JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
https://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Adiba, J. I. Q., Sarno, R., Sungkono, K. R., Haryono, A. T., Tjoa, A. Min, & Lee, S. S. (2026). Wasserstein Generative Adversarial Network with gradient penalty and threshold-enhanced for imbalanced panel data for financial fraud detection. *Journal of Information and Communication Technology, 25*(1), 39-63. https://doi.org/10.32890/jict2026.25.1.3

Wasserstein Generative Adversarial Network with Gradient Penalty and Threshold-Enhanced for Imbalanced Panel Data for Financial Fraud Detection

[1]Jordan Istiqlal Qalbi Adiba, [2]Riyanarto Sarno, [3]Kelly Rossa Sungkono,
[4]Agus Tri Haryono, [5]A Min Tjoa & [6]Sang-Seok Lee
[1,2,3&4]Department of Informatics, Institut Teknologi Sepuluh Nopember, Indonesia
[5]Faculty of Computer Science, University of Vienna, Austria
[6]Graduate School of Engineering, Tottori University, Japan

[1]6025231026@student.its.ac.id
[*2]riyanarto@its.ac.id
[3]kelly@its.ac.id
[4]70252310200@student.its.ac.id
[5]a.tjoa@tuwien.ac.at
[6]sslee@tottori-u.ac.jp
*Corresponding author

Received: 4/8/2025 Revised: 14/10/2025 Accepted: 6/1/2026 Published: 31/1/2026

ABSTRACT

Financial statement fraud detection is critical to maintaining trust among investors, regulators, and analysts. However, traditional audit procedures often fail to detect anomalies effectively because they occur infrequently but can result in significant economic losses. This study proposes an oversampling approach using a modified threshold in the Wasserstein Generative Adversarial Network with Gradient Penalty (WGANGP) to enhance synthetic data variance in financial fraud detection. The financial data were collected from the financial reports of companies listed on the Indonesia Stock Exchange and were labelled according to the Balanced Scorecard framework into four categories: normal, alarm, risky, and fraud. Given the severe class imbalance, this study introduces a WGANGP model with threshold optimisation in the generator and a gradient penalty to generate high-quality synthetic samples. This study conducted general and per-entity oversampling scenarios and evaluated them using Euclidean distance, Wasserstein distance, and classification metrics. In Scenario 1, the Generative Adversarial Network (GAN) outperformed the Synthetic Minority Oversampling Technique (SMOTE) and vice

versa in Scenario 2. However, in financial fraud detection, the WGANGP with enhanced thresholding improved the F1-score by 13% to 17% compared to SMOTE and five GAN-based models across thirteen classification models, including traditional, machine learning, and deep learning models. This finding suggests that optimising the threshold in WGANGP reduces variance and improves model performance. Furthermore, generating synthetic data that is very similar to actual data may not necessarily improve classification; therefore, it is necessary to test how oversampling affects subsequent stages.

Keywords: Fraud detection, generative adversarial network, imbalanced, oversampling, financial statement.

INTRODUCTION

Financial reports in the investment world are crucial for maintaining the trust of investors, regulators, and analysts, enabling them to make informed decisions based on the data they provide. However, many companies try to deceive investors by falsifying financial reports (Mohammed, 2022). Detecting financial statement falsification can increase investor confidence by stabilising the company's performance (B. Li et al., 2024). Existing fraud identification procedures rely on applicable regulations and routine auditor checks, which are less effective and may result in errors (Sokolenko, 2022). Developments in other sectors, such as artificial intelligence and big data, can be leveraged to identify fraud in financial statements through their integration. However, developing artificial intelligence for fraud detection in financial reports often faces challenges due to an imbalance in data distribution between normal and extreme values. Previous studies have developed oversampling models for fraud-detection datasets (Zhou et al., 2021), which are expected to overcome the limitations of ordinary statistical methods in detecting fraud in financial statements.

Data structure is also an important perspective for oversampling, as panel analysis is one of the most relevant approaches to uncovering the accuracy of fraud detection. Using panel data allows for the review of fiscal behaviour over time while accounting for the impact of factors across entities (Baesens et al., 2021). Other work based on these studies in panel data analysis has shown that this method enables the capture of patterns that would not be detected using static methods (Hashemi et al., 2023), for example, cyclical temporal patterns in reported changes or other unusual temporal patterns in financial figures. The previous study on fraud detection using financial statements and the application of artificial intelligence to predict fraud in the financial reports of officially registered companies from various sectors, such as banking and manufacturing, by applying textual intelligence recognition to improve prediction accuracy (Hashemi et al., 2023; X. Li et al., 2023).

The study applies data-balancing methods to address data imbalance using traditional models, such as the Synthetic Minority Oversampling Technique (SMOTE), as well as generative models, such as the Generative Adversarial Network (GAN). This data distribution helps balance the class distribution, enabling the model to learn more from the training data and better evaluate and predict financial statement fraud (Aftabi et al., 2023; Xiuguo & Shengyong, 2022). However, the oversampling side of SMOTE has the disadvantage of overfitting, which has already been tackled by GAN oversampling and can manage extreme imbalance.

Over the years, GAN models have emerged as an effective technique for intelligent data generation, especially for applications in anomaly detection, including financial fraud. However, various types of GANs, such as Wasserstein Generative Adversarial Networks (WGAN), still suffer from instability and difficulty learning the true data distribution. Among the various improvements made, one of the biggest is the WGAN with gradient penalty (WGANGP), which addresses instability in training. Although WGANGP provides greater stability to generative models, there are still opportunities for improvement, such as refining the distance between the discriminator and the actual distribution, as well as between the discriminator and the generated distributions.

This study proposes a method for identifying financial statement fraud by combining threshold optimisation of the generator value, discriminator loss, and output gradient in the WGANGP model with label-based oversampling, and comparing it with oversampling based on the panel data structure. It was applied to explore the effects of threshold settings on the quality of synthetic data, particularly in reducing mode collapse and improving the representation of data distributions. This study conducted two evaluation stages: measuring the similarity between synthetic and original data and testing oversampled data with various classification models. The evaluation aimed to assess how effectively the proposed method enhances the accuracy and stability of the classification model.

RELATED WORKS

This section summarises and discusses the literature relevant to the study's key steps. This section is divided into two main subsections, namely the discussion of the Generative Adversarial Network Model and the machine learning (ML) model.

Generative Adversarial Network

In fraud detection, particularly in financial statement studies, adopting an appropriate architectural approach or model can enhance prediction accuracy. Other works demonstrate the importance of data balancing in improvingetection performance, especially in fraud detection (Aftabi et al., 2023; X. Li et al., 2023). Complex deep learning (DL) structures also show greater potential of fraud detection than conventional ML models (Haryono et al., 2024; Xiuguo & Shengyong, 2022).

Many GAN architectures have been used for dataset balancing and oversampling methods in image and tabular datasets. It provides a way to generate more representative synthetic data, prevents overfitting, and is more effective at addressing extreme imbalances than traditional oversampling methods (Permataning Tyas et al., 2023; Shafqat & Byun, 2022). However, the GAN approach holds great promise for improving fraud detection performance (Hsin et al., 2022). The limitation in the use of GAN for oversampling is the inability of the model to handle various distribution data imbalances optimally, especially in terms of setting the output of a particular model to maintain statistical distribution on the gradient used in training (Dina et al., 2024; W. Li et al., 2023; Liao & Dong, 2022). Although GAN and their derivatives, such as WGANGP, which add gradient inputs, offer better training stability than traditional oversampling methods, they still struggle to generate representative synthetic samples under extreme imbalance conditions, such as fraudulent class ratios of less than 5% of the total data. These limitations prevent the model from learning the optimal minority distribution, thereby degrading fraud detection performance.

Although WGANGP offers improved training stability compared to classical GAN through the application of gradient penalty regularisation, challenges in generating quality synthetic samples remain, especially in highly imbalanced data conditions. In this study, further exploration is carried out by setting an output threshold. This approach allows the model to be more flexible and in-depth in regulating the generative process and evaluating how each component contributes to the quality and distribution of synthetic outputs (Chen et al., 2023; Sihwail et al., 2024). The results show that although threshold flexibility can expand the optimal parameter search space, the imprecision of the settings can cause an imbalance in the distribution of generated data—especially for minority classes—and disrupt overall training stability (Miftahushudur et al., 2024).

To address this, a more formal approach is needed to set the gradient penalty threshold and to more effectively regulate the roles of the generator and discriminator (Asokan & Seelamantula, 2023; Kim et al., 2022). The balance between these two components is crucial so that the generator can produce much more realistic and real samples. At the same time, the discriminator is more sensitive to the features of the minority class (Fernando & Tsokos, 2022). This change is expected to enhance oversampling quality, facilitate performance improvements in balancing imbalances, and increase the accuracy of error estimation in fraud detection.

The proposed model with threshold value adjustment is compared with several oversampling methods from traditional models, such as SMOTE and GAN, and their derivatives, including Wasserstein Generative Adversarial Network (WGAN), Conditional Generative Adversarial Network (CGAN), Relativistic Generative Adversarial Network (RGAN), and Packed Generative Adversarial Network (PacGAN). The WGAN utilises the generative distance, measured by the Wasserstein distance, to evaluate the quality of synthetic data relative to the original data (Man et al., 2022). CGAN models incorporate condition parameters, such as specific features like labels or other conditions, to provide additional information to the generator and discriminator (Naderi et al., 2021; Yin et al., 2023). In the RGAN model, the discriminator does not compare fake samples with real samples but determines how appropriate or unrealistic the fake samples are compared to the original samples (Hrishikesh et al., 2023). The pack will generate the PacGAN model to overcome the collapse mode that often occurs in oversampling models.

