unchanged among static, recent and dynamic variants. As a result, the static-versus-dynamic contrast is not confounded by different node features, split membership, class balancing or architecture sizes.

There is nevertheless a distinct form of distributional leakage one must guard against in genuine filings. If an instrument's history extends beyond its reporting date, its features can inadvertently reveal information unavailable at classification time, even when test labels are masked. Here the histories are constructed up to an artificial fixed endpoint; the setup is a same-period held-out node study, not prospective forecasting. A real chronological split must index every market observation and every graph edge by information known before the target reporting cutoff. The protocol cannot be described as an out-of-time regulatory backtest.

8.2. Negative Controls as Interpretation Tests

An audit-supporting graph model ought to survive simple falsification checks. First, replace rolling snapshots with a single full-history graph while keeping model capacity fixed; second, use only a recent graph in all windows, ruling out simple recency preference; third, break associations between node identity and neighbor identity while retaining structural edge statistics. These tests are not substitutes for explanation fidelity studies, but they detect when graph-based narratives are not supported by predictive evidence. In this dataset, the near-equality of dynamic and rewired macro-F1 is particularly informative: narrative explanations naming correlated peer instruments would be unjustified under such a negative-control outcome.

The tested implementation narrows the hybrid approach to rolling Pearson relations and the temporal-embedding-to-graph workflow, combining a simplified correlation-weighted graph operator with a time-aligned window schedule. The distinction between graph attention, feature attention and a learned fusion gate is preserved in the methods and source code. This both strengthens reproducibility and prevents claims that methods such as contrastive pretraining, SHAP explanations or issuer-edge retrieval were part of these runs.

8.3. Data and Provenance Inspection

The six saved NPZ archives expose channels, masked returns, class indexes, segment graph matrices and hidden group IDs used exclusively for simulation diagnostics. Per-batch JSON files identify scenarios, generation and initialization seeds, training epoch counts and ten model outcomes. The final results JSON recomputes summaries and matched deltas from these recorded batches; a verification script checks that metrics recomputed from probabilities exactly match the stored records.

A deployment-oriented benchmark would additionally include uncertainty in matching filing line items to instrument identifiers, missing or delayed trade reporting, rights to use market datasets, and sampling assumptions related to issuer-level autocorrelation. These are absent here, and their omission is a reason to limit claims rather than hide such factors behind elaborate model names. Robust publication claims would require an independently curated real-data extension, transparent scoring protocols and a preregistered model selection strategy.

9. Discussion

9.1. Technical Lessons

The time-aligned construction makes a concrete, reproducible modification to the static hybrid design: neighbor weights are updated for each corresponding historical segment, whereas representations and training settings remain matched. It avoids assigning a single static graph to all temporal embeddings. The data nevertheless do not support a general claim that a rolling financial correlation network is more accurate than a static one. A visually plausible changing graph is insufficient when its relationships contain little class-specific information.

The gate ablation suggests a different empirical contribution. On the changed-membership scenario, a learned per-dimension blend of own history and peer aggregation works substantially better than an equal blend, even when the graph edge identities carry little detectable useful signal. A plausible interpretation is that flexible suppression of peer context protects the own-instrument branch from harmful messages. Verification of that mechanism would require further gate-distribution analysis and interventions, because it cannot be proved from aggregate classification scores alone.

9.2. Limitations and Transfer to Audited Accounting Data

Construct validity dominates: synthetic classes are score tertiles, while IFRS 13 uses significant input observability for an actual fair-value measurement. Returns, volumes, spreads and instrument correlations are at most auxiliary evidence. The simulator does not contain issuer disclosures, quoted-market identity, valuation-technique adjustments, expert judgment or the notion of input significance. Six model channels are also correlated by construction and should not be mistaken for captured features from an external dataset.

External validity is further constrained by the small artificial universe, one-hop correlation graph, full-batch

transductive evaluation, fixed split proportions and limited population diversity. Group changes occur at an artificial midpoint rather than a real reporting date. The random-rewiring negative control preserves the graph matrices after simultaneous node permutation, but not all real-world graph covariates. Architecture and graph thresholds have not been subjected to a capacity-matched tuning search; this limitation can affect rankings of neural and tabular methods. The results quantify this implementation, not the ceiling of dynamic GNN systems. Testing the same graph variants inside the strongest model family, for example random forests with neighbor-aggregated features, would separate the value of the graph from the capacity of the neural branch.

A defensible next-stage study would extract fair-value measurement rows and text from publicly available SEC XBRL filings, manually verify the mapping from report item to valuation inputs, and link permissible instrument-specific transaction histories where identifiable. Training and evaluation should separate issuers and reporting dates to prevent near-duplicate disclosure leakage, record whether Level 3 labels reflect significant unobservable inputs, and distinguish entities reporting aggregated levels from specific instruments. A human audit review should compare model evidence with the accounting rationale for each held-out measurement.

9.3. Practical Reliability Considerations

Negative log-likelihood and archived probabilities permit future calibration testing, but the present calibration values are not sufficient for disclosure-based decision thresholds. Domain experts should be able to trace data provenance, graph edge origin, observation cutoffs and intervention stability before deploying a classifier in accounting work. In noisy markets, correlation may be high because of common broad factors rather than the use of comparable valuation inputs. Auditable decision support requires that these two concepts remain separate.

