Attention-based Conditional Random Field for Financial Fraud Detection

Xiaoguang Wang[1] , **Chenxu Wang**[*1,2,3,4] , **Luyue Zhang**[1] , **Xiaole Wang**[1] , **Mengqin Wang**[1] , **Huanlong Liu**[1] and **Tao Qin**[2]
[1]School of Software Engineering, Xi'an Jiaotong University
[2]MoE Key Lab of Intelligent Networks and Network Security (INNS), Xi'an Jiaotong University
[3]Interdisciplinary Research Center of Frontier science and technology, Xi'an Jiaotong University
[4]Shaanxi Joint (Key) Laboratory for Artificial Intelligence, Xi'an Jiaotong University
{cxwang, qin.tao}@mail.xjtu.edu.cn, {wangxg, zhangluyue, wmengqin, hlliu}@stu.xjtu.edu.cn, wangxiaole25@outlook.com

Abstract

Financial fraud detection is critical for market transparency and regulatory compliance. Existing methods often ignore the temporal patterns in financial data, which are essential for understanding dynamic financial behaviors and detecting fraud. Moreover, they also treat companies as independent entities, overlooking the valuable interrelationships. To address these issues, we propose ACRF-RNN, a Recurrent Neural Network (RNN) with Attention-based Conditional Random Field (CRF) for fraud detection. Specifically, we use an RNN with a sliding window to capture temporal dependencies from historical data, and an attention-based CRF feature transformer to model inter-company relationships. This transforms raw financial data into optimized features, fed into a multi-layer perceptron for classification. Besides, we also use the focal loss to alleviate the class imbalance problem caused by rare fraudulent cases. This work presents a real-world dataset to evaluate the performance of ACRF-RNN. Extensive experiments show that ACRF-RNN outperforms the state-of-the-art methods by 15.28% in *KS* and 4.04% in $Recall_m$. Data and code are available at: https://github.com/XNetLab/ACRF-RNN.git.

1 Introduction

Deceptions in financial statements such as inflating profits and understating liabilities can misguide investors and cause severe losses [Zhu *et al.*, 2021]. Thus, regulators mandate audits and public disclosure of annual financial statements. Financial statements comprise numerous accounting items that can be viewed as high-dimensional features. In such data, detecting financial statement fraud year by year is challenging because of the concealed nature of such fraud and the lack of labeled data. Recently, financial statement fraud detection has attracted much attention from both academics and industries. However, existing methods still face two drawbacks.

Drawback 1: Temporal correlations in financial data have long been neglected for fraud detection. Most existing methods rely on the financial statements in a single year to detect potential fraud. They overlook the strong temporal correlation inherent in financial data. The financial condition of a company tends to exhibit continuity and correlation in short periods, and many fraud cases have shown that ongoing financial distress often leads to financial fraud behaviors spanning multiple consecutive years [Bao *et al.*, 2020]. For instance, LeEco [Mao *et al.*, 2022a] committed consecutive fraud from 2016 to 2017, Luckin Coffee [Mao *et al.*, 2022b] from 2019 to 2020, and Enron [Bao *et al.*, 2020] from 1997 to 2001. Therefore, capturing temporal correlations in financial data aids in improving fraud detection.

Drawback 2: There is a lack of effective methods to capture behavioral homogeneity of fraudulent companies. Most existing methods treat companies as independent entities. Although a few studies have demonstrated that explicit relationships (e.g., related-party transactions and investments) can aid in financial fraud detection [Wang *et al.*, 2024; Wang *et al.*, 2025], the effect of implicit relationships on fraud detection remains under-explored. In practice, companies often exhibit behavioral homogeneity. For instance, fraudulent activities frequently involve similar deceptive tactics, resulting in similar behaviors. LeEco, Kangmei Pharmaceutical, and Luckin Coffee committed financial fraud by fabricating profits recorded in financial statements, while Hunan Chinasun Pharmaceutical Machinery, Hirisun and Xintai Electric artificially diminished their accounts receivable [Mao *et al.*, 2022a]. Implicit relationships can effectively model such behavioral homogeneity for fraud detection enhancement. Unfortunately, accurately and automatically capturing implicit relationships among companies presents a significant challenge. Furthermore, these real-world fraud cases also reveal diverse strategies, indicating that behavioral homogeneity varies across patterns. Consequently, relying on a single implicit relationship may fail to capture homogeneity under diverse patterns, underscoring the necessity of modeling multiple implicit relationships among companies.

To address these issues, this paper presents ACRF-RNN, a Recurrent Neural Network with Attention-based Conditional Random Field for financial fraud detection. To capture tem-

*Corresponding author.

poral patterns, ACRF-RNN treats a company's annual financial statements as multivariate time series data, and uses a recurrent neural network (RNN) to learn temporal correlations across years. It utilizes a sliding time window to extract short sub-sequences. The representation learned from each sub-sequence is then used as the temporal embedding for the most recent year within the window. Our model well captures the temporal correlations of financial statements.

To address the second drawback, we propose a novel multi-head attention-based Conditional Random Field (CRF) to capture complex implicit relationships among companies within a single year. Our method takes temporal embeddings from different years as input and employs the multi-head attention mechanism to capture implicit relationships under different patterns. The obtained similarity coefficients quantify the strength of these relationships. Subsequently, the CRF model recursively updates the company representations based on the learned similarity coefficients to ensure that companies with similar behaviors are embedded closer in the latent space. Finally, the company representations learned from all attention heads are aggregated through a pooling operation to produce the final representation, which is then fed into a multi-layer perceptron for classification.

Since fraudulent companies are far fewer than benign ones in practice, the model tends to predict test samples as benign, making it difficult to accurately identify fraud samples [Liu *et al.*, 2023]. To address this issue, we use the Focal Loss [Lin *et al.*, 2017] to optimize the model by assigning higher weights to fraudulent samples. The main contributions are as follows:

- We propose a novel Recurrent Neural Network with Attention-based Conditional Random Field to capture both temporal information and implicit relationships among companies for financial fraud detection.
- We propose a multi-head attention-based Conditional Random Field to model implicit relationships among companies, which captures various common patterns of fraudulent behaviors.
- We conduct extensive experiments based on a real-world financial statement dataset to verify the effectiveness of the proposed method. Experimental results demonstrate that our model outperforms state-of-the-art methods by more than 15.28% in *KS*.

