Corporate Financial Distress Prediction Using Machine Learning and GraphNeural Networks

Abstract. In an environment where financial distress and bankruptcy can have major consequences for firms and markets, early prediction of a company's financial condition is crucial for management and decision-making. However, financial statement data are often incomplete over time, and traditional prediction models may fail to exploit the relational structure among firms. In this paper, we formulate the task of predicting financial distress/bankruptcy for Nasdaq-100 companies by comparing conventional machine-learning models with graph neural network (GNN) approaches, and we further motivate the integration of a causal graph to improve predictive performance. We use a Kaggle dataset containing 45 financial indicators and ratios for Nasdaq-100 firms over 2017–2022 (seven annual periods); the data are cleaned, missing/infinite values are handled, and the dataset is transformed into a panel format before model training. Empirically, among traditional models, Gradient Boosting achieves an F1-score of 74.07%; among GNN models, GraphSAGE attains the best F1-score of 75.86%, suggesting that leveraging graph-based relational information can improve prediction quality. Future work will focus on optimizing graph construction and integrating the causal graph, as well as deploying the model as a decision-support module within enterprise management systems.

Keywords: Financial distress prediction; corporate bankruptcy prediction; machine learning; graph neural networks (GNN); graph-based learning; financial ratios; financial statement data; risk prediction.

1 Introduction

In today's business environment, competition is intensifying due to globalization and rapid technological innovation, while the macroeconomic landscape is continuously affected by geopolitical uncertainty, financial market volatility, and supply-chain disruptions, along with the consequences of tightened financial conditions and higher interest rates aimed at controlling inflation. These factors increase the variability of capital costs, make corporate cash flows less stable, and heighten the risk of deteriorating financial health-making early prediction of corporate financial distress and bankruptcy a topic of clear practical and academic importance. Bankruptcy rarely happens abruptly; it is typically the result of a prolonged deterioration process driven by multiple factors such as weak financial management, adverse market conditions, regulatory changes, and unexpected macroeconomic shocks [1]. Accurate early warnings can help investors, creditors, and corporate managers take preventive actions, reduce losses, and improve risk monitoring and decision-making [2, 3].

Methodologically, foundational studies show that early bankruptcy prediction primarily relied on financial ratios and statistical models: Beaver [4] proposed a ratio-based approach; Altman [5] introduced the Z-score using discriminant analysis; and Ohlson [6] developed a logit model to estimate bankruptcy probability [4-6]. However, these classical approaches often face several structural limitations: (i) linearity and distributional assumptions (or simple functional forms) make it difficult to capture nonlinear relationships and higher-order interactions among financial variables; (ii) sensitivity to industry heterogeneity and regime shifts over time; and (iii) heavy dependence on financial statement quality, which may be affected by earnings management. Accordingly, systematic reviews document a clear shift toward “intelligent techniques” and machine learning to better exploit high-dimensional data and complex patterns in bankruptcy prediction.

Within machine learning, ensemble methods are frequently reported to perform well because combining multiple base learners can reduce variance/overfitting and learn nonlinear decision boundaries, thereby improving classification performance under noisy and complex financial data [7]. A representative empirical study is Barboza, et al. [8] in Expert Systems with Applications, which reports that machine learning models, especially ensemble approaches often achieve better predictive performance than traditional techniques in bankruptcy detection tasks.

More recently, graph-based learning and graph neural networks (GNNs) have been explored to leverage relational structures among firms (e.g., ownership ties, board/interlocking directorates, guarantee networks, or financial similarity), enabling models to capture not only firm-level “intrinsic risk” but also spillover risk through network connections [9, 10]. Emerging work on graph representation learning for credit risk/default detection, as well as GNN and heterogeneous GNN models for bankruptcy prediction [11], suggests that incorporating meaningful relational information can further improve predictive performance when the underlying graph is informative.

