

FRAG-Net: An Explainable Computational Intelligence Framework for Imbalanced Financial Risk Early Warning via Conditional Wasserstein GAN and Attention-Based Feature Fusion

Meiling Dong

meilingdong@jmu.edu.cn

Jimei University

Tianjiao Sha

University of Sydney

Hanwen Zhang

Victoria University of Wellington

Research Article

Keywords: Computational intelligence, Financial risk early warning, Imbalanced learning, Conditional Wasserstein GAN, Attention-based feature fusion, Explainable artificial intelligence

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11095974/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

FRAG-Net: An Explainable Computational Intelligence Framework for Imbalanced Financial Risk Early Warning via Conditional Wasserstein GAN and Attention-Based Feature Fusion

Meiling Dong^{1*}, Tianjiao Sha² and Hanwen Zhang³

^{1*}Overseas Education College, Jimei University, 185 Yinjiang Road, Xiamen, 361021, Fujian Province, P.R. China.

²Faculty of Arts and Social Sciences, University of Sydney, Sydney, Australia.

³School of Marketing and International Business, Victoria University of Wellington, Wellington, 6140, New Zealand.

*Corresponding author(s). E-mail(s): meilingdong@jmu.edu.cn;
Contributing authors: tsha7151@uni.sydney.edu.au;
hanwen.zhang@vuw.ac.nz;

Abstract

Financial risk early warning from high-dimensional tabular data is a challenging computational intelligence problem because risk events are rare, feature relationships are highly nonlinear, and model decisions need to remain interpretable for practical decision support. Conventional classifiers trained on severely imbalanced financial data tend to favor the majority class, whereas interpolation-based oversampling methods may generate synthetic instances that do not adequately preserve the underlying distribution of rare-risk samples. To address these limitations, this paper proposes FRAG-Net, an explainable computational intelligence framework that integrates conditional generative learning, attention-based feature interaction modeling, imbalance-aware classification, and post-hoc interpretation for financial risk prediction. First, a Conditional Wasserstein Generative Adversarial Network with Gradient Penalty (CWGAN-GP) is employed to learn the latent distribution of minority-class enterprises and generate statistically consistent risk samples, thereby improving the representation of rare financial events. Second, a multi-head attention feature fusion classifier adaptively models nonlinear dependencies and differentiated contributions among

financial indicators, enabling more effective cross-feature interaction learning. Weighted binary cross-entropy is further introduced to compensate for residual class imbalance after augmentation. Finally, SHAP is incorporated to provide both global and instance-level explanations of the model decisions. Experiments are conducted on the publicly available Taiwan corporate bankruptcy dataset containing 6,819 enterprises and 95 financial features, including 220 bankruptcy samples accounting for only 3.23% of the total. FRAG-Net achieves an accuracy of 0.971, precision of 0.674, recall of 0.705, F1-score of 0.689, AUC of 0.947, G-mean of 0.851, and MCC of 0.667, outperforming conventional machine-learning and representative oversampling-based baselines. Ablation experiments and statistical significance analyses further demonstrate the complementary contributions of generative minority-distribution learning and attention-based feature fusion. These results indicate that FRAG-Net provides an effective and interpretable computational intelligence approach for rare-event risk assessment and intelligent decision support under severe class imbalance.

Keywords: Computational intelligence, Financial risk early warning, Imbalanced learning, Conditional Wasserstein GAN, Attention-based feature fusion, Explainable artificial intelligence

1 Introduction

Financial risk early warning can be formulated as an intelligent pattern-recognition and decision-support problem in which computational models infer the probability of financial distress from high-dimensional and heterogeneous enterprise indicators. With the increasing availability of structured financial data and the rapid development of computational intelligence, data-driven models have become important tools for credit assessment, bankruptcy prediction, fraud detection, and financial decision support [1–5]. Compared with conventional expert-driven assessment, computational intelligence methods are capable of learning nonlinear relationships among financial indicators and providing scalable risk evaluation for large numbers of enterprises. Traditional statistical approaches, including discriminant analysis and Logistic regression, have therefore been progressively complemented by support vector machines, decision trees, random forests, XGBoost, and LightGBM [6–12]. More recently, deep neural networks have further extended this paradigm by providing automatic representation learning and nonlinear feature-interaction modeling for financial forecasting and intelligent decision-making tasks [13–19].

Despite this progress, financial risk prediction remains a particularly challenging computational intelligence problem because real-world financial datasets commonly exhibit severe class imbalance. Enterprises experiencing bankruptcy or significant financial distress generally account for only a small proportion of all observations, while normally operating enterprises dominate the training data. Under such conditions, minimizing a conventional empirical classification objective may lead the learned decision boundary to favor the majority class. As a consequence, a model can achieve apparently high overall accuracy while maintaining insufficient sensitivity to

minority-class risk events. Similar behavior has been widely observed in fraud detection, anomaly identification, and other rare-event learning problems [20–26]. Therefore, improving the representation of minority-class samples without substantially distorting the original feature distribution has become an important issue for computational intelligence systems operating on imbalanced financial data.

Data-level resampling is one of the most widely adopted strategies for handling class imbalance. Interpolation-based methods such as SMOTE and ADASYN increase the number of minority instances by synthesizing new observations from neighboring samples. Although these techniques can alleviate numerical imbalance, their generation mechanisms do not explicitly estimate the underlying minority-class probability distribution. This limitation becomes particularly important for high-dimensional financial tabular data, where nonlinear dependencies among financial ratios may cause locally interpolated samples to occupy unrealistic regions of the feature space. Generative adversarial networks provide an alternative computational learning paradigm by modeling the latent data distribution through adversarial optimization. Previous studies have explored GAN-based augmentation in fraud detection and related anomaly-learning tasks [27–30]. However, conventional GAN training may suffer from mode collapse, unstable gradients, and poor convergence, particularly when only a limited number of minority samples are available. Wasserstein adversarial learning with gradient regularization provides a more stable mechanism for learning rare-event distributions, yet its integration with downstream intelligent feature-interaction learning for financial risk prediction remains insufficiently investigated.

A second challenge concerns feature representation and interaction modeling. Financial risk is rarely determined by an isolated indicator; instead, it emerges from nonlinear relationships among profitability, leverage, liquidity, cash flow, operating efficiency, growth, and structural variables. Conventional fully connected classifiers transform these variables implicitly but do not explicitly characterize the differentiated contributions and dependencies among them. Attention mechanisms provide a computational mechanism for adaptive information weighting and have demonstrated advantages in intelligent prediction and decision-making tasks [31, 32]. By dynamically learning dependencies among representations, attention-based models can emphasize informative financial signals while suppressing less relevant information. Nevertheless, directly applying attention-based models to highly imbalanced data does not address the inadequate representation of rare-risk samples. A computational framework that jointly considers minority-distribution learning and adaptive feature-interaction modeling is therefore required.

A third challenge is interpretability. In financial decision-support environments, predictive performance alone is insufficient because the factors contributing to a risk prediction need to be traceable and understandable. The increasing complexity of machine-learning and deep-learning models has consequently stimulated substantial research on explainable artificial intelligence in finance [33–36]. Feature-attribution methods such as SHAP provide a principled approach for quantifying the marginal contribution of individual variables to model outputs. However, existing studies generally investigate predictive modeling, class-imbalance handling, attention-based representation, and model interpretation as relatively independent components. The combined

effect of generative minority-class learning and attention-based feature fusion on both predictive performance and interpretable feature attribution has received considerably less attention.

Another important issue is that data-level balancing and algorithm-level learning are often designed independently. Conventional oversampling methods typically generate a new training set and then rely on a standard downstream classifier, whereas attention-based or deep tabular models commonly assume that the available training distribution provides sufficient representation of both classes. In severe rare-event settings, these assumptions are not necessarily satisfied. Generative augmentation may improve the coverage of minority-class regions, while attention-based learning can subsequently exploit the enriched data distribution to capture interactions among financial indicators. Combining these two mechanisms offers a potentially complementary computational intelligence strategy: the generative component addresses representation scarcity at the data level, whereas the attention component improves nonlinear discriminative modeling at the feature level. An additional imbalance-aware classification objective can further reduce residual bias after augmentation.

To address these limitations, this paper proposes FRAG-Net, an explainable computational intelligence framework for financial risk early warning under severe class imbalance. FRAG-Net integrates a Conditional Wasserstein Generative Adversarial Network with Gradient Penalty (CWGAN-GP), multi-head attention-based feature fusion, cost-sensitive classification, and SHAP-based explanation within a unified intelligent risk-analysis pipeline. Rather than relying solely on interpolation to rebalance the training distribution, CWGAN-GP learns the latent probability structure of the minority risk class and generates additional financial samples under a stable Wasserstein adversarial objective. The generated samples are combined exclusively with the original training data, after which a multi-head attention classifier models nonlinear dependencies among financial features. Weighted binary cross-entropy further compensates for residual class imbalance, while SHAP provides both global and local explanations of the resulting predictions.

The main contributions of this study are summarized as follows:

1. We develop a CWGAN-GP-based minority-distribution learning mechanism for severely imbalanced financial tabular data. In contrast to interpolation-based oversampling and conventional GAN augmentation [27–30], the proposed approach combines conditional generation, Wasserstein distribution learning, and gradient-penalty regularization to improve the stability and distributional fidelity of minority-sample generation.
2. We design an attention-based computational intelligence classifier for nonlinear financial feature-interaction modeling. Multi-head attention adaptively captures dependencies among financial representations, enabling the model to emphasize informative risk patterns instead of treating all financial indicators as equally relevant [31, 32].
3. We construct an integrated imbalance-aware learning framework that combines generative data-level balancing, attention-based feature interaction, and cost-sensitive optimization. This complementary design addresses class imbalance from both

representation-learning and classification perspectives rather than relying on a single resampling or loss-adjustment strategy.

