Journal of computer science and software applications
https://www.mfacademia.org/index.php/jcssa
ISSN:2377-0430
Vol. 4, No. 8, 2024

A Deep Fusion Framework for Financial Fraud Detection and Early Warning Based on Large Language Models

Jiangchuan Gong[1], Yuxi Wang[2], Weiyao Xu[3], Yiwei Zhang[4]
[1]Hebei Normal University, Shijiazhuang, China
[2]Carnegie Mellon University, Pittsburgh, USA
[3]Fordham University, New York City, USA
[4]Cornell University, Ithaca, USA
*Corresponding Author: Yiwei Zhang; zhangyiwei721@gmail.com

Abstract:

This study focuses on fraud detection and early warning in financial scenarios. A deep fusion architecture based on large language models is proposed to improve the modeling and classification of complex fraudulent behaviors. The method first applies a pre-trained language model to perform semantic embedding on multimodal input data. This captures deep information from both transaction texts and structured features. Then, a convolutional neural network (CNN) is used to extract local anomaly patterns, while a long short-term memory (LSTM) network models temporal dependencies in transaction sequences. Finally, a risk scoring function is used to determine the probability of fraudulent activity. To further enhance the model ’ s robustness and discriminative power, contrastive learning strategies and imbalance handling mechanisms are introduced. These components optimize detection performance from multiple perspectives. Experimental results on a public credit card fraud dataset show that the proposed model outperforms existing mainstream methods in terms of accuracy, precision, recall, and F1-score, demonstrating strong overall performance. In addition, a series of ablation studies and comparative experiments were designed. These tests validate the effectiveness and rationality of each sub-module in the proposed architecture.

Keywords:

Financial fraud, Large language model, Deep learning, Contrastive learning

1. Introduction

In recent years, with the rapid growth of the digital economy and the widespread adoption of financial technologies, financial services have become increasingly frequent, online, and intelligent. However, during this transformation, the complexity and concealment of financial fraud have also escalated, posing severe challenges to the stable operation of the financial system. From traditional credit card theft and identity forgery to emerging telecom fraud, fake transactions, and malicious money laundering, fraud methods are continuously evolving, with increasing technical sophistication and organizational coordination [1]. At the same time, the financial losses caused by such frauds are rising rapidly, becoming a significant risk that hampers the high-quality development of the financial sector. As a result, efficiently identifying and accurately predicting financial fraud has become a core issue that urgently needs to be addressed in the field of financial risk control.

Previous studies on financial fraud detection mainly relied on rule-based expert systems or traditional machine learning models. While these approaches can detect certain abnormal behaviors, they have clear limitations in handling high-dimensional, multi-modal, and unstructured data. In particular, they show weak generalization when confronted with novel and concealed fraud patterns. Furthermore, fraudulent activities are often hidden within a large number of normal transactions. Their expression tends to be semantically complex and highly context-dependent, making shallow-feature modeling insufficient to capture the underlying logic [2]. As financial data continues to grow in volume and complexity, there is a pressing need for more advanced intelligent algorithms to achieve dynamic detection and in-depth analysis of fraud behaviors.

In recent years, Large Language Models (LLMs) have made groundbreaking progress in natural language processing. They demonstrate strong capabilities in knowledge representation, contextual understanding, and multi-task generalization. Specifically, when dealing with a mix of structured and unstructured data, LLMs can perform semantic interpretation and logical reasoning to reach decisions with human-like cognition. This makes them a promising solution to challenges in financial fraud detection, such as high-dimensional modeling, semantic inconsistency, and fraud pattern migration [3]. In addition, LLMs support effective pre-training and fine-tuning mechanisms, enabling better adaptability to heterogeneous data sources. These features enhance their robustness and real-time performance in detecting dynamic fraud behavior.

Against this backdrop, developing a financial fraud detection and early warning system based on LLMs holds both theoretical and practical significance [4]. On the one hand, this research promotes the deep integration of natural language processing and financial risk management, expanding the application boundaries of LLMs in the financial domain. It also provides a new paradigm for balancing model generality and domain specificity. On the other hand, building an interpretable, efficient, and robust fraud detection system is expected to offer intelligent and automated tools for financial institutions. This would improve their ability to handle emerging fraud scenarios, reduce fraud-related costs, safeguard user assets, and contribute to a more secure and trustworthy digital finance ecosystem.