2 Related Work

2.1 Financial Fraud Detection

Existing studies often adopt classic machine learning algorithms, such as Logistic Regression [Lin *et al.*, 2015], Support Vector Machine [Papík and Papíková, 2022], Random Forest [Yao *et al.*, 2018], Decision Tree [Hajek and Henriques, 2017], XGBoost [Aftabi *et al.*, 2023], and Multi-Layer Perceptron [Wang *et al.*, 2021], to detect fraud behaviors. Craja *et al.* [Craja *et al.*, 2020] utilized a hierarchical attention network to extract text features from the Management Discussion and Analysis section of annual reports. Besides, companies' true financial conditions often show continuity, with many fraud cases revealing that ongoing financial distress can lead to fraud spanning multiple consecutive years [Bao *et al.*, 2020]. Chen *et al.* [Zhang *et al.*, 2021] detected corporate financial fraud through a two-stage mapping process within the combined temporal and financial feature domains. Besides, there are rich explicit relationships between companies, such as related-party transactions and investment relationships, which can aid in detecting financial fraud. Mao *et al.* [Mao *et al.*, 2022c] construct a related-party transaction knowledge graph and extract topological features to improve fraud detection performance. Unlike previous methods, we use an RNN model to effectively capture temporal correlations in financial statements. Moreover, our model captures richer implicit relationships among listed companies through a attention-based CRF.

2.2 Conditional Random Field

Traditional CRF models are often used as offline post-processing layers for label refinement, capturing dependency relationships between a reference sample and its context. They are widely utilized in image segmentation [Chen *et al.*, 2018b] and named entity recognition [Chen *et al.*, 2018a]. In recent years, researchers have combined CRF models with graph convolutional networks to ensure that similar nodes have similar representations [Gao *et al.*, 2019; Xu *et al.*, 2021]. Such similarity constraints are often established using similarity calculated using Gaussian functions [Gao *et al.*, 2019] or shared labels [Xu *et al.*, 2021]. Unlike previous works, this paper proposes a novel multi-head attention-based CRF to capture implicit relationships among companies under various fraudulent patterns.

3 Preliminaries

This paper formulates the financial fraud detection as a binary classification task. Consider a set of N companies. For a given company c_i, its multivariate time series financial data are denoted as $x_i = \left[\mathbf{x}_i^1, \mathbf{x}_i^2, \cdots, \mathbf{x}_i^T\right]$, where $\mathbf{x}_i^t \in \mathbb{R}^d$ is a d-dimensional feature vector extracted from its financial statement for the t-th year, and T is the length of the time series. Moreover, $y_i = \left[l_i^1, l_i^2, \cdots, l_i^T\right]$ is the label sequence of x_i, where l_i^t is the label of $\mathbf{x}_i^t$. $l_i^t = 1$ if company c_i engages in fraud at time t, and $l_i^t = 0$ if otherwise. To effectively capture temporal correlations from x_i, we set a sliding window with a fixed width w and slide it in x_i to extract T-w+1 sub-sequences, i.e., $\{x_i^{1\to w}, x_i^{2\to w+1}, \cdots, x_i^{T\text{-}w+1\to T}\}$, where $x_i^{t\text{-}w+1\to t} = \left[\mathbf{x}_i^{t\text{-}w+1}, \cdots, \mathbf{x}_i^t\right]$ and $w \le t \le T$. For simplicity, we denote $x_i^{t\text{-}w+1\to t}$ as x_i^t. And then the sub-sequence embedding $\mathbf{s}_i^t$ is learned from x_i^t. Through learning the implicit relationships across the N companies from $\mathbf{S}^t = \{\mathbf{s}_1^t, \cdots, \mathbf{s}_N^t\}$, the final embeddings $\mathbf{H}^t = \{\mathbf{h}_1^t, \cdots, \mathbf{h}_N^t\}$ are produced for classification. A classifier $\mathcal{C}$ is trained on $\{\mathbf{H}^1, \cdots, \mathbf{H}^{t-1}\}$ to predict the labels of $\mathbf{H}^t$.

4 Methodology

4.1 Overview

Figure 1 shows the framework of ACRF-RNN. It consists of three components: a Temporal Feature Extractor (TFE), an Attention-based CRF Feature Transformer (ACRF), and a

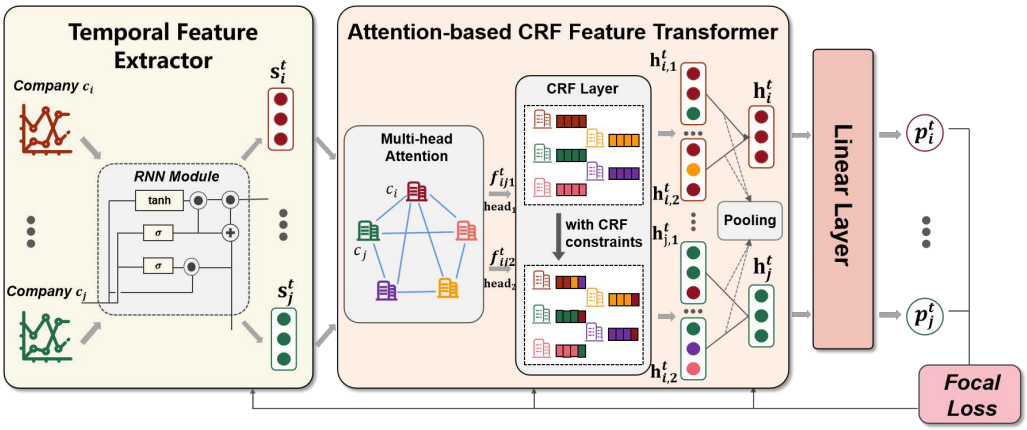


Figure 1: The framework of ACRF-RNN.

Classifier. TFE takes the multivariate time series x_i^t as input and produces a sub-sequence embedding $\mathbf{s}_i^t$ by learning temporal information from financial statements. Next, ACRF takes the sub-sequence embedding $\mathbf{s}_i^t$ as input and produces the final embedding $\mathbf{h}_i^t$ for company c_i by learning implicit relationships among behavior-similar companies. Finally, the final embedding $\mathbf{h}_i^t$ is fed into the Classifier for classification.