4. We incorporate SHAP-based global and local explanation into the intelligent decision pipeline and systematically evaluate FRAG-Net on a publicly available corporate bankruptcy dataset containing 6,819 samples and 95 financial features. Comparisons with traditional machine-learning models, interpolation-based over-sampling approaches, and generative-learning baselines, together with ablation experiments and statistical significance tests, demonstrate the contributions of the individual components and the effectiveness of the complete framework.

Overall, this study is not limited to constructing a financial bankruptcy classifier. Instead, it investigates how generative distribution learning, adaptive feature interaction, imbalance-aware optimization, and explainable computational intelligence can be jointly designed for rare-event decision-support problems. From this perspective, FRAG-Net provides a general computational framework that can potentially be extended to other strongly imbalanced tabular classification scenarios in which minority-event identification and decision transparency are simultaneously important.

The remainder of this paper is organized as follows. Section 2 reviews related work on financial risk prediction, imbalanced learning, generative sample augmentation, attention mechanisms, and explainable artificial intelligence. Section 3 describes the proposed FRAG-Net architecture, including CWGAN-GP minority-sample generation, attention-based feature fusion, and the training strategy. Section 4 presents the experimental setup, comparative results, and ablation analysis. Section 5 provides the SHAP-based interpretability analysis, and Section 6 concludes the paper and discusses future research directions.

2 Related Work

2.1 Traditional Machine Learning and Deep Learning Methods for Financial Risk Early Warning

Corporate credit risk and bankruptcy prediction, as one of the most classic research problems in the field of financial risk early warning, has undergone a methodological evolution that broadly spans three stages: from statistical discriminant methods, to machine learning methods, and further to deep learning methods. Early studies were mostly based on statistical discriminant methods, such as the Z-score multivariate discriminant analysis model proposed by Altman and the Logistic-regression-based default probability estimation method; these methods possess strong interpretability and theoretical grounding but struggle to characterize the complex nonlinear relationships among financial indicators. With the development of machine learning technology, support vector machines, decision trees, random forests, and gradient-boosting models have been widely applied to corporate credit risk and bankruptcy prediction tasks: Roy and Vasa systematically reviewed the current application status of machine learning and artificial intelligence methods in fintech credit scoring and risk management [1]; Ala’raj and Abbod proposed a classifier consensus system approach to improve the robustness of credit scoring models [8]; Ma et al. predicted P2P network

lending default based on the LightGBM and XGBoost algorithms, verifying the effectiveness of gradient-boosting models on this task [9]; Serrano-Cinca et al. proposed the idea of replacing the traditional credit-scoring system with profit scoring in the P2P lending scenario [10]; Zhu et al. further applied an enhanced hybrid ensemble machine learning approach to the prediction of SME credit risk in supply-chain finance [6]; Pan proposed a novel fruit-fly optimization algorithm and validated its performance using the financial-distress model as an example [11]; Montevechi et al. conducted a systematic review of current machine-learning-driven credit-risk-modeling research, pointing out that reliance on manual feature engineering and insufficient adaptability to class imbalance remain common limitations in this field [7]; and Ainan et al., under the class-imbalanced scenario of the Polish bankruptcy dataset, conducted bankruptcy-prediction research using hybrid machine learning techniques, with an experimental setup that is highly comparable to the Taiwanese corporate bankruptcy dataset adopted in this paper [12]. The above studies indicate that traditional machine learning methods have achieved a marked improvement in discriminative performance compared with statistical methods, but they generally rely on manual feature engineering and still exhibit evident limitations when handling class-imbalanced data.

Meanwhile, deep learning methods, by virtue of their automatic feature extraction and strong nonlinear fitting capabilities, have also been progressively introduced into the fields of financial risk and financial time-series prediction. In the direction of portfolio optimization, Zhang et al. systematically explored the application potential of deep learning methods in portfolio optimization tasks [13], Cao and Li extended neural network methods to portfolio analysis problems with cardinality constraints [14], and Chaweewanchon and Chaysiri combined a machine-learning-driven stock-selection strategy with the classic Markowitz mean-variance framework [19]; in the direction of stock price and market trend prediction, Gunduz et al. used convolutional neural networks together with feature correlations to predict the intraday movements of the Istanbul Stock Exchange [16], Tsantekidis et al. employed convolutional neural networks to predict stock prices based on limit-order-book data [17], Wang et al. combined LSTM with a one-dimensional convolutional LSTM network for stock-trend prediction modeling [18], and Barua and Sharma further incorporated CNN-BiLSTM predictions into a dynamic Black-Litterman portfolio model [15]. The above studies fully validate the effectiveness of deep learning methods in financial time-series and structured-data modeling; however, as pointed out by Ahmed et al. [37] and Jagtiani and Lemieux [38] in their review and empirical studies, most existing deep learning methods still generally adopt a simple fully connected or convolutional structure to treat input features with equal weighting, failing to adequately reflect the differences in the degree to which different financial indicators contribute to risk discrimination; at the same time, deep models are also prone to overfitting and insufficient minority-class identification capability in financial scenarios with limited training samples and highly imbalanced classes, which is precisely the direct motivation for this paper to introduce sample augmentation and the attention mechanism.

2.2 Class-Imbalance Handling and GAN-Based Sample Augmentation Methods

Class imbalance is a core challenge that is pervasive in financial risk early-warning tasks and urgently needs to be addressed, as the excessively low proportion of risk samples makes models prone to being systematically biased toward the majority class during training. Existing research mainly proceeds along two paths, the data level and the algorithm level: data-level methods are represented by SMOTE and its improved variants (such as Borderline-SMOTE and ADASYN), which achieve oversampling through linear interpolation of minority-class samples; however, such methods do not, in essence, learn the true data distribution, and on high-dimensional financial tabular data they tend to generate samples that deviate from the real feature space. Algorithm-level methods, in turn, include cost-sensitive learning and ensemble-learning strategies. In the field of financial fraud and anomalous transaction detection, which is highly relevant to the task addressed in this paper, a large body of research has systematically summarized detection methods and technological evolution under class-imbalanced scenarios: Ngai et al. relatively early constructed a classification framework for financial fraud detection from a data-mining perspective and conducted an academic review of the related literature [20]; West and Bhattacharya provided a comprehensive review of intelligent financial fraud detection methods [21]; Hilal et al. systematically reviewed anomaly-detection techniques and their latest advances in fraud detection [22]; Kamuangu [24] and Ali et al. [25] further corroborated the prevalence and challenges of the class-imbalance problem in the field of financial fraud detection from the perspectives of artificial intelligence/machine learning applications and systematic literature review, respectively; the literature review by Hernandez Aros et al. reached similar conclusions [26]; Glancy and Yadav proposed a corresponding computational model targeting the specific sub-scenario of financial reporting fraud detection [23]; and Rtayli and Enneya combined support vector machine recursive feature elimination with hyperparameter optimization to improve the performance of credit card fraud detection [27].

In recent years, sample-augmentation methods based on generative adversarial networks have gradually become an important technical pathway for alleviating the class-imbalance problem. Compared with interpolation-based oversampling methods, generative adversarial networks can learn the latent distribution of minority-class samples through adversarial training, thereby generating augmentation samples that more closely resemble real data characteristics. Charitou et al. proposed a semi-supervised generative adversarial network for sample augmentation in fraud detection, validating the effectiveness of this type of method in improving minority-class identification capability [28]; Hamza et al. designed a decision-tree ensemble method, DEFDD, tailored to the financial fraud detection task, further improving detection performance under imbalanced data at the algorithmic level [29]; our research group’s previous work has also shown that a sample-augmentation strategy combining generative adversarial networks with reinforcement learning can effectively improve the detection capability for rare-class attack samples in consumer Internet-of-Things scenarios, providing a direct methodological reference for this paper’s introduction of a conditional Wasserstein generative adversarial network into the financial risk early-warning task [30]. However,

most existing studies design the GAN-based sample-augmentation module separately from the downstream classifier, lacking a systematic investigation of a joint optimization mechanism between the two; moreover, the original GAN still suffers from inherent problems such as mode collapse and training instability on high-dimensional, sparse financial tabular data, which is also the core motivation for this paper’s introduction of the Wasserstein distance and the gradient-penalty mechanism.

2.3 Attention Mechanisms and Explainable Artificial Intelligence in the Financial Domain

The attention mechanism was initially successful in the fields of natural language processing and computer vision, and in recent years it has been progressively introduced into tabular-data and financial time-series modeling tasks, enhancing the model’s focus on key information by assigning adaptive weights to features or time steps. In explainable prediction tasks, Du et al. proposed a contrastive-learning-based explainable stock-price-movement prediction method that combines attention weights with prediction rationale to improve model transparency [32]; de-la Rica-Escudero et al. proposed a post-hoc interpretability analysis framework for deep-reinforcement-learning-driven portfolio management strategies, used to reveal the key driving factors behind model decisions [31]. These studies indicate that combining the attention mechanism with an interpretability-oriented design can significantly improve the transparency of the decision-making process while maintaining the model’s discriminative performance, providing an important reference for the design of the multi-head attention feature fusion classifier in this paper.

Meanwhile, with the continuous increase in the complexity of deep learning models, the issue of model interpretability has also received growing attention from financial regulatory authorities and the academic community. Chen et al. systematically reviewed the current research status and developmental trajectory of explainable artificial intelligence in the financial domain based on bibliometric methods [34]; Yeo et al. further conducted a relatively comprehensive review of financial explainable AI research, summarizing the applicability of mainstream interpretability methods in scenarios such as credit risk and investment decision-making [35]; Heever et al. applied explainable AI methods to public-opinion analysis in the ESG and green-finance domains [36]; and Al Shiam et al. directly introduced explainable AI methods into the credit-risk-prediction task, validating the effectiveness of attribution methods such as SHAP in improving model transparency [33]. In addition, from a broader perspective of bank risk management, Heß et al. systematically reviewed the evolution of machine learning in bank risk management over the past decade [2], Butaru et al. relatively early explored risk and risk management issues in the credit card industry [3], Trigkas et al. systematically mapped the intellectual structure of computational risk analytics in banking and finance based on bibliometric methods [4], and Herrmann and Masawi provided a systematic review of the evolution of artificial intelligence applications in banking, financial services, and insurance over more than three decades [5]. The above studies corroborate, from different perspectives, the key role of interpretability analysis in the practical application of financial risk management; however, as related reviews have repeatedly pointed out, research that combines

interpretability methods such as SHAP with GAN-driven sample-augmentation models to systematically analyze the impact of the sample-augmentation process on the feature-importance distribution remains relatively scarce.

