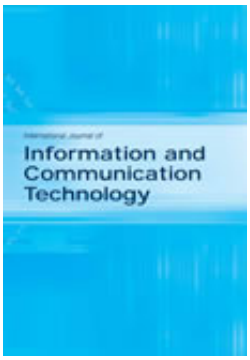
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Financial fraud prediction leveraging knowledge graphs and multimodal features

Jiangyan Cheng

DOI: 10.1504/IJICT.2025.10072942

Article History:

Received:	21 June 2025
Last revised:	14 July 2025
Accepted:	15 July 2025
Published online:	08 September 2025

Financial fraud prediction leveraging knowledge graphs and multimodal features

Jiangyan Cheng

School of Business Management,
Fujian Polytechnic of Information Technology,
Fuzhou 350002, China
Email: 18050280088@163.com

Abstract: Aiming at the problem that existing financial fraud prediction methods have single feature extraction and do not fully utilise multimodal financial features, this paper first constructs a knowledge graph based on multimodal financial data, and uses BERT and ResNet to extract features from text entities and image entities respectively. A bidirectional cross-modal attention mechanism is designed, and a bidirectional text and image element matching method is proposed. The consistency of all element pairs is comprehensively used to represent the global consistency features of text and images. Finally, a hybrid fusion method is designed to fuse text features, image features, and text-image consistency features. The final prediction result is obtained through a fully connected layer. Experimental results show that the proposed model has a prediction accuracy of 92.4%, which is better than the comparative methods and can effectively improve the accuracy of financial fraud prediction.

Keywords: financial fraud prediction; knowledge graph; multimodal features; feature fusion; attention mechanism.

Reference to this paper should be made as follows: Cheng, J. (2025) 'Financial fraud prediction leveraging knowledge graphs and multimodal features', *Int. J. Information and Communication Technology*, Vol. 26, No. 32, pp.37–52.

Biographical notes: Jiangyan Cheng received her Master's degree from Fuzhou University in 2009. She is currently an Associate Professor at the School of Business and Management, Fujian Polytechnic of Information Technology. Her research interests include financial accounting, financial management, and management studies.

1 Introduction

In the context of the global wave of the digital economy, financial data fraud poses a serious threat to the order of the capital market, investors' interests, and the rational allocation of social resources (Hilal et al., 2022). Traditional detection methods based on financial indicator analysis or statistical models face bottlenecks in prediction accuracy due to their inability to capture potential risks in complex relationship networks and unstructured data (Omidi et al., 2019). With the deep development of big data technology, knowledge graphs, with their ability to structurally express entity

relationships, can effectively describe complex network characteristics (Wen et al., 2022). Multimodal technology, by integrating multimodal data such as text and images, can uncover fraud clues in unstructured information. The integration of the two provides a new technical approach for financial data fraud prediction (Wang et al., 2023). However, current research mostly focuses on single-modal data or simple feature fusion, leading to insufficient feature extraction and unsatisfactory results in financial data fraud prediction (Ali et al., 2022). Therefore, building an efficient and accurate financial data fraud prediction model has become a research hotspot.

Early financial fraud prediction models mainly explored indicator variables and established classical statistical models based on these variables. Sylwestrzak (2016) selected representative financial statement indicators such as asset turnover ratio and accounts receivable turnover ratio according to the research hypothesis as influencing factors, and used logistic regression to test and build a financial fraud identification model. Liu (2021) comprehensively considered financial indicators related to financial fraud of listed companies, and proposed a financial fraud prediction model combining Benford's Law and the logistic model, but the prediction error was relatively large. Methods based on statistics often rely on logistic regression, and their predictive performance is usually low. Moreover, most studies use accuracy as the evaluation metric for model performance, which is not sufficient to evaluate financial fraud accurately.

The limitations of the above methods are that when dealing with such large-scale data, they are not only inefficient but also prone to errors. Machine learning algorithms can quickly scan, analyse, and process these data to uncover hidden patterns and regularities. Qin (2021) applied the SVM method to the identification of financial fraud and found that it achieved better identification results than traditional multivariate discriminant analysis on the sample data at that time. Duan et al. (2024) used various ensemble learning algorithms to detect financial fraud, including random forest (Biau, 2012), decision tree (Ying, 2015), and CatBoost algorithm (Ibrahim et al., 2020), and finally found that the random forest performed best among all models. However, Xu et al. (2022) found that AdaBoost performed best on the dataset in Greece. An and Suh (2020) selected data from the Bombay Stock Exchange, used data mining techniques to select important indicators, and studied the random forest model, improving the prediction accuracy. Sibaroni et al. (2020) developed a system text analysis framework for predicting financial fraud, which extracted word-level and document-level features as input for the Liblinear support vector machine, achieving a high prediction accuracy.

Although machine learning algorithms have made significant progress in processing text information, they usually have difficulty in deep semantic understanding when processing text, and text features are not fully utilised. Neural network-based methods can fully explore the deep features of text, improving the prediction effect. Riany et al. (2021) organised financial fraud factors and performed importance ranking, and the results showed that the accuracy of the artificial neural network was higher than that of models such as decision trees and logistic. Karthikeyan et al. (2023) found in their study that the textual information in management discussion and analysis (MD&A) could be used to distort financial statements. They extracted linguistic variables from the MD&A text information and used a convolutional neural network (CNN) to identify financial fraud. GR and P (2024) used BiLSTM to extract text features from MD&A. Financial fraud prediction models involve not only single-text modalities but also multi-modal data such as images. Yang et al. (2023) detected financial fraud by combining financial ratios and multi-modal information from annual reports, using a hierarchical attention network

to achieve modal feature concatenation fusion and obtain fused features, thereby improving the prediction effect. Xia et al. (2022) designed a CNN-LSTM financial fraud prediction model, using TextCNN and LSTM to extract text and image features of finance, respectively, thereby reducing the prediction error. Using a knowledge graph can connect information from different modalities to enhance the ability of knowledge expression. Gong et al. (2024) innovatively constructed a financial knowledge graph and performed embedding representation, combining the Transformer model to achieve financial fraud prediction, achieving good prediction results.