ML Model

Based on various previous studies (Hashemi et al., 2023; X. Li et al., 2023; Qin, 2021; Sarno et al., 2015), the development of artificial intelligence is an important factor in detecting fraud in financial statements. Traditional models, ML, and DL technology have been applied to overcome data complexity, requested assets, asset distribution, and the unclear nature of fraud signs. Many related studies apply several models that perform well at identifying fraud across various financial aspects, including XGBoost, Random Forest, and Decision Tree for fraud detection, demonstrating the robustness of these models in fraud studies (Aftabi et al., 2023).

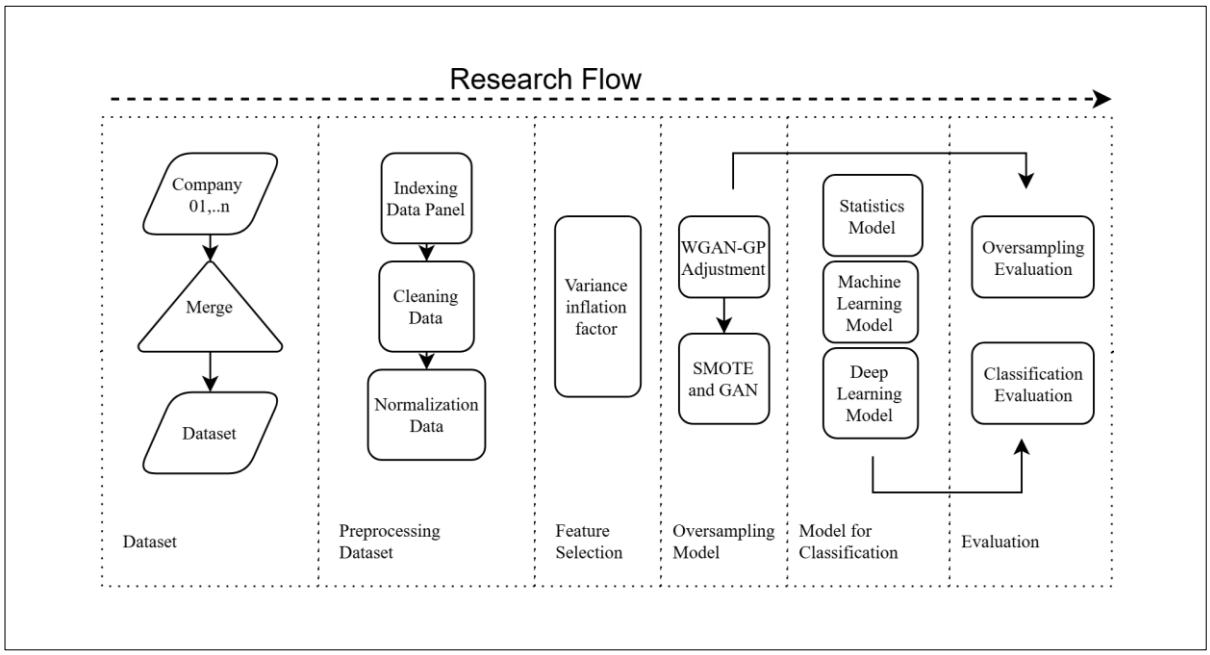
DL models are helpful for fraud detection because they address two main problems: non-linearities and temporal dynamics. A study on Chinese company data reveals that Long Short-Term Memory (LSTM) and GRU are more accurate than traditional machine-learning techniques (Xiuguo & Shengyong, 2022). The Artificial Neural Network (ANN) and Recurrent Neural Network (RNN) structures used in this study are designed to identify temporal patterns in companies' financial reports. In addition, a Convolutional Neural Network (CNN) enables the consideration of the spatial or non-linear characteristics of the data for deeper sensing (Widiantoro et al., 2024); CNN-LSTM combines both practical approaches.

THE PROPOSED METHOD

The proposed method comprises several steps, including dataset preparation, data labelling, dataset pre-processing, an oversampling model, a classification model, and evaluation, as shown in Figure 1.

Figure 1

Flowchart of the Proposed Method



Dataset

The data used in this study were obtained from the Stockbit investment portal, covering income statements, balance sheets, and cash flows. The scraping data-collection technique yielded several features in each report, including companies and quarters, which will be the primary aspects of panel data preparation. The time frame for data collection for each company will undoubtedly differ, but the earliest data for each company is from Quarter 1 in 2008 to Quarter 4 in 2024. The data collected includes 171 manufacturing companies. Based on the period applied, the data obtained includes 11,628 data points from 17 manufacturing company sectors, each with 68 quarters. This study used 77 qualified companies, each meeting 114 features in 68 quarterly data points. Thus, the total data obtained in this study is 5,236 from each company.

With the acquisition of this raw data, the labelling process using the Balanced Scorecard method was adjusted to the features available in the dataset so that the data that has been obtained has a label on each data point consisting of four labels, namely "*normal*", "*alarm*", "*risky*", and "*fraud*". This labelling was unbalanced, with some labels in the majority and others in the minority, study is available at *https://anonymous.4open.science/r/fraud-detection-using-oversampling-optimazation-3867*.

Labelling Data

The collected data lack default labels, so a representative, systematic labelling system is needed. In this study, the labelling method refers to the Balanced Scorecard (BSC) approach (Fadel et al., 2021; Saksono & Bernardus, 2023). BSC is a framework that evaluates company performance across four main perspectives: financial, customer, internal business processes, and learning and growth (Abueid et al., 2022). Each perspective is analysed using indicators that reflect the company's normal and anomalous performance conditions (Kaewninprasert et al., 2024). A total of 5,236 rows of data with 114 features were analysed and categorised into groups based on the anomalous patterns of each perspective. Each anomaly is determined based on statistical calculations, significant changes in values, and deviations from historical and industry averages.

The indicators used include Return on Assets (ROA), Return on Equity (ROE), financial average, and total revenue. Anomalous conditions from this perspective include a sudden drop, a decrease in ROA or ROE of more than 20% compared to the previous quarter. Deviation industry: a significant deviation from the company's average value. Unnatural Trend: ROA, ROE, or financial average values exceed the statistical threshold (z-score > 2.5). Incompatibility with Industry: ROA or ROE exceeds the company's historical average. From a financial perspective, the equation for each anomaly can be seen in Equation 1: x is the return on assets and equity, i is the depreciation value using the current value, and $i-1$ is the previous value.

$$Sudden\_drop_i = \begin{cases} \textbf{\textit{Anomaly}}, if\ \dfrac{x_i - x_{i-1}}{x_{i-1}} < -0.2 \\ \textbf{\textit{Normal}}, if\ \dfrac{x_i - x_{i-1}}{x_{i-1}} \geq -0.2 \end{cases}$$

$$Deviataion\_industry_i = \begin{cases} \textbf{\textit{Anomaly}}, if\ \dfrac{Sudden\_drop_i - \mu_{Sudden\_drop}}{\sigma_{Sudden\_drop}} > 2.0 \\ \textbf{\textit{Normal}}, \quad if\ \dfrac{Sudden\_drop_i - \mu_{Sudden\_drop}}{\sigma_{Sudden\_drop}} \leq 2.0 \end{cases} \quad (1)$$

$$\text{Unnatural tren}_i = \begin{cases} \textbf{\textit{Anomaly}}, if\ \left|\dfrac{x_i - \mu_x}{\sigma_x}\right| > 2.5 \\ \textbf{\textit{Normal}}, if\ \left|\dfrac{x_i - \mu_x}{\sigma_x}\right| \leq 2.5 \end{cases}$$

$$\text{Incompatibility industry}_i = \begin{cases} \textbf{\textit{Anomaly}}, if\ x_i > \mu_x \\ \textbf{\textit{Normal}}, if\ x_i \leq \mu_x \end{cases}$$

An anomaly from the customer perspective was identified as a *customer issue* when customer revenue fell below the historical average of financial and other comprehensive income. The equation for each anomaly from a customer perspective can be seen in Equation 2.

$$\text{Customer issue}_i = \begin{cases} \textbf{\textit{Anomaly}}, if\ x_i < \mu_x \\ \textbf{\textit{Normal}}, if\ x_i \geq \mu_x \end{cases} \quad (2)$$

$$Report\ quality_i = \begin{cases} \textbf{\textit{Anomaly}}, if\ \sum_{j=1}^{n} (x_{ij} = 0) > \mu_z \\ \textbf{\textit{Normal}}, if\ \sum_{j=1}^{n} \| (x_{ij} = 0) \leq \mu_z \end{cases} \quad (3)$$

$$\text{Number manipulation}_i = \begin{cases} \textbf{\textit{Anomaly}}, if\ x_{1i} > x_{1i} \mid x_{3i} < 0 \\ \textbf{\textit{Normal}}, if\ x_{1i} < x_{1i} \mid x_{3i} > 0 \end{cases}$$

$$\text{Significant discrepancies}_i = \begin{cases} \textbf{\textit{Anomaly}}, if \ \left|\frac{x_i - x_{i-1}}{x_{i-1}}\right| > 0.5 \\ \textbf{\textit{Normal}}, if \ \left|\frac{x_i - x_{i-1}}{x_{i-1}}\right| \leq 0.5 \end{cases}$$

Three anomalous conditions were analysed from the internal business perspective, including Report Quality: the number of variables not filled in the financial statements exceeds the average threshold. Number Manipulation: operating expenses exceed revenue, or net income is negative. Significant Discrepancies: a change of more than 50% in total revenue, operating expenses, or net income compared to the previous period. The equation can be seen in Equation 3.