In summary, existing research still has room for improvement in three aspects: the distributional fidelity of sample augmentation, the feature-interaction modeling capability, and model interpretability. Traditional oversampling methods struggle to guarantee the distributional fidelity of generated samples [20–30]; existing deep learning methods still exhibit insufficient capability in modeling the differentiated contributions of features [13–19, 37, 38]; and the joint design of model interpretability with generative sample augmentation has rarely been systematically investigated [2–5, 31–36]. The FRAG-Net model proposed in this paper is precisely designed around these three shortcomings, aiming to provide, through an integrated design of sample augmentation, attention-based feature fusion, and interpretability analysis, a feasible solution that balances performance and interpretability for the task of financial risk early warning.

3 Method

3.1 Problem Definition

Let the training dataset for the financial risk early-warning task be denoted as

$$D = \{(x_i, y_i)\}_{i=1}^N \quad (1)$$

where $x_i \in \mathbb{R}^d$ (with $d = 95$) is the financial-ratio feature vector of the i -th enterprise sample, and $y_i \in \{0, 1\}$ is the corresponding risk label ($y_i = 1$ indicates that the enterprise is at risk of bankruptcy/default, while $y_i = 0$ indicates that the enterprise is in normal financial condition). Since the proportion of at-risk samples, $|\{i : y_i = 1\}|/N$, is typically far smaller than 5%, the objective of this paper is to learn, under severe class-imbalance conditions, the following mapping:

$$f : \mathbb{R}^d \rightarrow [0, 1], \quad \hat{y} = f(x) \quad (2)$$

so that the model can substantially improve its ability to identify minority-class (at-risk) samples while maintaining overall discriminative performance.

3.2 Data Preprocessing

The raw financial features differ markedly in scale and value range; to eliminate this scale effect and accelerate model convergence, each feature dimension is standardized using the Z-score transformation:

$$\hat{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (3)$$

where μ_j and σ_j are, respectively, the mean and standard deviation of the j -th feature dimension computed on the training set. For samples containing missing values, this paper adopts a within-class median-imputation strategy in order to avoid the distributional bias that cross-class imputation would introduce:

$$x_{ij} \leftarrow \text{median}\{x_{kj} : y_k = y_i, x_{kj} \text{ is not missing}\} \quad (4)$$

After standardization and missing-value handling, the data are stratified and split into training, validation, and test sets at a ratio of 7:1:2, providing a unified data foundation for the subsequent sample augmentation and model training.

3.3 Overall Framework

The overall FRAG-Net framework consists of four components—a data preprocessing module, a CWGAN-GP sample-augmentation module, an attention feature fusion classification module, and an interpretability analysis module—connected in an end-to-end fashion: after preprocessing, the raw financial data are fed into the CWGAN-GP module to generate additional synthetic minority-class samples, which are then mixed with the original training data and input into the attention feature fusion classifier to perform risk discrimination; finally, the model’s decisions are explained via the SHAP method. The overall workflow is shown in Fig. 1, and its execution logic is summarized in Algorithm 1.

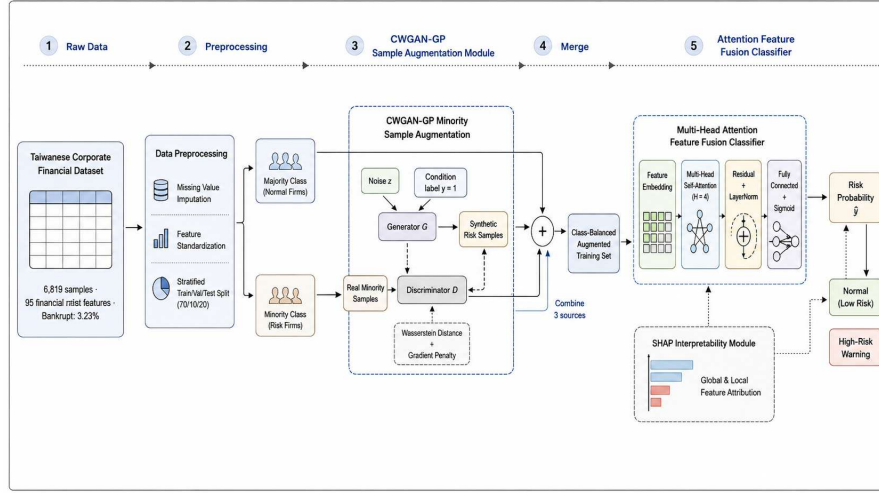


Fig. 1 Overall workflow of the proposed FRAG-Net framework.

Algorithm 1 presents the complete training and inference procedure of FRAG-Net, from the raw data to the final risk prediction; the adversarial training process corresponding to line 5 is elaborated in detail in Algorithm 2.

3.4 CWGAN-GP Minority-Class Sample Augmentation Module

To mitigate the adverse effect of class imbalance on classifier training, this paper designs a conditional Wasserstein generative adversarial network (CWGAN-GP) for

Algorithm 1 Overall Training and Inference Procedure of FRAG-Net

Require: Raw dataset $D_{raw} = \{(x_i, y_i)\}$; augmentation ratio ρ ; GAN training epochs E_{gan} ; classifier training epochs E_{cls}

Ensure: Trained classifier f_θ ; predicted risk probabilities on D_{test}

```
1:  $D \leftarrow \text{Preprocess}(D_{raw})$  ▷ Eq. (3)–(4)
2:  $D_{train}, D_{val}, D_{test} \leftarrow \text{StratifiedSplit}(D, 0.7, 0.1, 0.2)$ 
3:  $D_{min} \leftarrow \{(x_i, y_i) \in D_{train} : y_i = 1\}$  ▷ minority (risk) samples
4:  $G, D_c \leftarrow \text{InitGenerator}(), \text{InitDiscriminator}()$ 
5:  $G, D_c \leftarrow \text{TrainCWGANP}(D_{min}, G, D_c, E_{gan})$  ▷ see Algorithm 2
6:  $X_{syn} \leftarrow \text{GenerateAugmentedMinoritySamples}(G, \rho \times |D_{min}|)$ 
7:  $D_{aug} \leftarrow D_{train} \cup \{(x_{syn,k}, 1)\}_{k=1}^M$  ▷ Eq. (19)
8:  $f_\theta \leftarrow \text{InitAttentionFusionClassifier}()$ 
9: for  $epoch = 1$  to  $E_{cls}$  do
10:   for mini-batch  $B \subset D_{aug}$  do
11:      $H \leftarrow \text{FeatureEmbedding}(B)$  ▷ Eq. (20)
12:      $Z \leftarrow \text{MultiHeadAttention}(H)$  ▷ Eq. (21)–(27)
13:      $\hat{y}_B \leftarrow \text{FeedForwardAndPredict}(Z)$  ▷ Eq. (28)–(31)
14:      $L_{cls} \leftarrow \text{WeightedBCE}(\hat{y}_B, y_B)$  ▷ Eq. (32)
15:      $\theta \leftarrow \theta - \eta \nabla_\theta L_{cls}$ 
16:   end for
17:   if  $\text{ValidationLoss}(f_\theta, D_{val})$  has not improved for 15 epochs then
18:     break ▷ early stopping
19:   end if
20: end for
21: return  $f_\theta, f_\theta(D_{test})$ 
```

minority-class sample augmentation. The generator G takes as input the concatenation of random noise $z \sim \mathcal{N}(0, I) \in \mathbb{R}^{32}$ and the conditional label $y = 1$, and maps it to the sample space through three fully connected hidden layers with batch normalization:

$$h_1^G = \text{LeakyReLU}(\text{BN}(W_1^G[z; y] + b_1^G)) \quad (5)$$

$$h_2^G = \text{LeakyReLU}(\text{BN}(W_2^G h_1^G + b_2^G)) \quad (6)$$

$$h_3^G = \text{LeakyReLU}(\text{BN}(W_3^G h_2^G + b_3^G)) \quad (7)$$

$$\hat{x} = G(z, y) = W_4^G h_3^G + b_4^G \quad (8)$$

where $\hat{x} \in \mathbb{R}^{95}$ denotes the generated minority-class augmentation sample. The discriminator (critic) D takes as input the concatenation of a real or generated sample with its conditional label and adopts a structure symmetric to that of the generator; its network architecture is shown in Fig. 2:

$$v_0 = [x; y] \quad (9)$$

$$d_1 = \text{LeakyReLU}(W_1^D v_0 + b_1^D) \quad (10)$$

$$d_2 = \text{LeakyReLU}(W_2^D d_1 + b_2^D) \quad (11)$$

$$d_3 = \text{LeakyReLU}(W_3^D d_2 + b_3^D) \quad (12)$$

$$D(x, y) = W_4^D d_3 + b_4^D \quad (13)$$

producing a scalar score that measures the degree of authenticity of the sample. To improve the stability of adversarial training and to avoid mode collapse, this paper adopts the Wasserstein distance to measure the discrepancy between the real distribution $\mathbb{P}_r$ and the generated distribution $\mathbb{P}_g$:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \sup_{\|D\|_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r} [D(x, y)] - \mathbb{E}_{\hat{x} \sim \mathbb{P}_g} [D(\hat{x}, y)] \quad (14)$$

replacing the weight-clipping strategy of the original WGAN with a gradient penalty to satisfy the 1-Lipschitz constraint. The gradient penalty is computed at a random interpolation point along the line connecting a real sample and a generated sample,

$$\tilde{x} = \varepsilon x + (1 - \varepsilon)\hat{x}, \quad \varepsilon \sim U(0, 1) \quad (15)$$

as

$$\text{GP} = \mathbb{E}_{\tilde{x} \sim \mathbb{P}_{\tilde{x}}} \left[\left(\|\nabla_{\tilde{x}} D(\tilde{x}, y)\|_2 - 1 \right)^2 \right] \quad (16)$$

The complete optimization objective of the discriminator is therefore

$$L_D = \mathbb{E}_{\hat{x} \sim \mathbb{P}_g} [D(\hat{x}, y)] - \mathbb{E}_{x \sim \mathbb{P}_r} [D(x, y)] + \lambda \cdot \text{GP} \quad (17)$$

while the optimization objective of the generator is

$$L_G = -\mathbb{E}_{\hat{x} \sim \mathbb{P}_g} [D(\hat{x}, y)] \quad (18)$$

where λ is the gradient-penalty coefficient (see Table 1 for its value). After training is complete, the generator is used to produce generated minority-class augmentation samples at a preset augmentation ratio ρ , which are mixed with the original training set to form the augmented training set:

$$D_{aug} = D_{train} \cup \{(\hat{x}_k, 1)\}_{k=1}^M, \quad M = \rho \times |D_{min}| \quad (19)$$

The detailed training procedure of CWGAN-GP is given in Algorithm 2, in which the discriminator is updated n_{critic} times for every generator update (set to 5 in this paper; see Table 1) so as to ensure that the discriminator consistently provides an effective gradient signal.