This study develops and evaluates a binary classification framework for predicting financial distress/bankruptcy among Nasdaq-100 firms using a Kaggle dataset of 45 financial indicators and ratios over 2017–2022. The pipeline includes data cleaning, handling missing/infinite values, transforming the data into a panel structure, and constructing labels based on financial-risk criteria (e.g., the Altman Z-score) [5]. On the same dataset, GraphSAGE achieves the best F1-score (75.86%), while Gradient Boosting is the strongest traditional baseline (F1 = 74.07%). This outcome can be explained by a key property of GNNs: rather than treating each firm as an independent observation, GraphSAGE aggregates information from neighboring nodes in a graph (firms connected by similarity or economic relationships), thereby capturing spillover signals and latent relational structure that conventional models do not explicitly represent. Moreover, integrating a causal graph can constrain relationships among financial indicators and among firms to better reflect underlying financial mechanisms, reducing reliance on spurious correlations and improving consistency under noisy or time-varying data. Future work will optimize graph construction (edge criteria, edge weights, temporal dynamics) and more tightly incorporate causal-graph signals, with the goal of deploying the model as a practical decision-support module: (i) automated risk scoring on a monthly/quarterly basis, (ii) a risk dashboard highlighting key drivers behind elevated risk, and (iii) early intervention workflows (e.g., tightening credit limits, reviewing cash-flow policies, adjusting investment/cost plans) once predicted risk exceeds predefined thresholds.

2 Methodology

2.1 Background

The dataset is constructed based on the Nasdaq-100 Company Fundamental Data collected from Kaggle, where the original data were aggregated and scraped from wikipedia.de and gurufocus.com. It includes 45 financial indicators and ratios of companies listed in the Nasdaq-100 index, recorded over five consecutive fiscal years (up to September 2021). Due to differences in accounting periods, fiscal years may not fully align across companies.

The input feature set consists of fundamental financial ratios that comprehensively reflect a firm's financial condition, including operating efficiency, profitability, capital structure, liquidity, and growth, such as Debt to Equity, Current Ratio, Interest Coverage, Altman Z-score, Enterprise Value ratios, year-over-year growth indicators (Revenue, EBITDA, EPS), Piotroski F-Score, and Beneish M-Score...

The dataset presents several key challenges. First, missing values arise due to differences in financial reporting standards and firms' operating conditions across years; NAN and INF values were handled by imputing financially reasonable values to avoid bias or errors during model training. Second, the dataset exhibits class imbalance, as the number of firms in the high-bankruptcy-risk group is significantly smaller than that in the low-risk group, which may bias the model if not properly addressed. Third, financial ratios are strongly influenced by industry-specific characteristics and economic cycles, leading to substantial variation in data distributions across firms and time, thereby increasing the complexity ofthe prediction task.

Therefore, this study adopts a controlled data preprocessing pipeline to (i) consistently handle missing and abnormal values, (ii) mitigate the effects of class imbalance, and (iii) normalize/stabilize feature distributions to reduce the influence of industry heterogeneity and economic cycles.

A sense check was conducted to ensure the consistency and validity of the dataset before applying Machine Learning models. The original dataset contained 283 columns. First, four descriptive and categorical columns like symbol, company, sector, subsector, wereremoved, as they do not directly contribute to model learning. Next, six columnscontaining data for only a single year including financial distress latest, financial strengthlatest, free float percentage latest, goodwill to asset latest, predictability, profitability, were excluded to maintain temporal consistency.

For the target label, the problem was formulated as a binary classification of bankruptcy risk based on the Altman Z-score, a widely used indicator of corporate financial distress. According to the original study, firms with a Z-score < 1.8 are considered to have a high risk of bankruptcy and were assigned label 1, while firms with a Z-score ≥ 1.8 were classified as having a low bankruptcy risk and assigned label 0.

