Computational Economics
https://doi.org/10.1007/s10614-025-11114-9

Graph-Based Inductive Learning for Credit Risk Prediction with Imbalance Mitigation

Sogand Pourkhoshgoftar[1] · **Asadollah Shahbahrami**[1,3] · **Nima Esmi**[2,3]

Received: 30 April 2025 / Accepted: 1 September 2025

Abstract
Credit default prediction is crucial for financial institutions, as it enables more informed lending decisions. However, current methods face notable Limitations, especially in Handling extreme class imbalance and the scarcity of default instances. Additionally, research into uncovering hidden patterns in non-linear borrower data and improving model interpretability remains relatively underexplored. To mitigate these issues, we propose a hybrid approach that combines a conditional tabular generative adversarial network with a graph neural network, specifically the graph sample and aggregate, for inductive learning. This allows for generalization to new borrowers in large datasets and applies to real-world, dynamic financial environments. The generated default instances are added to the original dataset, creating a more balanced distribution between default and non-default classes. The data are then structured as a graph to capture the relationships among borrowers. This approach directly reduces class imbalance while maintaining non-linear relational context. Experimental results demonstrate that our approach achieves a 7.78% improvement in accuracy and a 13.35% increase in the area under the receiver operating characteristic curve over the non-augmented version, while outperforming baselines that use class balancing techniques. Moreover, Shapley additive explanations provide interpretability by highlighting key features influencing credit risk. These findings underscore the effectiveness of combining synthetic data augmentation with graph learning for robust credit risk prediction in imbalanced settings.

Keywords Credit default prediction · Graph neural network · Generative adversarial network · Decision support system · Graph representation learning · Shapley additive explanations

Extended author information available on the last page of the article

Published online: 11 September 2025

1 Introduction

Financial institutions are essential to economic stability, primarily through providing credit to businesses and individuals, which supports investment and stimulates consumption (Alvi et al., 2024; Giri et al., 2021). However, the sustainability of credit systems depends on accurately assessing and managing credit risk, as it facilitates accurate credit risk assessment, improves lending decisions, and strengthens risk management strategies to minimize potential economic losses. In this context, credit default, a borrower's failure to fulfill debt obligations, poses a significant challenge to lenders and threatens overall financial stability (Li et al., 2022; Runchi et al., 2023). In addition, the demand for credit often exceeds the available financial resources, making it imperative for lenders to adopt more precise risk assessment methodologies to allocate credit efficiently while mitigating default risks. Consequently, enhancing predictive performance requires advanced analytical models that can capture the complex nature of borrower characteristics. These models enable more accurate risk assessments, reduce the incidence of non-performing loans, and support better-informed lending decisions, ultimately strengthening the stability of financial institutions (Dumitrescu et al., 2022; Lenka et al., 2022; Alam et al., 2020). Although widely adopted, traditional credit risk assessment approaches typically depend on static and linear modeling techniques, failing to reflect borrowers' dynamic nonlinear characteristics and associated risk patterns. Additionally, these methods frequently struggle with class imbalance, a common limitation across most existing credit risk prediction studies, as non-default cases tend to dominate the dataset (Lee and Sohn, 2022; Rao et al., 2023). For example, in many real-world lending datasets, default rates range from moderate levels like the German Credit (GC) dataset (30%) to more extreme cases of 10% or less, meaning non-default cases typically represent over 90% of the data. Specifically, in the Give Me Some Credit (GMSC) dataset, the proportion of defaulters is approximately 1 in every 14 instances, resulting in a ratio of 1:14, which poses a substantial challenge for model training and evaluation. This disproportionate distribution can lead to biased predictive models that perform well on the majority class but poorly on the minority class, ultimately reducing the model's effectiveness in identifying high-risk borrowers and undermining the reliability of credit evaluation systems. This is further compounded by the limited availability of default data, given its lower frequency. To address the limitations inherent in traditional credit risk assessment methods, recent studies have proposed various methodological advancements. Among these, resampling techniques have been widely adopted to mitigate class imbalance by enhancing the representation of minority classes within the dataset (Wu et al., 2023; Feng et al., 2022). Building upon the increasing emphasis on relational information, graph-based approaches have emerged as a promising direction, offering a flexible and expressive framework capable of modeling complex interdependencies among borrowers, lenders, and financial attributes, which enables more context-aware and adaptive risk assessments (Gao et al., 2023). Correspondingly, deep learning models have shown considerable effectiveness in capturing the nonlinear and dynamic patterns inherent in borrower behavior and credit risk (Rao et al., 2023). In particular, Graph Neural Networks (GNNs) have become powerful tools for analyzing relational data, offering advanced capabilities in capturing inter-

actions and dependencies (Wu et al., 2020; Gkarmpounis et al., 2024). Their ability to model complex interactions makes them particularly suited for credit risk evaluation, where understanding borrower relationships is crucial (Shi et al., 2024; Huang et al., 2022b). In this context, borrower interactions may include similarities in demographic or socioeconomic attributes, such as income level, employment length, or geographic location, behavioral patterns, such as repayment history or loan usage, or shared credit characteristics, such as loan amount, interest rate, or credit score. Modeling these interactions allows GNNs to uncover latent structure in the data, enhancing the model's capacity to predict credit outcomes more effectively (Wu et al., 2020). This underscores the transformative potential of graph-based methodologies in enhancing the accuracy and reliability of credit risk assessments. Traditional, widely used risk assessment strategies often struggle to identify high-risk borrowers, primarily due to the overwhelming prevalence of non-default cases. These limitations are further exacerbated in large-scale credit risk prediction and graph-based learning models, where severe class imbalance continues to diminish performance. Despite numerous efforts, many existing studies have reported only moderate results, underscoring the need for more advanced methodologies to enhance predictive accuracy in real-world credit risk applications.

To overcome these issues, we propose a hybrid approach that integrates synthetic data generation with graph-based learning to enhance predictive accuracy and robustness, particularly in scenarios with limited minority class representation. Specifically, we employ Graph Sample and aggreGatE (GraphSAGE), a state-of-the-art graph neural network architecture designed for inductive learning on large graphs, which enables the model to capture complex relational patterns by aggregating information from neighboring borrowers. To complement this, we leverage Conditional Tabular Generative Adversarial Networks (CTGAN) to generate high-quality synthetic samples for the underrepresented class. This integration helps mitigate the adverse effects of data imbalance, supporting more informed credit allocation decisions. Model performance is evaluated using standard classification metrics, including the accuracy, Receiver Operating Characteristic - Area Under the Curve (ROC-AUC), and F1-score. The proposed approach demonstrates notable improvements over several baseline machine learning algorithms and prior approaches, achieving an accuracy of 0.8174, an AUC of 0.8796, and an F1-score of 0.8346 on the GC dataset. In addition, we attained an accuracy of 0.9694, an AUC of 0.9911, and an F1-score of 0.9698 on the highly imbalanced GMSC dataset, reflecting strong discriminative capability and balanced predictive performance. Furthermore, to improve transparency and interpretability, we apply Shapley Additive exPlanations (SHAP) with saliency analysis, enabling the quantification of each feature's contribution to the model's predictions and the identification of the primary factors driving decision-making. The key contributions of this paper are as follows:

- A hybrid approach is proposed to mitigate class imbalance in credit risk prediction, combining graph sampling and aggregation with a conditional generative adversarial network. Experimental results on the imbalanced datasets demonstrate that the approach outperforms baseline models, achieving significant improvements in accuracy, F1 score, and AUC.

- The relational patterns among borrowers, such as similar credit risk profiles and repayment behaviors, are captured, allowing the model to generalize to unseen borrowers while preserving the structural integrity of the graph. This is achieved through graph-based inductive learning, which enables dynamic and nonlinear aggregation of node features from local neighborhoods. Additionally, model interpretability is enhanced using SHAP, providing feature-level insights that support transparent credit risk assessment.

The remainder of this paper is organized as follows. Section 3 provides a comprehensive overview of the GraphSAGE and CTGAN models. Section 3 reviews existing literature on credit risk prediction, particularly emphasizing traditional machine learning approaches, graph-based methodologies, and strategies for handling data imbalance. Section 4 outlines the proposed framework in detail. Section 5 presents the experimental results, including dataset analysis, evaluation metrics, and comparative performance against baseline models. Finally, Section 6 concludes the paper by summarizing the key contributions and findings.

2 Background

This section provides a comprehensive overview of the two models employed: GraphSAGE, which is used to develop the predictive model by learning node representations through neighborhood aggregation, and CTGAN, which is applied to address data limitations and mitigate class imbalance by generating synthetic default instances. A summary of the notations used throughout this paper is provided in Table 1.

2.1 Graph Sampling and Aggregation

Graph-based machine learning offers a robust framework for modeling intricate relationships among entities, especially in scenarios where connections form an essential part of the data structure (Jin et al., 2024; Wu et al., 2022). In contrast to traditional machine learning approaches that primarily utilize tabular data, GNNs learn expressive node and edge representations by leveraging the underlying graph structure (Gkarmpounis et al., 2024). This capability allows GNNs to capture implicit dependencies and interactions unmodeled in independent feature sets, enhancing performance in tasks like classification, prompting the development of specialized architectures such as Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), and graph isomorphism networks to address domain-specific challenges (Wu et al., 2020).

