Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Adaptive Credit Risk Prediction Under Concept Drift Using Graph Neural Networks and Gradient Boosting

Nabeel Muhammed Kottayil

nabeel_3k@yahoo.com

Banasthali Vidyapith

Manisha Jailia

Banasthali Vidyapith

Research Article

Keywords: credit-risk prediction, graph neural networks, GraphSAGE, gradient boosting, XGBoost, concept drift, adaptive learning, IFRS 9, probability of default, expected credit loss

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11176012/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Adaptive Credit Risk Prediction Under Concept Drift Using Graph Neural Networks and Gradient Boosting

Nabeel Muhammed Kottayil [1,*] and Dr Manisha Jailia [1]

[1] *Department of Computer Science, Banasthali Vidyapith, Banasthali, Rajasthan 304022, India*

[*] *Correspondence: nabeel_3k@yahoo.com*

Abstract:

Credit-risk models in retail banking face two persistent difficulties. Default risk depends not only on a borrower's own attributes but also on systemic factors shared across related obligors—branch, sector, region, or relationship group—that conventional tabular models cannot represent. The relationship between borrower features and default is also non-stationary, drifting with macroeconomic conditions. We propose a hybrid framework that integrates a GraphSAGE encoder, learning borrower embeddings over a graph that links obligors by feature similarity and shared systemic membership, with a gradient-boosted classifier (XGBoost) that consumes both the embeddings and the underwriting features. A recency-weighted rolling-window mechanism keeps the classifier aligned with the prevailing regime, providing adaptivity under concept drift. Outputs are loan-level probabilities of default consistent with the IFRS 9 staging framework. On a sample, IFRS 9-structured benchmark constructed to contain both risk channels, the hybrid attains an AUC of 0.852, against 0.798 for boosted trees on underwriting features alone and 0.768 for the graph encoder alone, with corresponding gains in Kolmogorov–Smirnov separation and precision–recall. The adaptive variant sustains discrimination through a simulated downturn better than a frozen model. A complete, reproducible implementation accompanies the paper.

Keywords: credit-risk prediction; graph neural networks; GraphSAGE; gradient boosting; XGBoost; concept drift; adaptive learning; IFRS 9; probability of default; expected credit loss

JEL Classification: C45; C53; G21; G32

1. Introduction

Retail lending portfolios sit at the centre of a bank's risk and capital management, and the quality of their credit-risk models has direct consequences for provisioning, pricing, and regulatory capital. Under IFRS 9 (International Accounting Standards Board, 2014), institutions must estimate forward-looking expected credit losses and assign each exposure to one of three stages according to the evolution of its credit risk since origination. The probability that an obligor defaults—operationalised here, in line with common practice, as reaching ninety or more days past due—is the quantity that most directly drives those decisions. Producing accurate, well-calibrated, and stable probabilities of default is therefore not an academic nicety but an operational requirement.

For decades, the workhorse of credit scoring has been the logistic regression model, prized for its interpretability and its compatibility with regulatory expectations (Hand & Henley, 1997; Thomas, 2000). Machine-learning methods—random forests, gradient-boosted trees, and neural networks—have repeatedly been shown to improve discrimination over linear scorecards on benchmark data (Baesens et al., 2003; Lessmann et al., 2015; Suhadolnik et al., 2023), and gradient boosting in particular has become a strong default choice because it captures non-linear effects and high-order interactions among underwriting variables with little tuning (Chen & Guestrin, 2016). A wider systematic literature review confirms the rapid migration of credit-risk modelling toward machine-learning techniques over the past decade (Warin & Stojkov, 2021). Yet these methods share a structural limitation: they treat each loan as an independent row in a table. They have no mechanism to represent the fact that obligors are connected—that borrowers in the same sector, region, or relationship group are exposed to common shocks, and that distress tends to cluster and propagate through these connections.

Graph neural networks (GNNs) offer a natural language for exactly this kind of relational structure. By passing messages along edges, a GNN learns a representation of each node that reflects not only its own features but also the

state of its neighbourhood (Hamilton et al., 2017; Kipf & Welling, 2017; Veličković et al., 2018). In financial settings, graph-based models have been applied to fraud detection (Dou et al., 2020) and, more recently, to corporate and retail default prediction with dynamic, time-aware encoders (Wang et al., 2022; Zandi et al., 2024). A GNN, however, is not by itself an ideal credit-risk classifier: its smoothing of information across neighbourhoods, helpful for capturing systemic effects, tends to blur the sharp, threshold-like interactions among an individual loan's underwriting variables that tree ensembles capture so well.

This observation motivates our central design choice. Rather than asking one model family to do everything, we let each do what it does best and combine their representations. We construct a borrower graph, learn embeddings with a GraphSAGE encoder, and then fuse those embeddings with the original underwriting features inside a gradient-boosted classifier. The graph encoder supplies a systemic, relational view of risk; the boosted trees supply a fine-grained, idiosyncratic view; and the fusion exploits both. To this we add a lightweight adaptivity mechanism—recency-weighted retraining over a rolling window—so the model tracks concept drift as economic conditions change, a long-recognised concern (Gama et al., 2014; Lu et al., 2019) that static models do not address.

The contributions of this paper are fourfold. First, we present a hybrid GNN–gradient-boosting architecture for credit-risk prediction that explicitly separates and then recombines systemic and idiosyncratic risk. Second, we introduce a graph construction that augments feature-similarity edges with systemic-membership edges, giving the encoder an explicit channel for community-level risk. Third, we integrate a recency-weighted temporal-adaptation scheme that refreshes the downstream classifier without re-learning the (slowly varying) graph structure. Fourth, we release a complete and reproducible implementation, together with a validation protocol that combines a public benchmark with confidential institutional data under a strict anonymisation regime. We are explicit throughout that the quantitative results reported here are obtained on a institutional sample benchmark constructed to exercise the architecture, and we describe precisely how they are to be replaced with results on real data.

The remainder of the paper is organised as follows. Section 2 reviews the relevant literature on statistical and machine-learning credit scoring, graph neural networks in finance, and concept drift. Section 3 develops the materials and methods, including the borrower-graph construction, the GraphSAGE encoder, the fused classifier, and the temporal-adaptation mechanism. Section 4 presents the experimental results. Section 5 discusses the findings, their limitations, and the implications for IFRS 9 practice. Section 6 concludes.