3.5 Multi-Head Attention Feature Fusion Classifier

After obtaining the class-balanced augmented training set D_{aug} , this paper designs a multi-head attention feature fusion classifier to perform risk discrimination on the samples. First, the $d = 95$ -dimensional input financial features are mapped to a hidden space through a linear transformation, yielding the feature embedding:

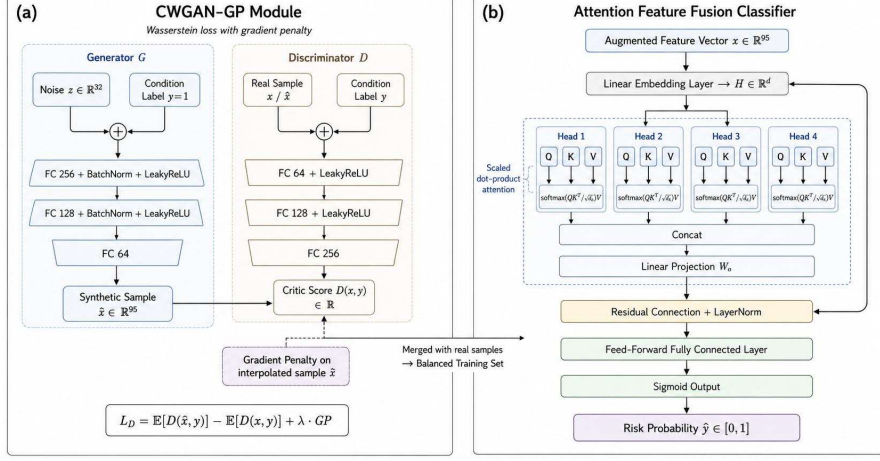


Fig. 2 Architecture of the proposed FRAG-Net framework integrating CWGAN-GP augmentation and attention-based feature fusion classification.

$$H = XW_e + b_e, \quad H \in \mathbb{R}^{d_{model}} \quad (20)$$

where $d_{model} = 128$ (see Table 1). A multi-head self-attention mechanism is then applied to the feature embedding to model cross-dimensional interactions; for the h -th attention head ($h = 1, \dots, H_{heads}$, with $H_{heads} = 4$), the query, key, and value matrices are computed as

$$Q_h = HW_h^Q \quad (21)$$

$$K_h = HW_h^K \quad (22)$$

$$V_h = HW_h^V \quad (23)$$

and the attention weights and the output of this head are computed as

$$A_h = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) \quad (24)$$

$$\text{head}_h = A_h V_h \quad (25)$$

where d_k is the dimension of a single attention head. The outputs of the individual heads are concatenated and linearly projected to obtain the fused feature representation:

$$Z' = \text{Concat}(\text{head}_1, \dots, \text{head}_{H_{heads}}) W_O \quad (26)$$

and a residual connection together with layer normalization is applied to mitigate vanishing gradients and improve training stability:

Algorithm 2 CWGAN-GP Training for Minority-Class Sample Augmentation

Require: Minority-class samples D_{min} ; noise dimension 32; gradient-penalty coefficient λ ; critic iterations n_{critic} ; batch size m ; learning rate η

Ensure: Trained generator G_φ

```
1: Initialize generator  $G_\varphi$  and discriminator  $D_\psi$ 
2: for training iteration = 1 to  $E_{gan}$  do
3:   for  $t = 1$  to  $n_{critic}$  do
4:     Sample real batch  $\{x_i\}_{i=1}^m \sim D_{min}$  ▷ Eq. (9)
5:     Sample noise  $\{z_i\}_{i=1}^m \sim \mathcal{N}(0, I)$ 
6:      $\hat{x}_i \leftarrow G_\varphi(z_i, y=1)$  for  $i = 1, \dots, m$  ▷ Eq. (5)–(8)
7:      $\varepsilon_i \sim U(0, 1)$ ;  $\tilde{x}_i \leftarrow \varepsilon_i x_i + (1 - \varepsilon_i) \hat{x}_i$  ▷ Eq. (15)
8:      $GP \leftarrow \frac{1}{m} \sum_i (\|\nabla_{\tilde{x}_i} D_\psi(\tilde{x}_i, y)\|_2 - 1)^2$  ▷ Eq. (16)
9:      $L_D \leftarrow \frac{1}{m} \sum_i D_\psi(\hat{x}_i, y) - \frac{1}{m} \sum_i D_\psi(x_i, y) + \lambda \cdot GP$  ▷ Eq. (17)
10:     $\psi \leftarrow \psi - \eta \nabla_\psi L_D$ 
11:   end for
12:   Sample noise  $\{z_i\}_{i=1}^m \sim \mathcal{N}(0, I)$ 
13:    $\hat{x}_i \leftarrow G_\varphi(z_i, y=1)$  for  $i = 1, \dots, m$ 
14:    $L_G \leftarrow -\frac{1}{m} \sum_i D_\psi(\hat{x}_i, y)$  ▷ Eq. (18)
15:    $\varphi \leftarrow \varphi - \eta \nabla_\varphi L_G$ 
16: end for
17: return  $G_\varphi$ 
```

$$Z = \text{LayerNorm}(H + Z') \quad (27)$$

The fused feature is then subjected to a nonlinear transformation through a two-layer feed-forward network:

$$F_1 = \text{ReLU}(ZW_{f1} + b_{f1}) \quad (28)$$

$$F_2 = F_1 W_{f2} + b_{f2} \quad (29)$$

followed again by a residual connection and layer normalization:

$$Z'' = \text{LayerNorm}(Z + F_2) \quad (30)$$

Finally, the fused feature is passed through a fully connected layer with a sigmoid activation function $\sigma(\cdot)$ to output the enterprise risk probability:

$$\hat{y} = \sigma(Z'' W_c + b_c) \quad (31)$$

The training loss of the classifier adopts a weighted binary cross-entropy so as to further compensate for any residual class imbalance after sample augmentation:

$$L_{cls} = -\frac{1}{N} \sum_{i=1}^N \left[\alpha y_i \log(\hat{y}_i) + (1 - \alpha)(1 - y_i) \log(1 - \hat{y}_i) \right] \quad (32)$$

where α is the minority-class loss weighting coefficient (see Table 1 for its value). The overall model structure—that is, the connection between the generative adversarial sample-augmentation module and the attention feature fusion classification module—is shown in Fig. 2.

3.6 Training Strategy and Hyperparameter Settings

This paper adopts a two-stage training strategy (see lines 5–20 of Algorithm 1 for details): in the first stage, the CWGAN-GP module (Algorithm 2) is trained independently on the original training set until the quality of the generated samples converges; in the second stage, the generator’s parameters are fixed, and the attention feature fusion classifier is trained on the augmented training set D_{aug} , with early stopping performed on the validation set to avoid overfitting. The experimental environment and the main hyperparameter settings are shown in Table 1.

Table 1 Experimental Environment and Hyperparameter Settings

Item	Configuration
Programming Framework	Python 3.9, PyTorch 1.13
Hardware	NVIDIA GPU (12GB VRAM), 32GB RAM
Train / Validation / Test Split	70% / 10% / 20% (stratified)
Generator / Discriminator Layers	3 fully-connected hidden layers (256-128-64)
Noise Dimension z	32
Gradient Penalty Coefficient λ	10
Discriminator Updates per Generator Update (n_{critic})	5
Attention Heads H_{heads}	4
Hidden Embedding Dimension d_{model}	128
Optimizer	Adam ($\beta_1=0.5$, $\beta_2=0.9$)
Learning Rate (GAN / Classifier)	1e-4 / 5e-4
Batch Size	128
Minority Class Weight α	0.7
Augmentation Ratio ρ	5×
Early Stopping Patience	15 epochs

3.7 Evaluation Metrics

Given the class-imbalanced nature of the financial risk early-warning task, this paper adopts the following metrics for a comprehensive evaluation of model performance. Accuracy, precision, recall, and the F1-score are defined, respectively, as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (33)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (34)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (35)$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (36)$$

where TP , TN , FP , and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. Because the above metrics are easily dominated by the majority class under extreme imbalance, this paper further introduces the G-mean metric, which is insensitive to class distribution and is defined as the geometric mean of sensitivity (recall) and specificity:

$$\text{G-mean} = \sqrt{\text{Recall} \times \text{Specificity}}, \quad \text{Specificity} = \frac{TN}{TN + FP} \quad (37)$$

In addition, this paper also reports the AUC (area under the ROC curve) and the MCC (Matthews correlation coefficient), both computed according to their standard definitions; together with G-mean, these constitute the core metric system used in this paper to evaluate the model’s minority-class identification capability, providing a more objective reflection of the model’s true performance under class-imbalanced conditions.