The indicators analysed in the learning and growth perspective are cash, fixed, and current assets. External Inspection Anomaly is determined if the indicator value is below the industry or company average (Akinbowale et al., 2023). The equation for each anomaly, from a learning-and-growth perspective, is shown in Equation 4.

$$\text{External Inspection}_i = \begin{cases} \textbf{\textit{Anomaly}}, if \ x_i < \mu_x \\ \textbf{\textit{Normal}}, if \ x_i \geq \mu_x \end{cases} \quad (4)$$

The labelling process was carried out after determining the condition status for each perspective, and the perspective value was calculated using Equation 5.

$$\text{f, c, i, lg} = 100 \times \frac{\text{normal condition}}{\text{number of perspective conditions}} \quad (5)$$

Next, the BSC value was determined using the weight of each perspective in Equation 6.

$$bsc_{value} = (f \times 0.15) + (c \times 0.35) + (i \times 0.20) + (lg \times 0.30) \quad (6)$$

Finally, the label was assigned based on the BSC value using the following rules from Equation 7.

$$bsc_{label} = \begin{cases} \textbf{\textit{Fraud}}, if \ bsc\_value < 0.05 \\ \textbf{\textit{Risky}}, if \ bsc\_value < 0.3 \\ \textbf{\textit{Alarm}}, if \ bsc\_value < 0.5 \\ \textbf{\textit{Normal}}, if \ bsc\_value \geq 0.5 \end{cases} \quad (7)$$

Pre-processing Dataset

Data pre-processing is a crucial step in the analysis process, as it helps prevent errors and ensures consistency with the model. In this study, the financial data were collected longitudinally and across entities. Consequently, pre-processing involved applying entity and time panel data indices to ensure each data point had a unique index. The cleaning stage involved selecting features with no more than a certain number of missing values to avoid bias. In the next stage, all financial data was standardised to a numerical format to align units, then normalised using the min-max method to equalise the scale across features.

Variance Inflation Factor

Features that have passed normalisation were selected to improve predictions using the variance inflation factor (VIF), which detects multicollinearity between features. Features with a VIF exceeding the specified threshold may indicate high multicollinearity, which can compromise the model (Seireg et al., 2022).

Oversampling Model

This study's main approach to handling data imbalance is based on the WGANGP method. The WGANGP method can address common issues in GAN models, including mode collapse and gradient problems that often arise with imbalanced data.

The WGANGP in this study was developed with higher thresholds for the gradient range, generator output, and discriminator. The configurations are detailed in Tables 1 (Generator), 2 (Discriminator), 3 (Loss Function), and 4 (Gradient). This adjustment aims to ensure that the resulting synthetic data accurately reflects the original data's characteristics and to improve the model's classification performance in the next stage. The WGANGP model threshold adjustment aims to reduce the dominance of penalties between original and synthetic samples, thereby making the training process more balanced. The generator and discriminator sections are modified by adding additional conditions that adjust to the panel data, namely, the issuer entity condition and the quarter's time. The threshold was set to 0.1-0.9, based on the gradient range, to address the shortcomings of the WGAN model's clip-weighting, so the penalty on the gradient can be measured and scaled as needed.

Table 1

WGANGP Generator

Layer	Input shape/value
Input layer	(n_features,)
Entity Embedding	(n_entities, n_features)
Time Embedding	(n_times, n_features)
Dense Layer 01	(n_times * n_features * 4, relu)
LSTM	(64)
Lambda layer	(activation_threshold(x), threshold)
Time Distributed > Dense Layer	(n_features,)
Dense Layer 02	(32, relu)
Lambda layer	((x: r0 + r1 – r0) *(x +1)/2)

Table 2

WGANGP Discriminator

Layer	Input shape/value
Input layer	(n_features,)
Entity Embedding	(n_entities, n_features)
Time Embedding	(n_times, n_features)
LSTM	-64
Dense Layer 03	(32, relu)
Dense layer	1

Table 3

WGANGP Loss Threshold

Wasserstein loss with threshold	
Discriminator loss real	mean (cv (real, 0, 1))
Discriminator loss fake	mean (cv (fake, 1, 0))

Table 4

WGANGP Gradient Configuration

Components	Explanation
Input	r_samples, f_samples, n_company, n_quarter, threshold
Thresholding	lowest=threshold, highest=1.0
Alpha	Penalty Calculation
Output	Gradient Penalty

The first modification introduces Equation 8, which introduces a threshold-referenced discriminator loss by truncating the discriminator output predictions for real and fake samples within a controlled threshold range with a value of τ. This loss function is written as Equation 8. With this adjustment, discriminator models are less likely to exhibit overconfidence, helping stabilise training and preventing gradient bias.

$$L_D^{threshold} = -(\mathbb{E}_{\varphi \sim P_{real}}[clip(D(\varphi), 1 - \tau, 1)] + \mathbb{E}_{\hat{\varphi} \sim P_{fake}}[clip(D(\hat{\varphi}), -1, -1 + \tau)] \quad (8)$$

Equation 9 fills in the essential part of the gradient penalty term by imposing a threshold on the gradient norm. This mechanism ensures that the discriminator gradient remains within the desired range while maintaining greater flexibility than standard WGANGP.

$$L_D^{threshold} = \mathbb{E}_{\hat{\varphi} \sim P_{interp}}[(\max(\| \nabla_{\hat{\varphi}} D(\hat{\varphi}) \|_2 - 1,0)^2 + \max(\tau - \| \nabla_{\hat{\varphi}} D(\hat{\varphi}) \|_2, 0)^2)] \quad (9)$$

The modification aligns with Equation 10, yielding a loss that provides robust regularisation and adapts to complex financial data patterns.

$$L_{WGAN-GP} = L_D^{threshold} + \lambda_{GP} - L_{GP}^{threshold} \quad (10)$$

Modifications were also applied to the generator structure. The generator is adjusted to accept entity (company) and time (quarter) information through embedding. In addition, threshold activation ensures that the output is more controlled, forwarding only important values. The output is limited to a specific range of values using a final scale-based activation function, as seen in Equation 11.

$$\varphi_{scaled} = a + (b - a).\frac{\varphi + 1}{2} \quad (11)$$

With this approach, the model not only focuses on improving the quality of synthetic data but also corrects class imbalances that can affect the accuracy of the fraud detection model. The results of this approach will be compared with conventional oversampling methods in the next chapter to demonstrate the advantages of the proposed generative approach.

Model for Classification

This study used conventional inferential statistics, ML, and DL classifiers and regressors to classify and predict from processed financial statement data. It aimed to capture more data patterns to increase the possibility of identifying anomalies or fraud. In addition, the problem of class imbalance in the data was addressed using a generative-based oversampling method, which can improve the predictive performance of the classification models used.

Traditional Method

The statistical techniques used in this study were Logit and Probit. Logit is based on logistic regression, connecting independent variables to probabilities for each feature, with the error distribution as the main measure in this model. Logit assesses how much change in independent variables increases or decreases the probability of the desired category. On the other hand, Probit uses the cumulative distribution function (CDF) to estimate probabilities, assuming that errors follow a normal distribution.

ML Method

ML methods are applied to learn irregular patterns and dependencies that are identified through statistical analysis. The models used are XGBoost and Random Forest. XGBoost has been chosen for its ability to iteratively improve and minimise errors when working with large, complex data. Random Forest is a type of ensemble learning that uses a combination of decision trees to minimise overfitting. The ML model will be applied with several parameters set equally across models, including 4 classes, a random state of 42, 10 estimators, a learning rate of 0.01, and the Gini criterion.

DL Method

The DL models used in this study are LSTM and CNN-LSTM. Building on RNN, a previous study developed an LSTM to overcome the problem of long-term dependencies in time series data (Sunaryono et al., 2023). The proposed CNN-LSTM was implemented to combine the strengths of CNNs in extracting local features with LSTM, leveraging temporal correlations and temporal patterns. The settings for each DL model are identical across all layers, with parameters including a seed of 42, an epoch of 1000, and a batch size of 32. The model utilised the Adam optimiser with a learning rate of 0.0001, the sparse categorical cross-entropy loss function, and the LeakyReLU activation function. The DL model settings are shown in Table 5.

Table 5

DL Layer Configuration

Layer	Input shape/value
Input layer	(n_features,)
Dense Layer	(64,)
Batch Normalization	
Dropout	(0.3)
Dense Layer	(128,)
Batch Normalization	
Dropout	(0.3)
Dense Layer	(256,)
Batch Normalization	
Dropout	(0.3)
Layer Output	(n_features,)

Evaluation

The research evaluated the effectiveness of the oversampling method and the model's classification performance. To measure the efficacy of synthetic data, the study employed Euclidean Distance (ED) and Wasserstein distance (WD) to assess the similarity between the original data and the synthetic distribution patterns and feature characteristics. The assessment model used accuracy, precision, recall, and F1-score from the error matrix to measure correct classification, including fraud detection.

Experiment Setup

The experiments in this study were designed to evaluate the effectiveness of various oversampling methods in improving classification performance for minority cases, particularly fraud indicators in financial statements. Two main approaches were applied in the oversampling process. First, the experiment was conducted with general oversampling based on the distribution of labels across the entire dataset, without oversampling per entity per label. This method was intended to obtain samples as close as possible to each class in aggregate. Testing of oversampling results was carried out by measuring the similarity between the original and synthetic distributions using ED and WD, and by evaluating the classification model using accuracy and F1-score.

The second experiment employed a more structured approach, oversampling per label for each entity to accommodate the panel data structure. In implementing this approach, the oversampling process was applied exclusively to each entity, ensuring each entity later has a balanced label distribution. Evaluation of oversampling data was also carried out using ED and WD metrics, and classification performance was assessed to evaluate the effectiveness of this approach in detecting minority classes more accurately and contextually.