GraphSAGE (Hamilton et al., 2017) is an inductive framework for generating node embeddings in large-scale graphs. In contrast to transductive GNNs, which require access to the entire graph during both training and inference, GraphSAGE learns a general aggregation function that can be applied to new, unseen nodes. This inductive learning approach enables the model to make predictions on nodes that were not present during training. Instead of aggregating information from all neighbors,

Table 1 Notations and their descriptions

Notation	Description
v	Target node in the graph
u	Neighbor node of target node v
$\mathcal{N}(v)$	Set of neighbors of node v
$\mathcal{N}_S(v)$	Sampled subset of neighbors of v
S	Number of sampled neighbors
k	Specific layer in the graph neural network
h_v^k	Embedding of node v at layer k
W_k, b_k	Weight matrix and bias vector at layer k
σ	Nonlinear activation function
$\|$	Vector concatenation operator
$c_{i,j}$	Continuous feature value for -th feature of instance j
p	Specific mode/component of the distribution
η_p, q_p	Mean and standard deviation of mode p
$\alpha_{i,j}$	Normalized scalar value of continuous feature
$\beta_{i,j}$	One-hot encoded mode indicator
$d_{i,j}$	One-hot encoded categorical feature
Y_j	Concatenated latent vector for instance j
m_i	Conditional mask for categorical feature
N_c, N_d	The number of continuous and categorical features
cv	Conditional vector guiding generation

GraphSAGE samples a fixed-size subset from each node's neighborhood, known as the neighborhood sampling mechanism, which significantly reduces computational overhead. For a node v at layer k, its embedding h_v^k is computed by aggregating information from the sampled neighbors and combining it with its representation:

$$h_v^k = \sigma\left(W_k \cdot \text{AGGREGATE}\left(\{h_u^{k-1} : u \in \mathcal{N}(v)\}\right) + b_k\right), \tag{1}$$

where W_k and b_k are learnable parameters for the layer, σ is a non-linear activation function, u represents a neighbor node of target node v from its neighborhood set $\mathcal{N}(v)$, and AGGREGATE is a permutation-invariant operation applied over the neighborhood. GraphSAGE supports several aggregation strategies to combine neighborhood information. A simple and effective method is the mean aggregation, where the embedding is calculated as the average of the neighbors' representations:

$$\text{AGGREGATE}\left(\{h_u^{k-1} : u \in \mathcal{N}(v)\}\right) = \frac{1}{|\mathcal{N}(v)|} \sum_{u \in \mathcal{N}(v)} h_u^{k-1}, \tag{2}$$

Alternative strategies include pooling aggregation, which applies element-wise max-pooling over the neighbors' embeddings, and Long Short Term Memory (LSTM) aggregation, where an LSTM processes the neighbors as a sequence to capture structural dependencies. Each method offers a different trade-off between expressiveness and computational complexity. An extended version of the node update rule includes the node's previous embedding concatenated with the aggregated neighborhood representation before applying the transformation:

$$h_v^k = \sigma\left(W_k \cdot \left[h_v^{k-1} \,\|\, \frac{1}{|\mathcal{N}(v)|} \sum_{u \in \mathcal{N}(v)} h_u^{k-1}\right] + b_k\right), \tag{3}$$

where $\|$ denotes vector concatenation. The operation $\frac{1}{|\mathcal{N}(v)|} \sum_{j \in \mathcal{N}(v)} h_u^{k-1}$ computes the mean aggregation of the embeddings from neighboring nodes, ensuring that information is propagated effectively across the borrower network. This formulation emphasizes both the node's historical state and its neighborhood context in the current representation. To improve scalability, GraphSAGE employs neighborhood sampling by selecting a fixed number of neighbors per node during training:

$$\mathcal{N}_S(v) \subseteq \mathcal{N}(v), \quad |\mathcal{N}_S(v)| = S, \tag{4}$$

where S is a predefined sample size. This sampling approach enables the model to scale to large graphs by limiting the computational burden at each layer. To maintain stable training dynamics, GraphSAGE normalizes the aggregated node embeddings using L2 normalization:

$$\hat{h}_v^k = \frac{h_v^k}{\|h_v^k\|_2}, \tag{5}$$

where $h_v^{(k)}$ is the aggregated embedding at layer k, $\|h_v^{(k)}\|_2 = \sqrt{\sum_i (h_v^{(k)})^2}$ denotes its L2 norm, and $\hat{h}_v^{(k)}$ is the corresponding normalized embedding. This normalization helps prevent representation explosion and ensures a more stable gradient flow during training. GraphSAGE enables efficient and generalizable learning across large-scale and evolving graphs through this design, providing a strong foundation for tasks involving complex relational structures. The primary distinction between GraphSAGE and other GNNs is its inductive learning capability, which enables it to generalize to unseen nodes and dynamically evolving graphs. This property enhances scalability for large datasets and makes GraphSAGE particularly suitable for real-world decision-making systems. In financial applications like credit risk prediction, where borrower data is frequently updated or new borrowers are continuously added, this can embed and evaluate new nodes (Motevallian and Hossein Hasheminejad, 2021). Additionally, GraphSAGE offers a range of customizable aggregation functions, allowing for adaptation to the unique characteristics of the graph data. Figure 1 visually summarizes the GraphSAGE aggregation process across k hops of sampled neighbors.

2.2 Conditional tabular generative adversarial network

In traditional tabular data representation, information is organized in a matrix format, where each row corresponds to a unique entity, and each column represents a specific feature. These features are typically classified into two primary types: numerical and categorical (Kim et al., 2024). Numerical features capture quantifiable properties expressed as continuous values, such as age, interest rates, or monthly expen-

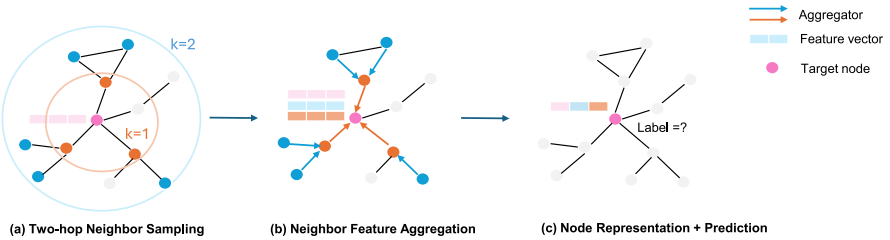


Fig. 1 Overview of the graph sampling and aggregation process. (a) A two-hop neighborhood is sampled around the target node, with one-hop neighbors shown in blue and two-hop neighbors in orange, while gray nodes indicate unsampled nodes of the graph. (b) Node features from the sampled neighbors are aggregated. (c) The resulting representation is used for node-level prediction

ditures, whereas categorical features consist of discrete values that fall into defined categories, including loan purpose, credit grade, or employment type. Effectively handling these heterogeneous feature types is essential for accurate modeling and robust decision-making. The generation of synthetic tabular data has emerged as a valuable approach to mitigating data-related limitations, particularly class imbalance, which frequently weakens model performance in real-world applications (Wang et al., 2024). By augmenting underrepresented classes, synthetic data generation can enhance both the generalizability and fairness of predictive models while preserving the statistical properties of the original dataset. Recent progress in generative modeling, especially through Generative Adversarial Networks (GANs) (Goodfellow et al., 2020), has enabled the creation of highly realistic synthetic data by leveraging a generator-discriminator architecture that learns to approximate complex data distributions. These models have shown considerable promise in addressing class imbalance challenges across various domains (Han et al., 2022a). Nevertheless, the application of GANs to tabular data introduces specific challenges due to the differing statistical characteristics of numerical and categorical data. Numerical features often exhibit skewed distributions, while categorical features may possess high cardinality and non-ordinal relationships, complicating the generation process and necessitating tailored architectural and preprocessing strategies (Borisov et al., 2024).

To address the complexities of modeling mixed-type tabular data, CTGAN (Xu et al., 2019), a type of deep generative model (Wang et al., 2024), adopts a conditional adversarial architecture. The overall procedure is composed of several stages designed to accommodate the unique characteristics of categorical and numerical feature distributions. An overview of CTGAN's data generation process is illustrated in Fig. 2. CTGAN begins by processing continuous features through mode-specific normalization. Unlike standard normalization, it applies a variational Gaussian mixture model to identify multiple modes in each continuous feature's distribution. Each value $c_{i,j}$ is normalized using the mean η_p and standard deviation q_p of its assigned mode, as shown in Eq. 6:

$$\alpha_{i,j} = \frac{c_{i,j} - \eta_p}{4q_p}, \tag{6}$$

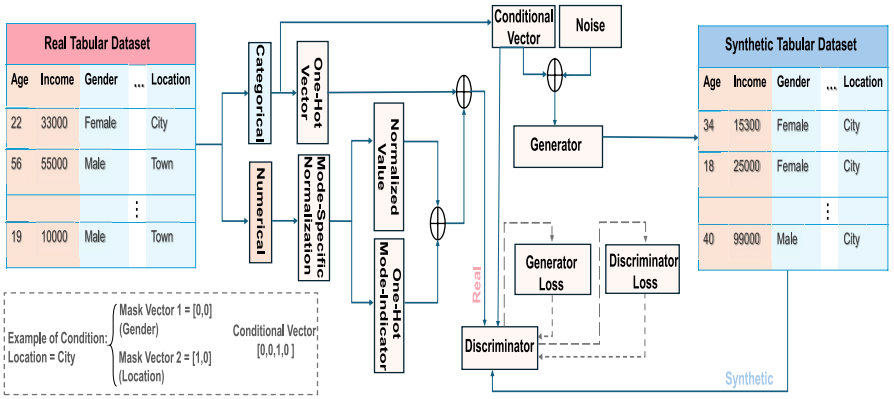


Fig. 2 Overview of the conditional tabular generative adversarial network model, illustrating the generation of synthetic data through the use of conditional vectors and feature transformations

where $\alpha_{i,j}$ is the normalized value. This approach stabilizes training by accounting for local density variations. Each normalized value is paired with a one-hot mode indicator to capture which Gaussian component the value originated from. For categorical features, CTGAN employs one-hot encoding to preserve the discrete nature of these variables. All feature representations are then concatenated into a unified latent vector, as formalized in Eq. 7:

$$Y_j = \alpha_{1,j} \oplus \beta_{1,j} \oplus \cdots \oplus \alpha_{N_c,j} \oplus \beta_{N_c,j} \oplus d_{1,j} \oplus \cdots \oplus d_{N_d,j}, \tag{7}$$

where $\oplus$ denotes concatenation, $\beta_{i,j}$ is the mode indicator, and $d_{i,j}$ is the one-hot encoded categorical feature. This integration harmonizes continuous and categorical features into a consistent input structure. To enhance categorical feature generation, CTGAN introduces a conditional vector cv, which guides the generator toward producing samples with specific category values. The conditional vector is expressed in Eq. 8:

$$cv = m_1 \oplus m_2 \oplus \cdots \oplus m_{N_d}, \tag{8}$$

where each mask vector m corresponds to a distinct categorical column. This strategy effectively handles class imbalance and sparsity, making the generator more robust and targeted. The generator receives both the latent vector and the conditional vector, along with a noise vector to inject variability. The output is evaluated by a discriminator that distinguishes real from synthetic instances. CTGAN effectively models multimodal, non-Gaussian, and imbalanced tabular data by integrating mode-aware normalization, conditional sampling, and adversarial training. Its ability to unify high-cardinality categorical and skewed numerical features enables the generation of high-fidelity synthetic data for real-world applications.

3 Related Work

This section provides a comprehensive review of existing approaches for predicting credit risk. The methods are grouped into two main categories: (1) traditional and graph-based credit risk models, and (2) data balancing techniques specifically developed to address the class imbalance in credit risk prediction.