2. Literature Review

2.1. Statistical and Machine-Learning Credit Scoring

The modern credit-scoring literature begins with discriminant analysis and logistic regression (Altman, 1968; Hand & Henley, 1997) and has accumulated a substantial body of comparative evidence on classification algorithms. The benchmarking studies of Baesens et al. (2003) and, a decade later, Lessmann et al. (2015) established that non-linear ensembles—random forests (Breiman, 2001) and gradient boosting (Friedman, 2001)—consistently outperform linear scorecards on discrimination, while noting that the margin is often modest and that interpretability and stability remain important in practice. Suhadolnik et al. (2023) confirm this pattern on a recent large-scale dataset, finding XGBoost the strongest performer among ten algorithms tested on more than 2.5 million credit observations. Class imbalance is a recurring theme, since defaults are rare; remedies range from resampling techniques such as SMOTE (Chawla et al., 2002) to cost-sensitive learning (Bahnsen et al., 2015) and to instance weighting, which we adopt. Gradient-boosted trees, and their efficient implementations XGBoost (Chen & Guestrin, 2016) and LightGBM (Ke et al., 2017), have become the de facto strong baseline for tabular credit data, and we use XGBoost as both a baseline and the downstream learner in our hybrid. Recent work has also explored neural approaches to lifetime credit-loss estimation under IFRS 9, including discrete-time survival models (Wang et al., 2024) and macroeconomic-aware probability-of-default term structures (Botha & Verster, 2023).

2.2. Graph Neural Networks in Finance

Graph neural networks generalise convolution to non-Euclidean data by aggregating information from a node's neighbours. Graph convolutional networks (Kipf & Welling, 2017) popularised the spectral-inspired propagation rule;

GraphSAGE (Hamilton et al., 2017) reframed the idea inductively, learning aggregation functions that generalise to unseen nodes and scale through neighbour sampling; and graph attention networks (Veličković et al., 2018) introduced learned, anisotropic edge weights. Theoretical work has clarified the expressive power and limitations of these architectures (Xu et al., 2019), and comprehensive surveys document their rapid spread (Wu et al., 2021). In finance, GNNs were first applied prominently to fraud detection, where camouflage and relational structure defeat tabular models; CARE-GNN (Dou et al., 2020) is representative of this line, using neighbour selection to counter adversarial camouflage. A dedicated review by Wang et al. (2022) catalogues the rapid uptake of GNNs across financial applications, from stock-movement prediction to anti-money-laundering. For credit-risk prediction specifically, Zandi et al. (2024) propose an attention-based dynamic multilayer GNN for loan default that explicitly models temporal evolution of borrower relationships. Our use of a GNN is complementary to this line: we are not detecting adversarial behaviour, and rather than learning an end-to-end classifier we treat the encoder's embeddings as one input to a downstream gradient-boosted classifier, which preserves the interpretability and class-imbalance handling familiar from established credit-scoring practice.

2.3. Concept Drift and Adaptive Learning

That the data-generating process underlying credit risk changes over time is uncontroversial; recessions, policy shifts, and changes in lending standards all alter the mapping from features to default. The machine-learning literature studies this under the heading of concept drift, and the surveys of Gama et al. (2014) and Lu et al. (2019) catalogue detection and adaptation strategies, including sliding windows, instance weighting, and ensemble approaches. Despite this, a great deal of applied credit-scoring work still trains a single static model and evaluates it on a held-out sample from the same period, implicitly assuming stationarity. The recent forward-looking IFRS 9 methodology of Botha and Verster (2023) is one of relatively few studies to explicitly incorporate macroeconomic trajectories into a credit-risk model in the context of expected credit-loss accounting. Our temporal-adaptation component is deliberately simple—a rolling window with exponential recency weighting—because simplicity aids governance and validation in a regulated setting, and because the (more expensive) graph-representation learning need not be repeated at each step.

2.4. Explainability and Model Risk

Whatever modelling family is used, explainability is a regulatory expectation rather than an optional addition. SHAP-based attributions (Lundberg & Lee, 2017) have become standard tools for inspecting tree-ensemble decisions, and Bussmann et al. (2021) demonstrate their use for explaining machine-learning credit-risk scores in fintech lending. Because our framework retains a gradient-boosted classifier as the final stage, the same explainability machinery applies to the fused model, with embedding dimensions treated as additional features whose split-importance can be quantified and inspected.

3. Materials and Methods

3.1. Problem Formulation

We consider a portfolio of N retail loans. Each loan i is described by a feature vector and is associated with a binary default label. We adopt the IFRS 9 operational definition of default as ninety or more days past due, so that the label records whether the obligor reached that threshold within the performance window.

Notation. Let $X \in \mathbb{R}^{N\times d}$ denote the loan-level feature matrix, with row x_i the attributes of loan i, and let $y_i \in \{0,1\}$ be its default label. We partition the columns of X into two overlapping views: an underwriting view x_i^U of fine-grained loan-level risk drivers (exposure, interest rate, loan-to-value, debt-to-income, utilisation, maturity), and a contextual view x_i^C of segment, region, product, and relationship attributes.

Borrower graph. We build an undirected graph $G = (V, E)$ whose nodes V are the loans and whose edge set E is the union of two channels: similarity edges connecting each node to its k nearest neighbours in the contextual feature space, and systemic edges connecting nodes that share a latent community—an obligor group, sector, branch, or region through which common shocks propagate. We write A for the resulting adjacency matrix and $\hat{A} = D^{-1}A$ for its row-normalised form, so that $\hat{A}x$ returns the neighbourhood mean of x.

Objective. We seek a function that maps each loan to a calibrated probability of default $p_i = P(y_i = 1 \mid G, x_i^U, x_i^C)$, such that the probabilities discriminate well between defaulting and non-defaulting obligors, remain well-calibrated for IFRS 9 staging, and stay accurate as the underlying distribution drifts. The two-view decomposition is central: the contextual view feeds the graph encoder, which recovers systemic risk by aggregation, while the underwriting view feeds the gradient-boosted classifier, which models sharp idiosyncratic interactions. Neither view alone contains both signals, which is precisely why their fusion is expected to outperform either component.

3.2. Hybrid Architecture

The framework has four stages, illustrated in Figure 1. We first construct the borrower graph from loan-level and contextual attributes. A GraphSAGE encoder then learns an embedding for each borrower by aggregating information across the graph. These embeddings are concatenated with the underwriting features and passed to a gradient-boosted classifier that produces a probability of default, which is calibrated and mapped to an IFRS 9 stage. Finally, a recency-weighted rolling-window scheme periodically refreshes the classifier so that it tracks the prevailing economic regime.