The general settings in the first test were different from those in the second experiment. The first test settings were determined by evaluating several parameters on one of the oversampling models to obtain good parameter values for the other models. The parameter settings obtained were a batch size of 64 and 1000 epochs per model. However, due to the complexity of the second experiment, the batch size was reduced to 16 and the number of epochs per model to 10 to enable the device to run the test. These two tests aimed to compare classification accuracy with the oversampling strategy, especially in detecting financial fraud, which is often a minority class.

RESULTS AND DISCUSSIONS

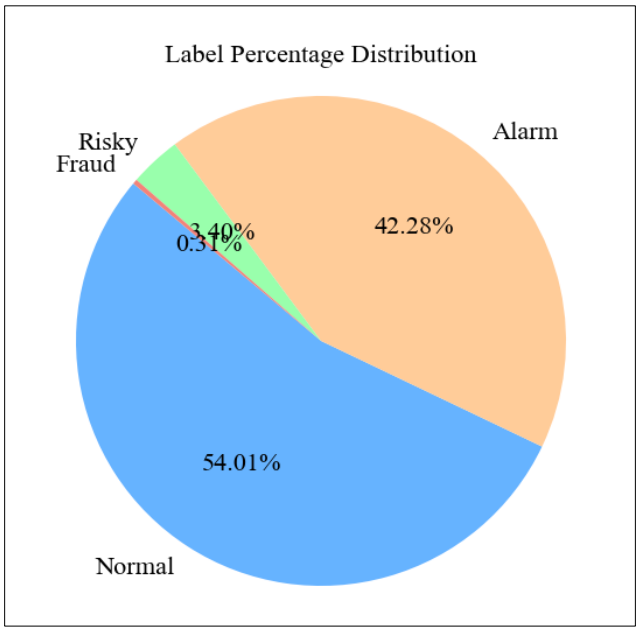
The results section is divided into several sections that explain the dataset, labelling, oversampling result evaluation, and fraud-identifying results.

Dataset Distribution

The dataset used in this study was labelled using the BSC approach, resulting in four label categories: *normal*, *alarm*, *risky*, and *fraud*. The distribution of each label shows a significant class imbalance. As shown in Figure 2, the normal label dominates, with more than 2,800 data points, followed by the alarm label, with about 2,200. Meanwhile, the risky label has a tiny distribution, with less than 200 data points, and the fraud label is almost invisible in the visualisation because the number is extremely low. The distribution of each label shows a significant class imbalance. The percentages for each label are shown in Figure 2.

Figure 2

Label Presentation Distribution



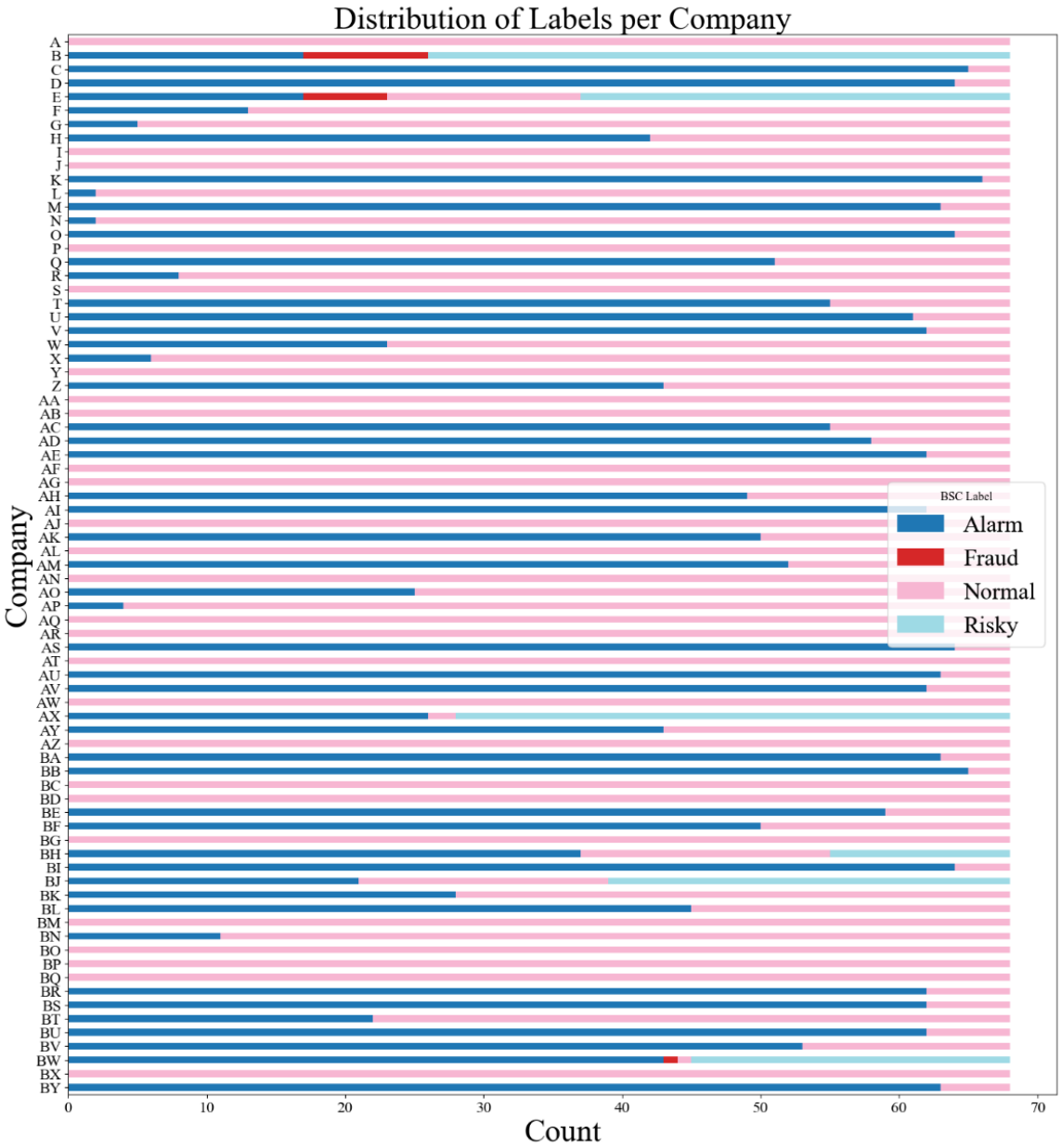
The labels used in this study were arranged based on their priority order. Data labelled as fraudulent also meets the criteria for the risky and alarm labels. Data labelled as risky also meets the criteria for the alarm label, while data labelled as alarm stands alone because it does not meet the criteria for the normal label or other labels. This labelling system reflects progressive risk escalation and allows ML models to understand that class transitions are gradual rather than mutually exclusive. This approach is relevant to the concept of ordinal classification and offers advantages in risk modelling by accounting for the semantic relationships among label distributions.

The distribution of these labels was also uneven between companies, as shown in Figure 3. In the graph, each row represents one company, and the colour indicates the number of each label. Many companies have a dominant share of normal labels, while fraud labels appear in only a few companies and in tiny numbers. Some companies exhibit better label diversity, yet still demonstrate the dominance of the

majority class. This imbalance affects the overall data distribution and the variation between entities and time, which is one of the challenges in panel data classification. Therefore, the analysis and experiments in the next stage will focus on oversampling strategies to address this imbalance, both at the aggregate and per-entity levels, to ensure the model learns equally from all classes.

Figure 3

Label Distribution for Each Entity



Oversampling Result

This study evaluated the oversampling results using several comparison models, namely SMOTE, GAN, WGAN, CGAN, RGAN, PACGAN, and WGANGP, with various threshold variations (0.1 to 0.9). The evaluation process was divided into two scenarios (i.e., Scenario 1 and Scenario 2). The first oversampling was generally performed per label, without considering the entity and time in the panel data. The second oversampling was then applied to each label and entity, accounting for the panel data structure. The performance evaluation was based on two main metrics, ED and WD, and on visualisation of the distributions using the t-Distributed Stochastic Neighbor Embedding (t-SNE) method to compare synthetic and real data. In Scenario 1, the results of ED and WD calculations are presented in Table 6, while the t-SNE visualisation is shown in Figure 4.