4.2 Temporal Feature Extractor

Financial condition of a company often exhibits continuity and correlation in a short period. Capturing such correlations is vital for fraud detection. Various enhanced RNN-based networks, such as LSTM and GRU, have been used to capture temporal information [Cheng and Li, 2021; Cheng *et al.*, 2020]. This paper utilizes a two-layer GRU model [Cho *et al.*, 2014] to capture temporal information due to the advantages of fewer parameters, faster convergence, and less overfitting. The GRU model takes the sub-sequence $x_i^t = [\mathbf{x}_i^{t-w+1}, \cdots, \mathbf{x}_i^t]$ of company c_i as input to learn the sub-sequence embedding $\mathbf{s}_i^t = GRU(x_i^t)$ as follows:

$$\mathbf{r}^t = \sigma(\mathbf{W}_{xr}\mathbf{x}_i^t + \mathbf{W}_{sr}\mathbf{s}_i^{t-1} + \mathbf{b}_\mathbf{r}), \quad (1)$$

$$\mathbf{z}^t = \sigma(\mathbf{W}_{xz}\mathbf{x}_i^t + \mathbf{W}_{sz}\mathbf{s}_i^{t-1} + \mathbf{b}_z), \quad (2)$$

$$\mathbf{n}^t = \tanh\left(\mathbf{W}_{xn}\mathbf{x}_i^t + \mathbf{r}^t \odot \mathbf{W}_{sn}\mathbf{s}_i^{t-1} + \mathbf{b}_n\right), \quad (3)$$

$$\mathbf{s}_i^t = \mathbf{z}^t \odot \mathbf{s}_i^{t-1} + (1 - \mathbf{z}^t) \odot \mathbf{n}^t, \quad (4)$$

where $W = \{\mathbf{W}_{xr}, \mathbf{W}_{xz}, \mathbf{W}_{xn}\}$ and $b = \{\mathbf{b}_r, \mathbf{b}_n, \mathbf{b}_z\}$ are learnable weights and biases, respectively, $\odot$ is the Hadamard product, $\mathbf{r}^t$, $\mathbf{z}^t$, $\mathbf{n}^t$ are the reset gate, update gate and candidate hidden state, respectively. The reset gate determines whether the network ignores the previous hidden state information, while the update gate determines whether the network remembers the previous hidden state information. The GRU produces $\mathbf{s}_i^t \in \mathbb{R}^d$ by aggregating temporal information from the sub-sequence x_i^t, where d is the dimension. $\mathbf{s}_i^{t-1}$ is derived from the sub-sequence x_i^{t-1}, and so on.

4.3 Attention-based CRF Feature Transformer

This module comprises two sub-modules: a Multi-head Attention component and a Conditional Random Field (CRF) model. The former aims to effectively learn inter-company implicit relationships, and the corresponding similarity coefficients indicate the strength of these relationships. The multiple attention heads are used to capture various behavioral patterns. The learned similarity coefficients are fed into the CRF model to optimize the sub-sequence embeddings under each attention head. The optimized sub-sequence embeddings are aggregated to produce the final embedding for each company.

Multi-head Attention Mechanism

In practice, fraudulent companies often exhibit various fraud patterns, such as inflating profits or understating liabilities. A notable example is LeEco [Mao *et al.*, 2022a]. Besides, the implicit relationships may vary under different patterns. Therefore, we introduce a multi-head attention mechanism to capture implicit relationships under different patterns, enabling more nuanced fraud detection. Specifically, the multi-head attention mechanism consists of a shared feed-forward network with m LeakyReLU gates [Maas *et al.*, 2013]. First, we define $R_{i,j,m}^t$ to measure the m-th implicit relationship between company c_i and c_j as follows:

$$R_{i,j,m}^t = LeakyReLU_m(\mathbf{W}_a[\mathbf{s}_i^t \oplus \mathbf{s}_j^t]), \quad (5)$$

where $LeakyReLU_m(\cdot)$ is the m-th gate to prevent information loss, $\oplus$ denotes concatenating operation, $\mathbf{W}_a \in \mathbb{R}^{d' \times 2d}$ is the trainable weight matrix of the feed-forward network, shared with all attention heads, and $\mathbf{s}_i^t$ and $\mathbf{s}_j^t$ are sub-sequence embeddings of company c_i and c_j, respectively. To ensure all $R_{i,j,m}^t$'s are in the same magnitude and comparable, we perform the softmax normalization operation to $R_{i,j,m}^t$ to obtain the similarity coefficient $f_{i,j,m}$:

$$f_{i,j,m}^t = \frac{\exp(R_{i,j,m}^t)}{\sum_{k \in N_t, k \neq i} \exp(R_{i,j,m}^t)}. \quad (6)$$

The larger the value of $f_{i,j,m}^t$, the stronger the implicit relationship between company c_i and company c_j under the m-th pattern.

Conditional Random Field Model

Based on the learned similarity coefficients, the CRF model further optimizes the sub-sequence embeddings by perceiving

the homogeneity among companies [Krähenbühl and Koltun, 2011]:

$$P\left(\mathbf{h}_i^t|\mathbf{s}_i^t\right) = \frac{1}{Z(\mathbf{s}_i^t)} \exp\left(-E\left(\mathbf{h}_i^t|\mathbf{s}_i^t\right)\right) \tag{7}$$

where $Z(\cdot)$ denotes the normalization operation, $\mathbf{h}_i^t$ denotes the optimized sub-sequence embedding, and $P\left(\mathbf{h}_i^t|\mathbf{s}_i^t\right)$ denotes the conditional probability of producing $\mathbf{h}_i^t$ given $\mathbf{s}_i^t$. For simplicity, the subscript m of the similarity coefficients is temporarily omitted. The goal is to maximize the conditional probability $P\left(\mathbf{h}_i^t|\mathbf{s}_i^t\right)$, which is equivalent to minimizing the energy function $E\left(\mathbf{h}_i^t|\mathbf{s}_i^t\right)$ [Gao *et al.*, 2019]:

$$\min E\left(\mathbf{h}_i^t|\mathbf{s}_i^t\right) = \alpha \left\|\mathbf{h}_i^t - \mathbf{s}_i^t\right\|_2 + \beta \sum_{j \in N} f_{i,j} \left\|\mathbf{h}_i^t - \mathbf{h}_j^t\right\|_2^2 \tag{8}$$