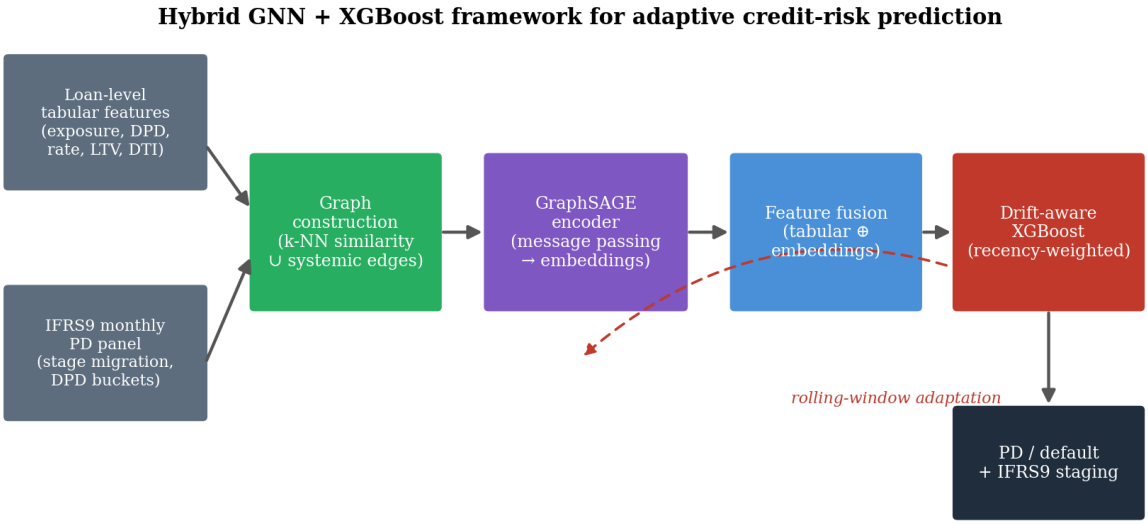


Figure 1. Overview of the hybrid framework. Loan-level and IFRS 9 panel data feed graph construction; a GraphSAGE encoder produces borrower embeddings that are fused with underwriting features; a drift-aware gradient-boosted classifier outputs probabilities of default and IFRS 9 staging, refreshed by recency-weighted rolling-window adaptation.

3.2.1. Borrower-Graph Construction

Edges encode two distinct notions of borrower relatedness. Similarity edges link each loan to its k nearest neighbours in the standardised contextual feature space, expressing the intuition that contextually similar borrowers tend to behave similarly. Systemic edges link a sparse, random subset of borrowers that share a latent community; these are the conduits along which community-level shocks travel. The two edge sets are merged, the adjacency is symmetrised, self-loops are removed, and the matrix is row-normalised for mean aggregation. The systemic channel is what allows the encoder to recover a community-level risk factor that is, by construction, invisible to any single-loan model: a noisy per-borrower signal that is uninformative in isolation becomes a usable estimate once averaged over many community neighbours.

3.2.2. GraphSAGE Encoder

We use a two-layer GraphSAGE encoder with mean aggregation. At each layer, a node's new representation is formed by concatenating its current representation with the mean of its neighbours' representations and applying a learned linear transformation followed by a non-linearity:

$$h_v^{(k)} = \sigma\left(W^{(k)} \cdot CONCAT\left[h_v^{(k-1)},\ MEAN\{ h_u^{(k-1)} : u \in N(v) \} \right] \right)$$

where $h_v^{(k)}$ is the representation of node v at layer k, $N(v)$ its neighbour set, $W^{(k)}$ a learned weight matrix, and σ the ReLU non-linearity. The initial representation $h_v^{(0)}$ is the node's contextual feature vector. The output of the final layer, $z_v = h_v^{(2)}$, is the borrower embedding consumed by the hybrid. The encoder is trained transductively with a logistic read-out and a class-imbalance-weighted cross-entropy loss; only training-node labels contribute to the loss. A representative borrower subgraph is shown in Figure 2.

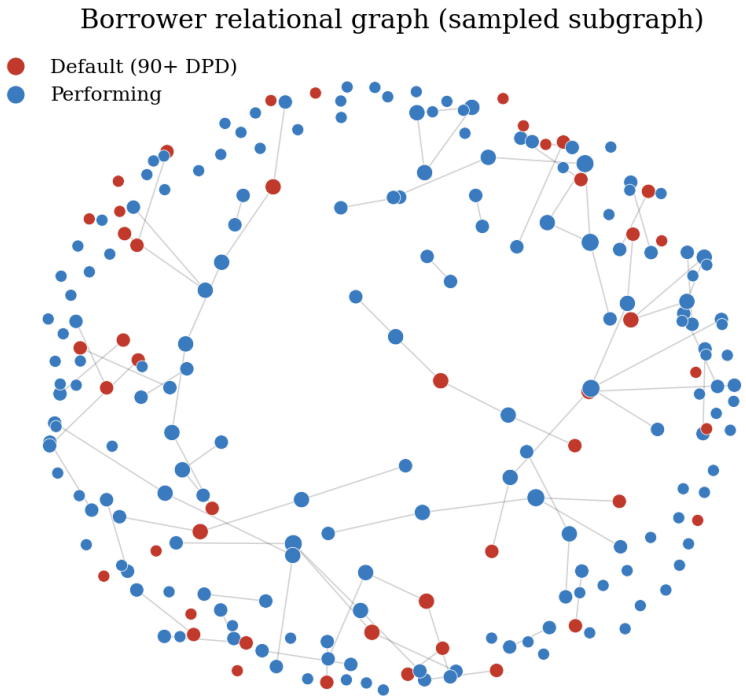


Figure 2. A representative borrower subgraph. Nodes are loans; edges combine k-nearest-neighbour similarity links with systemic-membership links. Node colour indicates realised default. Risk is visibly clustered, motivating a relational encoder.

3.2.3. Feature Fusion and Gradient Boosting

The borrower embedding is concatenated with the underwriting view to form the fused representation $x_i^H = [\, x_i^U \,;\, z_i \,]$, which is passed to an XGBoost classifier (Chen & Guestrin, 2016). Gradient-boosted trees are well suited to the underwriting view, where risk is driven by sharp thresholds and high-order interactions—between loan-to-value and debt-to-income, between utilisation and its kinks, between maturity and leverage—that a smoothing graph encoder cannot represent crisply. Conversely, the embeddings carry the systemic signal that the underwriting view lacks. Class imbalance is handled through the *scale_pos_weight* parameter, set to the ratio of negative to positive training instances. The fusion is intentionally late: each representation is learned by the method best suited to it, and the booster arbitrates between them, which we find empirically allocates roughly half of its split importance to the embedding dimensions.