2.2 Model Selection and Architecture

The model suite in this study was deliberately designed to encompass two main methodological directions in bankruptcy prediction: traditional machine learning on tabular financial data and graph-based deep learning to capture inter-firm relationships. The baseline models include Logistic Regression, Decision, and ensemble methods such as Random Forest, Gradient Boosting, and Balanced Bagging. In addition, an SVC with an RBF kernel was employed to learn non-linear decision boundaries in the feature space.

On the graph-learning side, the study implements GCN to aggregate information from related firms, GAT to assign differential attention weights to neighboring nodes in the network, and a hybrid graph–temporal architecture to jointly capture contemporaneous inter-firm relationships and multi-year financial dynamics. Overall, this methodological framework positions traditional machine learning models as reliable benchmarks, while GNN-based models provide a more advanced approach to integrating relational and temporal information for bankruptcy risk prediction.

2.3 Model Training

This section details the implementation-level training workflow for both classical machine learning baselines and graph neural network (GNN) models under a strict time-based evaluation protocol. After splitting the dataset into training (2017–2020), validation (2021), and test (2022) sets, all models were trained exclusively on the training subset and evaluated on the test subset to prevent temporal information leakage. For consistent comparison and threshold-based analysis, each model produced both hard class predictions and predicted probabilities for the distressed class.

For scale-sensitive models (Logistic Regression and SVC), a RobustScaler was applied within a pipeline to mitigate the influence of outliers and heavy-tailed financial ratios, with scaling parameters learned solely from the training data. Logistic Regression and SVC were configured with class-weighted loss functions to address class imbalance, while SVC employed an RBF kernel with probabilistic outputs enabled for ROC and PR analysis. Tree-based models include Decision Tree, Random Forest, Gradient Boosting, and Balanced Bagging were trained directly on the unscaled feature matrix, as they are invariant to monotonic scaling. Imbalance-aware mechanisms were incorporated at the algorithm level, including balanced class weights, balanced subsampling in Random Forest, and class-balanced bootstrap sampling in Balanced Bagging. A fixed random seed (42) and parallel execution were used to enhance reproducibility and efficiency.

For GNN training, reproducibility was further ensured by fixing random seeds across Python, NumPy, and PyTorch, and enabling deterministic cuDNN settings. Node features were constructed from the same financial indicators as the classical models; infinite values were converted to missing values and imputed with zeros. To avoid data leakage, RobustScaler was fitted only on training-year observations and then applied to the full dataset.

The graph was constructed under a strict no-leakage constraint: validation and test nodes could connect only to training nodes. Two types of edges were combined: (i) sector-based edges linking each firm to three randomly sampled training firms within the same industry with a small prior weight, and (ii) similarity-based edges linking each firm to its top-10 most similar training firms based on cosine similarity in the scaled feature space. Duplicate edges were merged by retaining the maximum weight, and the final graph was represented using a PyTorch Geometric Data object with edge indices, edge weights, and train/validation/test masks.

All GNN models (GCN, GraphSAGE, and GAT) output a binary logit per node, converted to probabilities via a sigmoid function. To handle class imbalance, training employed BCEWithLogitsLoss with a positive-class weight computed from the training subset. Models were optimized using Adam (learning rate = 0.005, weight decay = 1×10^{-4}) for up to 400 epochs. Early stopping with patience of 40 was applied based on validation PR-AUC, and the best checkpoint was restored prior to testing.

Instead of fixing the classification threshold at 0.5, an optimal threshold maximizing F1-score was selected on the validation set along the Precision–Recall curve and then applied unchanged to the 2022 test set. Reported performance metrics include Precision, Recall, F1 at the tuned threshold, threshold-independent PR-AUC, and the confusion matrix. All three GNN architectures were trained under identical protocols to ensure fair comparability, with results aggregated based on test-set performance.