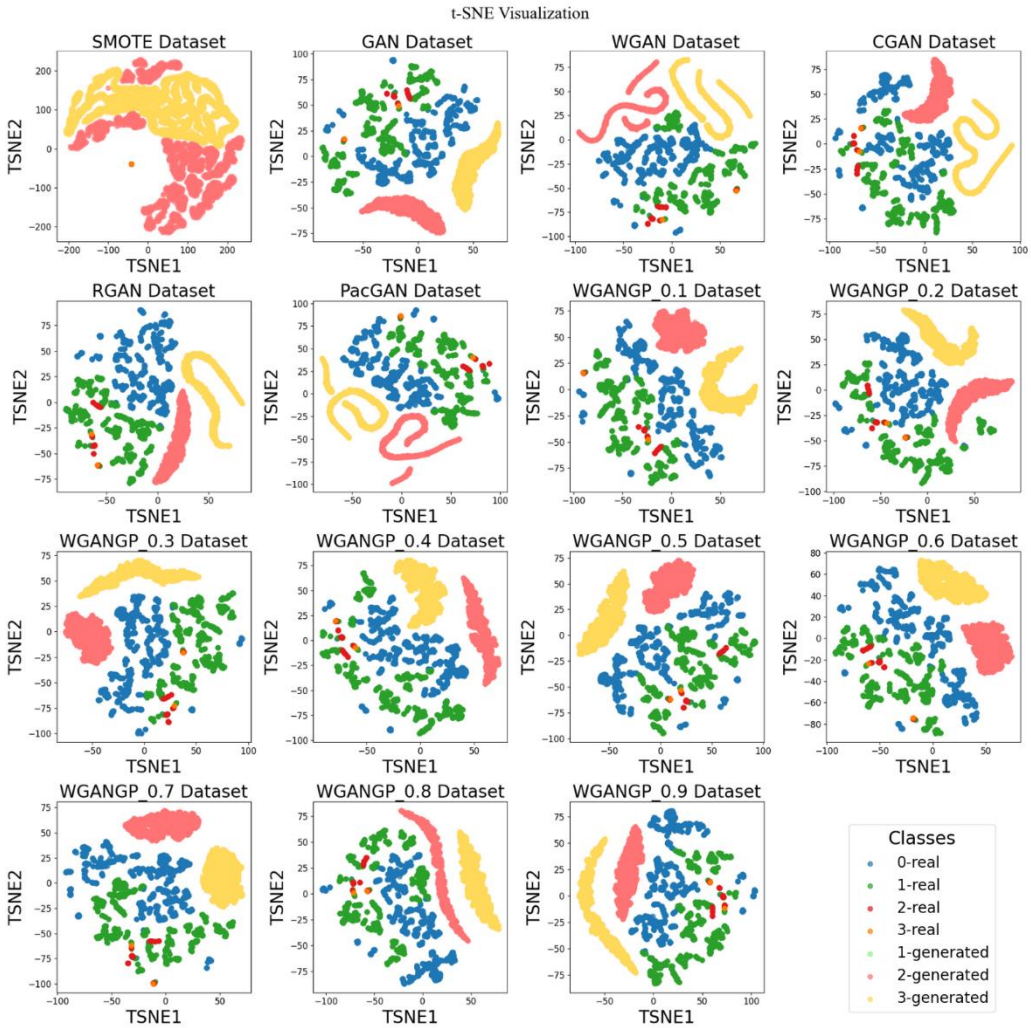
Table 6

Distribution Data Evaluation Scenario 1

Model	ED	WD
SMOTE	3.1581^{e+12}	3.8200^{e+14}
GAN	$\mathbf{1.8098^{e+01}}$	$\mathbf{1.8300^{e+03}}$
WGAN	2.3375^{e+02}	2.3400^{e+04}
CGAN	1.8012^{e+02}	2.1400^{e+04}
RGAN	2.4269^{e+01}	2.9100^{e+03}
PACGAN	1.5667^{e+02}	1.5700^{e+04}
WGANGP 0.1	2.9781^{e+01}	3.2400^{e+03}
WGANGP 0.2	2.9828^{e+01}	3.2100^{e+03}
WGANGP 0.3	4.0013^{e+01}	4.0600^{e+03}
WGANGP 0.4	5.2252^{e+01}	6.1200^{e+03}
WGANGP 0.5	4.7372^{e+01}	4.8700^{e+03}
WGANGP 0.6	3.8757^{e+01}	4.0500^{e+03}
WGANGP 0.7	3.5982^{e+01}	3.7000^{e+03}
WGANGP 0.8	4.4062^{e+01}	4.8600^{e+03}
WGANGP 0.9	3.7715^{e+01}	4.2600^{e+03}

Figure 4

Data Distribution Scenario 1



The results obtained are SMOTE, which shows an ED value of 3.1581^{e+12} and a WD of $1,8136^{e+02}$, much larger than other methods, indicating poor synthetic data quality. WGAN, CGAN, RGAN, and PacGAN yield smaller ED and WD values than SMOTE, but they are still relatively large. As shown in Table 6, WGAN achieves an ED value of 2.3375^{e+02} and a WD value of 2.3400^{e+04}, which are better than those of SMOTE. However, GAN shows an ED value of 1.8098^{e+01} and a WD value of 1.8300^{e+03}, which are the best evaluation results compared to other models. WGANGP at each threshold shows varying results, ranging from poor performance at the 0.4 threshold with ED 5.2252^{e+01} and WD $6,1200^{e+03}$ to satisfactory performance at the 0.1 threshold with ED 2.9781^{e+01} and WD $3,2400^{e+03}$.

In Figure 4, t-SNE visualisation reveals that SMOTE's data points are significantly separated from the actual data, resulting in unnatural clusters. GAN, WGAN, and CGAN exhibit clustering, but there is still uneven overlap. WGANGP, especially at thresholds of 0.1–0.5, produces a synthetic data distribution that closely resembles the real data distribution, as indicated by greater overlap and more natural clusters. Thus, WGANGP at thresholds of 0.2 to 0.5 is the best model candidate in Scenario 1 based on the combination of ED, WD, and t-SNE visualisation.

In Scenario 2, the oversampling approach considers more complex oversampling for each label and entity. The results of the distribution data evaluation are presented in the ED and WD evaluations, as shown in Table 7. Compared to Scenario 1, all models in Scenario 2 exhibit more stable performance than those in Scenario 1. Based on the evaluation results, the SMOTE model quantitatively has the closest distribution to the original data, with an ED value of 2.1333^{e+00} and a WD of $1,8136^{e+02}$. Followed by standard generative models such as GAN, WGAN, and RGAN, with a slight difference. Meanwhile, WGANGP shows a decrease in performance as the penalty threshold increases. From threshold 0.1 to 0.9, the ED and WD values show minimal increases, with the highest ED at 2.3970^{e+00} and the highest WD at 2.0875^{e+02} at threshold 0.1, indicating that the data distribution deviates most from the original data.

Table 7

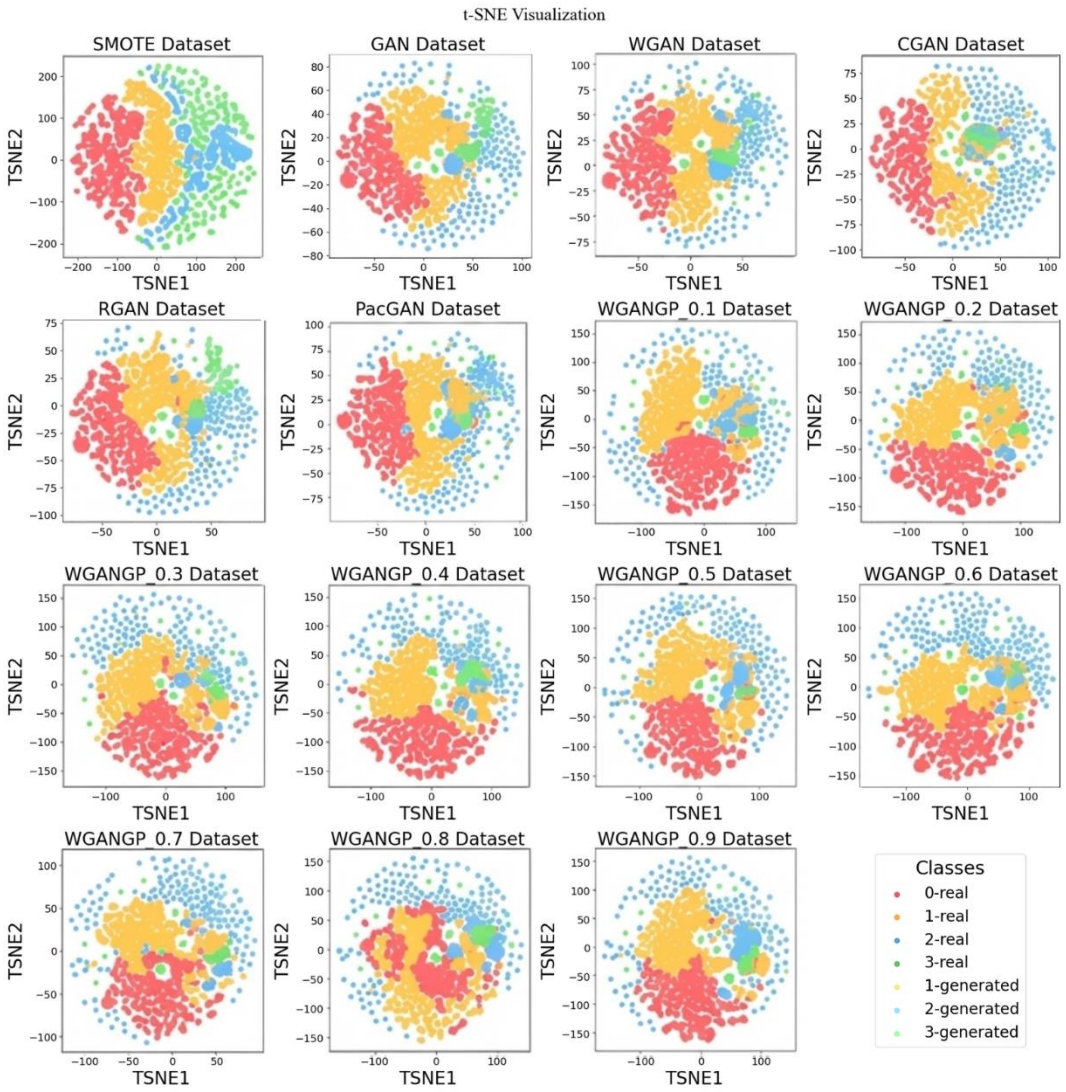
Distribution Data Evaluation Scenario 2

Model	ED	WD
SMOTE	**2.1333^{e+00}**	**1.8136^{e+02}**
GAN	2.1938^{e+00}	1.8785^{e+02}
WGAN	2.5855^{e+00}	2.5457^{e+02}
CGAN	2.4823^{e+00}	2.2427^{e+02}
RGAN	2.4344^{e+00}	2.1537^{e+02}
PACGAN	2.9995^{e+00}	3.2249^{e+02}
WGANGP 0.1	2.3970^{e+00}	2.0875^{e+02}
WGANGP 0.2	2.4005^{e+00}	2.1012^{e+02}
WGANGP 0.3	2.4005^{e+00}	2.1012^{e+02}
WGANGP 0.4	2.4005^{e+00}	2.1012^{e+02}
WGANGP 0.5	2.4006^{e+00}	2.1012^{e+02}
WGANGP 0.6	2.4006^{e+00}	2.1012^{e+02}
WGANGP 0.7	2.4005^{e+00}	2.1012^{e+02}
WGANGP 0.8	2.4005^{e+00}	2.1012^{e+02}
WGANGP 0.9	2.4005^{e+00}	2.1012^{e+02}

The t-SNE visualisation of the data distribution is shown in Figure 5. Each plot illustrates the distributions of original and synthetic samples, demonstrating that SMOTE, GAN, and WGAN exhibit clear class distinctions and effective integration between original and synthetic samples. The clusters look quite well-maintained, and the distribution is balanced. CGAN shows less stable distributional evaluation results, with the distribution spreading farther from the centre of the original data cluster. RGAN and PacGAN are more stable than CGAN but still perform worse than SMOTE and GAN. The WGAN with threshold settings shows that higher threshold values correspond to more unstructured synthetic distributions. Clusters overlap and are difficult to distinguish visually. Regarding visualisation, the WGANGP is less successful in maintaining the original distribution structure, especially at high penalty, which aligns with the increasingly deteriorating ED and WD values.