In summary, with the rapid development of financial technologies, there is an urgent need for novel approaches capable of handling complex data structures and dynamic fraud patterns. LLMs offer strong support for financial fraud detection through their strengths in semantic understanding, multi-modal data integration, and logical inference. This study focuses on "Financial Fraud Detection and Early Warning Based on Large Language Models," aiming to explore its feasibility, effectiveness, and scalability in the field of financial risk control. The research not only proposes a technological innovation for the financial industry but also provides theoretical and practical insights into the application of artificial intelligence in high-risk and sensitive scenarios.

2. Related work

As one of the core technologies in financial risk control systems, financial fraud detection has long attracted significant attention from both academia and industry. Early studies mainly relied on rule-based detection systems [5]. These systems constructed manually defined behavioral indicators to conduct static analysis and threshold-based judgment on transaction data. Such methods were effective for detecting known fraud patterns. However, they struggled to adapt to the evolving and dynamic nature of fraudulent activities. Later, traditional machine learning methods were gradually introduced into fraud detection. Models such as decision trees, support vector machines, random forests, and ensemble learning were trained on historical data to capture patterns of fraudulent behavior. Nonetheless, these models still showed clear limitations in handling high-dimensional sparse data, unstructured semantic information, and imbalanced sample distributions. Moreover, they typically depended on feature engineering and static inputs, making them inadequate for capturing complex temporal and contextual variations.

With the widespread adoption of deep learning, researchers began exploring neural network-based fraud detection models. Architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention-based models like Transformers were introduced. These models significantly improved the ability to model nonlinear relationships and sequential features. They showed strong adaptability in handling transaction sequences, user behavior trajectories, and textual records. For example, some studies used LSTM networks to model transaction time series, aiming to detect latent changes in user behavior and improve fraud detection accuracy. Other works applied graph neural networks (GNNs) to capture the relational structure between users and transactions, revealing hidden connections among fraud groups. However, despite these advances, deep learning models still face challenges in semantic comprehension, unified modeling of cross-modal data, and model interpretability. Especially in fast-changing fraud scenarios, their generalization and early warning capabilities remain insufficient.

In recent years, large language models (LLMs), such as GPT, BERT, and their variants, have attracted growing attention in the financial domain due to their remarkable abilities in language understanding and generation. Some studies have explored applying LLMs to tasks such as financial text analysis, sentiment monitoring, intelligent question answering, and contract review, achieving promising results. In fraud detection, initial attempts have used LLMs for semantic analysis of transaction behavior text, intent recognition from customer interaction data, or as embedding generators to support downstream detection models. By semantically encoding multimodal data through pre-trained language models, these approaches enhance model comprehension and partially alleviate problems of data sparsity and imbalance. Meanwhile, some works have combined LLMs with graph or time series models to form hybrid architectures, aiming to improve modeling of complex fraud patterns. However, current research remains in an early stage. There is still a lack of a systematic application framework that covers the entire fraud detection process. In particular, further exploration is needed in areas such as early warning mechanisms, real-time response, and model interpretability. Therefore, expanding the application of LLMs in financial fraud detection and early warning and building an intelligent risk control system that integrates semantic understanding, dynamic modeling, and causal reasoning hold significant research and practical value.

3. Method

Based on the semantic modeling capability of the large language model, this study builds an end-to-end financial fraud identification and early warning framework, which uses transaction behavior logs, text records, and structured account data as input and achieves accurate identification of potential fraudulent behaviors through unified vector encoding, dynamic modeling, and risk score generation. The model architecture is shown in Figure 1.

As shown in Figure 1, the model first uses BERT to perform a unified semantic representation conversion on multimodal input data and extract context-related embedded information. Subsequently, local features are extracted through the CNN module, and the recognition ability of short-term behavior patterns is enhanced by combining operations such as pooling and normalization. Finally, the LSTM network further models time series dependencies and combines Softmax output to achieve classification prediction of fraudulent behavior.