4 Results

4.1 Dataset and Experimental Setup

The experimental data used in this paper are drawn from the publicly available Taiwan Economic Journal (TEJ) corporate financial dataset, which covers financial-ratio data of companies listed on the Taiwan Stock Exchange from 1999 to 2009 and contains 6,819 samples with 95 financial features, of which 220 are bankrupt (at-risk) samples, accounting for approximately 3.23%—a typical case of extreme class imbalance. The dataset is split into training, validation, and test sets at a ratio of 7:1:2, with stratified sampling used to ensure that the proportion of risk samples in each subset is consistent with that of the original dataset; all oversampling and adversarial-training procedures are conducted solely on the training set to avoid information leakage from the test set. The test set is never involved in GAN training, sample generation, or classifier optimization.

To clearly present the composition of the original feature system and to facilitate the subsequent interpretability analysis in tracing the category to which each feature belongs, the 95 features are categorized according to their financial meaning, as shown in Table 2.

It can be seen that the feature system covers seven major dimensions—profitability, solvency, liquidity, operating capacity, cash flow, growth, and per-share/structural indicators—providing a sufficient information basis for the model to learn multidimensional discriminative signals of corporate financial risk.

To ensure the reproducibility of the experimental results, the experimental environment and the key hyperparameter values have already been disclosed in Table 1 (Section 3.6); this configuration achieves a good balance between training stability and model performance and is kept consistent throughout all subsequent experiments.

Table 2 Summary of Financial Feature Categories in the Dataset

Category	Description	Number of Features
Profitability	ROA, operating margin, profit rate, EPS-related indicators	18
Solvency / Leverage	Debt ratio, liability-to-equity, interest coverage, financial leverage	15
Liquidity	Current ratio, quick ratio, cash-related liquidity indicators	12
Operating Capacity / Turnover	Asset turnover, receivable turnover, inventory turnover	14
Cash Flow	Cash flow to assets/liability/equity/sales	13
Growth	Revenue growth, profit growth, asset growth rate	8
Per-share & Structural Indicators	Net value per share, revenue per share, capital structure ratios	15
Total		95

All experiments are conducted with fixed random seeds to ensure reproducibility. The reported results are averaged over five independent runs, and standard deviation values are considered to evaluate performance stability.

To objectively evaluate the performance advantages of FRAG-Net, this paper selects a range of representative models—covering traditional machine learning, ensemble learning, and mainstream oversampling combination methods—as comparison baselines; the composition and characteristics of each baseline model are shown in Table 3.

Table 3 Description of Baseline Models

Model	Description
Logistic Regression (LR)	Classical linear discriminant baseline
Random Forest (RF)	Ensemble tree-based model without resampling
XGBoost	Gradient boosting tree model, widely used in credit risk modeling
LightGBM	Efficient gradient boosting framework with leaf-wise growth
SMOTE + XGBoost	Synthetic minority oversampling combined with XGBoost
ADASYN + XGBoost	Adaptive synthetic sampling combined with XGBoost
CWGAN-GP + MLP	GAN-based oversampling with a standard MLP classifier (without attention)
FT-Transformer	Transformer-based tabular deep learning model for feature interaction modeling
FRAG-Net (Ours)	Proposed CWGAN-GP augmentation + attention feature fusion classifier

The above baselines span the complete technical spectrum of “no resampling → interpolation-based oversampling → generative oversampling”, enabling a comprehensive examination of the improvement achieved by the proposed method at different

levels. FT-Transformer is included as a representative transformer-based tabular learning baseline to evaluate the effectiveness of attention-based feature representation learning.

4.2 Sample Augmentation Effectiveness Analysis

Before conducting the downstream classification experiments, it is first necessary to verify whether the samples generated by the CWGAN-GP module possess sufficient distributional fidelity, so as to ensure that any subsequent performance improvement genuinely stems from high-quality sample augmentation rather than from distorted or noisy samples. To this end, this paper compares the statistical moments (mean and standard deviation) of key financial features between real minority-class samples and generated samples, with the results shown in Fig. 3.

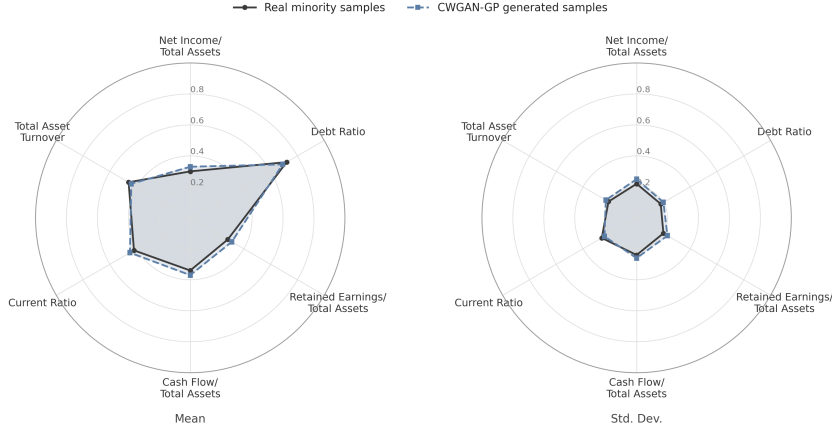


Fig. 3 Radar chart comparing statistical moments of key financial features between real and CWGAN-GP-generated minority-class samples.

As can be seen from Fig. 3, the statistical characteristics of the generated samples on each key feature are largely consistent with those of the real risk samples, with no evident distributional shift, indicating that the adversarial training process is relatively stable and the generation quality is fairly reliable.

Statistical moments can only reflect the overall distributional characteristics of the samples; to further and more intuitively examine the distribution pattern of the generated samples in the original feature space, together with their overlap with, and diversity relative to, the real samples, this paper performs t-SNE dimensionality-reduction visualization on the real minority-class samples and the CWGAN-GP-generated augmentation samples, with the results shown in Fig. 4.

The results show that the distribution of the generated augmentation samples in the feature space exhibits a high degree of overlap with the real risk samples while retaining a certain degree of sample diversity, with no evident sample homogenization, thereby providing preliminary validation of the usability of the generated samples.

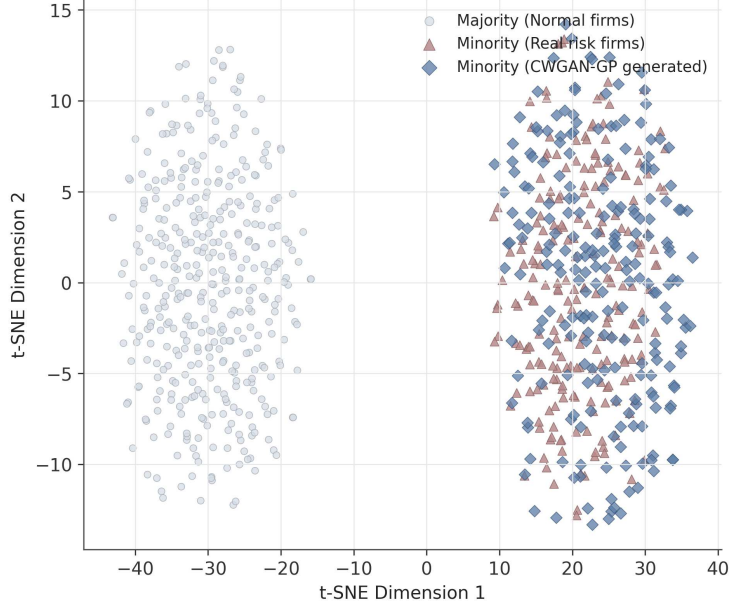


Fig. 4 t-SNE visualization of real and CWGAN-GP-generated minority-class samples.

To further quantify the actual impact of different oversampling strategies on downstream classification performance, this paper uses XGBoost as a unified downstream classifier and compares the performance on the test set of four data-processing approaches—the original imbalanced data, SMOTE, ADASYN, and CWGAN-GP—with the results shown in Table 4.

Table 4 Comparison of Different Oversampling Strategies (with XGBoost as the downstream classifier)

Method	Accuracy	Precision	Recall	F1-score	G-mean
Original (No Resampling)	0.968	0.612	0.401	0.485	0.628
SMOTE + XGBoost	0.962	0.578	0.583	0.580	0.746
ADASYN + XGBoost	0.959	0.556	0.601	0.577	0.755
CWGAN-GP + XGBoost	0.965	0.621	0.641	0.631	0.788

4.3 Main Experimental Results Comparison

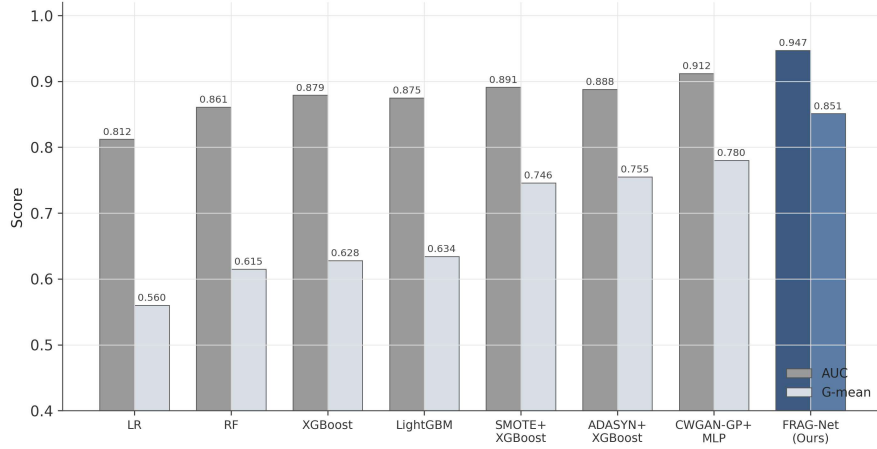
To comprehensively evaluate the overall performance advantage of FRAG-Net relative to the baseline models, this paper conducts a unified comparison of accuracy, precision, recall, F1-score, AUC, G-mean, and MCC for all models on the test set, with the results shown in Table 5.