where $\left\|\mathbf{h}_i^t - \mathbf{s}_i^t\right\|_2$, named the unary potential, measures the transformation loss from $\mathbf{s}_i^t$ to $\mathbf{h}_i^t$, $\sum_{j \in N} f_{i,j} \left\|\mathbf{h}_i^t - \mathbf{h}_j^t\right\|_2^2$, named the binary potential, perceives the homogeneity between company c_i and those measured by their weighted distances, α and β are learnable parameters adjusting their importance, respectively. The sub-sequence embedding maintains a certain degree of similarity before and after optimization by minimizing $\left\|\mathbf{h}_i^t - \mathbf{s}_i^t\right\|_2$. Moreover, when $f_{i,j}^t$ is large, $\mathbf{h}_j^t$ and $\mathbf{h}_i^t$ are highly similar, and $f_{i,j} \left\|\mathbf{h}_i^t - \mathbf{h}_j^t\right\|_2^2$ approaches 0. Conversely, when $f_{i,j}^t$ is small, $\mathbf{h}_j^t$ and $\mathbf{h}_i^t$ are discriminative and $f_{i,j} \left\|\mathbf{h}_i^t - \mathbf{h}_j^t\right\|_2^2$ also approaches 0. That is, $f_{i,j}^t$ is a supervisory signal promoting the optimized sub-sequence embeddings of companies sharing similar fraud patterns to be more similar. Equation (8) aims to achieve a balance between unary and binary potentials, striving to enhance the discriminative power of the optimized sub-sequence embeddings while preserving the information from the original subsequence embeddings.

The prerequisite for learning the optimized sub-sequence embedding $\mathbf{h}_i^t$ is to determine the true distribution of the conditional probability $P\left(\mathbf{h}_i^t|\mathbf{s}_i^t\right)$. Thus, we first formulate $P\left(\mathbf{h}_i^t|\mathbf{s}_i^t\right) \sim \mathbf{P}\left(\mathbf{H}^t|\mathbf{S}^t\right)$, denoting the distribution function of the conditional probability. However, the time complexity of computing exact $\mathbf{P}\left(\mathbf{H}^t|\mathbf{S}^t\right)$ will cause the whole model to barely work. Therefore, we learn the sample distribution $\mathbf{Q}\left(\mathbf{H}^t\right)$ and make it approximate the true distribution $\mathbf{P}\left(\mathbf{H}^t|\mathbf{S}^t\right)$, based on the mean-field approximation method [Gao *et al.*, 2019]. This sample distribution $\mathbf{Q}\left(\mathbf{H}^t\right)$ can be represented as a product of independent marginal distributions through variational inference, expressed as $\mathbf{Q}\left(\mathbf{H}^t\right) = \prod_i^N \mathbf{h}_i^t$. We minimize the KL divergence distance between $\mathbf{Q}\left(\mathbf{H}^t\right)$ and $\mathbf{P}\left(\mathbf{H}^t|\mathbf{S}^t\right)$ to obtain the optimal sample distribution $\mathbf{Q}^*\left(\mathbf{H}^t\right)$. For company c_i, according to Equations (7), it can be deduced that:

$$Q_i^*\left(\mathbf{h}_i^t\right) \sim \exp(-E\left(\mathbf{h}_i^t|\mathbf{s}_i^t\right)), \tag{9}$$

where $Q_i^*\left(\mathbf{h}_i^t\right)$ is a multi-dimensional Gaussian function with the kernel $E\left(\mathbf{h}_i^t|\mathbf{s}_i^t\right)$. Therefore, $\mathbf{h}_i^t$ can be updated by calculating the expectation of $Q_i^*\left(\mathbf{h}_i^t\right)$:

$$\left(\mathbf{h}_i^t\right)^{k+1} = \frac{\alpha \mathbf{s}_i^t + \beta \sum_{j \in N} f_{i,j}^t \left(\mathbf{h}_i^t\right)^k}{\alpha + \beta \sum_{j \in N} f_{i,j}^t}, k = 1, \ldots, K, \tag{10}$$

where K is a hyper-parameter indicating the epochs of iterations, initially, $\left(\mathbf{h}_i^t\right)^0 = \mathbf{s}_i^t$. Based on the learned similarity coefficient $f_{i,j}^t$, the optimized sub-sequence embedding $\mathbf{h}_i^t$ can be learned recursively. It should be emphasized that there are m learned similarity coefficients with m attention heads. Therefore, we learn the m-th optimized sub-sequence embedding $\mathbf{h}_{i,m}^t$ under m-th attention heads through Equation (10) for company c_i. And then we use a SUM pooling operation to obtain the final embedding:

$$\mathbf{h}_i^t = \sum\nolimits_{m=1}^{M} \mathbf{h}_{i,m}^t \tag{11}$$

where M is the total number of attention heads. Experimental results verify that summation pooling is the best aggregation in our fraud detection task.

4.4 Model Training

The optimized sub-sequence embedding $\mathbf{h}_i^t$ is then fed into a multi-layer perceptron (MLP) for classification:

$$p_i^t = \sigma(MLP\left(\mathbf{h}_i^t\right)) \tag{12}$$

where p_i^t denotes the probability that company c_i is predicted to be fraudulent at time t, $\sigma(\cdot)$ denotes the activation function. We employ the Focal Loss [Lin *et al.*, 2017] to optimize the model, which can alleviate the class imbalance problem by assigning higher weights to fraud companies:

$$FL = -\frac{1}{N} \sum\nolimits_{i=1}^{N} [-a_{\text{balance}}(1 - p_i^t) \log(p_i^t)] \tag{13}$$

where $a_{balance}$ is a hyperparameter balancing the weight of benign and fraud companies.

5 Experiments

5.1 Datasets

We collect a real-world dataset containing multivariate time series financial data from 491 Chinese listed companies from 2010-2020. In each year, each listed company has 208-dimensional features derived from its financial statement. To extract more comprehensive features for accurate fraud detection, we refer to the fraud triangle theory to select the following three types of features: i) 91-dimensional financial features such as profits, ownership interest and cash flow, ii) 7-dimensional basic features such as industry and listed time, and iii) 110-dimensional fraud features aligned with Skousen's framework [Skousen *et al.*, 2009]. Companies penalized by China Securities Regulatory Commission due to the violation of the accounting standards are labeled as fraud companies, and others are defined benign ones. All data are collected from the China Securities Regulatory Commission[1].