Figure 5

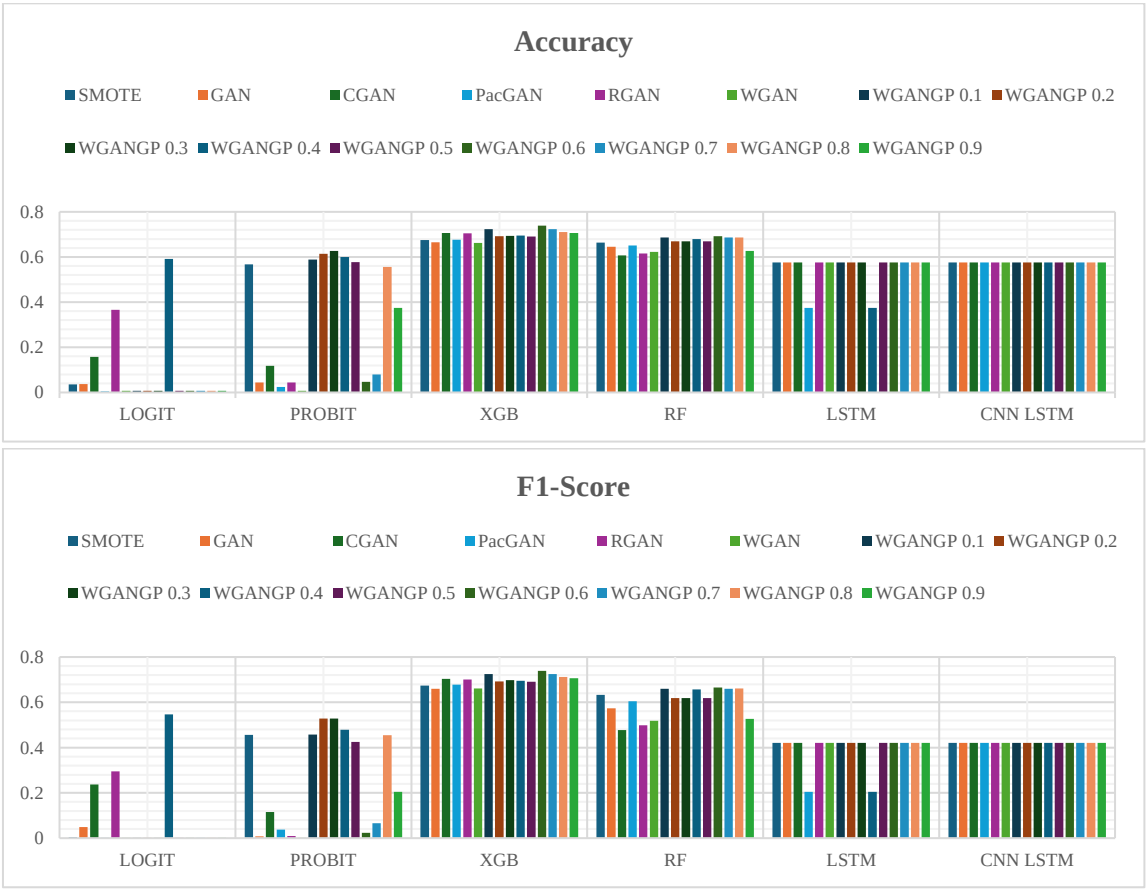
Data Distribution Scenario 2



Regarding evaluation, GAN appears superior in Scenario 1, but in Scenario 2, it is not superior to SMOTE. Meanwhile, WGANGP, with a threshold in evaluation, does not show superior performance in every scenario; however, comparing WGANGP with different thresholds produces varying evaluation values. Although WGANGP theoretically promises higher training stability, it performs poorly in both scenarios. An overly strong gradient penalty may distort the loss function, hindering the generator's learning.

Figure 6

Classification Result Scenario 1



Classification Results

This section presents the classification results of various ML and DL models after oversampling unbalanced data. This study employs the GAN approach, specifically WGANGP, with various threshold values. Other methods include SMOTE, GAN, WGAN, CGAN, RGAN, and PacGAN. The comparison value is the average classification result for each model across the oversampling methods, to assess the variability of the synthetic data when tested on the classification model.

In Scenario 1, the oversampling process was performed without accounting for the panel data structure. The accuracy and F1-score metrics shown in Figure 6 indicate that the XGBoost model outperformed other classification methods. In contrast, DL methods such as LSTM and CNN-LSTM remained stable at 0.56-0.60. However, it did not demonstrate superiority over ML methods, such as XGBoost and random forests, which achieved average scores above 0.60 across all proposed oversampling methods. On the other hand, statistical methods such as Logit and Probit did not demonstrate superiority or stability. Specifically, the Logit test showed a significant imbalance in WGANGP 0.4 compared to the others, and Probit was also imbalanced due to WGANGP oversampling.

In general, the classification results of Scenario 1 in this scenario show that the WGANGP approach with the use of thresholds between 0.4 and 0.6 has a significant impact on the average value in each classification model presented in Table 8 for accuracy which is outperformed by WGANGP 0.4 with an average of 5.8606^{e-01} and Table 9 for F1-score which is outperformed by WGANGP 0.4 with an

average of 5.0031^{e-01} and is still superior to the conventional GAN model. A comparison of the evaluation results of the data distribution between the thresholds applied to the WGANGP model shows a significant impact on the model's ability to generate a more representative synthetic data distribution. However, the performance of some models still varies across classification algorithms.

Table 8

Classification Accuracy Result Scenario 1

Model		Max	Min	Average
SMOTE	Accuracy	0.67493	0.03568	5.1553^{e-01}
GAN	Accuracy	0.66502	0.03667	4.2385^{e-01}
CGAN	Accuracy	0.70664	0.11794	4.5672^{e-01}
PacGAN	Accuracy	0.67691	0.00496	3.8470^{e-01}
RGAN	Accuracy	0.70466	0.04460	4.8051^{e-01}
WGAN	Accuracy	0.66204	0.00793	4.0866^{e-01}
WGANGP 0.1	Accuracy	0.72349	0.00793	5.2626^{e-01}
WGANGP 0.2	Accuracy	0.69277	0.00793	5.2263^{e-01}
WGANGP 0.3	Accuracy	0.69376	0.00793	5.2494^{e-01}
WGANGP 0.4	Accuracy	0.69475	**0.37463**	**5.8606^{e-01}**
WGANGP 0.5	Accuracy	0.69078	0.00793	5.1635^{e-01}
WGANGP 0.6	Accuracy	**0.73836**	0.00793	4.3938^{e-01}
WGANGP 0.7	Accuracy	0.72349	0.00793	4.4136^{e-01}
WGANGP 0.8	Accuracy	0.71061	0.00793	5.1883^{e-01}
WGANGP 0.9	Accuracy	0.70565	0.00793	4.7770^{e-01}

Table 9

Classification F1-Score Result Scenario 1

Model		Max	Min	Average
SMOTE	F1-score	0.67332	0.00347	4.3455^{e-01}
GAN	F1-score	0.65903	0.00731	3.5496^{e-01}
CGAN	F1-score	0.70332	0.11592	3.9587^{e-01}
PacGAN	F1-score	0.67751	0.00042	3.2426^{e-01}
RGAN	F1-score	0.70098	0.00942	3.9092^{e-01}
WGAN	F1-score	0.66158	0.00013	3.3707^{e-01}
WGANGP 0.1	F1-score	0.72512	0.00013	4.4731^{e-01}
WGANGP 0.2	F1-score	0.69259	0.00013	4.4693^{e-01}
WGANGP 0.3	F1-score	0.69853	0.00013	4.4778^{e-01}
WGANGP 0.4	F1-score	0.69523	**0.20420**	**5.0031^{e-01}**
WGANGP 0.5	F1-score	0.69117	0.00013	4.2944^{e-01}
WGANGP 0.6	F1-score	**0.73890**	0.00013	3.7818^{e-01}
WGANGP 0.7	F1-score	0.72512	0.00013	3.8205^{e-01}
WGANGP 0.8	F1-score	0.71127	0.00013	4.4478^{e-01}
WGANGP 0.9	F1-score	0.70617	0.00013	3.7982^{e-01}

Scenario 2 adopted an entity-based and time-based oversampling approach, which is better suited to panel data characteristics. The classification accuracy and F1-score results for this scenario, as shown in Figure 7, are similar to those of Scenario 1. Specifically, XGBoost and Random Forest outperform other models, offering a superior range of values compared to Scenario 1: 0.60-0.70. This improvement highlights the importance of maintaining the entity-time structure in synthetic data generation to preserve temporal and spatial patterns in panel data. DL in Scenario 2 is not as superior or stable as in Scenario 1, as LSTM and CNN-LSTM experience a significant decline in CGAN and RGAN performance. In

Table 10 and Table 11, SMOTE shows the best average performance, achieving an accuracy of 6.4255^{e-01}, but differs from the average accuracy results on the WGANGP F1-score metric, with a threshold of 0.8, showing the best average performance with an F1-score of 6.2880^{e-01}. The results in Scenario 2 show an improvement, reflecting the importance of maintaining the entity-time structure in the synthetic data generation process to preserve temporal and spatial patterns in panel data.