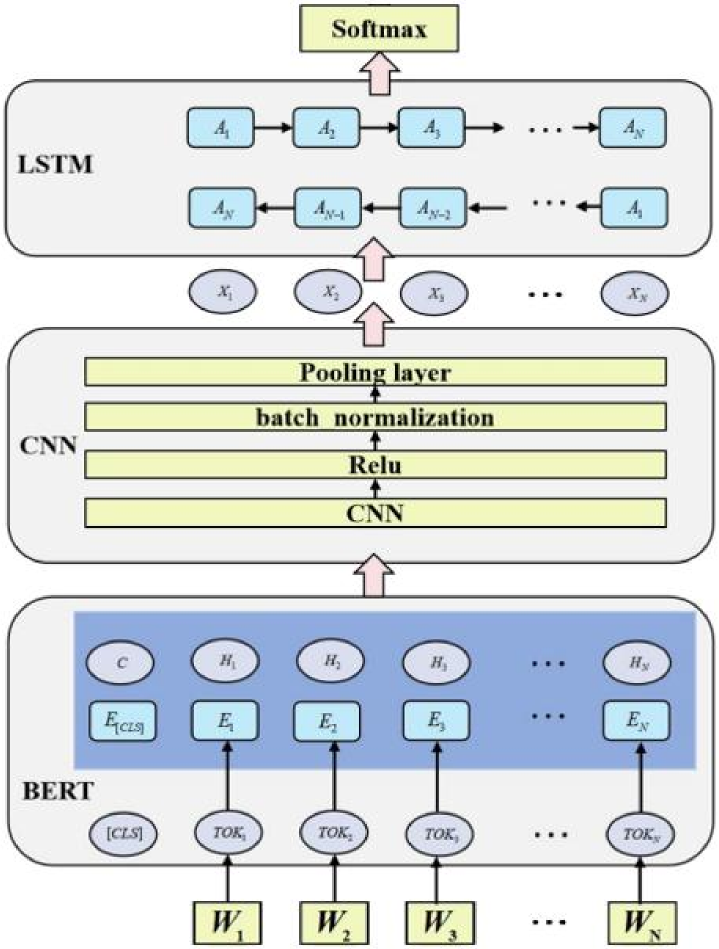

Figure 1. Model network architecture

First, a unified semantic representation conversion is performed on the multimodal input data. Let the original input data be $x = \{x_s, x_t, x_u\}$, which represents structured transaction data, text data, and user behavior sequence respectively. The pre-trained large language model M_θ is used for unified encoding, and its output semantic vector is represented as:

$$h = M_\theta(x) = M_\theta(x_s \oplus x_t \oplus x_u)$$

Where $\oplus$ represents the concatenation operation of multimodal data, and $h \in R^d$ is the high-dimensional semantic vector representation, which is used for subsequent fraud modeling.

Considering that financial fraud may have hidden pattern information in both local features and time series structures, this paper uses an encoder structure combining CNN and LSTM to model multimodal semantic vectors. First, for each transaction fragment representation $h \in R^d$ in the input sequence, we introduce a one-dimensional convolution operation to extract local context features to enhance the perception of short-term fraud signals. Let the convolution kernel be $W_c \in R^{k \times d}$ and the sliding window size be k, then the convolution representation of each time step is:

$$c_t = RELU(W_c \star h_{t-k+1:t} + b_c)$$

Where $\star$ represents a one-dimensional convolution operation, b_c is a bias term, ReLU is an activation function, and the output c_t is the extracted local pattern representation, which helps capture local abnormal features in transaction sequences[6].

After the convolution operation, in order to model the time dependency and dynamic change process of transaction behavior, the long short-term memory network (LSTM) is introduced to perform temporal encoding on the feature sequence. Assuming the CNN output sequence is $\{c_1, c_2, ..., c_T\}$, the state update formula of LSTM is as follows:

$$h_t^{(LSTM)} = LSTM(c_t, \quad h_{t-1}^{(LSTM)})$$

This formula indicates that at each time step t, the model updates the hidden state through the previous state $h_{t-1}^{(LSTM)}$ and the current input c_t, thereby modeling long-term dependent transaction features and outputting the final representation $h_t^{(LSTM)}$ as the semantic basis for fraud discrimination.

In order to achieve binary classification of fraudulent behavior, this paper introduces a risk scoring function f_ϕ to perform nonlinear transformation on the sequence representation, and the output is the fraud probability. Assuming the final representation is $z = h_T^{(LSTM)}$, the output probability is:

$$y' = f_\phi(z) = \sigma(W_o z + b)$$

Where W_o and b are the linear layer parameters, $y' \in [0,1]$ is the Softmax function, and the output value represents the probability that the transaction sequence is identified as fraudulent.