This paper proposes a financial data fraud prediction model based on knowledge graphs and multi-modal features. First, a financial knowledge graph is built based on multi-modal financial data, and text entity vectors and image entity vectors are defined for entities. Then, the Resnet18 and BERT models are used to extract the features of image entities and text entities, respectively. On this basis, a bidirectional cross-modal attention mechanism is designed to perform fine-grained element alignment, to obtain image-text and text-image element pairs, and calculate the image-text consistency features. Then, a hybrid fusion method is used to fuse the text features, image features, and image-text consistency features, and the final prediction result is obtained through the fully connected layer. At the same time, an improved focus loss function with reshaped focus weights was used during model training to mitigate the impact of data imbalance on model performance. Experimental results show that the proposed model achieved at least a 3.2% and 4.3% improvement in prediction accuracy on the training and testing sets, respectively, compared to the baseline method, effectively enhancing its ability to identify financial fraud.

2 Relevant technologies

2.1 Multimodal fusion theory

The value of multimodal information lies in its ability to provide richer and more complete data. Multimodal fusion is a means of promoting the collaborative use of complementary multimodal information. It can improve the model's understanding and decision-making capabilities regarding complex real-world problems, thereby enhancing its generalisation ability. Multimodal fusion strategies are also one of the key issues in multimodal model classification tasks (Poria et al., 2017).

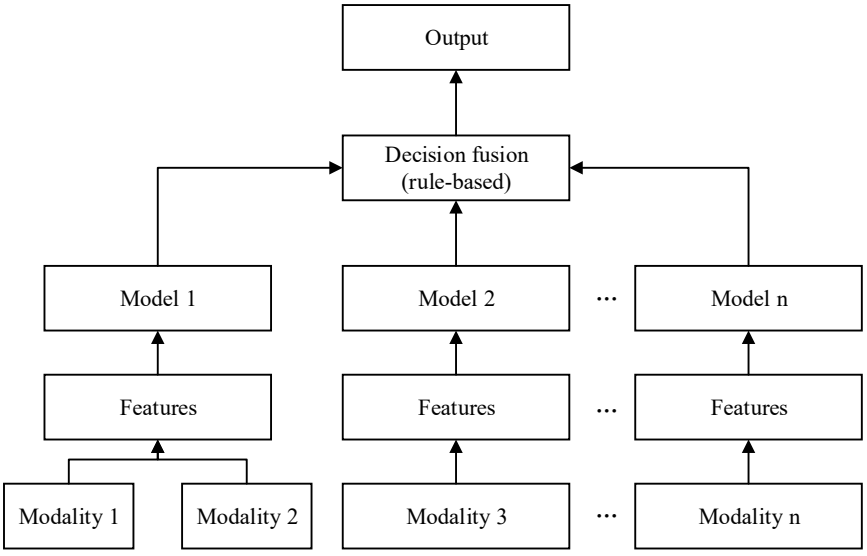
1 Feature layer fusion. Features are extracted using a multimodal feature extractor, and then a joint feature is formed through various fusion methods such as feature vector concatenation.

2 Decision layer fusion. After each single-mode classifier extracts modal features and completes classification independently, the classification results are integrated according to certain rules to achieve modal information fusion (Pinar et al., 2016). This method maintains the independence of each mode, but may amplify the errors of a particular mode classifier during training.

3 Hybrid fusion. After extracting features from certain modalities, they are concatenated and fed into a multimodal classifier. Meanwhile, single-modal classifiers are trained for some modalities, and the classification results are obtained

by applying decision layer fusion rules. This method combines the advantages of both fusion methods. The fusion process of this method is shown in Figure 1.

Figure 1 Blending and merging process



In summary, multimodal fusion strategies include feature-level fusion, decision-level fusion and hybrid fusion. Feature-level fusion takes the raw data of different modalities through preprocessing and directly performs splicing, weighting or other combination operations at the feature level to form a unified feature vector, which is then inputted into the subsequent model for processing. Decision-level fusion first processes and models the data of each modality separately and independently to obtain their respective decision results, and then fuses these decision results according to certain rules to obtain the final decision. Hybrid fusion combines the advantages of feature-level fusion and decision-level fusion to perform multimodal fusion at different levels of the model. For example, feature-level fusion is performed at part of the level, and then decision-level fusion is performed at a higher level. For complex multimodal tasks, where different levels of information need to be considered together, hybrid fusion often achieves better results.

2.2 Knowledge graph

The essence of a knowledge graph is a graph-based semantic network that uses graphs to express real-world concepts, entities, relationships, and events. The basic building block is a semantic triplet, and knowledge is expressed by connecting two nodes with edges. Nodes can be concepts, specific entities, and entity attributes, while edges represent the semantic relationships between two nodes.

The logical structure of a knowledge graph is divided into a schema layer and a data layer (Ji et al., 2021). The schema layer is the core of a knowledge graph. The key to constructing the schema layer lies in defining an ontology library that can represent domain or event knowledge, enabling administrators to standardise concepts and the

relationships between them. The data layer stores a triplet. The patterns for constructing technical architecture are mainly divided into top-down and bottom-up approaches. Top-down refers to first defining the pattern layer, constructing a relatively standardised conceptual hierarchy tree and relationships between concepts, and then filling in the data. This approach is suitable for constructing domain knowledge graphs. Bottom-up involves extracting all entities and relationships from massive amounts of data, organising and summarising them to form a pattern layer. This approach is suitable for constructing public knowledge graphs.