Figure 7

Classification Result Scenario 2

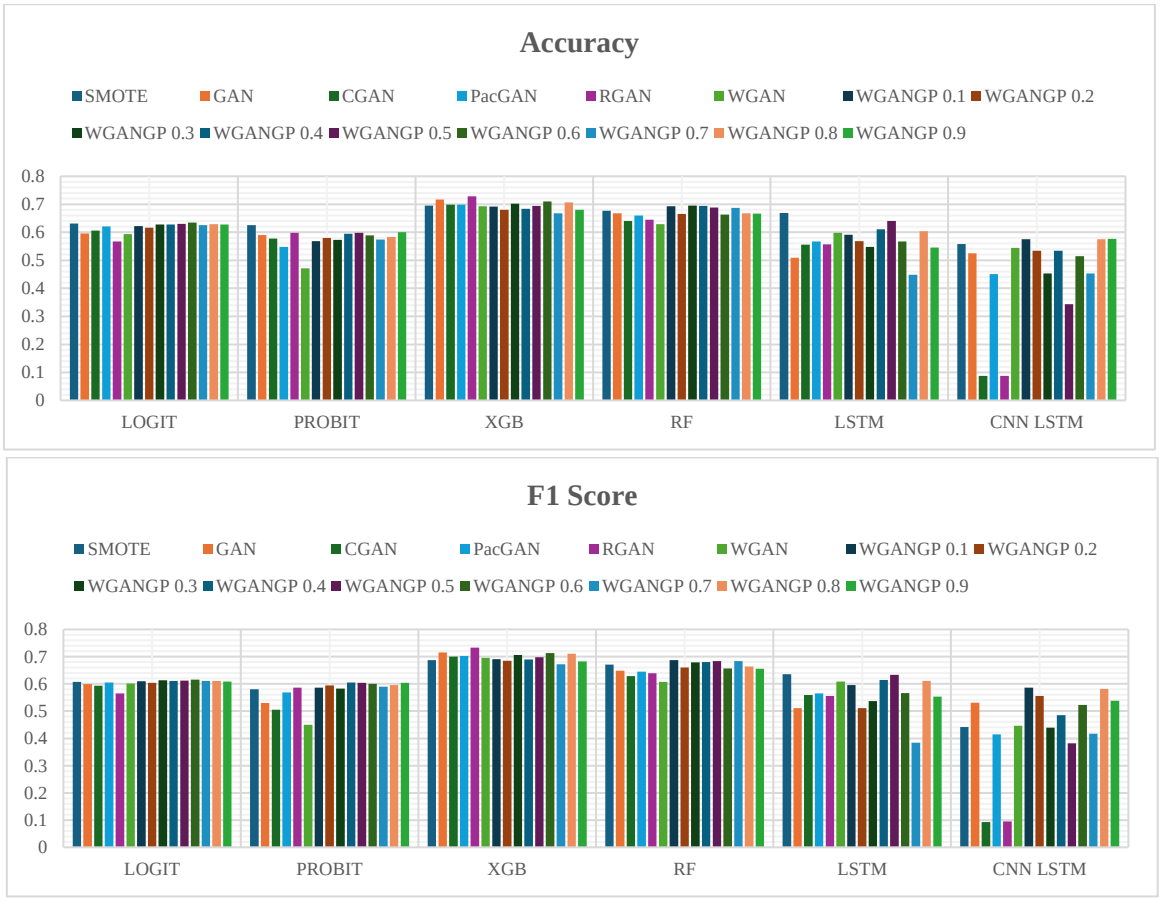


Table 10

Classification Accuracy Result Scenario 2

Model		Max	Min	Average
SMOTE	Accuracy	0.69475	0.55798	**6.4255^{e-01}**
GAN	Accuracy	0.71655	0.50942	6.0093^{e-01}
CGAN	Accuracy	0.69871	0.08722	5.2775^{e-01}
PacGAN	Accuracy	0.69871	0.45094	5.9118^{e-01}
RGAN	Accuracy	**0.72844**	0.08722	5.3056^{e-01}
WGAN	Accuracy	0.69277	0.47175	5.8821^{e-01}
WGANGP 0.1	Accuracy	0.69277	0.56888	6.2389^{e-01}
WGANGP 0.2	Accuracy	0.68087	0.53419	6.0753^{e-01}
WGANGP 0.3	Accuracy	0.70169	0.45292	5.9977^{e-01}
WGANGP 0.4	Accuracy	0.69376	0.53419	6.2405^{e-01}
WGANGP 0.5	Accuracy	0.69376	0.34291	5.9911^{e-01}
WGANGP 0.6	Accuracy	0.70961	0.51536	6.1298^{e-01}
WGANGP 0.7	Accuracy	0.68781	0.44896	5.7615^{e-01}
WGANGP 0.8	Accuracy	0.70664	**0.57582**	6.2785^{e-01}
WGANGP 0.9	Accuracy	0.67988	0.54609	6.1645^{e-01}

Table 11

Classification F1-Score Result Scenario 2

Model		Max	Min	Average
SMOTE	F1-score	0.68763	0.44133	6.0377^{e-01}
GAN	F1-score	0.71506	0.51058	5.8903^{e-01}
CGAN	F1-score	0.69974	0.09238	5.1315^{e-01}
PacGAN	F1-score	0.70208	0.41422	5.8337^{e-01}
RGAN	F1-score	**0.73261**	0.09449	5.2884^{e-01}
WGAN	F1-score	0.69514	0.44642	5.6823^{e-01}
WGANGP 0.1	F1-score	0.69119	**0.58572**	6.2598^{e-01}
WGANGP 0.2	F1-score	0.68467	0.51113	6.0162^{e-01}
WGANGP 0.3	F1-score	0.70587	0.43897	5.9285^{e-01}
WGANGP 0.4	F1-score	0.68960	0.48575	6.1442^{e-01}
WGANGP 0.5	F1-score	0.69850	0.38137	6.0228^{e-01}
WGANGP 0.6	F1-score	0.71290	0.52330	6.1254^{e-01}
WGANGP 0.7	F1-score	0.68365	0.38428	5.5972^{e-01}
WGANGP 0.8	F1-score	0.71100	0.58102	**6.2880^{e-01}**
WGANGP 0.9	F1-score	0.68209	0.53810	6.0701^{e-01}

The evaluation of the distance between synthesised and real data in WGANGP shows poorer evaluation compared to GAN in Scenario 1 and compared to SMOTE in Scenario 2. However, WGANGP shows strong, stable performance in classification evaluations compared to other models. These classification results contrast with the oversampling evaluation results because the WGANGP model is good at handling overfitting, enabling the classification model to deliver more stable performance. The classification results also show that the oversampling process, which considers entities and time, produces a more realistic synthetic data distribution and improves the classification model's generalisation. The behavioural pattern of a particular entity over time becomes an important factor in the model for detecting anomalies or fraud in financial data reports.

Overall, there are differences in performance between the two scenarios, consisting of oversampling and classification tests, which have different results. In Scenario 1, the GAN outperformed other models; however, in the actual test with GAN classification, its performance was subpar. On the other hand, WGANGP with a small threshold value showed superior performance. Based on the oversampling evaluation results, WGANGP 0.4 is not as far from the best unit as SMOTE, which is the worst unit in Scenario 1.

On the other hand, in Scenario 2, oversampling testing was dominated by the traditional SMOTE model, and in average value testing, SMOTE also outperformed in terms of accuracy metrics. However, when examined in detail, SMOTE testing outperformed the majority of statistical and DL models, followed by WGANGP 0.8. In terms of the F1-score metric, WGANGP 0.8 excelled, with scenario 2 experiencing a 12-17% increase in accuracy and F1-score compared to Scenario 1.

Based on these findings, combining WGANGP with generalised oversampling and temporal entities is the most effective approach for managing unbalanced data in panel time-series classification. This method produces more stable classification results, with performance variance shrinking from 0.14 in Scenario 1 to 0.08 in Scenario 2, and is consistent with financial studies based on quarterly published public reports.

CONCLUSION AND FUTURE STUDIES

This study found that the WGANGP method produced better data. This method appears to work well even when there is little real fraud data to learn from, and it maintains the data panel's structure (entity and time). This oversampling enables the classification model to identify suspicious financial patterns better. WGANGP successfully tested with various methods of synthetic data generation. However, its performance depends on the settings and classification methods used.

However, this study has limitations, particularly in terms of the size of the dataset, which does not cover all sectors, the data labelling process, which uses a Balanced Scorecard framework that contains elements of subjectivity and has not been confirmed by auditor reports or official regulations, and the classification evaluation, which tends to be small. Therefore, future studies can conduct further validation with external data. Additionally, developing more objective labelling techniques is necessary to enhance the accuracy of future models, and this can be achieved through a hybrid model approach. This approach also allows for investigating WGANGP's sensitivity to the threshold, developing adaptive threshold methods, and implementing automatic hyperparameter optimisation.

ACKNOWLEDGMENT

This research is funded by the National Research and Innovation Agency (BRIN) under RIIM Kompetisi Gelombang 10 Program; the Indonesian Endowment Fund for Education (LPDP) on behalf of the Indonesian Ministry of Higher Education, Science and Technology and managed under the EQUITY Program (Contract No. 4299/B3/DT.03.08/2025 & No 3029/PKS/ITS/2025; and Institut Teknologi Sepuluh Nopember under Penelitian Kemitraan A.