5.2 Baseline Methods

We compare eight baseline methods with our method, including two classical machine learning methods (Logistic Regression (LR) [Cucchiara, 2012] and Decision Tree (DT) [SONG and LU, 2015]), two ensemble learning methods (LightGBM [Ke *et al.*, 2017] and XGBoost [Chen and

[1]http://www.csrc.gov.cn/

Partitions	Training set (# Fraud / # Benign)	Testing set (# Fraud / # Benign)
Validation	2010-2016 (874/2653)	2017 (122/369)
Test 1	2010-2017 (996/2932)	2018 (127/364)
Test 2	2010-2018 (1123/3296)	2019 (110/381)
Test 3	2010-2019 (1233/3677)	2020 (86/405)

Table 1: Training, validation and testing data split by year.

Guestrin, 2016]), three GNN-based methods (GCN [Kipf and Welling, 2017], GAT [Veličković *et al.*, 2018] and GraphSAGE [Hamilton *et al.*, 2017]), and one state-of-the-art fraud detection model ADGAT [Cheng and Li, 2021]. For a fair comparison between ACRF-RNN and GNN-based methods, we construct a static audit-sharing graph, where the nodes represent the company-year pair, and the edges connect nodes audited by the same audit institution.

5.3 Evaluation Metrics

Four widely adopted metrics are used to measure the performance of different methods. Accuracy calculates the percentage of all correctly classified samples to assess classification performance. *Recall*$_m$ [Grandini *et al.*, 2020] averages recall across all classes. *G-mean* [Sun *et al.*, 2006] is defined as the geometric mean of the recalls over all classes. A low *G-mean* indicates poor identification ability of a model for at least one class. As a popular metric to assess the discrimination ability between fraud and benign samples, *KS* [Massey and Frank, 1951] measures the maximal difference between True Positive Rate and False Positive Rate as the classification threshold shifts from 0 to 1.

5.4 Experimental Settings

In real business scenarios, financial fraud detection is conducted once a year, using models trained on historical data to detect fraud companies for the current year. Therefore, we split the real-world dataset by year. Specifically, we use the data from 2010 to 2016 for training and use the data from 2017 as validation to adjust hyperparameters. Next, with fixed hyperparameters, we train the model on data from 2010 to T, and test it on data at the $(T{+}1)$-th year, where $T \in \{2017, 2018, 2019\}$. Table 1 details the partitions.

We implement ACRF-RNN based on PyTorch 1.12.1 with Python 3.8, and all the experiments are run on a Ubuntu 16.04 LTS server. Grid search is employed to select the optimal hyper-parameters based on the validation set. All parameters are initialized using the Kaiming initialization and are trained using the Adam optimizer with an initial learning rate of 0.01. The optimal time window size w is 5, the iteration of CRF K is set to 5, the sub-sequence embedding dimension d is set to 64. The attention layer consists of $M = 3$ attention heads and its hidden size d' is set to 32. And the penalty coefficient of fraud sample $a_{balance}$ is set to 0.15.

5.5 Results and Analysis

Table 2 shows the experimental results. ACRF-RNN significantly outperforms all baselines in all experimental settings. Among all the test results, ACRF-RNN achieves the highest accuracy score, indicating the best overall classification performance. Compared to all baselines, the improvement in *KS* is 15.28-25.34%, indicating that ACRF-RNN excels more in distinguishing between fraud and benign samples. Besides, *KS* exceed 40% over three consecutive testing years, demonstrating its powerful generalization and robustness. The improvement of ACRF-RNN in *Recall*$_m$ scores is 4.04-6.58%, indicating the best identification ability. However, such higher *Recall*$_m$ scores may also be due to excessive focus on fraud samples, which can lead to more benign samples being misclassified. Therefore, we use *G-mean* for auxiliary evaluation. ACRF-RNN achieves a high *G-mean* score of around 70% in all test results and outperforms all baselines. This finding indicates that our model steadily achieves high performance at different years.

Machine learning methods demonstrate suboptimal performance, indicating that the extracted rich raw features are effective in detecting financial fraud. Ensemble learning methods perform better than machine learning methods because of their stronger robustness and generalization ability. However, these methods have lower performance than ACRF-RNN because they fail to capture temporal information and implicit relationships.

GCN, GAT, and GraphSAGE have relatively poor performance, indicating that explicit relationships, such as audit-sharing, may not be suitable for the financial fraud detection task. This finding also demonstrates the importance of selecting proper explicit relationships, which are labor-intensive. On the contrary, our method can automatically explore various implicit relationships among companies for latent feature optimization and achieves optimal performance.

ADGAT achieves the second-best performance, trailing only ACRF-RNN. The key difference lies in how it model inter-company implicit correlations. ADGAT concatenates high-attention neighbor node vectors, whereas ACRF-RNN optimizes feature vectors by minimizing energy loss. Besides, ACRF-RNN employs an inductive feature transformation, enhancing generalization without relying on specific prior knowledge.

5.6 Ablation Study

We design three variants of ACRF-RNN to validate the effectiveness of three key components. ACRF-RNN (**w/o TFE**) removes TFE, utilizes a linear layer to transform the raw features of a single annual financial statement instead of subsequences, and feeds the output of the linear layer into ACRF. ACRF-RNN (**w/o FT**) excludes ACRF, feeding TFE outputs directly to the classifier. ACRF-RNN (**w/o FL**) replaces Focal Loss with cross-entropy to assess class imbalance handling. Table 3 shows the average experimental results of all variants on three test sets. The full ACRF-RNN outperforms all variants, demonstrating the effectiveness of all components. ACRF-RNN (**w/o TFE**) shows the worst performance on all metrics scores, demonstrating the effectiveness of learning temporal information. Compared to ACRF-RNN (**w/o FL**), the full model significantly boosts *Recall*$_m$, *G-mean*, and *KS* scores. This finding indicates that Focal Loss effectively enhances the model's ability to identify fraud samples but may misclassify benign samples.