2.3 Attention mechanism

The attention mechanism is a technology that mimics human visual and auditory attention mechanisms to enhance the performance of machine learning models. Its core idea is to enable models to automatically learn which parts of the input are most important, thereby better completing tasks (Niu et al., 2021). Specifically, in deep learning, normalised attention weights must first be obtained, and then the weights are multiplied by the feature vectors to obtain new feature vectors. The weighted sum obtained at this point reinforces important information and suppresses non-important information, allowing subsequent models to focus more on key parts of the information.

The attention mechanism mainly consists of three steps. First, the similarity between the query vector and the key vector is calculated. Then, the similarity is normalised using softmax. Finally, the weighted average of the value vector is the attention value. This is expressed by equation (1) and equation (2) as follows.

$$\alpha_n = \frac{\exp\left(s\left(q, k_n\right)\right)}{\sum_{j=1}^{N} \exp\left(s\left(q, k_j\right)\right)} \tag{1}$$

$$attention(q, k) = \sum_{n=1}^{N} \alpha_n k_n \tag{2}$$

where $s(q, k)$ is the similarity calculation function, α_n is the attention weight, $attention(q, k)$ is the attention vector after weighted summation, q and k are the query vector and key vector respectively.

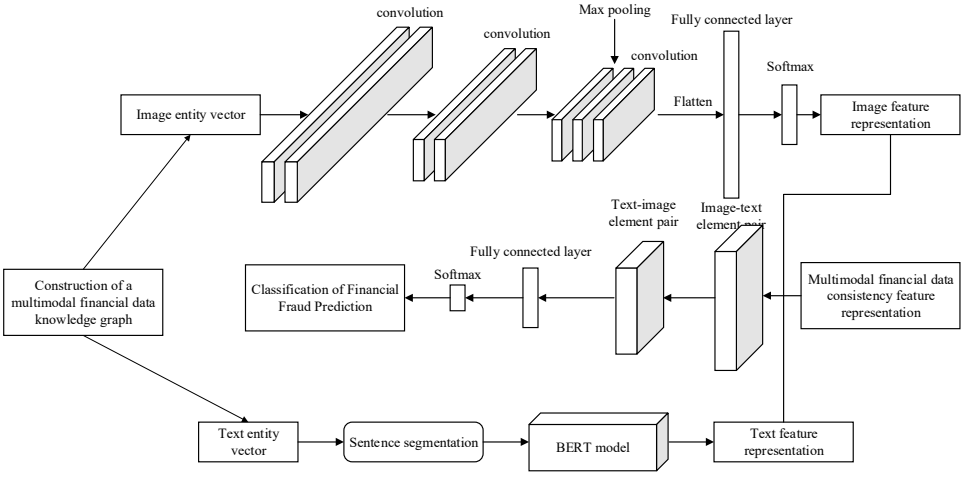
When processing input data, the model receives a large amount of information, but not all of it is equally important to the task at hand. The attention mechanism acts like an intelligent filter that automatically evaluates the importance of each part of the input data and assigns different weights to different parts. For example, in a natural language processing task, for each word in a sentence, the attention mechanism will determine which words are more critical to understanding the meaning of the sentence or accomplishing a specific task based on the context and the needs of the task at hand, so as to give these words higher attention weights. The attention weights generated by the attention mechanism visualise the model's focus of attention when processing the input data. By visualising these attention weights, we can understand how the model makes decisions, thus enhancing the interpretability of the model.

3 Construction of a financial fraud prediction model based on knowledge graphs and multimodal features

3.1 Construction of a multimodal financial data knowledge graph

As introduced in the basic knowledge section, the construction method of a knowledge graph is generally divided into two approaches: top-down and bottom-up, and the choice depends on the actual purpose of the graph construction. Since the confidence level of financial regulations is high and the professionalism is strong, it can be designed in a top-down manner, building the financial knowledge graph framework from the top. Currently, the construction of financial data knowledge graphs is mostly based on the text modality, with a single feature extraction. Therefore, this paper constructs a multimodal financial data knowledge graph.

Figure 2 Structure of financial fraud prediction model



Abstract the financial knowledge graph as $KG = (E, R, T)$, where E is the set of entities in the knowledge graph, R is the set of relations, and T is the set of fact triplets. Each triplet $(h, r, t) \in T$ indicates that there is a relation r between h and t. To introduce various modal information in the knowledge graph to enhance the financial knowledge representation, text feature vectors, image feature vectors, and multimodal feature vectors are defined for entities, and text feature vectors are defined for relations, as follows: h_d, t_d, and r_d represent the text feature vectors obtained by encoding the text descriptions of the head, tail entities, and relations, respectively. h_i and t_i represent the image feature vectors of the head and tail entities, obtained from the images corresponding to the entities through the image feature extractor. h_m and t_m represent the multimodal feature vectors of the head and tail entities, which are formed by fusing the text features and image features of the entities. The model in this paper first complements the entity descriptions through text features and image features, then establishes the connection between the head and tail entities through the text features of the relations, and finally uses the multimodal negative sampling strategy to generate negative samples, which are trained together with positive samples through contrastive learning, so that the model can better explore the semantic information in the correct triplets, thereby enhancing the model's ability to

predict missing entities or relations in the knowledge graph. The model in this paper mainly consists of three parts: multimodal feature extraction, image-text consistency feature representation, and multimodal feature fusion, as indicated in Figure 1.

3.2 Multimodal financial entity feature extraction

To extract features from multiple different modalities of information in order to enhance the representation of the financial knowledge graph and improve prediction performance, this paper selects different models for different modalities of financial information to extract text and image features.