3.1 Traditional and Graph-Based Models for Credit Risk Prediction

The ability to predict credit defaults with high accuracy is necessary for financial stability, causing extensive research into machine learning models for credit risk assessment (Bulut and Arslan, 2024). Unlike traditional statistical methods that depend on predefined assumptions, machine learning techniques autonomously identify patterns within datasets. Nevertheless, their effectiveness remains contingent on careful feature engineering and appropriate model selection (He and Fan, 2021). Traditional methods, including Logistic Regression (LR), Decision Trees (DT) (Safavian and Landgrebe, 1991), Support Vector Machines (SVM) (Cortes and Vapnik, 1995), K-nearest neighbors (KNN) (Cover and Hart, 1967), and other established algorithms, have been widely employed for credit risk prediction. These are valued for their interpretability and computational efficiency but often face challenges in modeling complex relationships within large, high-dimensional datasets, which can limit their predictive performance (Chen et al., 2021; Yıldırım et al., 2021; Lee and Sohn, 2022). Boosting techniques, such as gradient boosting, iteratively refine models by correcting previous errors, enhancing predictive accuracy, though they may still face challenges in maintaining model diversity (Shiam et al., 2024). Building on the advancements in machine learning that allow for more detailed financial risk assessment, considering factors such as income, repayment capacity, and credit history (Arora and Kaur, 2020), researchers have expanded the scope of evaluation to include not only predictive accuracy but also the interpretability of models. For instance, Aljadani et al. (2023) presented a comparative evaluation of algorithms such as the Light Gradient Boosting Model (LGBM) and eXtreme Gradient Boosting (XGBoost), alongside other relevant methods, with a focus on their interpretability. In addition, Dastile et al. (2022) introduced an optimization formulation, implemented via a custom genetic algorithm, to generate sparse and interpretable counterfactual explanations for predictions made by black-box machine learning models, including Random Forests (RF). Despite their benefits, these models alone may not always be adequate for handling large-scale, high-dimensional datasets. To overcome the limitations of existing methods, hybrid techniques have emerged. For instance, reduced minority KNN (Junior et al., 2020) enables dynamic selection techniques for imbalanced datasets by adjusting local region definitions. Similarly, stepwise multi-granular augmented gradient-boosting decision trees (Liu et al., 2021) incrementally expand feature sets to balance model diversity and variance. Despite these advancements, a persistent challenge is the trade-off between model accuracy and interpretability. Many high-performing models use complex, black-box architectures that can be difficult to explain, making them less favorable in financial decision-making contexts (Talaat et al., 2024). Interpretability remains crucial for financial

institutions, as transparent models help foster trust among stakeholders (Bussmann et al., 2021). The introduction of the Multi-Layer Perceptron (MLP) (Rumelhart et al., 1986) marked a significant advancement in deep learning, facilitating hierarchical feature representation through multiple interconnected layers of neurons. Building on this foundation, more advanced architectures have been developed to enhance predictive accuracy (Wang et al., 2018). Concurrently, research efforts have aimed at refining credit scoring models by addressing key challenges such as data imbalance and model interpretability. A notable example is the Non-Parametric Approach for Explainable Credit Scoring (NATE) (Han and Jung, 2024), which integrates oversampling strategies with tree-based classifiers to improve both predictive performance and transparency, ultimately supporting more balanced and interpretable credit risk assessment. Another advancement is the XGBoost-B-GHM model (Xia et al., 2024), which combines the Boruta algorithm for feature selection and the Gradient Harmonizing Mechanism (GHM) loss function. Boruta identifies relevant features by iteratively comparing their importance to randomized counterparts, while GHM mitigates class imbalance issues by weighting examples based on gradient density. Extracting latent features from nonlinear data while maintaining model interpretability remains a key challenge in credit scoring. Penalized Logistic Tree Regression (PLTR) (Dumitrescu et al., 2022) is another interpretable approach that improves logistic regression using decision tree-based features. Similarly, the logistic Balancing and Weighting Effects (logistic-BWE) model (Runchi et al., 2023) represents an ensemble approach based on logistic regression designed to handle sample imbalances by dynamically adjusting prediction weights. Furthermore, a study by (Qamar et al., 2023) compared various machine learning classifiers for credit scoring and bankruptcy prediction, highlighting the effectiveness of the XGBoost algorithm in determining credit eligibility based on applicants' payment behavior.

However, these machine learning methodologies assume independent data points, limiting their ability to capture complex borrowing dependencies, whereas advanced approaches aim to process large datasets, uncover hidden patterns, and improve default risk prediction (Liu et al., 2024). As a result, recent research has increasingly employed deep learning techniques tailored for graph-based analysis, leveraging their ability to model complex interconnections (Shi et al., 2024; Zandi et al., 2025). Applying GNNs in loan allocation has demonstrated significant advantages, particularly in enhancing credit default prediction. To address the limitations of relying solely on local neighborhood information, (Liu et al., 2024) introduced a GNN approach, which integrates local and global structural dependencies for improved risk assessment. Similarly, Li et al. (2022) proposed SaM-GNN, a model that combines self-attention, graph convolution, and multi-task learning while leveraging information from neighboring nodes to mitigate the impact of missing data. In the context of corporate loan prediction, Feng et al. (2022) developed CCR-GNN, which constructs individual graphs for each company and employs a graph attention network to capture intra-resource relationships. Furthermore, some advancements have incorporated temporal models into GNN frameworks, allowing for the simultaneous modeling of structural and temporal dynamics in credit data (Wu et al., 2023; Zandi et al., 2025).

3.2 Class Balancing Techniques in Credit Risk Prediction

Addressing class imbalance is crucial in credit risk prediction, as the disproportionate number of non-defaulting borrowers (majority class) compared to defaulting borrowers (minority class) can lead to biased models that favor the majority group (Namvar et al., 2018). Several strategies have been developed to mitigate this issue, including resampling techniques, such as oversampling and undersampling methods, along with synthetic data generation and ensemble learning approaches (Shi et al., 2022). Although ensemble methods like random forest with random under-sampling (Moscato et al., 2021) are frequently used to handle class imbalance in credit scoring, they can suffer from information loss, overfitting, and high computational costs. To address these issues, Yotsawat et al. (2021) proposed a Cost-Sensitive Neural Network Ensemble (CS-NNE) that leverages multiple class weights to create diverse classifiers without the drawbacks of traditional resampling methods. Beyond these methods for addressing class imbalance, one widely used technique is Generative Adversarial Networks (GANs), particularly in highly imbalanced datasets, which can create synthetic data samples that resemble real distributions, improving the representation of minority classes (Goodfellow et al., 2020), making them particularly suitable for highly imbalanced datasets. For instance, Li et al. (2021) demonstrated that GAN-generated data enhances the credit risk model by reducing classification bias. Similarly, Lei et al. (2020) introduced a GAN-driven framework incorporating user profiles and behavioral data, refining credit scoring models despite the inherent challenges of tabular datasets, such as sparse distributions and weaker feature correlations (Han et al., 2022a). Another widely-used method is the Synthetic Minority Oversampling Technique (SMOTE), which generates artificial minority samples to improve dataset balance (Chawla et al., 2002). Hybrid approaches, such as SMOTE combined with Tomek Links (SMOTE-TOMEK) and SMOTE with Edited Nearest Neighbor (SMOTE-ENN), further refine these synthetic samples by eliminating noise and reducing class overlap, ultimately improving predictive performance (Namvar et al., 2018). In addition, Lenka et al. (2022) proposed a framework integrating SMOTE with feature selection and multiple classifiers, achieving enhanced model performance. Another approach, adaptive synthetic sampling, generates synthetic samples adaptively, ensuring better representation of minority class instances (Owusu et al., 2023). Moreover, the Non-parametric Oversampling Technique for Explainable credit scoring (NOTE) was introduced (Han et al., 2024). This framework combines a non-parametric stacked autoencoder for capturing non-linear features, a conditional Wasserstein GAN (cWGAN) for class distribution balancing, and a classification process to improve explainability. Its effectiveness was evaluated by comparing baseline performance on imbalanced datasets with other oversampling techniques, such as ADSGAN (Yoon et al., 2020) for privacy-sensitive scenarios and the Deep-SMOTE (Dablain et al., 2023). Additionally, the framework incorporated LR, Extra Trees (ET), and other machine learning algorithms to further enhance its predictive capabilities. By incorporating these diverse techniques, credit risk prediction models can overcome class imbalance challenges, leading to unbiased assessments of borrowers' default risk. Table 2 summarizes recent methodological approaches to class

Table 2 Summary of selected graph-based and machine learning approaches for credit risk prediction, with their class imbalance mitigation strategies

Research	Model	Description	Imbalance Handling
Dumitrescu et al. (2022)	PLTR	Combines logistic regression with shallow decision tree rules to improve non-linear modeling while preserving interpretability.	-
Li et al. (2022)	GCN	Employs multi-task learning with self-attention enhanced feature representations to improve predictive performance.	OversamplingUndersampling
Qamar et al. (2023)	XGBoost	Compares bagging, logistic regression, gradient classifier, and random forest; XGBoost achieved the best performance.	SMOTE
Runchi et al. (2023)	logistic-BWE	Proposes an ensemble logistic regression model with dynamic weighting across sub-models trained on varying imbalance ratios while cleaning the data to enhance training set quality and maintaining interpretability.	SMOTE-ENN
Wu et al. (2023)	CD-GAT	Constructs a transaction network using a GAT-GRU architecture, optimizing for scalability and computational efficiency.	Augmentation-Undersampling
Han et al. (2024)	NOTE-RF	Integrates a non-parametric stacked autoencoder, a conditional Wasserstein GAN for synthetic data generation, and ensemble classifiers, such as random forest and extra trees, to enhance model explainability.	cWGAN
Han and Jung (2024)	NATE-GB	Integrates non-parametric oversampling with tree-based classifiers, such as gradient boosting, to address class imbalance and improve interpretability.	SMOTE
Xia et al. (2024)	XGBoost-B-GHM	Introduces an ensemble model using Boruta for feature selection and GHM loss to address feature redundancy and class imbalance. Enhances generalization and predictive performance in credit evaluation tasks.	GHM loss function
Ong and Lee (2025)	MLP	Conducts a comparative analysis of machine learning models with incremental modifications for credit scoring, finding that logistic regression performs best on smaller datasets, while random forests and multilayer perceptrons are well-suited for larger datasets.	SMOTE

imbalance handling in credit risk prediction literature, highlighting key techniques and their implementations.