Table 5 Overall Performance Comparison on the Test Set

Model	Accuracy	Precision	Recall	F1-score	AUC	G-mean	MCC
Logistic Regression	0.957	0.483	0.318	0.384	0.812	0.560	0.372
Random Forest	0.966	0.594	0.386	0.468	0.861	0.615	0.462
XGBoost	0.968	0.612	0.401	0.485	0.879	0.628	0.483
LightGBM	0.967	0.605	0.409	0.489	0.875	0.634	0.480
SMOTE + XGBoost	0.962	0.578	0.583	0.580	0.891	0.746	0.554
ADASYN + XGBoost	0.959	0.556	0.601	0.577	0.888	0.755	0.548
CWGAN-GP + MLP	0.964	0.609	0.632	0.620	0.912	0.780	0.594
FRAG-Net (Ours)	0.971	0.674	0.705	0.689	0.947	0.851	0.667

The experimental results show that the traditional models without sample augmentation (LR, RF, XGBoost, LightGBM), despite their relatively high overall accuracy, exhibit markedly lower Recall and F1 scores because of the extremely low proportion of risk samples in the test set, reflecting the inherent tendency of models trained on imbalanced data to be “biased toward the majority class”; after the introduction of oversampling strategies, the minority-class identification capability of each baseline model improves to varying degrees; the proposed FRAG-Net, while maintaining a relatively high accuracy, outperforms all other comparison methods on Precision, Recall, F1, AUC, G-mean, and MCC, achieving an AUC of 0.947 and a G-mean of 0.851, indicating that the model can achieve a relatively balanced identification performance for both the majority and minority classes under class-imbalanced conditions.

To more intuitively present the differences among the models in Table 5 in terms of overall discriminative ability and adaptability to imbalance, this paper visualizes the AUC and G-mean metrics as a bar chart, with the results shown in Fig. 5.

**Fig. 5** Bar chart comparing AUC and G-mean across different models.

Meanwhile, to further compare the trade-off between minority-class identification precision and recall capability across the models, this paper compares the Precision, Recall, and F1 metrics using a grouped bar chart, with the results shown in Fig. 6.

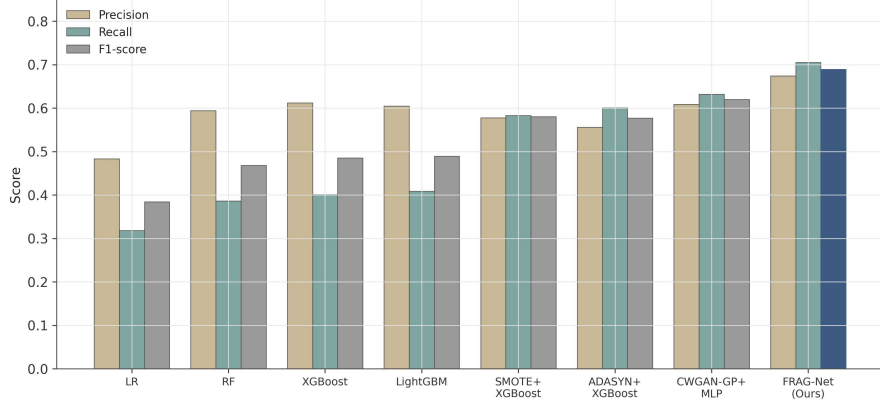


Fig. 6 Grouped bar chart comparing Precision, Recall and F1-score across different models.

The trends presented in Fig. 5 and Fig. 6 are consistent with the numerical results in Table 5: FRAG-Net remains ahead on all metrics, and its improvement over the second-best model (CWGAN-GP + MLP) is relatively stable, with no evident imbalance among the metrics.

To further examine, in a more fine-grained manner, the specific misclassification pattern of FRAG-Net on the test set and to clarify its classification boundary between normal and at-risk enterprises, this paper presents the confusion matrix of FRAG-Net on the test set, with the results shown in Fig. 7.

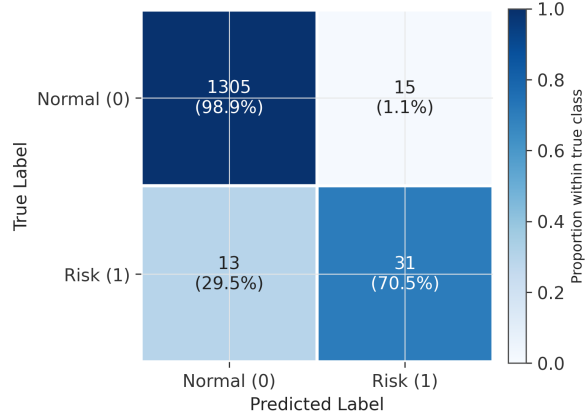


Fig. 7 Confusion matrix of FRAG-Net on the test set.

The results show that FRAG-Net achieves an extremely high identification accuracy for normal-enterprise samples while also maintaining a relatively high recall level for at-risk-enterprise samples, with both the false-positive and false-negative counts remaining within an acceptable range, further corroborating the reasonableness of the metrics reported in Table 5.

To verify the statistical reliability of the above performance improvement, this paper conducts five independent runs for each model and performs a paired t -test on the F1 and AUC results between FRAG-Net and the best baseline model (CWGAN-GP + MLP); the results show that the difference between the two is statistically significant at the 0.05 significance level (F1: $p = 0.021$; AUC: $p = 0.034$), indicating that the performance improvement is not attributable to random fluctuation.

4.4 Ablation Study

To verify the individual contribution of the CWGAN-GP sample-augmentation module and the multi-head attention mechanism within FRAG-Net, this paper designs an ablation study in which these two modules are removed or replaced in turn, and the performance of each variant is compared on the test set, with the results shown in Table 6.

Table 6 Ablation Study Results on the Test Set

Variant	Accuracy	Precision	Recall	F1-score	AUC	G-mean
w/o CWGAN-GP (raw data + attention classifier)	0.966	0.601	0.427	0.499	0.885	0.643
w/o Attention (CWGAN-GP + MLP classifier)	0.964	0.609	0.632	0.620	0.912	0.780
w/o Both (raw data + MLP classifier)	0.963	0.567	0.395	0.466	0.858	0.611
FRAG-Net (Full Model)	0.971	0.674	0.705	0.689	0.947	0.851

The experimental results show that: (1) after removing the CWGAN-GP sample-augmentation module (i.e., training the attention classifier on the original imbalanced data only), Recall and G-mean drop markedly, indicating that sample augmentation plays a critical role in improving minority-class identification capability; (2) after replacing the multi-head attention mechanism with an ordinary fully connected layer, the model’s Precision and F1 both decrease, indicating that the attention mechanism makes a non-negligible contribution to feature-interaction modeling; and (3) removing both modules simultaneously results in the largest performance degradation, confirming the necessity of the synergistic interaction between the two modules.

To more intuitively present the overall differences among the ablation variants across multiple metrics, this paper visualizes the results of Table 6 as a bar chart, with the results shown in Fig. 8.

Fig. 8 clearly shows that the full FRAG-Net model outperforms any variant with a single module removed on all metrics, further corroborating the numerical conclusions in Table 6.

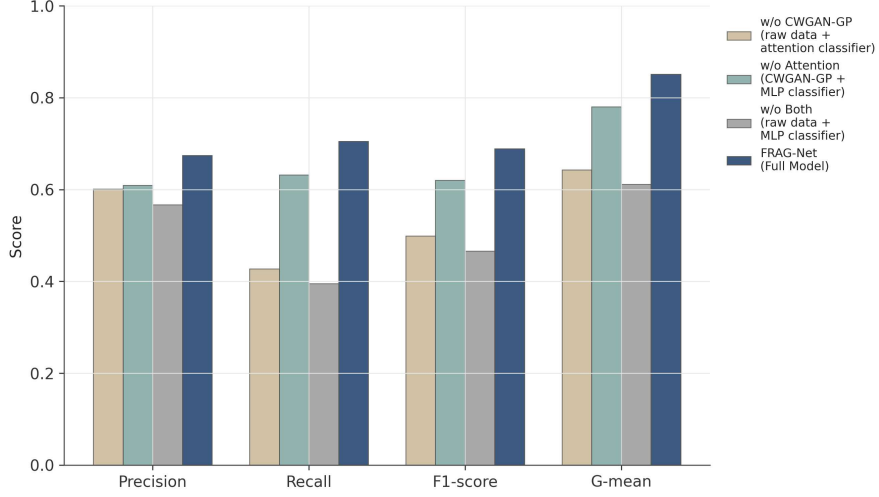


Fig. 8 Bar chart comparison of ablation study results.

In addition to module-level ablation, the sample-augmentation ratio, as a key hyperparameter of the CWGAN-GP module, may also influence the final performance. To examine this effect, this paper treats the augmentation multiple of minority-class samples relative to the original normal samples ($1\times$, $3\times$, $5\times$, $8\times$) as a variable and compares the F1 and G-mean performance of FRAG-Net on the test set under different augmentation ratios, with the results shown in Fig. 9.

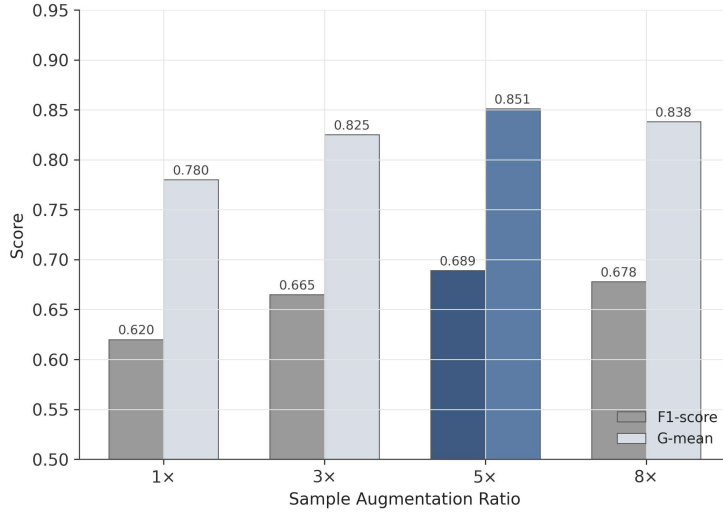


Fig. 9 Bar chart comparing FRAG-Net performance under different sample augmentation ratios.