For text information, the BERT pre-trained language model is used as the financial text feature extractor to obtain feature vectors of entities and relations based on text descriptions. BERT uses a bidirectional Transformer structure to explore context information in the text and can effectively complete text feature extraction. For image information, the residual network (ResNet) is selected as the feature extractor to obtain the image features of entities.

1 Text feature representation of financial entities. Compared to general word vector methods, the BERT model considers the information interaction between word vectors in the same sentence when converting randomly encoded word sequences into word vectors with semantic information allowing the model to capture the relationship between each word and the context on both sides, thereby making the text feature representation more semantic, more expressive, and enhancing the model's generalisation ability. Therefore, it is used as the base model for deep semantic feature extraction. To capture global information, BERT first uses an m-head self-attention level to convert each position in the input sequence into a weighted sum of the input level. Specifically, for the i^{th} head of attention, the input is the text of the public welfare crowdfunding project T, which is transformed based on the dot product attention mechanism. Then, the outputs of the m attention mechanisms are concatenated and subjected to a linear transformation. Based on the output of the self-attention level, BERT adds a residual connection from the input to the output, as well as a normalisation level. In addition, a standard feed-forward network with GeLU as the activation function and another residual connection with a normalisation level are stacked on top to generate the output of the first layer of BERT. Finally, the entire BERT model stacks L such layers. This paper assumes that given the financial entity description text $T = \{t_1, t_2, \ldots, t_n\}$, n is the length of the text, and after inputting it into BERT, a deep semantic text representation $E = \{e_1, e_2, \ldots, e_n\}$ is finally obtained, as shown below, where f_{BERT} is the BERT model, whose parameters remain fixed during training, and e_i, $i \in 1, 2, \ldots, n$ is a word vector of 768 dimensions.

$$E = f_{BERT}(T) = \{e_1, e_2, ..., e_n\} \tag{3}$$

2 Financial entity image feature representation. In the financial fraud prediction model, the relevant statistical images in the financial reports contain a large number of financial indicator features. The growth or decline trend of normal enterprise financial indicators (such as revenue, profit, etc.) is usually stable or in line with industry rules. If there is a sudden significant change in the slope, it may indicate fraud. Since the depth of the network affects the performance of many visual

recognition tasks, the deeper the depth, the better the model performance but also more difficult to train. Based on the structure of VGG, scholars have proposed ResNet (Xu et al., 2023), which adds a residual learning module, making it easier to achieve cross-layer information transmission and solving the problem of gradient disappearance. To simplify the model complexity, and considering that the text feature dimension is 768, this paper selects the pre-trained Resnet18 with lower output feature vector dimension to extract image features. Specifically, for an image in a financial report, the purpose of this paper is to use a set of image features $V = \{v_1, v_2, \ldots, v_m\}$ to encode various regions in the image. It is input into Resnet18 for image processing, after a 7 × 7 convolution module, then batch normalisation and Relu activation function are used to improve the model's fitting ability, and finally max pooling operation is performed. Subsequently, the output feature map enters the subsequent residual modules, continuously performs convolution, and finally outputs the feature map, as shown below, where f_{Resnet17} is the pre-trained Resnet18 model, and v_i is the visual region vector.

$$V = f_{Resnet18}(I) = \{v_1, v_2, ..., v_m\} \tag{4}$$

To make the final output dimension of the visual region feature vector consistent with the dimension of the text feature, this paper adds a fully linked level after the last convolutional level of Resnet18. Here, F_v is the image feature vector finally extracted after the feature dimension expansion through the fully linked level. W is the two-dimensional weight matrix parameter of the fully linked level, and b is the bias term. The activation function is LeakyRELU, which allows the network to converge quickly and prevents gradient disappearance.

$$F_v = LeakyRELU(WV + b) \tag{5}$$

4 Multimodal financial data feature fusion and fraud result prediction

4.1 Multimodal financial data consistency feature representation

Aiming at the problem that the current financial fraud model has a single feature extraction, leading to low prediction accuracy, this paper, based on the previous section's multimodal feature extraction, designs a bidirectional image-text and text-image element matching method using the attention mechanism. It uses a fully connected neural network to calculate the consistency of each element pair, and finally integrates all element pairs' consistency to represent the global consistency feature of the image and text. Finally, a hybrid fusion method is used to combine the text features, image features, and image-text consistency features. During the model training process, an improved focal loss function is used to improve the impact of data imbalance on the model, thereby enhancing the prediction effect.

Through the feature extraction steps of the previous section, financial image-text feature vectors of the same dimension have been obtained, and the similarity between them can be calculated. The calculation of image-text consistency in financial data is divided into three stages in this paper. First, a bidirectional attention mechanism is used for image-text element alignment, making the information with higher correlation between modalities more prominent during alignment. Then, fine-grained image-text

consistency is calculated using the aligned image-text and text-image element pairs. Finally, the global consistency feature of financial data is obtained by leveraging local consistency features.

For the attention element alignment of image-text in financial reports, the following steps are taken. For an image I with m regions and a financial text description E with n words, this paper first calculates the cosine similarity matrix of all image-text pairs, as shown in equation (6). Empirically, it is found that setting the similarity threshold to 0 yields better results. Therefore, a normalisation operation is performed on the calculated similarity matrix, as shown in equation (7).

$$s_{ij} = \frac{v_i^T e_j}{\|v_i\| \|e_j\|} \tag{6}$$

$$\overline{s_{ij}} = \frac{S'_{ij}}{\sqrt{\sum_{i=1}^{m} s'^2_{ij}}} \tag{7}$$

where s_{ij} is the similarity between image v_i and text word e_j, m is the number of image visual regions, n is the number of words, s'_{ij} is the similarity after converting negative values of s_{ij} to 0, and $\overline{s_{ij}}$ is the normalised similarity.