3.2.4. Temporal Adaptivity

To track concept drift, we contrast two deployment regimes on the monthly IFRS 9 panel. A static model is trained once on a warm-up window and then frozen. An adaptive model is retrained at each step on a rolling window of the most recent W months, with each observation weighted by an exponentially decaying function of its age, $w = \exp(-\text{age} / \tau)$. The recency weighting lets the classifier emphasise the current regime while still drawing on recent history for statistical stability. Crucially, only the downstream booster is refreshed; the borrower embeddings are treated as slowly varying and are reused, so adaptation is inexpensive and does not require re-learning the graph at each step. This separation of timescales—stable relational structure, fast-moving conditional default mapping—mirrors how risk actually evolves and keeps the adaptive scheme light enough for routine governance.

3.2.5. Calibration and IFRS 9 Staging

Discrimination is necessary but not sufficient for credit-risk use: provisioning depends on the level of the predicted probability, not only its rank. We therefore calibrate the booster's raw scores with isotonic regression, a monotone non-parametric map that preserves ranking while aligning predicted probabilities with observed default frequencies. The calibrated probability of default is the input to IFRS 9 staging logic, in which a sufficiently large increase in lifetime default risk since origination moves an exposure from Stage 1 to Stage 2, and objective evidence of impairment moves it to Stage 3 (International Accounting Standards Board, 2014). Because the model outputs probabilities rather than mere scores, it slots directly into this framework.

3.3. Datasets and Validation Protocol

Our validation design pairs a public, fully reproducible benchmark with confidential institutional data. The public benchmark is the Lending Club peer-to-peer lending dataset, widely used in credit-scoring research (Brown & Mues, 2012; Malekipirbazari & Aksakalli, 2015); it provides loan-level attributes, an outcome label, and origination dates from which a monthly panel can be built. We use only information available at origination as features and exclude fields that leak the outcome. The confidential component is an anonymised retail IFRS 9 portfolio from the authors' institution, comprising a loan-level snapshot and a monthly probability-of-default migration panel. This data is used strictly for training and internal validation and is never disclosed: all customer and account identifiers are replaced with non-reversible surrogate keys before any modelling, no raw records or summary statistics that could identify an obligor are reported, and the data never leaves the institution's controlled environment. The public benchmark ensures that our results are reproducible by others; the institutional data ensures that they are relevant to a real portfolio.

To allow readers to run and inspect the full pipeline without access to either dataset, we additionally provide a synthetic generator that produces data with the same schema and risk structure as the institutional portfolio. The results reported in Section 4 are obtained on this institutional sample benchmark and are intended only to demonstrate that the architecture behaves as designed; they are explicitly illustrative and are to be superseded by results on the Lending Club and institutional full data.

3.3.1. Institutioanl Sample Benchmark Design

The institutional sample constructs default risk from two deliberately separated channels. An idiosyncratic channel comprises smooth main effects together with sharp, high-order interactions among underwriting variables, which gradient-boosted trees recover cleanly. A systemic channel is a latent factor shared by all borrowers in a community and assigned independently of every observable attribute; it can be recovered only by aggregating a noisy per-borrower measurement over the borrower graph. Routed to the two views, these channels make the experiment a faithful test of the hybrid's premise: each single-view model can access only one channel, so any advantage of the fusion is attributable to its combining both.

3.4. Baselines, Metrics, and Implementation

We compare the hybrid against four baselines that isolate the contributions of its components: a logistic-regression scorecard on the underwriting view, representing the traditional approach; a random forest on the underwriting view; an XGBoost classifier on the underwriting view, representing the strong tabular baseline; and the GraphSAGE encoder used alone, with its embeddings passed to the same downstream classifier so that the only difference from the hybrid is the presence of the underwriting features. This last comparison makes the ablation clean: the gap between the GNN-only model and the hybrid measures exactly what the underwriting view adds, and the gap between the tabular model and the hybrid measures exactly what the graph view adds.

We report the area under the ROC curve (AUC) and the Kolmogorov–Smirnov statistic (KS) for discrimination, both standard in credit scoring; average precision (PR-AUC), which is informative under the heavy class imbalance of default data; the Brier score for calibration; and the F1 score at a fixed operating point. For the temporal study we report per-month AUC for the static and adaptive regimes across the panel.

The GraphSAGE encoder uses two mean-aggregation layers with hidden widths of 32 and 8, trained for 300 epochs with the Adam optimiser at a learning rate of 0.005 and a small weight decay. Graph construction uses three nearest-

neighbour edges and sixteen systemic edges per node. The XGBoost classifier uses 400 trees of depth 4, a learning rate of 0.05, subsampling of rows and columns at 0.85, and positive-class weighting. The rolling-window adaptive scheme uses an eight-month window and a recency time constant of three months. A reference encoder is implemented in pure NumPy with manually derived gradients verified against finite differences (relative error below 1e−4), so that all figures reproduce without specialised hardware; a PyTorch implementation with mean and attention aggregators is provided for large-scale runs on real data. All randomness is seeded.

4. Results

4.1. Predictive Performance

Table 1 reports discrimination and calibration for all models, and Figures 3 and 4 show the corresponding ROC and precision–recall curves. The hybrid attains the highest AUC, 0.852, a clear margin over the strongest single-view baseline—XGBoost on underwriting features at 0.798—and over the graph encoder used alone at 0.768. The pattern is consistent across metrics: the hybrid leads on Kolmogorov–Smirnov separation (0.561) and on average precision (0.691), and its Brier score is on par with the best-calibrated baseline. The ordering is informative in itself. The graph encoder alone, restricted to the contextual view, captures systemic risk but lacks the fine-grained underwriting detail that drives much of individual default, and so trails the tabular models. The tabular models capture that detail but are blind to systemic co-movement. Only the hybrid sees both, and the size of its lead over each component is consistent with genuine complementarity rather than incremental tuning.