	Test 1 (2018)				Test 2 (2019)				Test 3 (2020)			
Methods	**Acc**	**Recall$_m$**	**KS**	**G-mean**	**Acc**	**Recall$_m$**	**KS**	**G-mean**	**Acc**	**Recall$_m$**	**KS**	**G-mean**
LR	61.71	60.33	20.67	60.27	55.40	60.27	20.53	59.62	53.16	57.87	15.73	57.41
DT	57.43	60.01	20.02	59.77	55.19	56.90	13.81	56.82	60.08	58.40	16.80	58.34
LightGBM	63.14	65.65	31.31	65.45	62.32	64.08	28.17	64.00	63.54	65.99	31.99	65.89
XGBoost	61.71	61.87	23.74	61.87	62.12	61.36	22.73	61.35	66.40	66.35	32.70	66.35
GCN	54.18	57.56	15.12	57.13	56.42	55.75	11.50	55.74	52.55	57.50	14.99	56.99
GAT	61.91	57.39	14.79	56.63	49.69	56.92	13.83	55.39	65.38	56.57	13.15	54.93
GraphSAGE	47.05	47.37	5.27	47.36	63.34	50.51	1.03	44.85	68.02	53.14	6.28	47.95
ADGAT	76.17	68.04	36.07	65.91	79.23	72.07	44.13	70.89	77.39	67.52	35.04	65.79
ACRF-RNN	**77.60**	**70.79**	**41.58**	**69.37**	**80.45**	**75.44**	**50.88**	**74.89**	**79.43**	**71.96**	**43.92**	**71.04**
Impr.	1.88%	4.04%	15.28%	5.25%	1.54%	4.68%	15.30%	5.64%	2.64%	6.58%	25.34%	7.07%

Table 2: Comparison results (%) of ACRF-RNN with eight baseline methods.

Methods	**Acc**	**Recall$_m$**	**KS**	**G-mean**
ACRF-RNN (w/o TFE)	56.75	48.65	4.37	46.42
ACRF-RNN (w/o FT)	78.95	70.28	41.35	69.05
ACRF-RNN (w/o FL)	**79.63**	70.95	41.91	69.25
ACRF-RNN	79.16	**72.73**	**45.46**	**71.77**

Table 3: Performance (%) of different variants of ACRF-RNN.

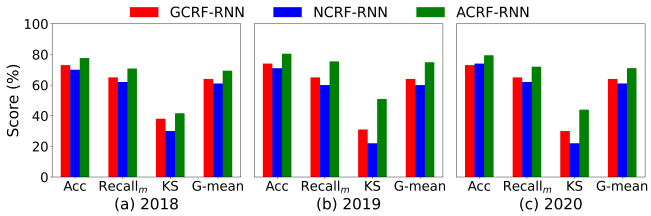


Figure 2: Performance of ACRF, GCRF and NCRF.

To further verify the effectiveness of the multi-head attention mechanism in capturing implicit relationships, we implement the CRF-Gaussian and CRF-NN layers proposed by [Gao *et al.*, 2019] in our model, creating two variants named GCRF-RNN and NCRF-RNN. Specifically, GCRF-RNN computes $f_{i,j}^t$ by a Gaussian function, and NCRF-RNN computes $f_{i,j}^t$ by a flexible neural network. Figure 2 shows their performance on three test sets. ACRF-RNN outperforms NCRF and GCRF on all metrics, indicating the effectiveness of the multi-head attention mechanism in learning similarity coefficients for capturing multiple similar behaviors among companies. The Gaussian function is highly sensitive to outliers, resulting in imprecise similarity assessments. Neural networks, albeit more adaptable than Gaussian functions, capture only a general sense of similarity. The multi-head attention mechanism is adept at discerning implicit relationships in different subspaces and thus achieves higher performance.

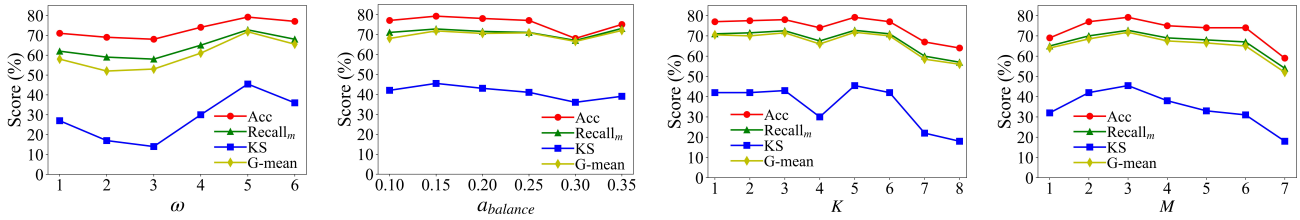
5.7 Sensitivity Analysis

Figure 3(a) illustrates the effect of the width of the time window w. As w increases, all metrics first decrease and then increase. A small w makes the model only observe a small part of the whole sequence, leading to insufficient understanding of the overall temporal correlation. The optimal performance is achieved when $w = 5$. However, further increasing w leads to a decline in performance because long-term correlation may harm the learning of recent characteristics. Figure 3(b) shows the effect of $a_{balance}$. *KS* initially rises and then decrease as $a_{balance}$ increases, reaching a peak at $a_{balance} = 0.15$. A larger value of $a_{balance}$ directs the model's attention towards fraud samples, aiding in improving the recognition capability for the fraud class. However, as $a_{balance}$ continues to increase, the model might overly emphasize fraud samples, neglecting crucial information from benign samples. This could lead to a decrease in overall performance. Figure 3(c) demonstrates the model's performance as the number of iterations K varies. As K gradually increases, the model's effectiveness experiences slight fluctuations. As K continues to rise, the model's performance begins to deteriorate, potentially due to the overfitting caused by excessive iterations. Figure 3(d) demonstrates the effect of the number of attention heads. When M is small, the model may fail to fully capture the complexity of various patterns, leading to insufficient representational capability. Conversely, too many heads might cause overfitting.