To calculate the new text vector v_i after adjusting the word vector weights through the attention mechanism, the attention distribution of v_i for the text vector E must be calculated first. This paper inputs $\overline{s_{ij}}$ into the softmax function to obtain α_{ij}, as shown in equation (8), which describes the relevance between v_i and e_j. Based on α_{ij}, this paper obtains a new reorganised text vector a_i^t through the weighted combination of financial text word vectors, as shown in equation (9).

$$\alpha_{ij} = softmax\left(\overline{s_{ij}}\right) = \frac{\exp\left(\overline{s_{ij}}\right)}{\sum_{j=1}^{n} \exp\left(\overline{s_{ij}}\right)} \tag{8}$$

$$a_i^t = \sum_{j=1}^{n} \alpha_{ij} e_j \tag{9}$$

Similarly, for the attention element alignment of text-image in financial reports, the steps are consistent with the attention element alignment of image-text. This paper first calculates the cosine similarity matrix of all image-text pairs, resulting in s_{ij}. Then, a normalisation operation is performed on the calculated similarity matrix, resulting in the normalised similarity $\overline{s_{ij}'}$. To calculate the new image visual vector e_j after processing through the attention weights, the attention distribution of e_j for the visual vector F_v must be calculated first, followed by the weighted combination. The visual vector is used to represent the reorganised financial visual vector through the attention mechanism. For the attention distribution, this paper inputs $\overline{s_{ij}'}$ into the softmax function to obtain α'_{ij}, as shown in equation (10). Based on α'_{ij}, this paper obtains a new reorganised visual vector a_j^v through the weighted combination of visual vectors, as shown in equation (11).

$$\alpha'_{ij} = softmax\left(\overline{s_{ij}'}\right) = \frac{\exp\left(\overline{s_{ij}'}\right)}{\sum_{i=1}^{m} \exp\left(\overline{s_{ij}'}\right)} \tag{10}$$

$$a_j^v = \sum_{i=1}^{m} \alpha'_{ij} v_i \tag{11}$$

On the image-text side, this paper concatenates v_i and a_i^t one by one. Subsequently, the concatenated vector is input into a fully connected network, and the network finally performs a nonlinear mapping through the Sigmoid activation function to obtain the image-text matching pair's consistency between 0 and 1, as shown in equation (13). m image-text matching pairs $\{v_i, a_i^t\}$ can obtain the similarity matrix S^{vt}.

$$s_i^{vt} = \sigma\left(W_l\left(v_i \oplus a_i^t\right)\right) \tag{12}$$

where $\oplus$ is the concatenation operation, W_l is the weight coefficient of the fully connected network, σ is the Sigmoid activation function, s_i^{vt} is the consistency calculated by the fully connected network for $\{v_i, a_i^t\}$.

Similarly, on the text-image side, this paper concatenates e_j and a_j^v one by one. Subsequently, the concatenated vector is input into a fully connected network, and the network finally performs a nonlinear mapping through the Sigmoid activation function to obtain the text-image matching pair's consistency between 0 and 1, as shown in equation (14). n text-image matching pairs $\{e_j, a_j^v\}$ can obtain the similarity matrix S^{tv}.

$$s_j^{tv} = \sigma\left(W_2\left(e_j \oplus a_j^v\right)\right) \tag{13}$$

where $\oplus$ is the concatenation operation, W_2 is the weight coefficient of the fully connected network, σ is the Sigmoid activation function, s_j^{tv} is the consistency calculated by the fully connected network for $\{e_j, a_j^v\}$.

If the text-image consistency matrices S^{vt} and S^{tv} obtained from the two-branch network are concatenated, the final text-image consistency representation S is obtained. S is the global text-image consistency feature of the financial report, which is obtained by combining the modality information of the attention mechanism through bidirectional interaction at the fine-grained level.

4.2 Multimodal feature fusion

The hybrid fusion method combines the advantages of feature-level fusion and decision-level fusion. Therefore, this paper uses two methods, feature fusion and decision fusion, to integrate the above-mentioned text features, image features, and text-image consistency features.

First, the three feature vectors, text features E, image features F_v, and text-image consistency features S, are directly concatenated to obtain a fused vector. This vector is then input into a fully connected layer with a softmax activation function to predict the

single-modal financial fraud result, as shown in equation (15), where W_3 is the weight of the last fully connected network, and P is the prediction result of a single modality.

$$P = Softmax\left(W_3\left(E \oplus F_v \oplus S\right)\right) \tag{14}$$

Then, a fully connected network is used to consider the weights of each single-modal classifier for the final financial fraud prediction. First, E, F_v, and S from the financial report are input into the fully connected layer classifier with softmax to obtain the results P_e, P_v, and P_s, respectively. Then, the average of the classification results is calculated, and the results are combined into a multimodal classification prediction vector P, which represents the fraudulence of the public crowdfunding project, as shown in equation (16).

$$P = \frac{\left(P_e + P_v + P_s\right)}{3} \tag{15}$$

4.3 Loss function

Previous studies often use the cross-entropy loss function (Li et al., 2020). Since the dataset used in this paper is an imbalanced dataset, the issue of data imbalance needs to be considered. The focal loss function (Tian et al., 2022), which evolves from the cross-entropy loss function, is suitable for class imbalance and can adjust the proportion of samples of each class. At the same time, the focal loss function can also improve the classification accuracy of difficult-to-distinguish samples, making the loss function training focus on difficult-to-distinguish samples and ignore easy-to-distinguish samples. This paper proposes an adaptive focal loss function (AFL), as shown in equation (16). Compared to the focal loss function, AFL further increases the weights of difficult and easy samples, making difficult-to-distinguish samples take up more weight, while the loss contribution of easy-to-distinguish samples decreases, to adapt to extremely imbalanced situations.