4 Methodology

The proposed approach is illustrated in Fig. 3. Our approach aims to improve credit default prediction through class imbalance mitigation and graph-based borrower representations. To mitigate the class imbalance issue, CTGAN is employed to generate synthetic samples for the minority class. CTGAN learns the distributions of the original data, producing realistic and statistically representative synthetic instances that enhance class balance. By augmenting the dataset with these additional samples, the model is exposed to a more representative distribution of borrowers, improving its ability to identify default patterns and enhancing predictive performance. Following data augmentation, a KNN-based borrower graph is constructed to capture feature-driven relationships among borrowers. In this graph, each borrower is represented as a node, while edges are established based on borrower similarity in the feature space. Specifically, an edge is formed between two borrowers if one falls within the K-nearest neighbors of the other. This approach ensures that borrowers with similar financial and demographic profiles are connected, enabling the model to leverage structural dependencies for more robust credit risk assessment. To learn meaningful and higher-order borrower representations, a two-layer GraphSAGE model is employed to capture implicit relational patterns among borrowers. GraphSAGE employs a hierarchical aggregation mechanism, iteratively updating borrower

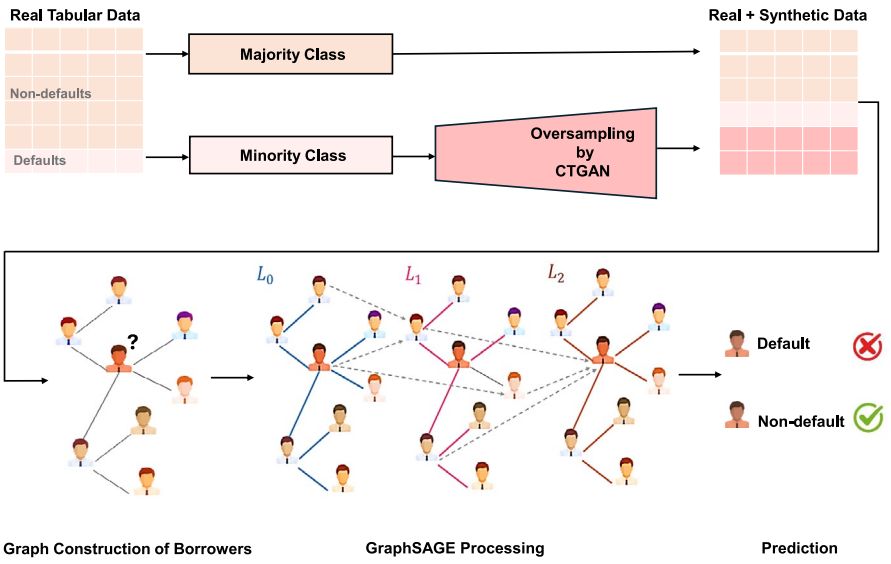


Fig. 3 Overview of the proposed approach where the dataset is balanced by oversampling the minority class of defaults with a conditional tabular generative adversarial network, followed by borrower graph construction and graph sampling and aggregation processing for credit default risk prediction

embeddings at layers L_0, L_1, and L_2. At each layer k, node embeddings are updated by aggregating information from their local neighborhood, as defined in Eq. 3. By incrementally aggregating information from neighboring borrowers, GraphSAGE refines the embeddings to capture both local and higher-order dependencies within the borrower network. The final borrower embeddings are then utilized for credit default prediction, where an MLP classifier is trained to distinguish between defaulting and non-defaulting borrowers. Incorporating graph-based features enhances the model's ability to recognize intricate borrower relationships that traditional machine learning models may overlook. By leveraging GraphSAGE for representation learning, this approach provides a scalable and data-efficient solution, particularly beneficial for financial institutions dealing with highly imbalanced datasets. To enhance model interpretability, we employed the SHAP framework (Lundberg and Lee, 2017) alongside saliency analysis, providing a comprehensive understanding of the contribution and relative importance of individual node features.

5 Experimental results

This section demonstrates the results of the experiments and provides a thorough evaluation of the proposed model's performance across multiple assessment criteria.

5.1 Data Preparation

We utilized two widely recognized credit risk datasets: the GMSC dataset (Freshcorn, 2011) and the GC dataset (Hofmann, 1994).

The GMSC dataset consists of 150,000 records, characterized by 10 features encompassing demographic, financial, and behavioral attributes. A detailed summary of the dataset's features is provided in Table 3. The dataset exhibits extreme class imbalance, with 10,026 instances (approximately 6.7%) representing defaults and 139,974 instances (93.3%) representing non-defaults, resulting in an imbalance ratio of approximately 1:14, as shown in Fig. 4, which makes the dataset particularly suitable for evaluating data generation techniques. To mitigate the issue of class imbalance, CTGAN, as detailed in Sections 2.2 and 4, was employed to generate synthetic samples of the minority class (defaults), resulting in a more balanced class distribution. Data preprocessing involved handling missing values appropriately, ensuring data integrity. Additionally, numerical features were standardized to maintain uniform scaling across the dataset, facilitating consistent model performance.

We additionally applied our model to the GC dataset. Unlike our primary dataset, GMSC, which is highly imbalanced and well-suited for evaluating GAN-based data augmentation techniques, the GC dataset exhibits a moderate imbalance with a ratio of 3:7 (300 defaults and 700 non-defaults). This provides a complementary evaluation scenario with a different class distribution and feature composition. The detailed description of the features included in the GC dataset is presented in Table 4. The comparison between the distribution of the original default instances and the CTGAN-generated samples on the GMSC and GC datasets is illustrated in Fig. 5.

Table 3 Description of features in the GMSC dataset

Feature	Description	Type
RevolvingUtilizationOfUnsecuredLines	Percentage of available unsecured credit (credit cards, personal lines) currently being utilized	Percentage
Age	Applicant's age in years	Integer
NumberOfTime30-59DaysPastDueNotWorse	Number of 30-59 day late payments (without more severe delinquency) in last 2 years	Integer
Debt Ratio	Ratio of total monthly debt obligations to gross monthly income	Percentage
MonthlyIncome	Applicant's total monthly earnings	Real
NumberOfOpenCreditLinesAndLoans	Total number of active credit accounts (installment loans, credit lines)	Integer
NumberOfTime90DaysLate	Number of times where payments were 90+ days late	Integer
NumberRealEstateLoansOrLines	Quantity of mortgage and real estate-related credit accounts	Integer
NumberOfTime60-89DaysPastDueNotWorse	Number of 60-89 day late payments (without more severe delinquency) in last 2 years	Integer
NumberOfDependents	Number of family members financially dependent on applicant (excluding self)	Integer

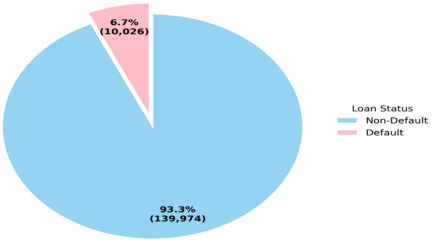


Fig. 4 Class distribution of default and non-default instances on the GMSC dataset

5.2 Experimental Setup and Model Configuration

The proposed approach was implemented in Python, utilizing PyTorch (Paszke et al., 2019) as the primary deep learning framework. To facilitate graph-based computations, PyTorch Geometric (Fey and Lenssen, 2019) was incorporated. The experimental setup included an NVIDIA Tesla K80 GPU with 11.17 GB of memory, coupled with an Intel Xeon Platinum 8260 CPU operating at 2.40 GHz. Both datasets were initially partitioned into training and testing sets at a ratio of 80:20, respectively. To evaluate the performance of both the proposed method and baseline models, several widely used metrics were employed, such as accuracy, F1-score, and ROC-AUC. The utilized loss function was the binary cross-entropy. The ROC-AUC metric assesses the model's ability to distinguish between positive and negative instances, with higher values indicating better classification performance. This metric is particularly useful for imbalanced classification tasks, such as credit default prediction, as it offers a comprehensive assessment of model performance without being influenced by class distribution. To enhance the model's performance, we implemented

Table 4 Description of features in the GC dataset

Feature	Description	Type
Status_existing_checking_account	Checking account status	Categorical
Duration_month	Loan duration	Integer
Credit_history	Record of previous credit performance	Categorical
Purpose	Intended use of the loan	Categorical
Credit_amount	Total amount of credit requested	Integer
Savings_account_bonds	Level of savings or investment holdings	Categorical
Present_employment_since	Employment length in current job	Categorical
Installment_rate_in_percentage_of_disposable_income	Installment as a percentage of income	Integer
Personal_status_and_sex	Marital status and gender	Categorical
Other_debtors_guarantors	Presence of co-applicants or guarantors	Categorical
Present_residence_since	Years at current residence	Integer
Property	Type of owned property or assets	Categorical
Age_in_years	Age of the individual	Integer
Other_installment_plans	Participation in other installment plans	Categorical
Housing	Type of housing situation	Categorical
Number_of_existing_credits_at_this_bank	Existing loans with the bank	Integer
Job	Employment category	Categorical
Number_of_people_being_liable_to_provide_maintenance_for	Number of financial dependents	Integer
Telephone	Whether a telephone is available	Categorical
Foreign_worker	Whether the applicant is a foreign worker	Categorical

architectural refinements and optimization strategies specifically tailored to the task. The GraphSAGE architecture employed in our study consists of two layers, each followed by batch normalization to stabilize and accelerate training. Table 5 presents a comprehensive overview of the hyperparameters specific to the GraphSAGE architecture, including neighbor sampling size, embedding dimensions, and other architectural settings.