The weighted binary cross entropy loss function is used in the model training phase to alleviate the imbalance problem caused by the scarcity of fraud samples. Assuming that the training sample is $\{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, the main loss function is defined as follows:

$$L_{bce} = -\frac{1}{N}\sum_{i=1}^{N}[y^{(i)} \log y'^{(i)} + (1 - y^{(i)} \log(1 - y'^{(i)}))]$$

This loss is balanced in the probability prediction of positive and negative samples, and gives higher learning sensitivity to abnormal patterns to improve recognition accuracy.

In addition, in order to enhance the discriminability and robustness of the model for transaction representation, this paper introduces a contrastive learning mechanism based on sample semantic perturbation to strengthen the feature distinction between fraudulent and non-fraudulent behaviors. Assume that the positive sample pair is the original representation z^+ and its perturbed version z, and the negative sample set is $\{z_j\}$, then the contrast loss function is defined as:

$$L_{contrast} = -\log \frac{\exp(sim(z^+, z)/\tau)}{\sum_j \exp(sim(z^+, z_j)/\tau)}$$

$sim(\cdot)$ represents the cosine similarity function, and τ is the temperature hyperparameter. This loss encourages the model to bring similar behaviors closer and separate heterogeneous behaviors in the representation space, which helps to improve the stability of the model under boundary samples. The final total loss is a weighted combination of cross entropy loss and contrast loss, which is used to jointly optimize recognition accuracy and semantic discrimination ability.

4. Experiment

4.1 Datasets

The experimental dataset used in this study is the publicly available Credit Card Fraud Detection dataset from the Kaggle platform. It was collected by a European credit card company in 2013 and contains over 280,000 credit card transaction records over a two-day period. In total, the dataset includes 284,807 transactions, among which 492 are labeled as fraudulent. Fraudulent samples account for only 0.172%, indicating a highly imbalanced distribution. This data pattern realistically reflects the sparsity of fraud events in real-world financial scenarios and provides an effective validation platform for building robust fraud detection models.

In terms of feature design, most of the original features in this dataset were processed using Principal Component Analysis (PCA)[7] to preserve user privacy. The final dataset retains 30 input features, including 28 anonymized variables (V1 to V28), one feature representing the transaction amount (Amount), and one indicating the transaction time (Time). Each record is labeled as either “0” or “1”, representing normal and fraudulent transactions, respectively, making it suitable for binary classification tasks. Although the anonymized features are not interpretable, their statistical distributions still preserve key information from the original transactions, making the dataset well-suited for representation learning-based fraud detection research.

This dataset presents typical challenges found in financial risk control tasks, such as high-dimensional sparsity, class imbalance, and anonymized features. Therefore, it is widely used for evaluating the performance of machine learning and deep learning models in the domain of financial fraud detection. In combination with the large language model-based embedding structure and the CNN+LSTM joint modeling strategy used in this study, this dataset enables effective testing of model accuracy and robustness under realistic financial conditions. It also lays a solid data foundation for future model generalization and real-world deployment.

4.2 Experimental Results

This paper first conducts comparative experiments with other deep learning models, and the experimental results are shown in Table 1.

Table 1: Table 1 Performance comparison of different deep learning models in financial fraud identification tasks

Model	ACC	Precision	Recall	F1-Score
GRU[8]	0.970	0.851	0.796	0.822
CNN[9]	0.968	0.842	0.791	0.815
BILSTM	0.973	0.864	0.805	0.833
Transformer	0.978	0.889	0.841	0.864

Ours(Bert+LSTM+CNN)	0.985	0.912	0.873	0.892

As shown in Table 1, the proposed Bert+LSTM+CNN fusion model achieves the best overall performance in the financial fraud detection task. The model reaches an accuracy (ACC) of 0.985, significantly outperforming other baseline models. This indicates stronger stability in overall classification ability. In addition, the model obtains a precision of 0.912 and a recall of 0.873. These results demonstrate that the model can effectively identify fraudulent samples while capturing more true fraud cases with a low false positive rate.