$$AFL\left(p_t\right) = -\beta_t \cdot w' \cdot log\left(p_t\right) = \begin{cases} -\beta_t \cdot \left(1 - \frac{1}{2}\left(\frac{p_t}{1-p_t}\right)^{\gamma}\right) \cdot log\left(p_t\right), & p_t < 0.5 \\ -\beta_t \cdot \left(\frac{1}{2}\left(\frac{1-p_t}{p_t}\right)^{\gamma}\right) \cdot log\left(p_t\right), & p_t \geq 0.5 \end{cases} \tag{16}$$

where p_t is the probability of each sample being assigned to the true class, β_t is the class sample proportion adjustment coefficient, w' represents the improved focal weight in AFL, w' is a piecewise symmetric function, γ is used to scale the adjustment effect of the focal weight. Moreover, $\alpha_t \in [0, 1]$, $\gamma > 0$, γ increase, the weight of difficult-to-distinguish samples increases to a greater extent, while also reducing the attention of the loss function to easy-to-distinguish samples to a greater extent.

5 Experimental results and analyses

Since financial fraud reviews are usually lagging, financial fraud behaviours in recent years may not have been discovered yet. Therefore, this paper selects the financial fraud

behaviours of listed companies in the CSMAR database (Wang and Wang, 2022) from 2015 to 2020 as the dataset. For the financial annual reports of listed companies, web crawlers are used to scrape and integrate text and image data from securities information websites such as Eastmoney. After data preprocessing, the dataset available for experiments contains a total of 12,504 rows of annual reports of listed companies, among which 12,326 rows have no financial fraud behaviours, while 178 rows have financial fraud behaviours. This paper sets the ratio of training set, validation set, and test set as 7:2:1. The software configuration environment for the experiments in this chapter is Python 3.7, where the Resnet18 network used is provided by the TorchVision library, and BERT is provided by the Transformer library of Huggingface. The learning rate of the model is set to 0.001, the batch size is set to 4, and the optimiser selects the Adam optimiser. The Pytorch framework is used for the overall model construction.

Figure 3 The scatter plot of the true values and predicted values in KG-HAM (see online version for colours)

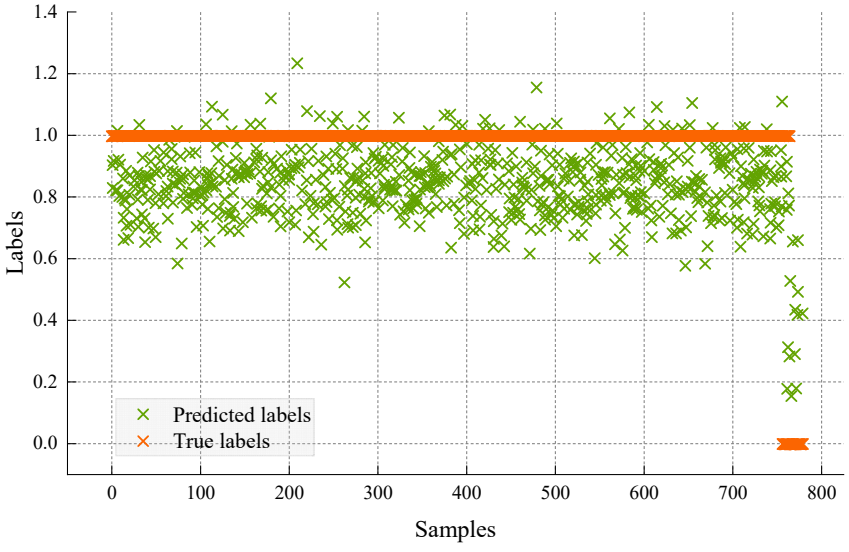
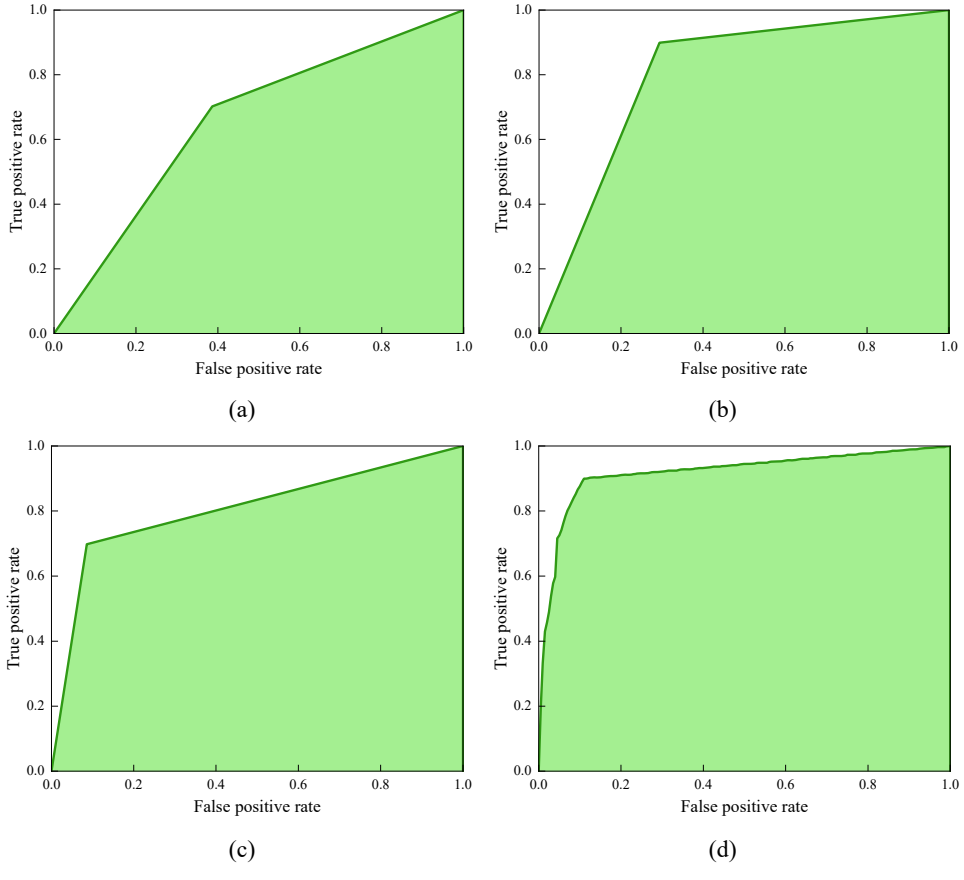