REFERENCES

Abueid, R., Rehman, S., & Nguyen, N. T. (2022). The impact of balanced scorecard in estimating the performance of banks in Palestine. *EuroMed Journal of Business.* https://doi.org/10.1108/emjb-03-2021-0047

Aftabi, S. Z., Ahmadi, A., & Farzi, S. (2023). Fraud detection in financial statements using data mining and GAN models. *Expert Systems with Applications, 227.* https://doi.org/10.1016/j.eswa.2023.120144

Akinbowale, O. E., Mashigo, P., & Zerihun, M. F. (2023). *Application of balance scorecard as a strategic management and performance measurement tool for cyberfraud mitigation.* 7th International Conference on Business and Information Management, ICBIM 2023, 60–69. https://doi.org/10.1109/ICBIM59872.2023.10303219

Asokan, S., & Seelamantula, C. S. (2023). Euler-lagrange analysis of generative adversarial networks. *Journal of Machine Learning Research, 24.* http://jmlr.org/papers/v24/20-1390.html.

Baesens, B., Höppner, S., & Verdonck, T. (2021). Data engineering for fraud detection. *Decision Support Systems, 150.* https://doi.org/10.1016/j.dss.2021.113492

Chen, C., Shen, W., Yang, C., Fan, W., Liu, X., & Li, Y. (2023). A new safe-level enabled borderline-smote for condition recognition of imbalanced dataset. *IEEE Transactions on Instrumentation and Measurement, 72*. https://doi.org/10.1109/TIM.2023.3289545

Dina, A. B., Sarno, R., Anggraini, R. N. E., Haryono, A. T., & Septiyanto, A. F. (2024). *Comparison of oversampling techniques in prediction judicial decisions of divorce trials in family courts.* Proceeding - 2024 International Conference on Information Technology Research and Innovation, ICITRI 2024, 13–18. https://doi.org/10.1109/ICITRI62858.2024.10699016

Fadel, S., Rouaski, K., Challal, M., & Bouaicha, H. (2021). The balanced scorecard (BSC) as a multidimensional performance measurement system tool: Case the company of algeria post. *Financial Markets, Institutions and Risks, 5*(4). https://doi.org/10.21272/fmir.5(4).87-105.2021

Fernando, K. R. M., & Tsokos, C. P. (2022). Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems, 33*(7), 2940–2951. https://doi.org/10.1109/TNNLS.2020.3047335

Haryono, A. T., Sarno, R., Esti Anggraini, R. N., & Sungkono, K. R. (2024). *Imputation missing stock prices using generative adversarial networks and attention mechanism.* 2024 2nd International Conference on Technology Innovation and Its Applications (ICTIIA), 1–6. https://doi.org/10.1109/ICTIIA61827.2024.10761233

Hashemi, S. K., Mirtaheri, S. L., & Greco, S. (2023). Fraud detection in banking data by machine learning techniques. *IEEE Access, 11,* 3034–3043. https://doi.org/10.1109/ACCESS.2022.3232287

Hrishikesh, P. S., Puthussery, D., Akhil, K. A., & Jiji, C. V. (2023). *Relativistic GAN using receptive field block for single image super-resolution with improved perceptual quality.* 2023 11th International Symposium on Electronic Systems Devices and Computing, ESDC 2023. https://doi.org/10.1109/ESDC56251.2023.10149876

Hsin, Y. Y., Dai, T. S., Ti, Y. W., Huang, M. C., Chiang, T. H., & Liu, L. C. (2022). Feature engineering and resampling strategies for fund transfer fraud with limited transaction data and a time-inhomogeneous modi operandi. *IEEE Access, 10,* 86101–86116. https://doi.org/10.1109/ACCESS.2022.3199425

Kaewninprasert, K., Chai-Arayalert, S., & Yamaqupta, N. (2024). The perspective classification of balanced scorecard with ontology technique. *Journal of Information and Communication Technology, 23*(3), 465–494. https://doi.org/10.32890/jict2024.23.3.4

Kim, C., Park, S., & Hwang, H. J. (2022). Local stability of Wasserstein GANS with abstract gradient penalty. *IEEE Transactions on Neural Networks and Learning Systems, 33*(9), 4527–4537. https://doi.org/10.1109/TNNLS.2021.3057885

Li, B., Yen, J., & Wang, S. (2024). Uncovering financial statement fraud: A machine learning approach with key financial indicators and real-world applications. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3520249

Li et al. (2023). EID-GAN: Generative adversarial nets for extremely imbalanced data augmentation. *IEEE Transactions on Industrial Informatics, 19*(3), 3208–3218. https://doi.org/10.1109/TII.2022.3182781

Li, X., Da, X., Shi, W., & Liu, W. (2023, November 30). *Manufacturing companies financial fraud detection based on interpretable machine learning.* Proceedings of the 2nd International Conference on Public Management. https://doi.org/10.4108/eai.1-9-2023.2338768

Liao, C., & Dong, M. (2022). ACWGAN: An auxiliary classifier Wasserstein GAN-based oversampling approach for multi-class imbalanced learning. *International Journal of Innovative Computing, Information and Control, 18*(3), 703–721. https://doi.org/10.24507/ijicic.18.03.703

Man, C. K., Quddus, M., Theofilatos, A., Yu, R., & Imprialou, M. (2022). Wasserstein generative adversarial network to address the imbalanced data problem in real-time crash risk prediction. *IEEE Transactions on Intelligent Transportation Systems, 23*(12), 23002–23013. https://doi.org/10.1109/TITS.2022.3207798

Miftahushudur, T., Grieve, B., & Yin, H. (2024). Permuted KPCA and SMOTE to guide GAN-based oversampling for imbalanced HSI classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17*, 489–505. https://doi.org/10.1109/JSTARS.2023.3326963

Mohammed, R. M. H. (2022). The impact of audit committee on financial reporting quality. *Journal of Global Economics and Business, 3*(11), 81–94. https://doi.org/10.31039/jgeb.v3i11.91

Naderi, M., Nabizadeh, Z., Karimi, N., Shirani, S., & Samavi, S. (2021). *MSGDD-CGAN: Multi-scale gradients dual discriminator conditional generative adversarial network.* ArXiv:2109.05614

Permataning Tyas, S. M., Sarno, R., Haryono, A. T., & Rossa Sungkono, K. (2023). *A robustly optimised BERT using random oversampling for analysing imbalanced stock news sentiment data.* ICCoSITE 2023 - International Conference on Computer Science, Information Technology and Engineering: Digital Transformation Strategy in Facing the VUCA and TUNA Era, 897–902. https://doi.org/10.1109/ICCoSITE57641.2023.10127725

Qin, R. (2021). Identification of accounting fraud based on support vector machine and logistic regression model. *Complexity, 2021*. https://doi.org/10.1155/2021/5597060

Saksono, L. A., & Bernardus, D. (2023). Design of balanced scorecard as a school's performance measurement. *Binus Business Review, 14*(2), 171–183. https://doi.org/10.21512/bbr.v14i2.8901

Sarno, R., Dewandono, R. D., Ahmad, T., Naufal, M. F., & Sinaga, F. (2015). Hybrid association rule learning and process mining for fraud detection. *IAENG International Journal of Computer Science*.

Seireg, H. R., Omar, Y. M. K., El-Samie, F. E. A., El-Fishawy, A. S., & Elmahalawy, A. (2022). Ensemble machine learning techniques using computer simulation data for wild blueberry yield prediction. *IEEE Access, 10,* 64671–64687. https://doi.org/10.1109/ACCESS.2022.3181970

Shafqat, W., & Byun, Y. C. (2022). A hybrid GAN-based approach to solve imbalanced data problem in recommendation systems. *IEEE Access, 10,* 11036–11047. https://doi.org/10.1109/ACCESS.2022.3141776

Sihwail, R., Ghamri, M. Al, & Ibrahim, D. (2024). An enhanced model of whale optimization algorithm and k-nearest neighbors for malware detection. *International Journal of Intelligent Engineering and Systems, 17*(3), 606–621. https://doi.org/10.22266/ijies2024.0630.47

Sokolenko, L. F. (2022). Detection and evaluation of fraud during the public sector audit. *Statistics of Ukraine, 97*(2), 95–103. https://doi.org/10.31767/su.2(97)2022.02.10

Sunaryono et al. (2023). Optimised one-dimension convolutional neural network for seizure classification from EEG signal based on whale optimization algorithm. *International Journal of Intelligent Engineering and Systems, 16*(3), 310–322. https://doi.org/10.22266/ijies2023.0630.25

Widiantoro, A. D., Mustafid, M., & Sanjaya, R. (2024). Model analytic in fintech user comment features using LDA-CNN on imbalanced data. *International Journal of Intelligent Engineering and Systems, 17*(4), 1079–1098. https://doi.org/10.22266/IJIES2024.0831.80

Xiuguo, W., & Shengyong, D. (2022). An analysis on financial statement fraud detection for Chinese listed companies using deep learning. *IEEE Access, 10,* 22516–22532. https://doi.org/10.1109/ACCESS.2022.3153478

Yin et al. (2023). Imbalanced working states recognition of sucker rod well dynamometer cards based on data generation and diversity augmentation. *SPE Journal.* https://doi.org/10.2118/214661-pa

Zhou, H., Sun, G., Fu, S., Wang, L., Hu, J., & Gao, Y. (2021). Internet financial fraud detection based on a distributed big data approach with node2vec. *IEEE Access, 9,* 43378–43386. https://doi.org/10.1109/ACCESS.2021.3062467