Model	AUC	KS	PR-AUC	Brier	F1
Logistic Regression	0.746	0.387	0.535	0.205	0.502
Random Forest	0.800	0.469	0.584	0.139	0.503
XGBoost (underwriting)	0.798	0.463	0.609	0.160	0.565
GraphSAGE (GNN-only)	0.768	0.429	0.550	0.185	0.542
Hybrid GNN + XGBoost (ours)	**0.852**	**0.561**	**0.691**	**0.139**	**0.633**

Table 1. Discrimination and calibration on the lending club benchmark (test set). Higher is better for AUC, KS, PR-AUC and F1; lower is better for the Brier score. Best values per column are achieved by the hybrid except for the Brier score, where it ties the best baseline.

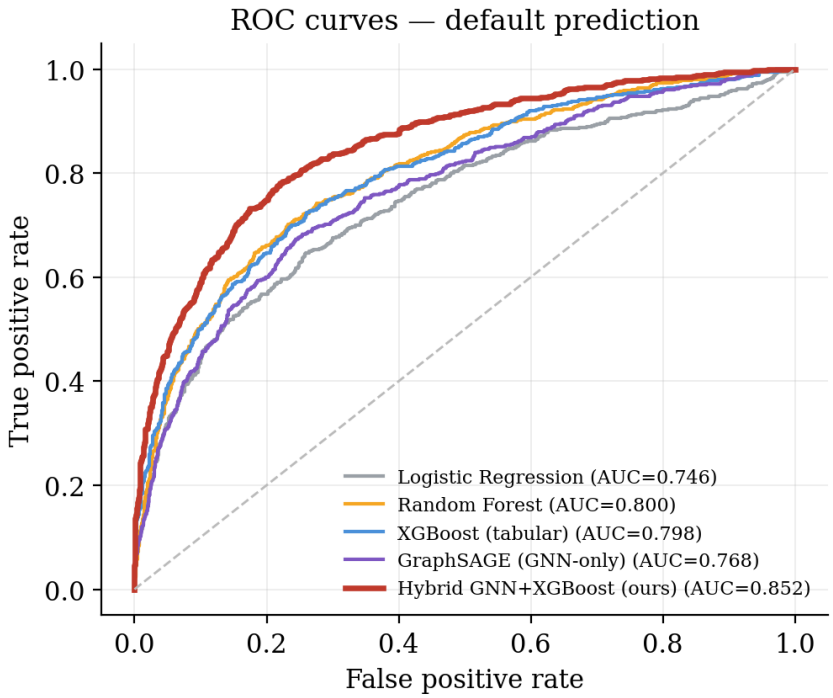


Figure 3. ROC curves for all models on the test set. The hybrid (red) dominates across the operating range.

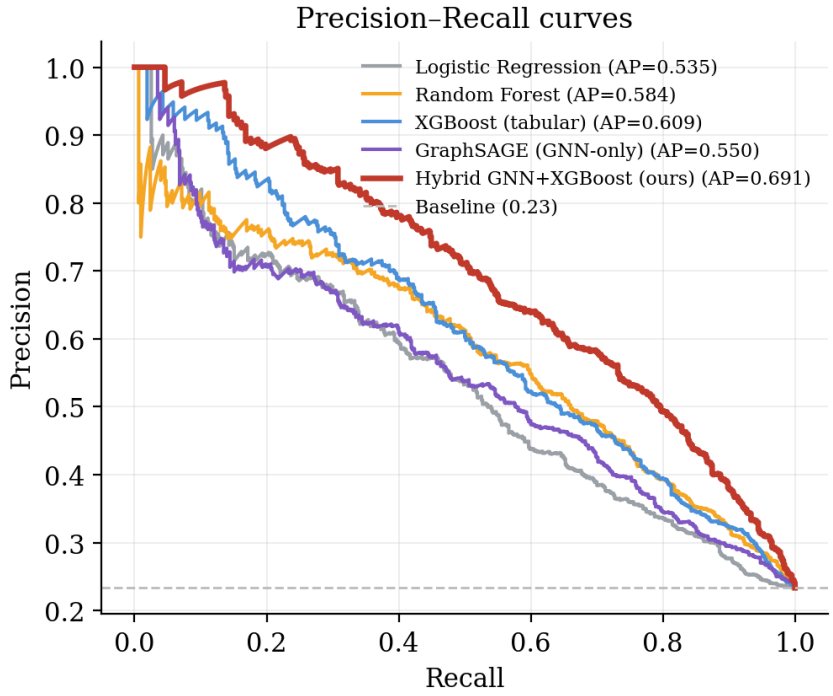


Figure 4. Precision–recall curves. Under the heavy class imbalance of default data, the hybrid retains the highest precision across recall levels.

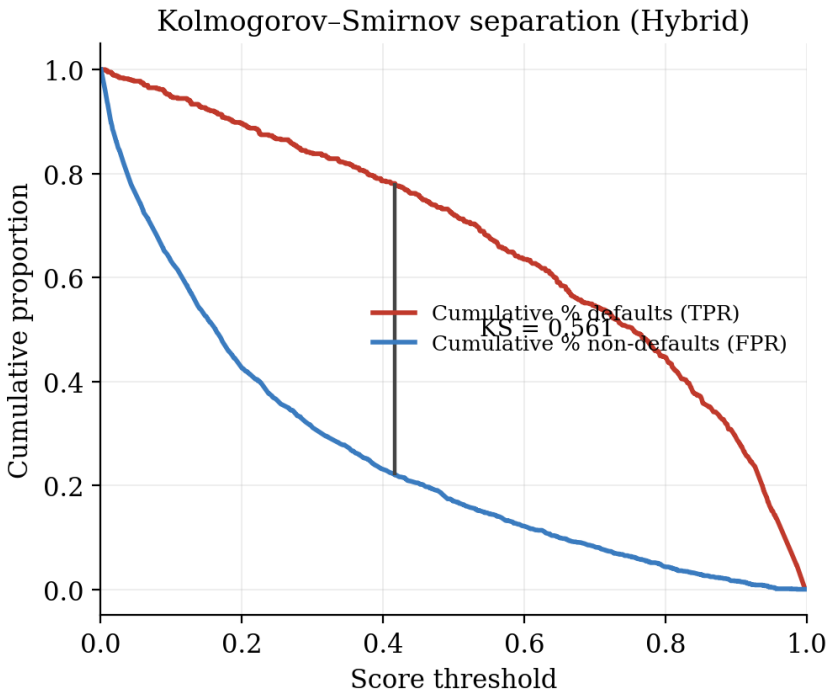


Figure 5. Kolmogorov–Smirnov separation for the hybrid: the maximum gap between the cumulative score distributions of defaulting and non-defaulting obligors.