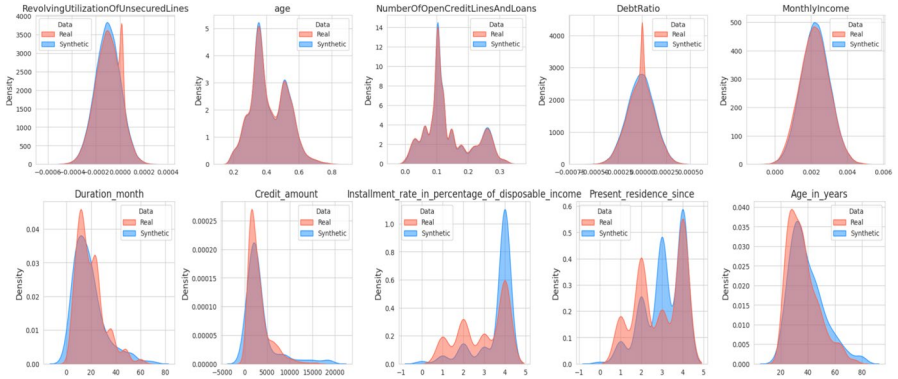
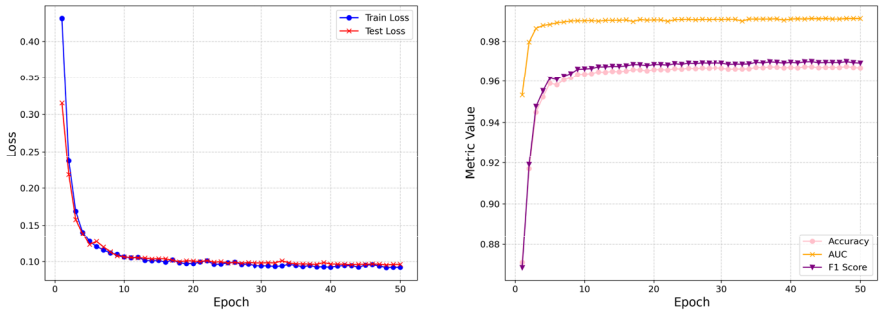


Fig. 5 Comparison of real and synthetic default instances distributions for key numerical features in the GMSC and GC datasets, with the top five features corresponding to GMSC, and the bottom five to GC

Table 5 Hyperparameter settings and selected values for our proposed approach

Dataset	Parameters	Parameter Space	Selected Value
GMSC	GraphSAGE hidden size	{32, 64, 128, 256}	128
	Learning rate	{1e-3, 1e-4, 5e-4 }	5e-4
	MLP hidden size	{32, 64, 128, 256}	256
	Weight decay	{1e-3, 1e-4, 5e-4 }	1e-3
	GraphSAGE dropout	{0.1, 0.2, 0.3, 0.4, 0.5}	0.3
	MLP dropout	{0.1, 0.2, 0.3, 0.4, 0.5}	0.5, 0.3
	Optimizer	{Adam, AdamW}	AdamW
	Activation Function	{Relu}	Relu
	Batchsize	{32, 64, 128, 256, 512, 1024, 2048}	1024
	Epoch	{30, 50, 100}	100
	K	{10, 15, 20, 25, 30}	20
	First order sampling	{2, 3 , 5, 10, 15}	5
	Second order sampling	{2, 3 , 5, 10, 15}	10
GC	GraphSAGE hidden size	{16, 32, 64}	16
	Learning rate	{1e-3, 1e-4, 5e-4 }	5e-4
	MLP hidden size	{16, 32, 64}	32
	Weight decay	{1e-3, 1e-4, 5e-4 }	5e-4
	GraphSAGE dropout	{0.1, 0.2, 0.3, 0.4, 0.5}	0.5
	MLP dropout	{0.1, 0.2, 0.3, 0.4, 0.5}	0.5, 0.3
	Optimizer	{Adam, AdamW}	AdamW
	Activation Function	{Relu}	Relu
	Batchsize	{8, 16, 32, 64, 128, 256}	128
	Epoch	{30, 50, 100}	100
	K	{10, 15, 20, 25, 30}	20
	First order sampling	{2, 3 , 5, 10, 15}	3
	Second order sampling	{2, 3 , 5, 10, 15}	10



(a) Training and test loss. (b) Metric values.

Fig. 6 Training and test loss and performance metrics of the proposed model on the GMSC dataset over the first 50 epochs

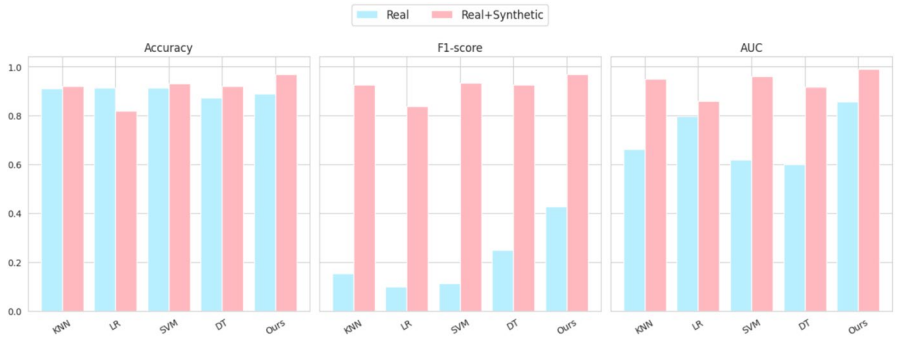


Fig. 7 Performance comparison of baselines and proposed approach on real and real+synthetic data of the GMSC dataset, evaluated using accuracy, F1-score, and AUC

5.3 Evaluation Results

Generating synthetic samples may introduce data points that resemble real ones, which could increase the risk of data leakage and overfitting if not carefully controlled. To mitigate such concerns, we systematically tracked training and validation losses on the GMSC dataset throughout the training process, observing stable convergence patterns that indicate an absence of overfitting. We also ensured that synthetic data was derived only from training splits and validated model performance on held-out real data. Figure 6 (a) shows the training and test loss curves, as well as the evaluation performance metrics in Fig. 6 (b), including accuracy, F1 score, and AUC, over the first 50 epochs of the model's training process.

The KNN, SVM, LR, and DT models were selected as baselines. These algorithms were applied to both the original datasets and an augmented version combining real and synthetic data. Figures 7 and 8 present a comparative analysis of various models across three key evaluation metrics: accuracy, F1-score, and area under the ROC curve on the GMSC and GC datasets, respectively. Across all baselines as well as the proposed approach, the integration of synthetic data results in substantial per-

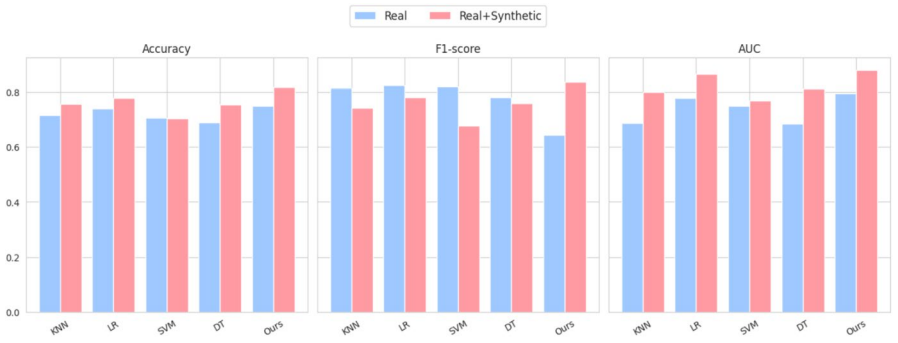


Fig. 8 Performance comparison of baselines and proposed approach on real and real+synthetic data of the GC dataset, evaluated using accuracy, F1-score, and AUC

formance improvements, particularly in terms of F1-score and AUC on the GMSC dataset. Notably, the proposed approach achieved the highest performance across all metrics, with synthetic data augmentation significantly enhancing the F1-score and improving class imbalance Handling. Specifically, when comparing the GraphSAGE model trained on the augmented dataset to the model trained solely on real data, accuracy improved by 7.78%, the F1-score increased by 54.23%, and AUC rose by 13.35%. These results highlight the effectiveness of synthetic data augmentation in enhancing model robustness and predictive performance, particularly in credit risk assessment tasks. On the GC dataset, baselines exhibited a decline in F1-score after class-balancing, likely due to the introduction of some noisy synthetic samples and the dataset's limited size constraining their capacity to learn robust decision boundaries. In contrast, our model demonstrated consistent robustness, improving its F1-score from 0.6429 to 0.8346. This suggests that our GraphSAGE approach effectively mitigates overfitting to synthetic artifacts while leveraging balanced distributions to enhance minority class recall on the GC dataset's small sample size.

To assess the robustness and reliability of our results, each experiment with the proposed approach and baselines was repeated five times, conducted on both the GC and GMSC datasets. To determine whether the observed differences in performance between our approach and the baselines were statistically significant, we employed a paired two-tailed t-test to find t-statistics. Specifically, the test examines whether the mean difference between paired observations significantly deviates from zero. The resulting t-statistic quantifies the strength of the evidence against the null hypothesis (H_0), which assumes no performance difference between the models. From this, a corresponding p-value is calculated to assess the Likelihood that the observed performance difference occurred by chance. A p-value below the conventional threshold of 0.05 is considered statistically significant, indicating strong evidence in favor of the proposed model's superiority. The results of these statistical tests, including the average performance scores, standard deviations (SD), t-statistics (t-stat), and p-values, are reported in Table 6. On the GC dataset, the proposed method on balanced data achieved an AUC of 0.8759, along with the highest F1-score and accuracy. Notably, standard deviations remain low across metrics, indicating model stability. The application of CTGAN-based class balancing yielded statistically significant improvements

Table 6 Comparison of average performance metrics and statistical significance tests between baselines and our approach on balanced and imbalanced versions of the GC and GMSC datasets