5.8 Further Research

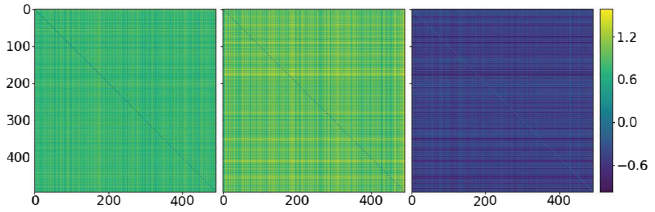
To further explore the role of the multi-head attention mechanism in capturing fraud pattern diversity, we visualize the learned similarity coefficients from different attention heads in experiment Test 2 (2019), as shown in Figure 4. The heatmaps illustrate the attention coefficients from three distinct attention heads. The lightness of color in these heatmaps corresponds to the attention weights, with lighter hues indicating stronger connections between companies. Each heatmap reveals unique attention patterns, signifying that each head specializes in a distinct feature subspace to infer similarity relationships among companies. Such diversity enables the model to capture broader information from the training data.

Additionally, within the same attention head, we randomly select 50 benign companies and 50 fraud companies to visualize their similarity coefficients, as shown in Figure 5. Fraudulent companies tend to have relatively higher similarity, suggesting that they share more common behavioral characteristics, such as overstatement of revenue, cost manipulation,



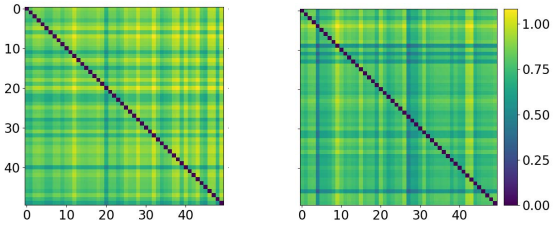
(a) The effects of w (b) The effects of $a_{balance}$ (c) The effects of K (d) The effects of M

Figure 3: The effects of the hyper-parameter w, $a_{balance}$, K and M.



(a) Head 1 (b) Head 2 (c) Head 3

Figure 4: Attention coefficients of different attention heads.



(a) Fraud companies (b) Benign companies

Figure 5: Similarity between benign and fraud companies.

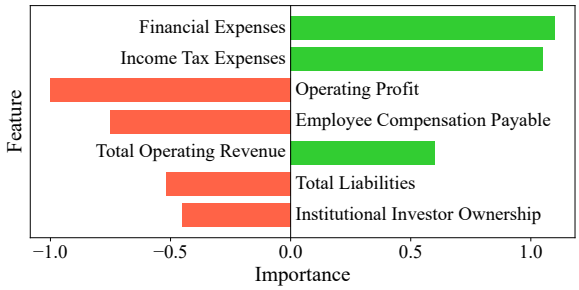
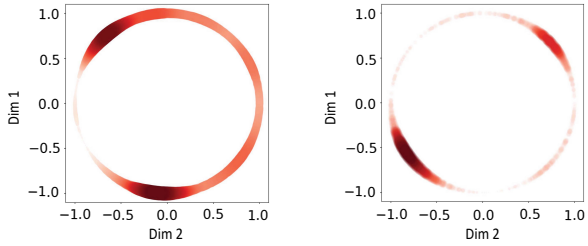


Figure 6: Explainability analysis.



(a) Fraud companies (b) Benign companies

Figure 7: Distribution differences of learned embeddings.

and inflation of reported financial outcomes. A comparison of Figure 5(a) and 5(b) reveals that benign companies exhibit more diverse behavioral patterns than fraudulent ones.

Moreover, we conduct experiments using the LIME [Ribeiro *et al.*, 2016] interpretability framework to identify key features aiding in detecting fraud companies. Figure 6 shows results for a randomly selected fraudulent case. "Financial Expenses" and "Income Tax Expenses" are identified as the top two features contributing to the classification of this sample as fraudulent. Both of them directly impact a company's cash flow. Fraudulent activities might aim to misrepresent the company's cash flow situation by manipulating these expenses. For instance, understating income tax expenses can temporarily inflate earnings, presenting an unrealistically positive financial performance.

In order to verify the performance of ACRF-RNN, the raw features and the learned final embeddings of companies are visualized by t-SNE [Rauber *et al.*, 2016; Lv *et al.*, 2025] in Figure 7. The shade of the color and the change in the size of the points together represent the density distribution of the data points, with darker and larger points having higher density. The distribution of raw features is more uniform, while the learned final embeddings are more discriminative and exhibit obvious cluster characteristics. It indicates that ACRF-RNN learns critical information during the training, improving the distinction between fraud and benign samples.

6 Conclusion

This paper presents ACRF-RNN, a novel recurrent neural network (RNN) with an attention-based conditional random field (CRF) for financial fraud detection. The RNN captures critical temporal patterns in multivariate financial time series for each company, producing the sub-sequence embeddings. Besides, the attention-based CRF models inter-company implicit relationships to refine these sub-sequence embeddings for improved classification. Experiments on real-world datasets show that ACRF-RNN outperforms state-of-the-art methods, achieving average gains of 2.02% in accuracy, 5.1% in *Recall*$_m$ and 18.64% in *KS*. Future work will focus on the effect of implicit relationships between different companies across adjacent years on financial fraud detection.

Acknowledgements

The research presented in this paper is supported in part by the National Natural Science Foundation of China (No. 62272379, T2341003, 62102310), the Shaanxi Science Fund for Outstanding Young Scientists (2025JC-JCQN-081), the Key R&D in Shaanxi Province (2023-YBGY-269), and the Fundamental Research Funds for the Central Universities (xzy012023068).

References

[Aftabi *et al.*, 2023] Seyyede Zahra Aftabi, Ali Ahmadi, and Saeed Farzi. Fraud detection in financial statements using data mining and gan models. *Expert Systems with Applications*, 227:120144, 2023.

[Bao *et al.*, 2020] Yang Bao, Bin Ke, Bin Li, Y Julia Yu, and Jie Zhang. Detecting accounting fraud in publicly traded us firms using a machine learning approach. *Journal of Accounting Research*, 58(1):199–235, 2020.

[Chen and Guestrin, 2016] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco California USA, August 2016. ACM.

[Chen *et al.*, 2018a] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group Consistent Similarity Learning via Deep CRF for Person Re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, June 2018.

[Chen *et al.*, 2018b] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, April 2018.

[Cheng and Li, 2021] Rui Cheng and Qing Li. Modeling the Momentum Spillover Effect for Stock Prediction via Attribute-Driven Graph Attention Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):55–62, May 2021.