In comparison, the Transformer model performs relatively well across metrics, achieving an F1-score of 0.864. However, its performance remains slightly lower than that of the proposed method. This suggests that relying solely on the self-attention mechanism may be insufficient for capturing local temporal features. Traditional sequential models such as BiLSTM and GRU retain some degree of temporal dependency. Yet, they lack the capacity to deeply model the semantics of multimodal input features. As a result, their overall detection performance is limited, especially in recall, indicating a higher risk of missing fraudulent cases.

Overall, the proposed model integrates a pre-trained language model (Bert) for semantic embedding, CNN for extracting local patterns, and LSTM for modeling sequential dependencies. This combination enables multi-level and multi-perspective feature representation. Such an architecture enhances the model's generalization and practical applicability in financial fraud scenarios characterized by high-dimensional sparsity, semantic complexity, and dynamic patterns. It offers a more reliable technical foundation for financial risk control.

This paper also compares the effects of different language model embedding methods, and the experimental results are shown in Figure 2.

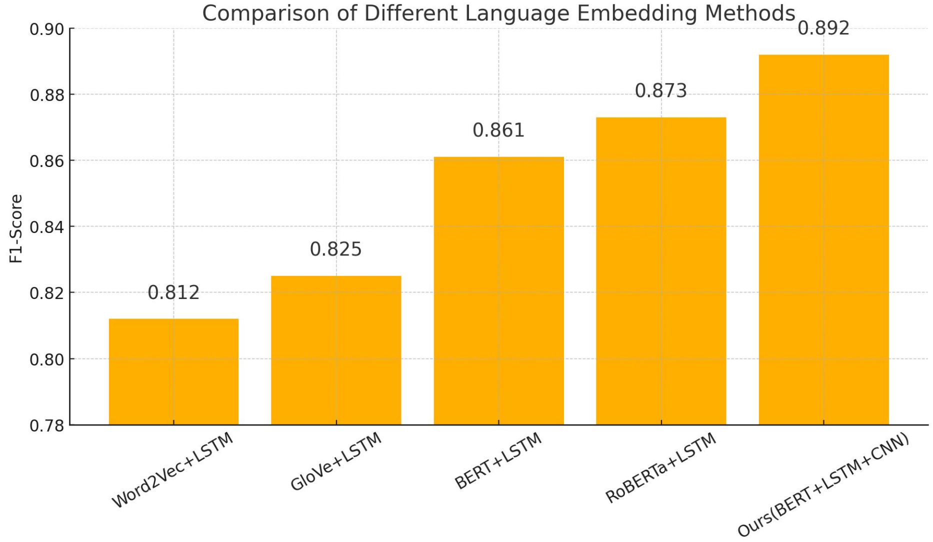

Figure 2. Comparing the effects of different language model embedding methods

As shown in the experimental results in Figure 2, the choice of language model has a significant impact on the performance of financial fraud detection. Traditional embedding methods such as Word2Vec and

GloVe achieve F1-scores of 0.812 and 0.825, respectively. While these methods have certain representation capabilities, they fail to capture dynamic contextual semantics, limiting their generalization in identifying complex fraud behaviors.

With the introduction of pre-trained language models, performance improves significantly. BERT+LSTM and RoBERTa+LSTM reach F1-scores of 0.861 and 0.873, respectively. This validates the effectiveness of context-aware embeddings in understanding financial texts and behavioral data. It indicates that large language models have a natural advantage in modeling semantic dependencies and handling abnormal features, enabling more accurate fraud detection.

Furthermore, the proposed fusion model (BERT+LSTM+CNN) achieves the highest F1-score of 0.892. This shows that incorporating a local convolutional feature extraction module on top of language modeling helps capture subtle anomaly patterns. It enables a synergistic enhancement between semantic understanding and structural modeling. These results demonstrate that multi-module fusion structures offer superior representation and discrimination capabilities when dealing with complex financial data.

This paper also tests the robustness of the model under noise disturbance, and the experimental results are shown in Figure 3.