Dataset	Model	Metric	Balanced by CTGAN				Imbalanced			
			Mean	SD	T-stat	P-value	Mean	SD	T-stat	P-value
GC	KNN	Accuracy	0.7479	0.0255	3.83	0.0100	0.7410	0.0208	0.692	0.5086
		F1-score	0.7199	0.0337	4.20	0.0034	0.8293	0.0134	-12.586	<0.001
		AUC	0.8084	0.0122	8.12	<0.001	0.7096	0.0256	5.041	<0.001
	SVM	Accuracy	0.7043	0.0097	9.43	<0.001	0.7330	0.0175	1.638	0.1400
		F1-score	0.6833	0.0133	8.65	<0.001	0.8279	0.0095	-12.903	<0.001
		AUC	0.7583	0.0123	11.55	<0.001	0.7840	0.0278	0.025	0.9804
	LR	Accuracy	0.7779	0.0118	11.12	<0.001	0.7480	0.0081	0.124	0.9044
		F1-score	0.7783	0.0131	10.45	<0.001	0.8272	0.0044	-13.201	<0.001
		AUC	0.8624	0.0081	9.87	<0.001	0.7890	0.0110	-0.503	0.6286
	DT	Accuracy	0.7300	0.0266	2.84	0.0080	0.7080	0.0211	3.614	0.0068
		F1-score	0.7312	0.0364	3.02	0.0040	0.8092	0.0195	-10.843	<0.001
		AUC	0.7982	0.0166	4.19	0.0002	0.7121	0.0337	3.920	0.0044
	Ours	Accuracy	**0.8047**	**0.0084**	–	–	**0.7487**	**0.0078**	–	–
		F1-score	**0.8218**	**0.0124**	–	–	**0.5860**	**0.0363**	–	–
		AUC	**0.8759**	**0.0056**	–	–	**0.7844**	**0.0149**	–	–
GMSC	KNN	Accuracy	0.9400	0.0012	29.98	<0.001	0.9087	0.0003	0.020	0.9847
		F1-score	0.9342	0.0012	29.53	<0.001	0.0061	0.0019	17.358	<0.001
		AUC	0.9592	0.0006	47.88	<0.001	0.5154	0.0044	58.571	<0.001
	SVM	Accuracy	0.8852	0.0006	227.42	<0.001	0.9309	0.0005	-1.065	0.3180
		F1-score	0.8922	0.0006	226.31	<0.001	0.0426	0.0069	15.246	<0.001
		AUC	0.9100	0.0018	61.20	<0.001	0.6300	0.0499	67.433	<0.001
	LR	Accuracy	0.8299	0.0005	226.30	<0.001	0.9314	0.0004	-1.344	0.2157
		F1-score	0.8378	0.0005	226.91	<0.001	0.0818	0.0047	13.351	<0.001
		AUC	0.8582	0.0013	179.53	<0.001	0.7985	0.0040	5.668	<0.001
	DT	Accuracy	0.9493	0.0010	20.11	<0.001	0.8922	0.0013	17.275	<0.001
		F1-score	0.9352	0.0009	18.14	<0.001	0.2571	0.0074	4.273	0.0027
		AUC	0.9382	0.0010	73.94	<0.001	0.6039	0.0052	40.671	<0.001
	Ours	Accuracy	**0.9643**	**0.0011**	–	–	**0.9287**	**0.0040**	–	–
		F1-score	**0.9674**	**0.0010**	–	–	**0.3409**	**0.0385**	–	–
		AUC	**0.9903**	**0.0005**	–	–	**0.8284**	**0.0097**	–	–

in discriminative performance across multiple models. The LR exhibited a 7.3% increase in AUC, from 0.7890 to 0.8624, while the DT showed an 8.6% improvement, from 0.7121 to 0.7982. Our GraphSAGE model achieved the highest absolute AUC of 0.8759, representing a 9.1% enhancement over the imbalanced baseline. Although GraphSAGE achieved a superior F1-score of 0.8218 on the GC dataset, all baseline models exhibited statistically significant reductions in performance. This suggests that while CTGAN effectively expands the minority class manifold, the synthetic samples may perturb the decision boundaries of conventional classifiers through the introduction of geometrically ambiguous instances. The consistent outperformance of our graph-based approach underscores its inherent robustness to such perturbations, Likely attributable to its capacity for preserving topological relationships during feature learning. However, on the larger GMSC dataset, our approach achieves an AUC of 0.9903, outperforming all baselines ($p<0.001$), while maintain-

ing the highest F1-score. The scalability advantages of our approach become particularly evident when examining performance on the larger GMSC dataset through its hierarchical feature learning mechanism, where graph layers first aggregate localized node features, then iteratively propagate these to higher layers. Furthermore, our approach maintained the highest F1-score of 0.9674, outperforming baselines. The consistent superiority of our method in both discrimination and classification metrics demonstrates its robustness for real-world credit risk assessment, generalizing effectively across datasets of varying scales while maintaining statistically significant improvements.

In addition, Table 7 presents the best-performing results of our proposed approach alongside various machine learning and deep learning methods, demonstrating its effectiveness in handling class imbalance through techniques like SMOTE, cWGAN, and other advanced data augmentation or reweighting strategies on GMSC and GC datasets. The proposed GraphSAGE-augmented model outperforms all baseline methods in terms of Accuracy (0.9674), F1-score (0.9698), and AUC (0.9911) on the GMSC dataset, and achieves Accuracy (0.8174), F1-score (0.8346), and AUC

Table 7 Performance metrics comparison across various research studies and the proposed approach on the GC and GMSC datasets, along with their imbalance mitigation methods

Dataset	Research	Model	Imbalance Handling	Accuracy	F1-score	AUC
GC	Yotsawat et al. (2021)	CS-NNE	Cost-sensitive	0.7440	–	0.8011
	Dastile et al. (2022)	RF	–	0.7400	–	0.7600
	Aljadani et al. (2023)	LGBM	–	0.7870	0.5782	0.7332
	Runchi et al. (2023)	logistic-BWE	SMOTE-ENN	–	–	0.8568
		GraphSAGE	–	0.7500	0.6429	0.7943
	Ours	**GraphSAGE**	**CTGAN**	**0.8174**	**0.8346**	**0.8796**
GMSC	Dumitrescu et al. (2022)	PLTR	–	–	–	0.8568
	Qamar et al. (2023)	XGBoost	SMOTE	0.9400	0.9200	0.8600
	Runchi et al. (2023)	logistic-BWE	SMOTE-ENN	–	0.7840	0.7970
		RF	ADSGAN	–	–	0.9766
		ET	DeepSMOTE	–	–	0.8588
	Han et al. (2024)	Note-RF	cWGAN	–	–	0.9837
		Nate-GB	–	0.9343	0.1579	0.8542
	Han and Jung (2024)	Nate-GB	SMOTE	–	0.9072	0.9649
	Xia et al. (2024)	XGBoost-B-GHM	GHM loss function	0.9433	0.3616	0.9163
	Ong and Lee (2025)	MLP	SMOTE	0.8379	0.8363	0.8378
		GraphSAGE	–	0.8896	0.4275	0.8576
	Ours	**GraphSAGE**	**CTGAN**	**0.9674**	**0.9698**	**0.9911**

(0.8796) on the GC dataset. These results demonstrate the effectiveness of incorporating synthetic data augmentation via CTGAN in enhancing predictive performance, particularly in imbalanced credit scoring scenarios.

For a deeper understanding of the interpretability and feature importance of our proposed approach, saliency analysis was conducted on the GMSC dataset, as demonstrated in Fig. 9, where the analysis is shown for a single node as an example. Additionally, to provide a more comprehensive evaluation of the impact of various node features, Fig. 10 (a) presents a SHAP-based feature importance analysis, providing an overview of the impact of various features on the model's predictions on the GMSC dataset. Figure 10 (b) compares these features' relative significance in real and real+synthetic data. The SHAP values provide deeper interpretability by illustrating the contribution of each feature to the model's predictions. The color gradient represents different attribute levels, while the horizontal spread indicates their influence on the output. The visualizations shown in Fig. 10 (b) quantify the impact of various attributes based on model training results on the GMSC dataset, demonstrating that a financial stability metric (Debt Ratio) holds the highest importance in the real+synthetic dataset, while a credit utilization factor (Revolving Utilization of Unsecured Lines) is the second most influential, with the number of late payments (NumberOfTime30-59DaysPastDueNotWorse) ranked third in terms of influence. However, in the real dataset alone, a demographic attribute, Age, is the second most important factor, with Revolving Utilization of Unsecured Lines identified as the most important. The differences in feature importance between the real and real+synthetic

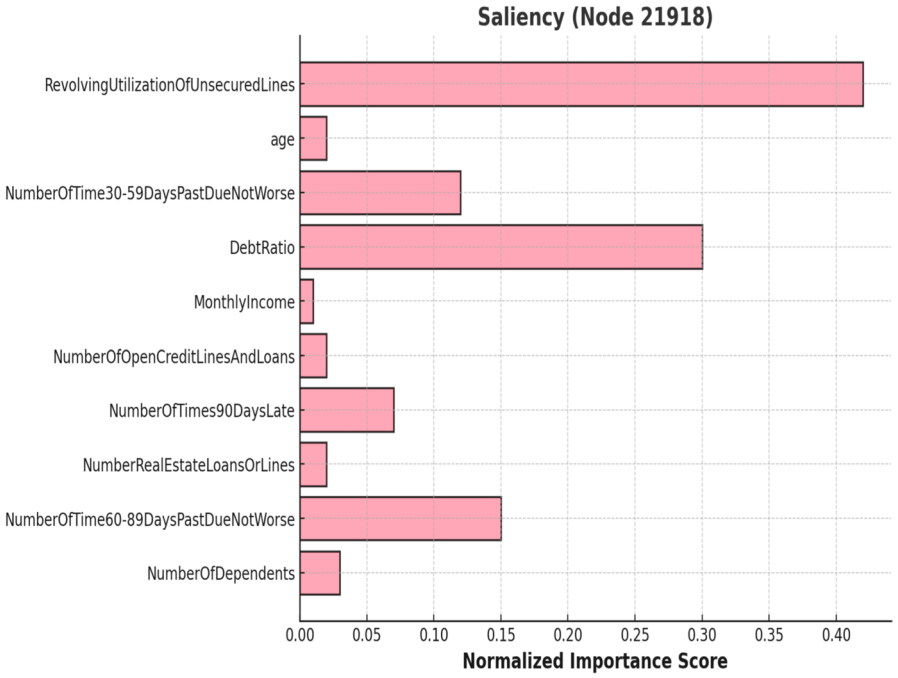
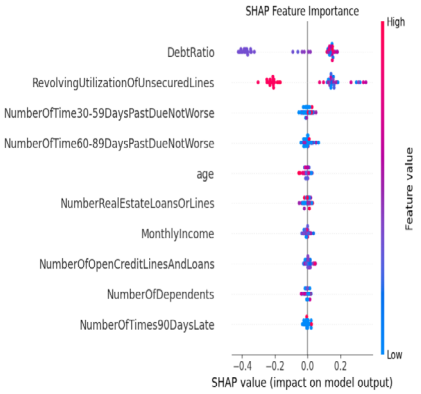
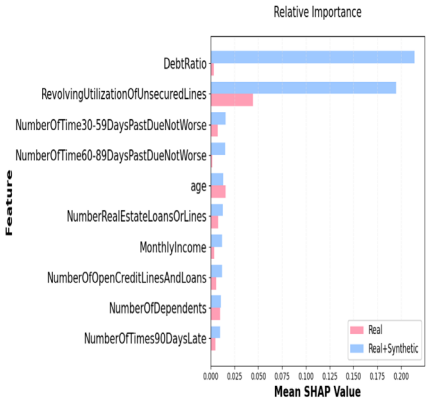


Fig. 9 Saliency-based feature importance for a representative node on the GMSC dataset



(a) Shapley value of our proposed model.



(b) Shapley importance on real and augmented data.