2.4 Hyperparameter Tuning

In this study, hyperparameter and operational tuning concentrates on two validation-driven decisions that are critical for deployment under severe class imbalance: (i) selecting the best model checkpoint through validation-based early stopping, and (ii) calibrating the classification threshold on the validation set rather than assuming a default value of 0.5. All tuning procedures were conducted exclusively on the 2021 validation subset, while the 2022 test subset was strictly reserved for final evaluation to ensure an unbiased assessment of generalization performance.

During training, class imbalance was mitigated using a weighted binary cross-entropy loss. The positive-class weight was computed solely from the training subset to avoid temporal leakage and to accurately reflect the imbalance encountered during model fitting. This weight was then incorporated into BCEWithLogitsLoss (pos_weight=…), increasing the penalty for misclassifying distressed firms without altering the data distribution itself.

For GNN models (GCN, GraphSAGE, and GAT), early stopping was implemented based on validation PR-AUC (Average Precision), which is more informative than accuracy in imbalanced settings. At each epoch, the model was updated using only training nodes, after which validation PR-AUC was computed. The model state yielding the highest validation PR-AUC was stored, and training was terminated when performance failed to improve for a predefined patience window. Prior to final inference, the best checkpoint was restored to ensure optimal validation performance.

After restoring the best model, a decision-threshold tuning step was performed on the validation set. Instead of fixing the threshold at 0.5, the threshold that maximized the F1-score along the Precision–Recall curve was selected. This tuned threshold was then held constant when evaluating the 2022 test set, allowing the study to report threshold-dependent metrics (Precision, Recall, F1) alongside threshold-independent PR-AUC.

For the temporal graph model (TGCN), the same principle was applied: the model was trained on historical time slices, validated on a later year, and its decision threshold was optimized on validation predictions to maximize F1. Model selection was therefore jointly guided by validation performance and the tuned operating point.

Overall, this tuning framework clearly separates (1) model learning driven by weighted loss on the training subset from (2) model selection and decision-threshold calibration performed on the validation subset. Such a design is particularly suitable for bankruptcy prediction, where the optimal threshold must carefully balance the risk of false negatives (missing distressed firms) against the cost of false positives (raising unnecessary alarms).

3 Results

This section presents the evaluation results of traditional machine learning models appliedto 45 financial metrics and ratios of every company included in the Nasdaq-100 stockmarket index (as of 09/2021) for the last five fiscal years. All models were trained andevaluated on the same dataset, with F-score (F1) serving as the primary performancemetric.

Table 1. Test results traditional Machine Learning Model

Model	Precision	Recall	F1	PR-AUC	Type	Graph
Gradient Boosting	0.9091	0.625	0.7407	0.8566	ML	None
Balanced Bagging	0.9091	0.625	0.7407	0.8137	ML	None
Logistic Regression	0.7857	0.6875	0.7333	0.847	ML	None
Ramdom Forest	0.8333	0.625	0.7143	0.7495	ML	None
Decision Tree	0.9	0.5625	0.6923	0.7013	ML	None
SVC (RBF)	0.875	0.4375	0.5833	0.6938	ML	None
GraphSAGE	6.25	6.25	6.25	6.899	GNN	Similarity (Sector + Cosin
GAT	9	5.625	6.923	7.007	GNN	Similarity (Sector + Cosin
GCN	7.5	3.75	5	7.017	GNN	Similarity (Sector + Cosin
GraphSAGE	8.462	6.875	7.586	7.654	Causal-GNN	Causal Similarity + Sector I
GCN	9.091	6.25	7.407	7.395	Causal-GNN	Causal Similarity + Sector I
GAT	7.5	5.625	6.429	7.151	Causal-GNN	Causal Similarity + Sector I

3.1 Performance of Traditional Machine Learning Models

The Gradient Boosting model was implemented with 400 estimators (n_estimators=400) and trained on the dataset. Gradient Boosting is an ensemble technique that builds decision trees sequentially, with each tree attempting to correct the errors made by the previous one. This method allows Gradient Boosting to focus on the harder-to-classify instances, improving its predictive power and accuracy, especially for complex data patterns. F1-score: Gradient Boosting achieved an F1-score of 0.7407, the highest among all traditional machine learning models. With a precision of 0.9091 and a recall of 0.6250, the model is highly reliable when predicting companies with high financial risk, meaning that most firms flagged as distressed are truly in distress. At the same time, it captures a substantial portion of distressed companies, making it a strong model for financial risk screening.