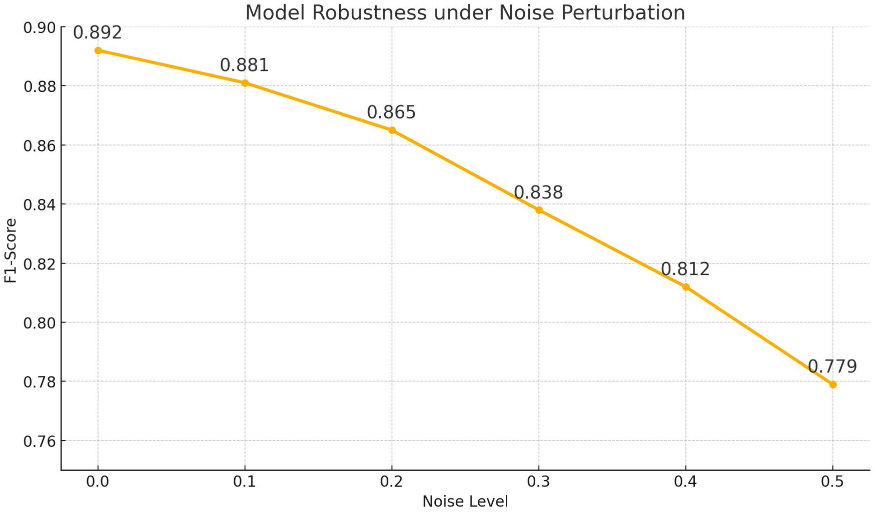

Figure 3. Model robustness test under noise disturbance

As shown in the experimental results in Figure 3, the model's F1-score shows a clear downward trend as the noise level increases. This indicates that noise interference negatively affects the model's ability to distinguish fraudulent transactions. Under noise-free conditions, the model achieves an F1-score of 0.892, which is the best performance. When the noise level increases to 0.5, the F1-score drops to 0.779, representing a decrease of over 11%.

This result suggests that although the model performs well on clean data, the quality of input data still plays a crucial role in real-world applications. In financial scenarios, input data often contain missing values, anomalies, or artificial disturbances. Therefore, the model's stability under high-noise conditions is particularly important. From the trend of the curve, the model maintains relatively high discriminative power under low to moderate noise levels, demonstrating a certain degree of robustness.

Further analysis reveals that this robustness is due to two factors: the multi-layer structure's buffering effect on local abnormal features, and the language model embedding's tolerance to semantic disturbances. However, once the noise level exceeds a certain threshold (e.g., noise ≥ 0.4), the model's performance declines more rapidly. This suggests that in the presence of complex perturbations, misclassifications may still occur. Thus, future work may consider strategies such as data augmentation, robust optimization, or uncertainty modeling to enhance the model's reliability and robustness in deployment scenarios. Finally, this paper gives the comparative experimental results of different sample imbalance processing methods, as shown in Figure 4.

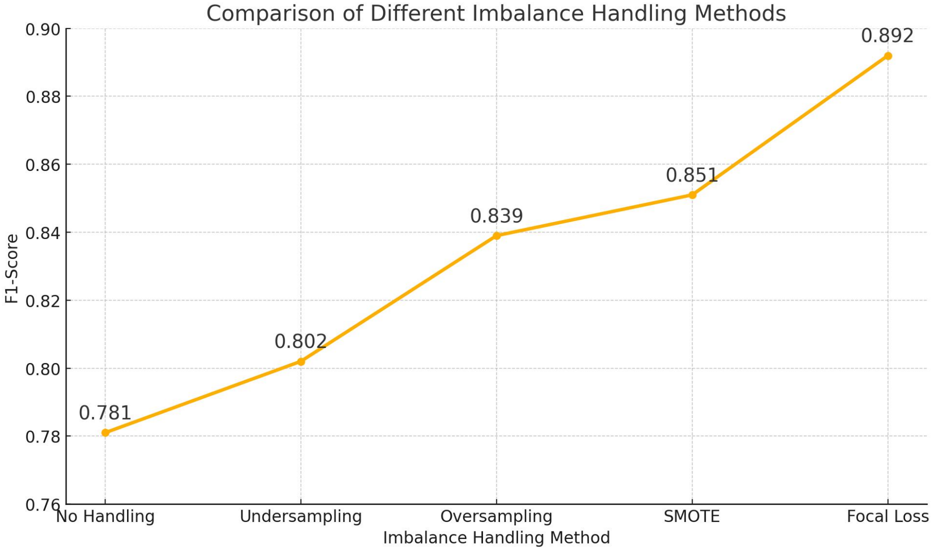

Figure 4. Comparison of Different Imbalance Handling Methods

As shown in Figure 4, sample imbalance handling methods have a significant impact on model performance, especially in terms of F1-score. The baseline model without any handling performs the worst, with an F1-score of only 0.781. This indicates that in financial fraud detection tasks with severe class imbalance, directly training the model fails to effectively identify the minority class of fraud samples, leading to a high risk of missed detection.