Fig. 10 Overview of the significance of node features on the GMSC dataset

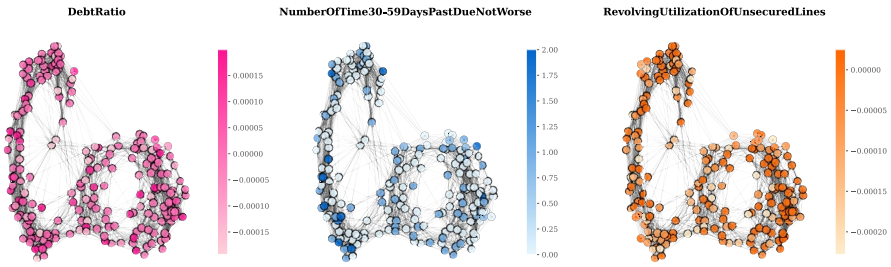
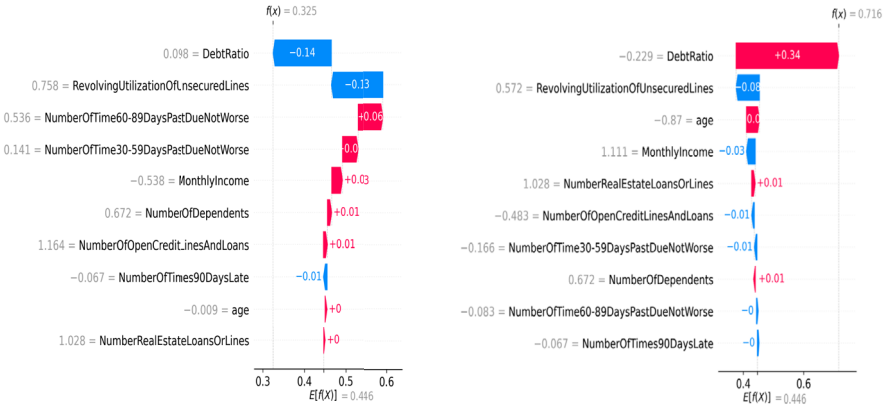


Fig. 11 Subgraphs of borrowers corresponding to the top three most influential features based on mean SHAP values on the GMSC dataset

datasets arise from the added variability of synthetic data. As shown in both Figs. 9 and 10 (a), Debt Ratio and the credit utilization factor (Revolving Utilization of Unsecured Lines) remain two of the most significant attributes, demonstrating a high relative impact compared to other features in a single node through saliency analysis and in the broader context via SHAP analysis. The subgraphs for the top three most important features, identified by SHAP values for our proposed model on the GMSC dataset, are shown in Fig. 11, illustrating how financial behaviors such as high debt ratio, past-due payments, and high revolving credit utilization are distributed within the borrower network.

We further demonstrate the interpretability of our model's outputs by presenting waterfall plots for two representative samples on the GMSC dataset, one belonging to the default class and the other to the non-default class, as shown in Fig. 12. These demonstrate the model's output $f(x)$ into additive contributions from each feature relative to the expected value $E[f(x)]$, enabling localized interpretability. In the non-default case, Fig. 12a, high values for features such as DebtRatio and RevolvingUtilizationOfUnsecuredLines reduce the predicted risk, with negative con-



(a) SHAP waterfall plot for a representative non-default sample.

(b) SHAP waterfall plot for a representative default sample.

Fig. 12 Visual explanation of feature contributions using SHAP waterfall plots for two representative nodes from the GMSC dataset, highlighting how individual features increase or decrease the model's prediction toward default or non-default outcomes

tributions of -0.14 and -0.13, respectively. In contrast, in the default case, Fig. 12b, a low DebtRatio contributes positively by $+0.34$, significantly increasing the predicted probability of default. Other features, such as MonthlyIncome and NumberOfOpenCreditLinesAndLoans, show reversed or negligible contributions across the two examples, indicating their context-dependent influence. These individualized explanations enhance the transparency and reliability of the model's credit risk predictions. Additionally, we evaluate the stability of the selected features by analyzing their consistency across different sampling sizes and perturbations, presented in Fig. 13. The mean absolute difference of SHAP values for each feature when training on 30%, 50%, 70%, and 90% of the dataset, compared to the full sample, is presented in Fig. 13a. The small differences across all sample sizes indicate high consistency of SHAP explanations even with reduced data. In addition, Fig. 13b presents an evaluation of the robustness of SHAP-based feature attributions under three perturbation strategies applied to the input features. The strategies include: (1) additive Gaussian noise with a standard deviation of 0.1, dropout perturbation, in which 10% of feature values were randomly set to zero, and permutation perturbation, involving independent random shuffling of feature values within each column. The values report the Pearson correlation between the original and perturbed SHAP values for each feature, providing a quantitative measure of attribution stability. Higher correlation values indicate greater robustness to the respective perturbation. However, for the third strategy, which involves permutation perturbation, lower correlation values are preferable, as they reflect a more significant change in the model's output and thus greater sensitivity to such perturbations. The results indicate that features such as Revolving Utilization of Unsecured Lines and Age exhibit consistently high stability across all perturbation types, whereas features like Monthly Income demonstrate lower robustness.

(a) SHAP value variation across different training set proportions.

(b) Feature importance robustness under feature perturbation using SHAP.

Fig. 13 Stability analysis of SHAP-based feature attributions on the GMSC dataset. (a) Examines the consistency of SHAP values when the training data size is varied. (b) Evaluates the sensitivity of feature importance under input perturbations

6 Conclusions

Credit default prediction is a necessary component of financial decision-making and poses significant challenges, primarily due to extreme class imbalance, Limited availability of default cases, and insufficient exploration of latent patterns and model interpretability in non-linear borrower data. This study improved these challenges by leveraging a graph neural network, specifically graph sampling and aggregation, to capture borrower relationships inductively for unseen borrowers, particularly in real-world credit risk prediction scenarios, while employing a conditional tabular generative adversarial network to generate synthetic data and mitigate class imbalance. By incorporating synthetic samples, this research aimed to enhance the robustness of predictive models. The significant superiority of the proposed approach over existing state-of-the-art models was demonstrated through extensive experiments. The results showed substantial improvements in key performance metrics for the highly imbalanced Give Me Some Credit dataset, achieving an AUC of 0.9911, while demonstrating improvements on the German Credit dataset. Additionally, a Shapley additive explanation analysis was conducted to enhance interpretability by quantifying the contribution of each feature to the model's predictions, deriving insights into the key drivers of credit risk, and promoting stakeholder confidence in the decision-making process. These comprehensive findings highlight the potential of combining synthetic data generation with graph-based modeling to improve prediction performance in domains with limited and imbalanced data.

Funding The authors did not receive support from any organization for the submitted work.

Data Availability The datasets are publicly available. The Give Me Some Credit dataset can be downloaded from https://www.kaggle.com/datasets/brycecf/give-me-some-credit-dataset and the German Credit datas et can be downloaded from https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data.

Declarations

Competing Interests The authors declare no competing interests.

Ethics approval and consent to participate Not applicable.

References

Aljadani, A., Alharthi, B., Farsi, M.A., Balaha, H.M., Badawy, M., Elhosseini, M.A. (2023). Mathematical modeling and analysis of credit scoring using the lime explainer: A comprehensive approach. Mathematics *11*(19). https://doi.org/10.3390/math11194055

Alvi, J., Arif, I., & Nizam, K. (2024). Advancing financial resilience: A systematic review of default prediction models and future directions in credit risk management. *Heliyon, 10*(21), 39770. https://doi.org/10.1016/j.heliyon.2024.e39770

Arora, N., Kaur, P.D. (2020). A bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing86*, 105936. https://doi.org/10.1016/j.asoc.2019.105936

Alam, T.M., Shaukat, K., Hameed, I.A., Luo, S., Sarwar, M.U., Shabbir, S., Li, J., Khushi, M. (2020) An investigation of credit card default prediction in the imbalanced datasets. *IEEE Access8*, 201173–201198. https://doi.org/10.1109/ACCESS.2020.3033784

Bulut, C., & Arslan, E. (2024). Comparison of the impact of dimensionality reduction and data splitting on classification performance in credit risk assessment. *Artificial Intelligence Review, 57*(9), 252. https://doi.org/10.1007/s10462-024-10904-1

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

Borisov, V., Leemann, T., Sebler, K., Haug, J., Pawelczyk, M., Kasneci, G. (2024). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems35*, 7499–7519. https://doi.org/10.1109/TNNLS.2022.3229161

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research16*, 321–357. https://doi.org/10.1613/jair.953

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Chen, Y.-R., Leu, J.-S., Huang, S.-A., Wang, J.-T., Takada, J.-I. (2021). Predicting default risk on peer-to-peer lending imbalanced datasets. *IEEE Access9*, 73103–73109 https://doi.org/10.1109/ACCESS.2021.3079701

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Dastile, X., Celik, T., Vandierendonck, H. (2022). Model-agnostic counterfactual explanations in credit scoring. *IEEE Access10*, 69543–69554. https://doi.org/10.1109/ACCESS.2022.3177783

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research, 297*(3), 1178–1192. https://doi.org/10.1016/j.ejor.2021.06.053

Dablain, D., Krawczyk, B., & Chawla, N. V. (2023). Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems, 34*(9), 6390–6404. https://doi.org/10.1109/TNNLS.2021.3136503

Fey, M., Lenssen, J. (2019). Fast graph representation learning with pytorch geometric. https://doi.org/10.48550/arXiv.1903.02428

Freshcorn, B. (2011) Give Me Some Credit. Kaggle. Accessed: 19/03/2025. https://www.kaggle.com/datasets/brycecf/give-me-some-credit-dataset

Feng, B., Xu, H., Xue, W., Xue, B. (2022). Every corporation owns its structure: corporate credit rating via graph neural networks. In: Chinese Conference on Pattern Recognition and Computer Vision, pp. 688–699. https://doi.org/10.1007/978-3-031-18907-4_53

Giri, P. K., De, S. S., Dehuri, S., & Cho, S.-B. (2021). Biogeography based optimization for mining rules to assess credit risk. *Intelligent Systems in Accounting, Finance and Management, 28*(1), 35–51. https://doi.org/10.1002/isaf.1486

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139–144. https://doi.org/10.1145/3422622

Gkarmpounis, G., Vranis, C., Vretos, N., Daras, P. (2024). Survey on graph neural networks. *IEEE Access12*, 128816–128832. https://doi.org/10.1109/ACCESS.2024.3456913