Figure 3 shows the scatter plot of the true values and predicted values in the proposed model KG-HAM. In the figure, the horizontal axis represents the sample number, and the vertical axis represents the predicted label and the true label. Blue dots represent the predicted labels of the model, and orange dots represent the true labels. From the scatter plot, it can be observed that the predicted scores of positive class samples are mostly concentrated at 0. Between 0.8 and 1, indicating that the model's prediction effect on positive class samples is good, but it also has cases where it identifies them as fraud samples. This indicates that KG-HAM sacrifices some prediction effect on positive class samples to gain better prediction effect on fraud samples. Most negative class samples can be accurately predicted, but the prediction scores of negative class samples are more dispersed. Overall, although the number of negative class samples is small, their prediction score distribution is more scattered, and some negative class samples have higher scores. However, the model's prediction scores on the samples accurately reflect its good prediction ability for various types of samples.

Figure 4 The ROC curves of various models, (a) BiLSTM model (b) CNN-LSTM model (c) KG-TRANS (d) KG-HAM (see online version for colours)



To further verify the prediction performance of KG-HAM, this paper selects BiLSTM (GR and P, 2024), CNN-LSTM (Xia, et al., 2022), and KG-TRANS (Gong et al., 2024) as baseline methods. Evaluation metrics include accuracy (ACC), F1 value, and AUC value. The ACC and F1 of different methods on the training set and test set are shown in Table 1. On the training set and test set, the ACC of KG-HAM are 0.947 and 0.924, respectively, which are at least 3.2% and 4.3% higher than the baseline methods. The F1 of KG-HAM are 0.962 and 0.905, respectively, which are at least 6.1% and 3.1% higher than the baseline methods.KG-HAM matches the elements with strong text-image correlation in the financial reports, achieves sufficient inter-modal information interaction through the attention mechanism, and calculates the text-image consistency based on this, which can more accurately and comprehensively measure the differences between text and image, helping to distinguish between normal and fraudulent samples. At the same time, for financial reports that contain both text and images, the model that integrates all financial features achieves the best performance among all experimental models. The experiments in this paper prove that the semantic consistency features of text and image in financial reports are effective, and the use of multi-modal ideas can effectively improve the performance of financial fraud prediction models.

Table 1 Predictive performance of different models

Model	*Training set*		*Test set*	
	ACC	*F1*	*ACC*	*F1*
BiLSTM	0.836	0.819	0.801	0.792
CNN-LSTM	0.883	0.861	0.859	0.848
KG-TRANS	0.915	0.901	0.881	0.874
KG-HAM	0.947	0.962	0.924	0.905

The ROC curves of various models are shown in Figure 4. BiLSTM shows a robust performance on the false positive rate (FPR), but performs poorly on the true positive rate (TPR), which indicates that although the model can provide a high prediction accuracy, it has insufficient ability to identify fraudulent samples and cannot effectively predict fraudulent samples. Similarly, although the TPR of CNN-LSTM exceeds 90%, it shows its shortcomings in the FPR metric, which indicates that although the model can predict fraudulent samples well, it will also identify a large number of non-fraudulent samples as fraudulent samples, which will also lead to serious problems. In KG-TRANS, TPR and FPR show a more balanced result, and its overall AUC value is higher, which indicates that the model performs the best among several basic models. KG-HAM performs excellently in the financial fraud task, with an AUC value of 0.9117. This means that the model can effectively classify in different contexts and distinguish between fraudulent and non-fraudulent samples.

6 Conclusions

Financial data fraud not only leads to distorted accounting information but also harms the healthy development of the economy. Therefore, predicting financial fraud has important practical value. To address the current research issue of insufficient utilisation of multimodal financial features, resulting in suboptimal prediction accuracy, this paper proposes a financial data fraud prediction model based on knowledge graphs and multimodal features. A financial knowledge graph was constructed based on multimodal financial data, and text entity vectors and image entity vectors were defined for entities. Text entity features were extracted using the pre-trained language model BERT, while Resnet18 was used to extract regional visual features of image entities. Based on this, a bidirectional cross-modal attention mechanism is designed to perform fine-grained feature alignment, thereby obtaining image-text and text-image feature pairs, and calculating image-text consistency features. Then, a hybrid fusion method is used to fuse text features, image features, and image-text consistency features, and a fully connected layer is used to obtain the final prediction results. During model training, an improved focus loss function is used to mitigate the impact of data imbalance on the model. The experimental results show that the prediction accuracy of the proposed model is 94.7% and 92.4% on the training set and test set, respectively, significantly improving the prediction accuracy. It provides new technical means and theoretical support for corporate financial supervision and investor decision-making, and has important practical significance for maintaining the order of the financial market.

Declarations

All authors declare that they have no conflicts of interest.

References

Ali, A., Abd Razak, S., Othman, S.H., Eisa, T.A.E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H. and Saif, A. (2022) 'Financial fraud detection based on machine learning: a systematic literature review', *Applied Sciences*, Vol. 12, No. 19, pp.37–45.