[Cheng *et al.*, 2020] Dawei Cheng, Xiaoyang Wang, Ying Zhang, and Liqing Zhang. Risk Guarantee Prediction in Networked-Loans. In *IJCAI International Joint Conference on Artificial Intelligence*, page 7, 2020.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, September 2014.

[Craja *et al.*, 2020] Patricia Craja, Alisa Kim, and Stefan Lessmann. Deep learning for detecting financial statement fraud. *Decision Support Systems*, 139:113421, 2020.

[Cucchiara, 2012] Andrew Cucchiara. Applied Logistic Regression. *Technometrics*, 34:358–359, March 2012.

[Gao *et al.*, 2019] Hongchang Gao, Jian Pei, and Heng Huang. Conditional Random Field Enhanced Graph Convolutional Neural Networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 276–284, Anchorage AK USA, July 2019. ACM.

[Grandini *et al.*, 2020] Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for Multi-Class Classification: An Overview, August 2020.

[Hajek and Henriques, 2017] Petr Hajek and Roberto Henriques. Mining corporate annual reports for intelligent detection of financial statement fraud–a comparative study of machine learning methods. *Knowledge-Based Systems*, 128:139–152, 2017.

[Hamilton *et al.*, 2017] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Ke *et al.*, 2017] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks, February 2017.

[Krähenbühl and Koltun, 2011] Philipp Krähenbühl and Vladlen Koltun. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[Lin *et al.*, 2015] Chi-Chen Lin, An-An Chiu, Shaio Yan Huang, and David C Yen. Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowledge-Based Systems*, 89:459–470, 2015.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[Liu *et al.*, 2023] Yajing Liu, Zhengya Sun, and Wensheng Zhang. Improving fraud detection via hierarchical attention-based graph neural network. *Journal of Information Security and Applications*, 72:103399, 2023.

[Lv *et al.*, 2025] Xiangwei Lv, Jingyuan Chen, Mengze Li, Yongduo Sui, Zemin Liu, and Beishui Liao. Grasp the key takeaways from source domain for few shot graph domain adaptation. In *Proceedings of the ACM on Web Conference 2025*, pages 2330–2340, 2025.

[Maas *et al.*, 2013] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectiper Nonlinearities Improve Neural Network Acoustic Models. In *Proceedings of the 30 Th International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013.

[Mao *et al.*, 2022a] Xuting Mao, Mingxi Liu, and Yinghui Wang. Using gnn to detect financial fraud based on the related party transactions network. *Procedia Computer Science*, 214:351–358, 2022.

[Mao *et al.*, 2022b] Xuting Mao, Hao Sun, Xiaoqian Zhu, and Jianping Li. Financial fraud detection using the related-party transaction knowledge graph. *Procedia Computer Science*, 199:733–740, 2022.

[Mao *et al.*, 2022c] Xuting Mao, Hao Sun, Xiaoqian Zhu, and Jianping Li. Financial fraud detection using the related-party transaction knowledge graph. *Procedia Computer Science*, 199:733–740, 2022.

[Massey and Frank, 1951] J Massey and J Frank. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[Papík and Papíková, 2022] Mário Papík and Lenka Papíková. Detecting accounting fraud in companies reporting under us gaap through data mining. *International Journal of Accounting Information Systems*, 45:100559, 2022.

[Rauber *et al.*, 2016] Paulo E Rauber, Alexandre X Falcão, and Alexandru C Telea. Visualizing time-dependent data using dynamic t-sne. In *Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers*, pages 73–77, 2016.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[Skousen *et al.*, 2009] Christopher J. Skousen, Kevin R. Smith, and Charlotte J. Wright. Detecting and predicting financial statement fraud: The effectiveness of the fraud triangle and SAS No. 99. In Mark Hirschey, Kose John, and Anil K. Makhija, editors, *Advances in Financial Economics*, volume 13, pages 53–81. Emerald Group Publishing Limited, January 2009.

[SONG and LU, 2015] Yan-yan SONG and Ying LU. Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27(2):130–135, April 2015.

[Sun *et al.*, 2006] Yanmin Sun, Mohamed S. Kamel, and Yang Wang. Boosting for Learning Multiple Classes with Imbalanced Class Distribution. In *Sixth international conference on data mining (ICDM'06)*, pages 592–602. IEEE, 2006.

[Veličković *et al.*, 2018] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks, February 2018.

[Wang *et al.*, 2021] Yurou Wang, Ruixue Li, and Yanfang Niu. A deep neural network based financial statement fraud detection model: Evidence from china. In *Proceedings of the 2021 4th Artificial Intelligence and Cloud Computing Conference*, pages 145–149, 2021.

[Wang *et al.*, 2024] Chenxu Wang, Mengqin Wang, Xiaoguang Wang, Luyue Zhang, and Yi Long. Multi-relational graph representation learning for financial statement fraud detection. *Big Data Mining and Analytics*, 7(3):920–941, 2024.

[Wang *et al.*, 2025] Xiaoguang Wang, Chenxu Wang, Huanlong Liu, Mengqin Wang, Tao Qin, and Pinghui Wang. Figraph: A dynamic heterogeneous graph dataset for financial anomaly detection. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 813–816, 2025.

[Xu *et al.*, 2021] Bingbing Xu, Huawei Shen, Bingjie Sun, Rong An, Qi Cao, and Xueqi Cheng. Towards consumer loan fraud detection: Graph neural networks with role-constrained conditional random field. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4537–4545, 2021.

[Yao *et al.*, 2018] Jianrong Yao, Jie Zhang, and Lu Wang. A financial statement fraud detection model based on hybrid data mining methods. In *2018 international conference on artificial intelligence and big data (ICAIBD)*, pages 57–61. IEEE, 2018.

[Zhang *et al.*, 2021] Yanci Zhang, Tianming Du, Yujie Sun, Lawrence Donohue, and Rui Dai. Form 10-Q Itemization. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4817–4822, Virtual Event Queensland Australia, October 2021. ACM.

[Zhu *et al.*, 2021] Xiaoqian Zhu, Xiang Ao, Zidi Qin, Yanpeng Chang, Yang Liu, Qing He, and Jianping Li. Intelligent financial fraud detection practices in post-pandemic era. *The Innovation*, 2(4):100176, November 2021.