Gao, C., Zheng, Y., Li, N., Li, Y., Qin, Y., Piao, J., Quan, Y., Chang, J., Jin, D., He, X., et al. (2023). A survey of graph neural networks for recommender systems: Challenges, methods, and directions. *ACM Transactions on Recommender Systems, 1*(1), 1–51. https://doi.org/10.1145/3568022

He, H., Fan, Y. (2021). A novel hybrid ensemble model based on tree-based method and deep learning method for default prediction. *Expert Systems with Applications176*, 114899. https://doi.org/10.1016/j.eswa.2021.114899

Han, S., & Jung, H. (2024). Nate: Non-parametric approach for explainable credit scoring on imbalanced class. *PloS one, 19*(12), 0316454. https://doi.org/10.1371/journal.pone.0316454

Han, S., Jung, H., Yoo, P. D., Provetti, A., & Cali, A. (2024). Note: non-parametric oversampling technique for explainable credit scoring. *Scientific Reports, 14*(1), 26070. https://doi.org/10.1038/s41598-024-78055-5

Han, G., Liu, S., Chen, K., Yu, N., Feng, Z., Song, M. (2022). Imbalanced sample generation and evaluation for power system transient stability using ctgan. In: Intelligent Computing & Optimization, pp. 555–565. https://doi.org/10.1007/978-3-030-93247-3_55

Huang, C., Li, M., Cao, F., Fujita, H., Li, Z., & Wu, X. (2022). Are graph convolutional networks with random weights feasible? *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(3), 2751–2768. https://doi.org/10.1109/TPAMI.2022.3183143

Hofmann, H. (1994). Statlog (German Credit Data). https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data. *UCI Machine Learning Repository*, accessed on 1 June 2025

Hamilton, W.L., Ying, R., Leskovec, J. (2017). Inductive representation learning on large graphs. In: the 31st International Conference on Neural Information Processing Systems, pp. 1025–1035

Jin, M., Koh, H. Y., Wen, Q., Zambon, D., Alippi, C., Webb, G. I., King, I., & Pan, S. (2024). A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(12), 10466–10485. https://doi.org/10.1109/TPAMI.2024.3443141

Junior, L., Nardini, F.M., Renso, C., Trani, R., Macedo, J.A. (2020). A novel approach to define the local region of dynamic selection techniques in imbalanced credit scoring problems. *Expert Systems with Applications152*, 113351. https://doi.org/10.1016/j.eswa.2020.113351

Kim, D.-K., Ryu, D., Lee, Y., Choi, D.-H. (2024). Generative models for tabular data: A review. *Journal of Mechanical Science and Technology38*, 1–17. https://doi.org/10.1007/s12206-024-0835-0

Lenka, S.R., Bisoy, S.K., Priyadarshini, R., Sain, M. (2022). Empirical analysis of ensemble learning for imbalanced credit scoring datasets: A systematic review. *Wireless Communications and Mobile Computing1*, 6584352. https://doi.org/10.1155/2022/6584352

Liu, W., Fan, H., Xia, M. (2021). Step-wise multi-grained augmented gradient boosting decision trees for credit scoring. *Engineering Applications of Artificial Intelligence97*, 104036. https://doi.org/10.1016/j.engappai.2020.104036

Lundberg, S.M., Lee, S.-I. (2017). A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777. https://doi.org/10.1007/978-3-030-93247-3_55

Li, J., Liu, H., Yang, Z., Han, L. (2021). A credit risk model with small sample data based on g-xgboost. *Applied Artificial Intelligence35*, 1550–1566. https://doi.org/10.1080/08839514.2021.1987707

Lee, J. W., & Sohn, S. Y. (2022). Evaluating borrowers' default risk with a spatial probit model reflecting the distance in their relational network. *PLOS ONE, 16*(12), 1–11. https://doi.org/10.1371/journal.pone.0261737

Liu, Y., Wang, X., Meng, T., Ai, W., Li, K. (2024). Lg-gnn: Local and global information-aware graph neural network for default detection. *Computers and Operations Research169*, 106738. https://doi.org/10.1016/j.cor.2024.106738

Li, Z., Wang, X., Yao, L., Chen, Y., Xu, G., Lim, E.-P. (2022). Graph neural network with self-attention and multi-task learning for credit default risk prediction. In: International Conference on Web Information Systems Engineering, pp. 616–629. https://doi.org/10.1007/978-3-031-20891-1_44

Lei, K., Xie, Y., Zhong, S., Dai, J., Yang, M., Shen, Y. (2020). Generative adversarial fusion network for class imbalance credit scoring. *Neural Computing and Applications32*, 8451–8462. https://doi.org/10.1007/s00521-019-04335-1

Motevallian, S.N., Hossein Hasheminejad, S.M. (2021). Using trust statements and ratings by graphsage to alleviate cold start in recommender systems. In: 2021 12th International Conference on Information and Knowledge Technology, pp. 139–143. https://doi.org/10.1109/IKT54664.2021.9685137

Moscato, V., Picariello, A., Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications165*, 113986. https://doi.org/10.1016/j.eswa.2020.113986

Namvar, A., Siami, M., Rabhi, F., & Naderpour, M. (2018). Credit risk prediction in an imbalanced social lending environment. *International Journal of Computational Intelligence Systems, 11*(1), 925–935. https://doi.org/10.2991/ijcis.11.1.70

Ong, J.C., Lee, L.S. (2025). Credit scoring: A comparison of machine learning models and their modifications. *Applied Mathematics and Computational Intelligence14*(1), 57–78. https://doi.org/10.58915/amci.v14i1.1362

Owusu, E., Quainoo, R., Mensah, S., Appati, J.K. (2023). A deep learning approach for loan default prediction using imbalanced dataset. *International Journal of Intelligent Information Technologies19*, 1–16 https://doi.org/10.4018/IJIIT.318672

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In: the 33rd International Conference on Neural Information Processing Systems

Qamar, S., Durad, M. H., Islam, F. U., Saleha, S. R., Hamza, M., Urooj, A. H., & Akber, S. M. A. (2023). Ai credit: Machine learning based credit score analysis. *Journal of Computing & Biomedical Informatics, 5*(01), 217–229.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Learning representations by back-propagating errors. nature, 323*(6088), 533–536. https://doi.org/10.1038/323533a0

Runchi, Z., Liguo, X., Qin, W. (2023). An ensemble credit scoring model based on logistic regression with heterogeneous balancing and weighting effects. *Expert Systems with Applications212*, 118732. https://doi.org/10.1016/j.eswa.2022.118732

Rao, S., Verma, A., Bhatia, T. (2023). Hybrid ensemble framework with self-attention mechanism for social spam detection on imbalanced data. *Expert Systems with Applications217*, 119594. https://doi.org/10.1016/j.eswa.2023.119594

Shiam, S., Hasan, M., Pantho, M., Shochona, S., Nayeem, M.B., Choudhury, M.T.H., Nguyen, T. (2024). Credit risk prediction using explainable ai. *Journal of Business and Management Studies6*, 61–66. https://doi.org/10.32996/jbms.2024.6.2.6

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics, 21*(3), 660–674. https://doi.org/10.1109/21.97458

Shi, Y., Qu, Y., Chen, Z., Mi, Y., & Wang, Y. (2024). Improved credit risk prediction based on an integrated graph representation learning approach with graph transformation. *European Journal of Operational Research, 315*(2), 786–801. https://doi.org/10.1016/j.ejor.2023.12.028

Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications, 34*(17), 14327–14339. https://doi.org/10.1007/s00521-022-07472-2

Talaat, F. M., Aljadani, A., Badawy, M., & Elhosseini, M. (2024). Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction. *Neural Computing and Applications, 36*(9), 4847–4865. https://doi.org/10.1007/s00521-023-09232-2

Wang, A.X., Chukova, S.S., Simpson, C.R., Nguyen, B.P. (2024). Challenges and opportunities of generative models on tabular data. *Applied Soft Computing166*, 112223. https://doi.org/10.1016/j.asoc.2024.112223

Wang, C., Han, D., Liu, Q., Luo, S. (2018). A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism lstm. *IEEE Access7*, 2161–2168. https://doi.org/10.1109/ACCESS.2018.2887138

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems, 32*(1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386

Wu, S., Sun, F., Zhang, W., Xie, X., & Cui, B. (2022). Graph neural networks in recommender systems: a survey. *ACM Computing Surveys, 55*(5), 1–37. https://doi.org/10.1145/3535101

Wu, J., Zhao, X., Yuan, H., & Si, Y.-W. (2023). Cdgat: a graph attention network method for credit card defaulters prediction. *Applied Intelligence, 53*(10), 11538–11552. https://doi.org/10.1007/s10489-022-03996-1

Xia, Y., Jiang, S., Meng, L., & Ju, X. (2024). Xgboost-b-ghm: An ensemble model with feature selection and ghm loss function optimization for credit scoring. *Systems, 12*(7), 254. https://doi.org/10.3390/systems12070254

Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN,. https://doi.org/10.5555/3454287.3454946

Yoon, J., Drumright, L. N., & Schaar, M. (2020). Anonymization through data synthesis using generative adversarial networks (ads-gan). *IEEE Journal of Biomedical and Health Informatics, 24*(8), 2378–2388. https://doi.org/10.1109/JBHI.2020.2980262

Yıldırım, M., Okay, F.Y., Özdemir, S. (2021). Big data analytics for default prediction using graph theory. *Expert Systems with Applications176*, 114840. https://doi.org/10.1016/j.eswa.2021.114840

Yotsawat, W., Wattuya, P., Srivihok, A. (2021). A novel method for credit scoring based on cost-sensitive neural network ensemble. *IEEE Access9*, 78521–78537. https://doi.org/10.1109/ACCESS.2021.3083490

Zandi, S., Korangi, K., Óskarsdóttir, M., Mues, C., & Bravo, C. (2025). Attention-based dynamic multilayer graph neural networks for loan default prediction. *European Journal of Operational Research, 321*(2), 586–599. https://doi.org/10.1016/j.ejor.2024.09.025

Authors and Affiliations

Sogand Pourkhoshgoftar[1] · Asadollah Shahbahrami[1,3] · Nima Esmi[2,3]

✉ Asadollah Shahbahrami
shahbahrami@guilan.ac.ir

✉ Nima Esmi
n.esmi.rudbardeh@rug.nl

Sogand Pourkhoshgoftar
pourkhoshgoftar@msc.guilan.ac.ir

[1] Department of Computer Engineering, University of Guilan, Rasht, Iran

[2] Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, Faculty of Science and Engineering, University of Groningen, Groningen, The Netherlands

[3] Intelligent Systems Research Center, Khazar University, Baku, Azerbaijan