An, B. and Suh, Y. (2020) 'Identifying financial statement fraud with decision rules obtained from modified random forest', *Data Technologies and Applications*, Vol. 54, No. 2, pp.235–255.

Biau, G. (2012) 'Analysis of a random forests model', *The Journal of Machine Learning Research*, Vol. 13, pp.1063–1095.

Duan, W., Hu, N. and Xue, F. (2024) 'The information content of financial statement fraud risk: an ensemble learning approach', *Decision Support Systems*, Vol. 182, pp.11–24.

Gong, J., Wang, Y., Xu, W. and Zhang, Y. (2024) 'A deep fusion framework for financial fraud detection and early warning based on large language models', *Journal of Computer Science and Software Applications*, Vol. 4, No. 8, pp.25–31.

GR, J. and P, A.I. (2024) 'Attention layer integrated BiLSTM for financial fraud prediction', *Multimedia Tools and Applications*, Vol. 83, No. 34, pp.80613–80629.

Hilal, W., Gadsden, S.A. and Yawney, J. (2022) 'Financial fraud: a review of anomaly detection techniques and recent advances', *Expert Systems with Applications*, Vol. 193, pp.16–29.

Ibrahim, A.A., Ridwan, R.L., Muhammed, M.M., Abdulaziz, R.O. and Saheed, G.A. (2020) 'Comparison of the CatBoost classifier with other machine learning methods', *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 11, pp.738–748.

Ji, S., Pan, S., Cambria, E., Marttinen, P. and Yu, P.S. (2021) 'A survey on knowledge graphs: Representation, acquisition, and applications', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 33, No. 2, pp.494–514.

Karthikeyan, T., Govindarajan, M. and Vijayakumar, V. (2023) 'An effective fraud detection using competitive swarm optimization based deep neural network', *Measurement: Sensors*, Vol. 27, pp.32–44.

Li, L., Doroslovački, M. and Loew, M.H. (2020) 'Approximating the gradient of cross-entropy loss function', *IEEE Access*, Vol. 8, pp.111626–111635.

Liu, X. (2021) 'Empirical analysis of financial statement fraud of listed companies based on logistic regression and random forest algorithm', *Journal of Mathematics*, Vol. 20, No. 1, pp.38–47.

Niu, Z., Zhong, G. and Yu, H. (2021) 'A review on the attention mechanism of deep learning', *Neurocomputing*, Vol. 452, pp.48–62.

Omidi, M., Min, Q., Moradinaftchali, V. and Piri, M. (2019) 'The efficacy of predictive methods in financial statement fraud', *Discrete Dynamics in Nature and Society*, Vol. 2, No. 1, pp.49–56.

Pinar, A.J., Rice, J., Hu, L., Anderson, D.T. and Havens, T.C. (2016) 'Efficient multiple kernel classification using feature and decision level fusion', *IEEE Transactions on Fuzzy Systems*, Vol. 25, No. 6, pp.1403–1416.

Poria, S., Cambria, E., Bajpai, R. and Hussain, A. (2017) 'A review of affective computing: from unimodal analysis to multimodal fusion', *Information Fusion*, Vol. 37, pp.98–125.

Qin, R. (2021) 'Identification of accounting fraud based on support vector machine and logistic regression model', *Complexity*, Vol. 7, No. 1, pp.55–63.

Riany, M., Sukmadilaga, C. and Yunita, D. (2021) 'Detecting fraudulent financial reporting using artificial neural network', *Journal of Accounting Auditing and Business*, Vol. 4, No. 2, pp.60–69.

Sibaroni, Y., Ekaputra, M.N. and Prasetiyowati, S.S. (2020) 'Detection of fraudulent financial statement based on ratio analysis in Indonesia banking using support vector machine', *Jurnal Online Informatika*, Vol. 4, pp.185–194.

Sylwestrzak, M. (2016) 'Application of logistic regression to detect the fraudulent financial statements', *European Management Studies*, Vol. 14, No. 63, pp.89–102.

Tian, Y., Su, D., Lauria, S. and Liu, X. (2022) 'Recent advances on loss functions in deep learning for computer vision', *Neurocomputing*, Vol. 497, pp.129–158.

Wang, G., Ma, J. and Chen, G. (2023) 'Attentive statement fraud detection: distinguishing multimodal financial data with fine-grained attention', *Decision Support Systems*, Vol. 167, pp.13–27.

Wang, J. and Wang, D. (2022) 'Corporate fraud and accounting firm involvement: evidence from China', *Journal of Risk and Financial Management*, Vol. 15, No. 4, pp.12–21.

Wen, S., Li, J., Zhu, X. and Liu, M. (2022) 'Analysis of financial fraud based on manager knowledge graph', *Procedia Computer Science*, Vol. 199, pp.773–779.

Xia, H., Zhou, Y. and Zhang, Z. (2022) 'Auto insurance fraud identification based on a CNN-LSTM fusion deep learning model', *International Journal of Ad Hoc and Ubiquitous Computing*, Vol. 39, Nos. 1–2, pp.37–45.

Xu, H., Fan, G. and Song, Y. (2022) '[Retracted] application analysis of the machine learning fusion model in building a financial fraud prediction model', *Security and Communication Networks*, Vol. 20, No. 4, pp.15–32.

Xu, W., Fu, Y-L. and Zhu, D. (2023) 'ResNet and its application to medical image processing: research progress and challenges', *Computer Methods and Programs in Biomedicine*, Vol. 24, pp.51–63.

Yang, J., Chen, K., Ding, K., Na, C. and Wang, M. (2023) 'Auto insurance fraud detection with multimodal learning', *Data Intelligence*, Vol. 5, No. 2, pp.388–412.

Ying, L. (2015) 'Decision tree methods: applications for classification and prediction', *Shanghai Archives of Psychiatry*, Vol. 27, No. 2, pp.13–20.