After introducing under-sampling and over-sampling strategies, model performance improves to some extent. Among them, over-sampling performs better than under-sampling, with the F1-score increasing to 0.839. This suggests that under the current data distribution, increasing the frequency of fraud samples is more effective for enhancing detection capability. Furthermore, the SMOTE method, based on synthetic data generation, further improves the model's learning on minority classes. Its F1-score rises to 0.851, indicating that introducing more discriminative information while maintaining data distribution stability helps strengthen the model's ability to classify borderline samples.

Notably, the model using the Focal Loss function achieves the highest F1-score of 0.892, outperforming all sampling-based methods. This shows that dynamically adjusting the loss weight of hard and easy samples during training allows the model to focus more on fraud cases. It also suppresses gradient updates dominated by normal samples, thereby achieving better performance in imbalanced scenarios. This result highlights the effectiveness of optimizing imbalance at the loss function level and provides valuable guidance for the design of robust models.

5. Conclusion

This study proposes a deep learning framework for financial fraud detection and early warning, integrating large language model embeddings, convolutional neural networks (CNN), and long short-term memory (LSTM) networks. By fully extracting semantic features and temporal dependencies from transaction data, the proposed model outperforms mainstream baselines across several key performance metrics. It significantly improves both detection accuracy and recall for fraudulent behaviors. Experimental results show that the model demonstrates strong discriminative power and robustness when handling complex, high-dimensional, and imbalanced financial data.

In a series of experiments, including different embedding strategies, architectural comparisons, robustness tests, and class imbalance handling, the proposed approach shows clear advantages. Especially with the introduction of pre-trained language models and contrastive learning mechanisms, the model better understands the semantic logic behind abnormal transactions and captures latent fraud patterns. Meanwhile, CNNs provide complementary modeling of local features. This enhances the model ’ s sensitivity to local anomalies while preserving overall semantic representation, improving its generalization performance. Despite the promising results, there are still some limitations. The model may experience performance degradation under extreme conditions with high semantic and structural noise. Additionally, the current approach relies on static training data and lacks mechanisms to adapt to real-time data streams. Furthermore, the interpretability and visualization of the model remain to be improved. For practical deployment in financial applications, it is still a challenge to provide risk decisions that are auditable and understandable. Future research can be extended in several directions. First, incorporating more efficient incremental learning mechanisms could improve the model ’ s adaptability to emerging fraud patterns. Second, combining graph neural networks to model transaction networks may help uncover hidden relationships in group fraud behaviors. Third, exploring federated learning and privacy-preserving techniques could enable cross-institution fraud detection while protecting user data. These directions may offer more forward-looking solutions for building a trustworthy, intelligent, and efficient financial risk control system.

References

[1] Fanai, H., & Abbasimehr, H. (2023). A novel combined approach based on deep autoencoder and deep classifiers for credit card fraud detection. Expert Systems with Applications, 217, 119562.

[2] Singh, A., & Jain, A. (2020). Cost-sensitive metaheuristic technique for credit card fraud detection. Journal of Information and Optimization Sciences, 41(6), 1319–1331.

[3] Tang, H. (2022). Mixture of experts models for fraud detection: A comprehensive review. arXiv preprint arXiv:2201.12345.

[4] Asudani, D. S., Nagwani, N. K., & Singh, P. (2023). Impact of word embedding models on text analytics in deep learning environment: a review. Artificial Intelligence Review, 56, 10345–10425.

[5] Zhang, Y., & Li, X. (2023). Robust fraud detection via supervised contrastive learning. arXiv preprint arXiv:2308.10055.

[6] Jurgovsky, J., Granitzer, M., Ziegler, K., et al. (2018). Sequence classification for credit-card fraud detection. Expert Systems with Applications, 100, 234–245.

[7] Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research, 61, 863–905.

[8] Zhang, Y., & Li, X. (2023). Credit card fraud detection using advanced transformer model. arXiv preprint arXiv:2406.03733.

[9] Deng, H., & Li, X. (2022). Anomaly detection via reverse distillation from one-class. arXiv preprint arXiv:2201.12345.