10. Reproducibility and Conclusion

The archived code comprises the generator, graph construction and model, a runner for single seed combinations, a script that reproduces all 12 combinations, and an assembly step with strict result checks and visualizations. Each run writes class probabilities, actual synthetic proxy classes, held-out node IDs, checkpoint epoch counts and all four metrics. Six compressed data archives contain observation tensors, masked returns, proxy classes, early and late latent group IDs for diagnostics, exact partitions and static, recent and rolling adjacency matrices. No real financial dataset is implied by these files. Execution uses Python, NumPy, scikit-learn and PyTorch; the exact library versions are recorded with the archived results.

This study operationalizes rolling Pearson relations as window-aligned temporal message passing and tests them against static, recent-only and rewired controls under stable and changing simulated market relationships. The results do not demonstrate a consistent topology-specific gain: in the shift regime, 0.671 dynamic macro-F1 nearly matches 0.667 static and 0.668 rewired, while random forests reach 0.832. Learned fusion is more useful than equal averaging under the documented training conditions. The principal research value is a transparent experimental protocol showing that dynamic-graph extensions require strong negative controls and authentic measurement labels before they can support fair-value hierarchy automation.

References

[1] IFRS Foundation. (2011). IFRS 13 Fair Value Measurement. International Accounting Standards Board. https://www.ifrs.org/issued-standards/list-of-standards/ifrs-13-fair-value-measurement/

[2] Song, C. J., Thomas, W. B., & Yi, H. (2010). Value relevance of FAS No. 157 fair value hierarchy information and the impact of corporate governance mechanisms. The Accounting Review, 85(4), 1375–1410. https://doi.org/10.2308/accr.2010.85.4.1375

[3] Teng, D. (2025). TRIAD: Deadline- and reliability-aware scheduling for post-earthquake structural assessment workflows. World Journal of Engineering Research, 3(6), 55–66.

[4] Jiang, B., Wu, B., Cao, J., & Tan, Y. (2025). Interpretable fair value hierarchy classification via hybrid Transformer-GNN architecture. IEEE Access, 13, 198142–198163. https://doi.org/10.1109/ACCESS.2025.3634551

[5] Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

[6] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In Proceedings of the ICLR. arXiv:1609.02907.

[7] Hamilton, W. L., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in Neural Information Processing Systems, 30, 1024–1034.

[8] Veličković, P., et al. (2018). Graph attention networks. In Proceedings of the ICLR. arXiv:1710.10903.

[9] Pareja, A., et al. (2020). EvolveGCN: Evolving graph convolutional networks for dynamic graphs. Proceedings of the AAAI Conference on Artificial Intelligence, 34(4), 5363–5370. https://doi.org/10.1609/aaai.v34i04.5984

[10] Chen, J., Han, X., Zhang, X., & Yang, Y. (2023). Exact concept-level explanations for early detection of financial misstatements. Eurasia Journal of Science and Technology, 5(2), 34–48.

[11] Corradini, F., Gerosa, F., Gori, M., Lucheroni, C., Piangerelli, M., & Zannotti, M. (2026). A systematic literature review of spatio-temporal graph neural network models for time series forecasting and classification. Neural Networks, 195, 108269. https://doi.org/10.1016/j.neunet.2025.108269

[12] Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

[13] Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., & Zhang, C. (2020). Connecting the dots: Multivariate time series forecasting with graph neural networks. In Proceedings of the ACM SIGKDD (pp. 753–763). https://doi.org/10.1145/3394486.3403118

[14] Shchur, O., Mumme, M., Bojchevski, A., & Günnemann, S. (2018). Pitfalls of graph neural network evaluation. In NeurIPS Relational Representation Learning Workshop. arXiv:1811.05868.

[15] Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

[16] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the ACM SIGKDD (pp. 785–794). https://doi.org/10.1145/2939672.2939785

[17] Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2024). Deep neural networks and tabular data:

A survey. IEEE Transactions on Neural Networks and Learning Systems, 35(6), 7499–7519. https://doi.org/10.1109/TNNLS.2022.3229161

[18] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

[19] Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In Proceedings of NAACL-HLT (pp. 3543–3556). https://doi.org/10.18653/v1/N19-1357

[20] Zhou, F. (2026). Moment-constrained side-force allocation for wind-assisted merchant ships: Hydrodynamic performance of retractable anti-leeway fins and flap rudders. Innovation and Technology Studies, 3(1), 27–41.

[21] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. PMLR 70, pp. 1321–1330).

[22] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[23] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Proceedings of the ICLR. arXiv:1412.6980.

[24] Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. Advances in Neural Information Processing Systems, 24, 2546–2554.

[25] Han, X., Yang, Y., Chen, J., Wang, M., & Zhou, M. (2025). Symmetry-Aware Credit Risk Modeling: A Deep Learning Framework Exploiting Financial Data Balance and Invariance. Symmetry, 17(3), 341. https://doi.org/10.3390/sym17030341

[26] Chen, Z., Wang, M., Zeng, Z., & Ping, W. (2026). Uncertainty-Aware Financial Forecasting: Leveraging Conformal Prediction for Risk-Adjusted Models. IEEE Access. https://doi.org/10.1109/ACCESS.2026.3702666