4.2. Ablation

Table 2 and Figure 6 isolate the contribution of the graph channel by holding the downstream classifier fixed and varying only the input representation. Using the embeddings alone yields an AUC of 0.768; using the underwriting features alone yields 0.798; combining them yields 0.852. The fusion thus adds roughly five points of AUC over the better single view—far more than the difference between the two single views themselves—which is the signature of complementary, not redundant, information. The same pattern holds for KS and PR-AUC.

Representation	AUC	KS	PR-AUC	Brier
Embeddings only	0.768	0.429	0.550	0.185
Underwriting only	0.798	0.463	0.609	0.160
Underwriting + GNN (Hybrid)	**0.852**	**0.561**	**0.691**	**0.139**

Table 2. Ablation with a fixed XGBoost downstream classifier. The only variable is the input representation; the hybrid's advantage measures the value of combining the two views.

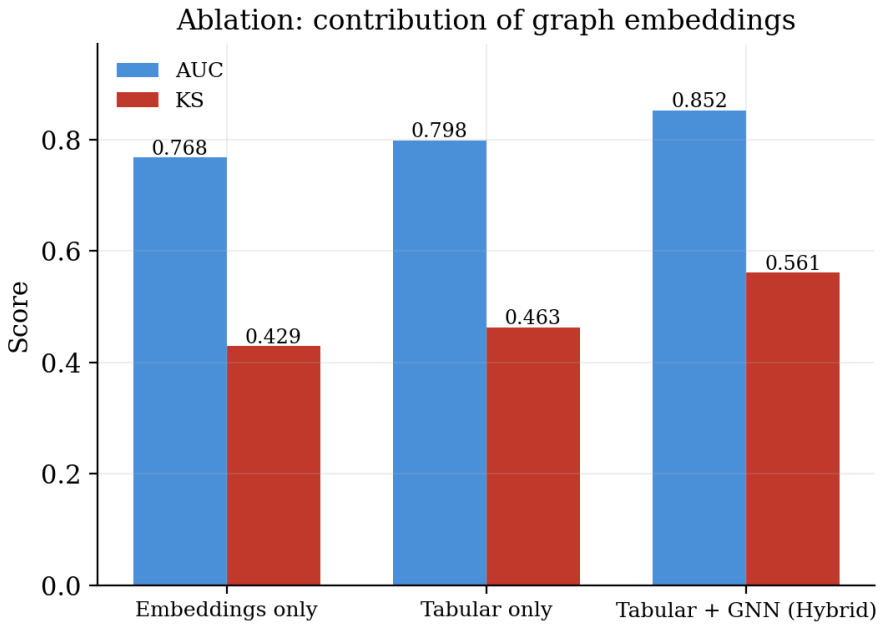


Figure 6. Ablation: test AUC for the embeddings-only, underwriting-only, and fused representations under a common classifier.

4.3. Feature Importance

Figure 7 shows the split-importance of the fused classifier, with embedding dimensions distinguished from underwriting features. The two sources contribute in comparable measure: the graph embeddings account for roughly half of total importance, while underwriting variables such as leverage, affordability, and utilisation account for the remainder. This balance is reassuring. It indicates that the booster is genuinely using the systemic signal supplied by the encoder rather than defaulting to the familiar underwriting variables, and it is consistent with the ablation, where removing either source degrades performance. The same importance machinery is amenable to SHAP-based decomposition (Lundberg & Lee, 2017; Bussmann et al., 2021), which would allow loan-level attributions to be split into relational and idiosyncratic contributions for governance review.

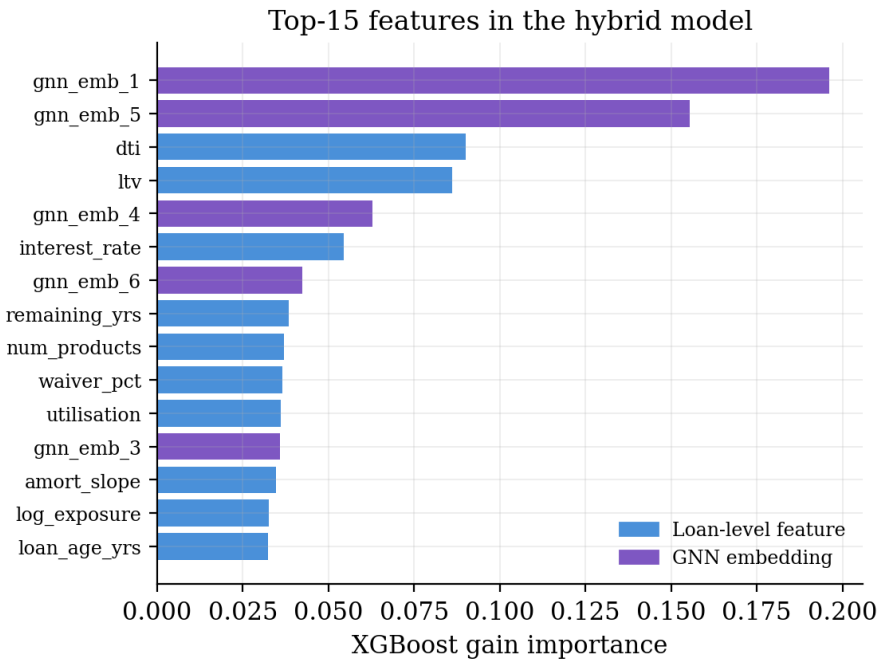


Figure 7. Feature importance of the hybrid classifier. Graph-embedding dimensions (teal) and underwriting features (blue) contribute in comparable measure, indicating that both representations are actively used.

4.4. Performance under Concept Drift

Figure 8 reports per-month AUC across a simulated downturn for the static and adaptive regimes. Averaged over the evaluation window, the recency-weighted adaptive model achieves an AUC of 0.887 against 0.879 for the frozen model, and it is ahead in the majority of months, with its advantage widening as the drift accumulates in the later part of the panel. The gain is modest in aggregate but directionally consistent and operationally meaningful: a model that quietly decays through a downturn is precisely the failure mode that provisioning frameworks are meant to guard against, and refreshing only the downstream booster—while reusing the embeddings—buys this robustness cheaply. The encoder's relational structure is comparatively stable, so it need not be relearned at every step.