The Balanced Bagging model was implemented using 300 decision trees (n_estimators=300) and trained on the dataset. Balanced Bagging is an ensemble method that applies bootstrapped sampling while addressing class imbalance by creating balanced sub-samples of the dataset. Each tree in the ensemble is trained on a balanced dataset, allowing the model to capture minority class instances more effectively. F1-score: Balanced Bagging obtained an F1-score of 0.7407, matching the performance of Gradient Boosting. With a precision of 0.9091 and a recall of 0.6250, the model shows a similar balance between accurately identifying distressed firms and limiting false alarms. Its ensemble structure helps it remain robust in the presence of class imbalance, making it effective for financial distress prediction.

3.2 Graph Construction & GNN, Causal GNN Performance

This section presents the performance results of three graph-based deep learning models like GraphSAGE, GCN, and GAT together with their causal-enhanced variants (Causal GNN). All models were trained on the dataset under a strict temporal setting for binary classification of financial distress.

3.2.1 Graph Construction and Financial Network Representation

To capture relational dependencies among firms, three types of graphs were constructed to model both economic structure and financial similarity.

Sector-based Graph.

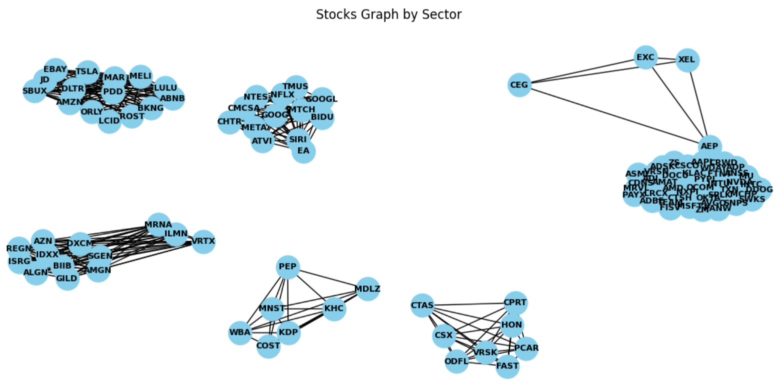

Figure 1. Stocks Graph by Sector

In this graph, each node represents a firm, and edges connect companies belonging to the same industry sector. This structure reflects shared exposure to macroeconomic conditions, regulatory environments, and demand cycles. Dense clusters observed in the graph indicate that firms within the same sector form tightly connected communities, suggesting that sectoral linkages play a crucial role in systemic risk transmission. In practice, industry-level shocks, such as downturns in technology or energy often propagate across firms within the same sector, making this graph an important representation of systematic risk.

Cosine Similarity Graph.