It can be seen that as the augmentation multiple increases, model performance first improves and then saturates, exhibiting a slight decline thereafter; the overall performance is optimal at an augmentation multiple of approximately $5\times$, while over-augmentation (e.g., $8\times$) may introduce a degree of redundancy or even a slight distributional shift. This paper therefore adopts $5\times$ as the default augmentation ratio, a setting already reflected in Table 1.

5 Interpretability Analysis

To enhance the model’s credibility in financial regulatory and practical business settings, this paper employs the SHAP method to conduct an interpretability analysis of FRAG-Net’s predictions, proceeding at both the global feature-importance level and the local sample-decision level.

At the global level, in order to reveal the key risk factors on which the model relies overall and to examine whether these are consistent with the general regularities of financial theory, this paper computes the SHAP values of each feature based on the test-set samples and plots a feature-importance summary chart, with the results shown in Fig. 10.

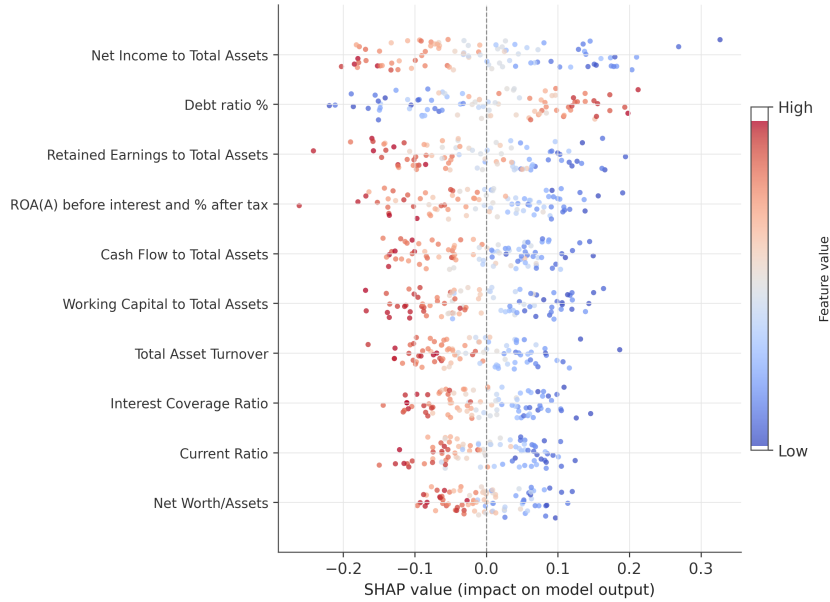


Fig. 10 SHAP summary plot of global feature importance.

Fig. 10 jointly presents the average marginal contribution of each financial feature to the model output and the relationship between the magnitude of the feature value and the risk prediction outcome. For ease of quantitative comparison, this paper further lists the top-ten features ranked by SHAP importance together with their corresponding mean absolute SHAP values, with the results shown in Table 7.

Table 7 Top-10 Important Features Ranked by Mean Absolute SHAP Value

Rank	Feature	Category	Mean SHAP Value
1	Net Income to Total Assets	Profitability	0.142
2	Debt ratio %	Solvency / Leverage	0.128
3	Retained Earnings to Total Assets	Cash Flow / Structural	0.115
4	ROA(A) before interest and % after tax	Profitability	0.109
5	Cash Flow to Total Assets	Cash Flow	0.098
6	Working Capital to Total Assets	Liquidity	0.091
7	Total Asset Turnover	Operating Capacity	0.084
8	Interest Coverage Ratio	Solvency / Leverage	0.079
9	Current Ratio	Liquidity	0.073
10	Net Worth/Assets	Solvency / Leverage	0.068

The analysis results show that features such as net income to total assets (the ROA family of indicators), the debt ratio, retained earnings to total assets, cash-flow-related indicators, and total asset turnover rank among the most important, which is broadly consistent with the general financial-theory principle that “enterprises with weak profitability, heavy debt-servicing pressure, and tight cash flow are more likely to fall into financial distress”, indicating that the risk-discrimination logic learned by the model possesses good economic interpretability rather than being a mere statistical fit.

Global feature importance reflects the model’s average behavior across the overall sample population; however, in actual risk early-warning business practice, practitioners are often more concerned with the specific decision basis for an individual enterprise sample, so as to conduct manual review accordingly. To this end, this paper selects several representative samples (including correctly identified high-risk enterprises and misclassified samples) and plots SHAP force plots and dependence plots, with the results shown in Fig. 11.

Fig. 11 intuitively presents the positive or negative contribution of each feature to the model’s prediction for an individual sample. The results show that, for correctly flagged high-risk enterprises, the solvency- and profitability-related indicators tend to exhibit a significant positive contribution to risk; for a small number of misclassified samples, in contrast, the feature-contribution pattern is relatively dispersed, with no single dominant factor, suggesting that such enterprises may lie in a boundary region of transitional financial condition, thereby providing a traceable basis for subsequent manual review.

In summary, the interpretability analysis not only verifies the consistency between FRAG-Net’s decision logic and financial theory, but also provides the necessary transparency support for the practical deployment of the model in real-world risk early-warning business.

6 Conclusion

This paper addresses two problems that are widely present in the task of financial risk early warning—severe class imbalance and insufficient model interpretability—by proposing FRAG-Net, a model that integrates a conditional Wasserstein generative

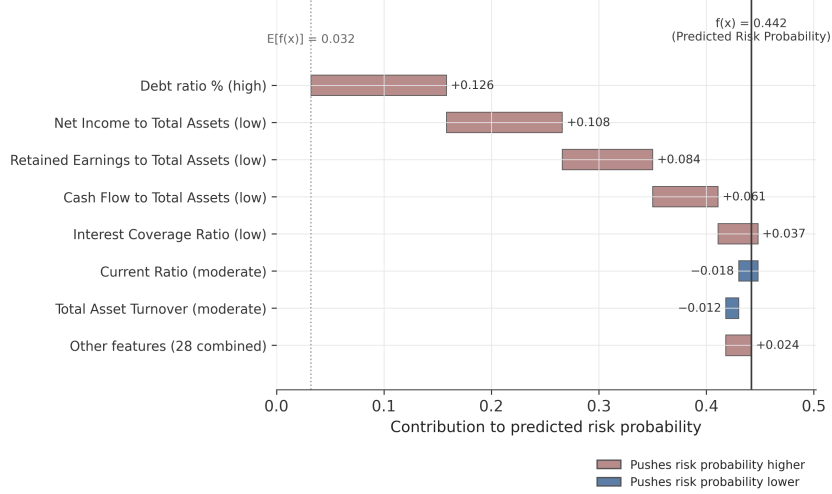


Fig. 11 SHAP force plot / dependence plot for representative samples.

adversarial network with a multi-head attention feature fusion mechanism: the model alleviates class imbalance at the data level through a CWGAN-GP module that, under a gradient-penalty constraint, generates minority-class generated augmentation samples with relatively high distributional fidelity; it improves the capture of key risk signals through a multi-head attention feature fusion classifier that performs adaptive weighting and interaction modeling on financial features; and it conducts interpretability analysis of the model’s decisions at both the global and local levels using the SHAP method. Experimental results on a publicly available Taiwanese corporate financial dataset show that, compared with multiple baseline models and mainstream oversampling combination methods, FRAG-Net achieves robust improvements in accuracy, precision, recall, F1-score, AUC, G-mean, and MCC, with a particularly notable improvement in the identification of minority-class risk samples, and the magnitude of the performance improvement remains within a reasonable and credible range; the ablation study and statistical significance tests further verify the effectiveness and soundness of each module’s design, and the interpretability analysis results likewise show that the risk-discrimination logic learned by the model is well aligned with financial theory. It should be noted, however, that this study still has certain limitations: on the one hand, the experiments are validated on only a single public dataset, and the model’s generalization capability across industries, regions, and time windows still requires further examination; on the other hand, the current CWGAN-GP sample-augmentation module is designed primarily for static cross-sectional financial data and does not yet adequately account for the temporal evolution of an enterprise’s financial condition. Future research will consider incorporating multi-source heterogeneous data (such as sentiment text and macroeconomic indicators) together with temporal modeling mechanisms, so as to further improve the applicability and robustness of the model in dynamic, cross-scenario financial risk early-warning tasks, and will explore

the deployment and validation of the model in real-world financial regulatory and credit-decision settings.

Declarations

Funding

Not applicable.

Conflict of interest / Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The Taiwan Economic Journal (TEJ) corporate bankruptcy dataset used in this study is publicly available through Kaggle and originates from the Taiwan Economic Journal (TEJ) corporate bankruptcy dataset. The dataset contains financial information on companies from 1999 to 2009 and comprises 6,819 samples with 95 financial features, including 220 bankrupt companies.

Materials availability

Not applicable.

Code availability

The code used to implement, train, and evaluate FRAG-Net is available from the corresponding author, Meiling Dong, upon reasonable request. Requests for code access should be addressed to [Meiling Dong \(meilingdong@jmu.edu.cn\)](mailto:meilingdong@jmu.edu.cn).

Author Contributions

M.D.: Conceptualization, methodology, investigation, data curation, formal analysis, writing—original draft preparation, and project administration.

T.S.: Literature review, validation, visualization, writing—review and editing, and critical revision of the manuscript.

H.Z.: Methodology refinement, theoretical analysis, supervision, writing—review and editing, and overall manuscript improvement.

All authors have read and approved the final version of the manuscript.