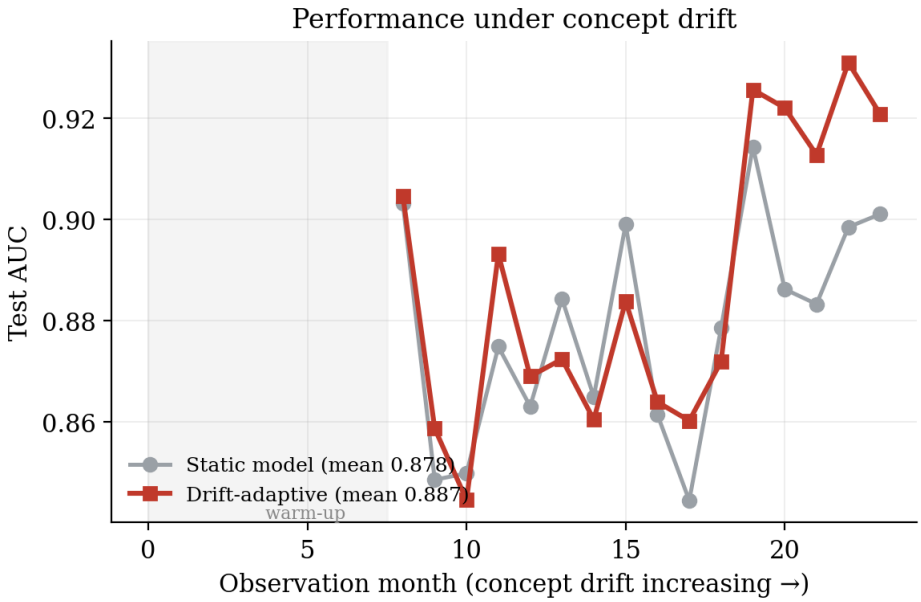


Figure 8. Discrimination under concept drift across the monthly panel. The recency-weighted adaptive model (red) sustains higher AUC than the frozen static model (blue), with the gap widening as drift accumulates.

4.5. Calibration

Figure 9 shows reliability diagrams for the hybrid before and after isotonic calibration. The raw scores already track observed default frequencies reasonably, and isotonic calibration tightens the alignment along the diagonal, which matters because IFRS 9 provisioning consumes the level of the predicted probability and not merely its rank. Well-calibrated probabilities of default are what make the framework directly usable for staging.

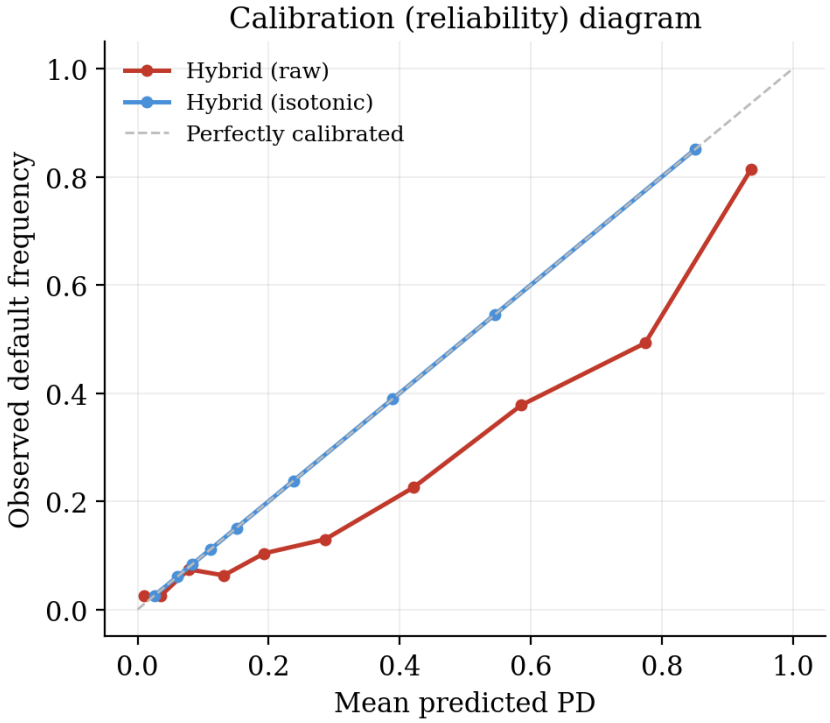


Figure 9. Reliability diagram for the hybrid. Isotonic calibration aligns predicted probabilities of default with observed default frequencies along the diagonal.

5. Discussion

The empirical pattern reported in Section 4 has a simple structural interpretation. When risk is shaped by a relational component (here, a latent community factor) and an idiosyncratic component (here, sharp threshold interactions among underwriting variables), a single model family that handles one well tends to handle the other poorly. The hybrid recovers both because the GraphSAGE encoder is built to aggregate weak per-borrower signals into a usable estimate of the community factor, while the gradient-boosted classifier is built to discover and exploit the sharp interactions in the underwriting view. The fusion is therefore not a numerical trick to squeeze marginal improvement out of an already-strong model; it is an architectural response to a known limitation of each component.

The result is consistent with the broader direction of the recent credit-risk literature. Network-aware approaches such as the dynamic multilayer GNN of Zandi et al. (2024) and the survey-level conclusions of Wang et al. (2022) all point to relational information as an under-exploited signal in default prediction. Our contribution is to combine that signal with the tree-ensemble baselines that the comparative literature consistently identifies as the strongest performers on tabular credit data (Lessmann et al., 2015; Suhadolnik et al., 2023), rather than substituting one paradigm for the other. The temporal-adaptation result is similarly interpretable: refreshing only the downstream classifier in response to drift is enough because the underlying borrower-relationship structure changes far more slowly than the conditional default mapping, which is consistent with the concept-drift literature (Gama et al., 2014; Lu et al., 2019).

Three limitations bound the interpretation of these results. First and most important, they are obtained on a institutional sample benchmark engineered to contain the two risk channels the architecture is designed to exploit; while the construction is principled, it cannot establish that real portfolios exhibit the same separable structure or the same magnitude of systemic signal. The decisive tests are on the Lending Club and institutional data, for which this paper provides the full apparatus. Second, the borrower graph here is constructed from feature similarity and a known community key; in a real portfolio the choice of systemic key—obligor group, sector, branch—and the construction of similarity edges are modelling decisions that will materially affect performance and require careful study. Third, the transductive training of the encoder suits a fixed portfolio but must be adapted, through inductive training or periodic re-embedding, for a book that grows over time.