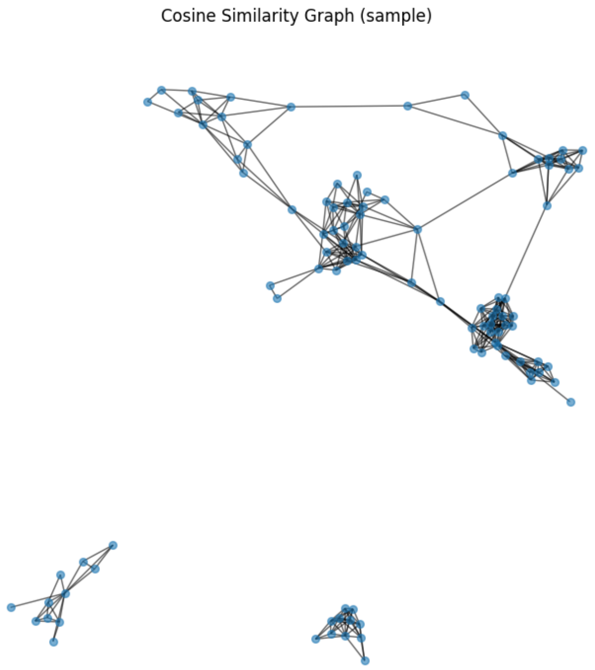

Figure 2. Consine Similarity Graph

This graph is built based on cosine similarity between firms' financial ratios, including profitability, leverage, liquidity, and cash-flow indicators. Firms with highly similar financial profiles are connected, even if they belong to different sectors. This captures latent financial relationships that are not visible from industry membership alone. For instance, companies from distinct industries may still share similar liquidity risk or debt structures, making them vulnerable to the same macro-financial shocks. Thus, the cosine similarity graph models behavioral financial resemblance rather than structural industry ties.

Merged Graph (Sector + Cosine Similarity).

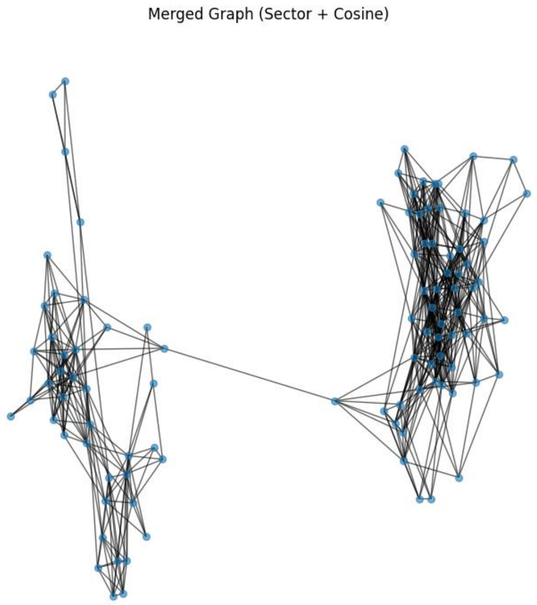

Figure 3. Merged Graph

The final graph integrates both sector-based relationships and financial similarity into a single network by assigning weighted edges from both sources. This hybrid structure better reflects real-world risk propagation, where distress can spread through both industry linkages and shared financial characteristics. By combining these perspectives, the merged graph allows GNN models to capture both industry-driven and balance-sheet-driven contagion effects.

3.2.2 GNN Performance on Baseline Graph

GCN, GraphSAGE, and GAT were trained using a temporally consistent graph, where all features were scaled based only on training data, and validation/test nodes were connected exclusively to training-period neighbors to prevent information leakage. Final evaluation was conducted on the 2022 test set.

Among the baseline GNNs, GAT achieved the strongest overall performance with a test F1-score of 0.692 and precision of 0.90. This indicates that GAT is particularly effective at minimizing false alarms, a critical requirement in financial decision-making contexts such as credit screening and portfolio risk management. The advantage of GAT stems from its attention mechanism, which enables the model to assign different importance to different neighboring firms rather than aggregating all information uniformly.

GraphSAGE produced a moderate F1-score of 0.625 with balanced precision and recall, suggesting that while neighborhood aggregation is useful, the model lacks explicit mechanisms to differentiate the relevance of neighbors.

GCN exhibited the weakest performance (F1 = 0.50, recall = 0.375), missing more than 60% of truly distressed firms. This behavior aligns with the well-known over-smoothing effect in GCNs, where node embeddings become too similar in dense graphs, reducing the model's ability to distinguish high-risk from low-risk firms.

Overall, these results indicate that attention-based GNNs (GAT) are more suitable than aggregation-only models (GCN, GraphSAGE) for financial distress prediction in relational networks.