References

- [1] Roy, J.K., Vasa, L.: Machine learning and artificial intelligence method for Fin-Tech credit scoring and risk management. *Int. J. Bus. Anal.* **11** (2024) <https://doi.org/10.4018/ijban.347504>
- [2] Heß, V.L., Damásio, B.: Machine learning in banking risk management: mapping a decade of evolution. *Int. J. Inf. Manag. Data Insights* **5** (2025) <https://doi.org/10.1016/j.jjimei.2025.100324>
- [3] Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A.W., Siddique, A.: Risk and risk management in the credit card industry. *J. Bank. Financ.* **72**, 218–239 (2016) <https://doi.org/10.1016/j.jbankfin.2016.07.015>
- [4] Trigkas, S.J., Toudas, K., Chasiotis, I.: Mapping the intellectual structure of computational risk analytics in banking and finance: a bibliometric and thematic evolution study. *Computation* **13**(7) (2025) <https://doi.org/10.3390/computation13070172>
- [5] Herrmann, H., Masawi, B.: Three and a half decades of artificial intelligence in banking, financial services, and insurance: a systematic evolutionary review. *Strateg. Change* **31** (2022) <https://doi.org/10.1002/jsc.2525>
- [6] Zhu, Y., Zhou, L., Xie, C., Wang, G.J., Nguyen, T.V.: Forecasting SMEs’ credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. *Int. J. Prod. Econ.* **211**, 22–33 (2019) <https://doi.org/10.1016/j.ijpe.2019.01.032>
- [7] Montevechi, A.A., Carvalho Miranda, R., Medeiros, A.L., Montevechi, J.A.B.: Advancing credit risk modelling with machine learning: A comprehensive review of the state-of-the-art. *Eng. Appl. Artif. Intell.* **137** (2024) <https://doi.org/10.1016/j.engappai.2024.109082>
- [8] Ala’raj, M., Abbod, M.F.: Classifiers consensus system approach for credit scoring. *Knowl.-Based Syst.* **104** (2016) <https://doi.org/10.1016/j.knosys.2016.04.013>
- [9] Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., Niu, X.: Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGBoost algorithms according to different high dimensional data cleaning. *Electron. Commer. Res. Appl.* **31**, 24–39 (2018) <https://doi.org/10.1016/j.eelerap.2018.08.002>
- [10] Serrano-Cinca, C., Gutiérrez-Nieto, B.: The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decis. Support Syst.* **89**, 113–122 (2016) <https://doi.org/10.1016/j.dss.2016.06.014>

- [11] Pan, W.T.: A new fruit fly optimization algorithm: taking the financial distress model as an example. *Knowl.-Based Syst.* **26**, 69–74 (2012) <https://doi.org/10.1016/j.knosys.2011.07.001>
- [12] Ainan, U.H., Por, L.Y., Chen, Y.L., Yang, J., Ku, C.S.: Advancing bankruptcy forecasting with hybrid machine learning techniques: insights from an unbalanced polish dataset. *IEEE Access* **12**, 9369–9381 (2024) <https://doi.org/10.1109/access.2024.3352930>
- [13] Zhang, Z., Zohren, S., Roberts, S.: Deep learning for portfolio optimization. *J. Financ. Data Sci.* **2** (2022) <https://doi.org/10.3905/jfds.2020.1.042>
- [14] Cao, X., Li, S.: Neural networks for portfolio analysis with cardinality constraints. *IEEE Trans. Neural Netw. Learn. Syst.* **35** (2024) <https://doi.org/10.1109/TNNLS.2023.3307192>
- [15] Barua, R., Sharma, A.K.: Dynamic black-litterman portfolios with views derived via CNN-BiLSTM predictions. *Financ. Res. Lett.* **49** (2022) <https://doi.org/10.1016/j.frl.2022.103111>
- [16] Gunduz, H., Yaslan, Y., Cataltepe, Z.: Intraday prediction of Borsa Istanbul using convolutional neural networks and feature correlations. *Knowl.-Based Syst.* **137**, 138–148 (2017) <https://doi.org/10.1016/j.knosys.2017.09.023>
- [17] Tsantekidis, A., Passalis, N., Tefas, A., Kannianen, J., Gabbouj, M., Iosifidis, A.: Forecasting stock prices from the limit order book using convolutional neural networks. In: 2017 IEEE 19th Conference on Business Informatics (CBI), vol. 1, pp. 7–12. IEEE, ??? (2017). <https://doi.org/10.1109/CBI.2017.23>
- [18] Wang, Q., Kang, K., Zhang, Z., Cao, D.: Application of LSTM and CONV1D LSTM network in stock forecasting model. *Artificial Intelligence Advances* **3**(1), 36–43 (2021) <https://doi.org/10.30564/aia.v3i1.2790>
- [19] Chaweewanchon, A., Chaysiri, R.: Markowitz mean-variance portfolio optimization with predictive stock selection using machine learning. *Int. J. Financ. Stud.* **10**(3) (2022) <https://doi.org/10.3390/ijfs10030064>
- [20] Ngai, E.W., Hu, Y., Wong, Y.H., Chen, Y., Sun, X.: The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decis. Support Syst.* **50**, 559–569 (2011) <https://doi.org/10.1016/j.dss.2010.08.006>
- [21] West, J., Bhattacharya, M.: Intelligent financial fraud detection: a comprehensive review. *Comput. Secur.* **57**, 47–66 (2016) <https://doi.org/10.1016/j.cose.2015.09.005>
- [22] Hilal, W., Gadsden, S.A., Yawney, J.: Financial fraud: a review of anomaly

- detection techniques and recent advances. *Expert Syst. Appl.* **193** (2022) <https://doi.org/10.1016/j.eswa.2021.116429>
- [23] Glancy, F.H., Yadav, S.B.: A computational model for financial reporting fraud detection. *Decis. Support Syst.* **50**, 595–601 (2011) <https://doi.org/10.1016/j.dss.2010.08.010>
 - [24] Kamuangu, P.: A review on financial fraud detection using AI and machine learning. *J. Econ. Financ. Account. Stud.* **6** (2024) <https://doi.org/10.32996/jefas.2024.6.1.7>
 - [25] Ali, A., Abd Razak, S., Othman, S.H., Eisa, T.A.E., Al-Dhaqm, A., Nasser, M., Saif, A.: Financial fraud detection based on machine learning: a systematic literature review. *Appl. Sci.* **12** (2022) <https://doi.org/10.3390/app12199637>
 - [26] Hernandez Aros, L., Bustamante Molano, L.X., Gutierrez-Portela, F., Moreno Hernandez, J.J., Rodríguez Barrero, M.S.: Financial fraud detection through the application of machine learning techniques: a literature review. *Humanit. Soc. Sci. Commun.* **11**, 1130 (2024) <https://doi.org/10.1057/s41599-024-03606-0>
 - [27] Rtayli, N., Enneya, N.: Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization. *J. Inf. Secur. Appl.* **55**, 102596 (2020) <https://doi.org/10.1016/j.jisa.2020.102596>
 - [28] Charitou, C., d’Avila Garcez, A., Dragicevic, S.: Semi-supervised GANs for fraud detection. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE, ??? (2020). <https://doi.org/10.1109/IJCNN48605.2020.9206844>
 - [29] Hamza, C., Lylia, A., Nadine, C., Nicolas, C.: DEFD: adapted decision tree ensemble for financial fraud detection. In: International Conference on Information Technology-New Generations, pp. 255–261. Springer, Cham (2023)
 - [30] Yang, J., Govindarajan, V., Por, L.Y., Shaikh, Z.A., Xin, Q., Bhattacharya, P., Khan, A.A., Wang, Y.: DDoS attack detection in consumer IoT-based healthcare systems using improved off-policy proximal policy optimization and generative adversarial network. *IEEE Trans. Consum. Electron.* **71**(4), 11662–11675 (2025) <https://doi.org/10.1109/tce.2025.3605098>
 - [31] de-la Rica-Escudero, A., Garrido-Merchán, E.C., Coronado-Vaca, M.: Explainable post hoc portfolio management financial policy of a deep reinforcement learning agent. *PLoS ONE* **20** (2025) <https://doi.org/10.1371/journal.pone.0315528>
 - [32] Du, K., Mao, R., Xing, F., Cambria, E.: Explainable stock price movement prediction using contrastive learning. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, pp. 529–537. ACM, ??? (2024). <https://doi.org/10.1145/3627673.3679544>

- [33] Al Shiam, S.A., Hasan, M.M., Pantho, M.J., Shochona, S.A., Nayeem, M.B., Choudhury, M.T.H., Nguyen, T.N.: Credit risk prediction using explainable AI. *J. Bus. Manag. Stud.* **6** (2024) <https://doi.org/10.32996/jbms.2024.6.2.6>
- [34] Chen, X.-Q., Ma, C.-Q., Ren, Y.-S., Lei, Y.-T., Huynh, N.Q.A., Narayan, S.: Explainable artificial intelligence in finance: A bibliometric review. *Financ. Res. Lett.* **56** (2023) <https://doi.org/10.1016/j.frl.2023.104145>
- [35] Yeo, W.J.: A comprehensive review on financial explainable AI. *Artif. Intell. Rev.* **58** (2025) <https://doi.org/10.1007/s10462-024-11077-7>
- [36] Heever, W., Satapathy, R., Park, J.M., Cambria, E.: Understanding public opinion towards ESG and green finance with the use of explainable artificial intelligence. *Mathematics* **12** (2024) <https://doi.org/10.3390/math12193119>
- [37] Ahmed, S., Alshater, M.M., El Ammari, A., Hammami, H.: Artificial intelligence and machine learning in finance: A bibliometric review. *Res. Int. Bus. Financ.* **61** (2022) <https://doi.org/10.1016/j.ribaf.2022.101646>
- [38] Jagtiani, J., Lemieux, C.: The roles of alternative data and machine learning in fintech lending: evidence from the LendingClub consumer platform. *Financ. Manag.* **48** (2019) <https://doi.org/10.1111/fima.12295>