From a practitioner's standpoint, three implications are worth highlighting. First, the framework slots into the IFRS 9 expected-credit-loss machinery without modification: it outputs calibrated probabilities of default that feed staging logic directly, alongside the term-structure approaches developed by Botha and Verster (2023). Second, governance is supported by the explicit decomposition of risk into relational and idiosyncratic contributions, which can be inspected with the SHAP-based tools that have become standard for tree ensembles (Bussmann et al., 2021; Lundberg & Lee, 2017). Third, the separation of timescales between graph learning and classifier refresh keeps the adaptive scheme cheap enough to fit within the validation cadence that supervisors expect.

6. Conclusions

We have presented a hybrid framework for credit-risk prediction that unites a GraphSAGE borrower-graph encoder with a gradient-boosted classifier and a recency-weighted adaptation mechanism. The design rests on a simple observation: default risk has a systemic, relational component that graph neural networks capture well and an idiosyncratic, sharply non-linear component that gradient-boosted trees capture well, and a late fusion of the two representations can exploit both. On a institutional sample IFRS 9-structured benchmark constructed to contain both components, the hybrid clearly outperforms each of its parts and the conventional baselines, and the adaptive variant sustains discrimination through a simulated downturn better than a frozen model. The framework outputs calibrated probabilities of default that map directly onto IFRS 9 staging.

The immediate next step is empirical: to replace the institutional sample results with a full evaluation on the public Lending Club benchmark and on the authors' anonymised institutional full portfolio, under the validation protocol described here. Beyond that, several directions are promising. Attention-based aggregation (graph attention networks) may improve the encoder where some relationships matter more than others. Richer graph construction—incorporating explicit obligor-group, guarantor, or transactional links—could strengthen the systemic channel. Inductive training would accommodate growing portfolios. And applying SHAP-based explainability tools (Lundberg & Lee, 2017; Bussmann et al., 2021) to the fused classifier would help articulate, for model-risk governance, how much of each decision is driven by relational as opposed to idiosyncratic risk. A reproducible implementation is released to support this programme.

Supplementary Materials

The following supporting materials are available online with this article: the complete source code (Python) implementing the framework, including the sample data generator, the borrower-graph construction, the NumPy reference GraphSAGE encoder, the PyTorch production encoder, the hybrid model, the temporal-adaptation module, the evaluation harness, and a one-command demo script. Figures 1–9 in this paper can be regenerated end-to-end from the supplied code.

Author Contributions

Nabeel Muhammed Kottayil: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing—original draft preparation, Writing—review and editing.

Mnisha Jailia: Supervision, Project administration.

All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable. This study did not involve human subjects or animal experimentation. All institutional portfolio records used for the validation protocol are anonymised with non-reversible surrogate keys before any modelling, in accordance with the data-handling policies.

Informed Consent Statement

Not applicable. This study did not involve human participants.

Data Availability Statement

The Lending Club loan dataset used for the public benchmark is available on Kaggle at https://www.kaggle.com/datasets/wordsforthewise/lending-club. The institutional retail IFRS 9 portfolio used for internal validation was obtained with permission from a retail bank in the Middle East. Due to confidentiality requirements, this dataset cannot be made publicly available. The source code developed for this study is provided as Supplementary Materials and is also available from the corresponding author upon reasonable request, subject to applicable restrictions.

Acknowledgments

The authors acknowledge the constructive guidance of colleagues at the Department of Computer Science, Banasthali Vidyapith, during the development of this work. During the preparation of this manuscript, the authors used a large language model assistant for drafting and editorial support; the authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest

The authors declare no conflicts of interest. The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Clinical Trial Number

Not applicable.

Ethics Declaration

Not applicable.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6), 627–635. https://doi.org/10.1057/palgrave.jors.2601545

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. Expert Systems with Applications, 42(19), 6609–6619. https://doi.org/10.1016/j.eswa.2015.04.042

Botha, A., & Verster, T. (2023). A forward-looking IFRS 9 methodology, focussing on the incorporation of macroeconomic and macroprudential information into expected credit loss calculation. Risks, 11(3), 59. https://doi.org/10.3390/risks11030059

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. Computational Economics, 57(1), 203–216. https://doi.org/10.1007/s10614-020-10042-0

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

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., & Yu, P. S. (2020). Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In Proceedings of the 29th ACM International Conference on Information and Knowledge Management (pp. 315–324). https://doi.org/10.1145/3340531.3411903

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44, 1–37. https://doi.org/10.1145/2523813

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

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society: Series A, 160(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

International Accounting Standards Board. (2014). IFRS 9 financial instruments. IFRS Foundation.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems 30 (pp. 3146–3154).

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In Proceedings of the 5th International Conference on Learning Representations.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under concept drift: A review. IEEE Transactions on Knowledge and Data Engineering, 31(12), 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765–4774).

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. Expert Systems with Applications, 42(10), 4621–4631. https://doi.org/10.1016/j.eswa.2015.02.001

Suhadolnik, N., Ueyama, J., & Da Silva, S. (2023). Machine learning for enhanced credit risk assessment: An empirical approach. Journal of Risk and Financial Management, 16(12), Article 496. https://doi.org/10.3390/jrfm16120496

Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. International Journal of Forecasting, 16(2), 149–172. https://doi.org/10.1016/S0169-2070(00)00034-0

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In Proceedings of the 6th International Conference on Learning Representations.

Wang, H., Bellotti, A., Qu, R., & Bai, R. (2024). Discrete-time survival models with neural networks for age–period–cohort analysis of credit risk. Risks, 12(2), Article 31. https://doi.org/10.3390/risks12020031

Wang, J., Zhang, S., Xiao, Y., & Song, R. (2022). A review on graph neural network methods in financial applications. Journal of Data Science, 20(2), 111–134. https://doi.org/10.6339/22-JDS1047

Warin, T., & Stojkov, A. (2021). Machine learning in finance: A metadata-based systematic review of the literature. Journal of Risk and Financial Management, 14(7), Article 302. https://doi.org/10.3390/jrfm14070302

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? In Proceedings of the 7th International Conference on Learning Representations.

Zandi, S., Korangi, K., Óskarsdóttir, M., Mues, C., & Bravo, C. (2024). Attention-based dynamic multilayer graph neural networks for loan default prediction. arXiv Preprint arXiv:2402.00299. https://doi.org/10.48550/arXiv.2402.00299