3.2.3 Causal GNN Performance

To address limitations of similarity-based graph construction, a causal-enhanced graph framework was introduced. Instead of relying purely on cosine similarity and sector proximity, edge formation was guided by causal feature weighting to emphasize financial variables more directly linked to distress risk.

The causal graph substantially improved performance across all GNN architectures. GraphSAGE exhibited the largest gain, with its F1-score rising from 0.625 in the baseline graph to 0.7586 in the causal graph, accompanied by higher precision and recall. Causal GCN achieved very high precision (0.9091), indicating strong reliability in identifying truly distressed firms.

These improvements suggest that performance gains arise not from changing the GNN architecture, but from improving the quality of the graph itself. When the network structure better reflects underlying risk mechanisms rather than superficial similarity, message passing becomes more meaningful, leading to better balance between precision and recall.

4 Conclusion

The findings show that the effectiveness of financial distress prediction is shaped not only by algorithm selection but, more decisively, by how relationships among firms are represented. Standard GNN models that rely purely on financial similarity reveal limitations due to noise in graph structure, meaning that aggregating information from "neighbors" does not always reflect true risk dynamics. In contrast, Causal GNNs, particularly Causal GraphSAGE and Causal GCN demonstrate that integrating causal feature weighting and industry-based constraints enables the model to better capture risk propagation across firms, achieving performance competitive with traditional machine learning models. Nevertheless, the study remains constrained by its data scope limited to Nasdaq-100 firms, the length of the time series, and the handling of missing data, suggesting that future research should extend to smaller firms, employ dynamic time-evolving graphs, and incorporate macroeconomic variables, market data, and textual information from financial disclosures.

Acknowledgement

On behalf of the research group, we would like to express our sincere gratitude to Dr. Trương Công Đoàn for his valuable guidance, insightful feedback, and continuous support throughout the research process. We are also deeply grateful to the International School – Vietnam National University, Hanoi for providing academic resources, a supportive research environment, and institutional encouragement that made this study possible. Finally, we would like to thank all members of our research group for their dedication, teamwork, and meaningful contributions, which were essential to the successful completion of this study.

5 References

[1] E. Kirkos, "Assessing methodologies for intelligent bankruptcy prediction," *Artificial Intelligence Review,* vol. 43, no. 1, pp. 83-123, 2015.

[2] Y. Shi and X. Li, "An overview of bankruptcy prediction models for corporate firms: A systematic literature review," *Intangible capital,* vol. 15, no. 2, pp. 114-127, 2019.

[3] M. Wang, X. Zheng, M. Zhu, and Z. Hu, "P2P lending platforms bankruptcy prediction using fuzzy SVM with region information," in *2016 IEEE 13th International Conference on E-Business Engineering (ICEBE)*, 2016: IEEE, pp. 115-122.

[4] W. H. Beaver, "Financial ratios as predictors of failure," *Journal of accounting research,* pp. 71-111, 1966.

[5] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The journal of finance,* vol. 23, no. 4, pp. 589-609, 1968.

[6] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of accounting research,* pp. 109-131, 1980.

[7] T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Comparing boosting and bagging techniques with noisy and imbalanced data," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans,* vol. 41, no. 3, pp. 552-568, 2010.

[8] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," *Expert systems with applications,* vol. 83, pp. 405-417, 2017.

[9] S. Tewari and A. Chitnis, "Leveraging Graph Based Machine Learning to Analyze Complex Enterprise Data Relationships," 2021.

[10] Y. Ren, Y. Zhu, H. Zhu, and J. Jiang, "Prediction of Financial Crisis Events Using Graph Neural Network Model: Based on Inter-Industry Spillover Information," *Available at SSRN 4390760,* 2023.

[11] Y. Zheng, V. C. Lee, Z. Wu, and S. Pan, "Heterogeneous graph attention network for small and medium-sized enterprises bankruptcy prediction," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2021: Springer, pp. 140-151.