Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 221 (2023) 609–616

Procedia Computer Science

www.elsevier.com/locate/procedia

Information Technology and Quantitative Management (ITQM 2023)

HY-RISE: Towards Risk Identification Learning from Massive Scientific Economic Activities

Weiwei Zhang[a,b], Wenjing Chang[a,b], Jianjun Yu[a,*], Fangyu Liao[a]

[a] *Computer Network Information Center, Chinese Academy of Sciences, Beijing,100083, China*
[b] *University of Chinese Academy of Sciences, Beijing,100080, China*

Abstract

Scientific economic activities document the utilization of research funds, which form a critical component of scientific research. Detecting potential risk behaviors from scientific economic activities is crucial to risk management for research institutions. Most of the existing attempts, however, tackle the problem with traditional machine learning algorithms, which rely on the manual feature extraction. Undoubtedly, these methods cannot extract complex semantic features or fuse information from hybrid data for effective risk identification. To overcome these challenges, in this paper, we propose a novel **R**isk **I**dentification model for **S**cientific **E**conomic activities from **HY**brid data (**HY-RISE**), which incorporates both textual and structured data. Firstly, we use a pretrained BERT module to capture the semantic information from textual data. After that, we introduce a BiGRU module to augment the contextual information in semantic embeddings. Finally, we use a shallow neural network to fuse the augmented semantic representation with other discrete features to obtain the final representation. Experimental results on the real reimbursement dataset demonstrate that HY-RISE apparently outperforms existing models in terms of effectiveness and robustness for risk identification.

Keywords: BERT, BiGRU, feature fusion, risk identification

1. Introduction

In China, with the rising investment in research funding, the government has become more concerned on the supervision of research funding. In light of the ”reform of government functions”, research institutions are encountering new challenges in risk management. How to design risk identification methods to analyze data generated by scientific economic activities and further identify risk behaviors that deviate from expected patterns have become an urgent concern for research institutions.

Currently, risk identification methods that have been widely used include rule engine-based [1], statistical learning-based [2] and machine learning-based [3] methods. Nevertheless, real-world scientific economic activities generate diverse unstructured data, such as text, images, and graph-structured data, which surpass the representation capability of aforementioned methods. Several examples of risks identified in real-world scientific economic activities are shown in Table 1. Considering the comprehensiveness of risk identification, it is crucial to extract deep information from unstructured data by combining it with other structured data.

*Corresponding author
E-mail address: yujj@cnic.ac.cn.

10.1016/j.procs.2023.08.029

Table 1. Examples of risks found in real-world scientific economic activities.

Risk Types	Basis for being identified as risk data	Risk cases
Authenticity risk	The ”Reason for reimbursement” is inconsistent with other reimbursement information.	The user completed the labor expense reimbursement form with ”Wan Wu and Li Si’s mid-term defense evaluation fee” in the ”Reason for Reimbursement” column. However, they selected ”Other” instead of the expected selection of ”Expert” in the ”Personnel type” column. This discrepancy needs to be addressed.
	Inconsistency between ”Reason for reimbursement” and ”Type of reimbursement”.	The user incorrectly selected the reimbursement type as ”General Reimbursement” instead of ”Consumables Reimbursement”, even though they filled in the ”Reason for Reimbursement” column with ”Functional Department 2020 Office Supplies Reimbursement”.
Compliance risk	The amount that an individual can be reimbursed does not comply with the rules.	In a certain hospitality expense reimbursement form, the number of guests is 1 and the number of accompanying persons is 3. However, the catering expenses are ¥1200 , which poses a risk of exceeding personal meal expenses.
	Exceeding the deadline for reimbursement.	The conference expense reimbursement receipt indicates that the date of reimbursement is August 21, 2019, while the conference ended on March 28, 2019. This violates the rule that requires reimbursements to be made within three months after the meeting end date.

To improve the representation ability in complex scenarios, numerous risk identification methods based on deep learning have been proposed and shown impressive progress in various domains, including social network analysis [4] and spam classification [5]. However, when applying these methods to risk identification for scientific economic activities, two problems arise: firstly, these methods mainly focus on a single type of data at a time, which may result in overlooking risks hidden in other types of data; secondly, they have insufficient ability to extract the complex semantic information from textual data. As illustrated by Case 1 in Table 1, the identification of risk requires a comprehensive understanding of the content in the ”Reason for reimbursement” column in conjunction with the ”Personal type” column. Nevertheless, traditional risk identification methods are not effective in extracting deep semantic relations from textual data. In recent years, deep learning methods, such as Bidirectional Encoder Representation from Transformers (BERT) and Long-Short Term Memory network (LSTM), have been applied for risk identification with promising results. For example, Mohamed [6] created a novel financial distress prediction model to determine whether a company encounters financial distress. Oluwaseun [7] proposed a framework that detects fake news messages from Twitter posts using a hybrid of convolutional neural networks (CNN) and LSTM models. Additionally, deep learning methods are also widely used in financial fraud detection [8] and credit scoring [9]. These methods exhibit superior performance in processing hybrid data and modeling textual data, which provides research ideas for risk identification in scientific economic activities.

Motivated by the aforementioned studies, in this paper, we propose a novel **R**isk **I**dentification model for massive **S**cientific **E**conomic activities from **HY**brid data (**HY-RISE**). Firstly, we collect 18,708 reimbursement receipts from research institutions containing information about reasons for reimbursement, receipt type, personnel type and amount (risk or no risk). By combining these data characteristics, we design the HY-RISE, which utilizes the strengths of BERT and BiGRU to obtain augmented semantic representations and shallow neural networks to fuse hybrid data. The combination of BERT and BiGRU aims to obtain augmented semantic representations after the semantic extraction by BERT that are more effective in identifying risks. Then, we fuse these augmented semantic representations with other structured data to obtain more comprehensive representations. Finally, we feed the fused representations into a neural network, and the classification result is output through a sigmoid function.

The main contribution of this paper can be summarized as follows:

- We categorize risks in scientific economic activities and identify risks in real-world scenarios, which are essential for effective risk management in research institutions.
- We propose HY-RISE, a novel risk identification method for scientific economic activities from hybrid data, which fuses the augmented semantic information and multiple structural features to obtain comprehensive representations.
- Extensive experiments are conducted on a real-world reimbursement dataset to verify the effectiveness of the proposed model. Experimental results demonstrate that HY-RISE outperforms compared baselines by a large margin for risk identification.

2. Related work

This work is closely related to risk identification based on deep learning, which is typically conducted in three dimensions: text-based, series-based, and graph-based risk identification. In this section, we provide a brief review of related works in these areas.

2.1. Text-based risk identification

Text mining is the process of extracting important information and topics from textual data. In the past, risk identification methods based on text mining used language models to identify patterns in the corpus and differentiate between positive and negative samples. However, Sumei's approach utilizes a BERT pre-training model for word embedding of financial texts, followed by a cascading attention mechanism that extracts relevant information at both word and sentence levels [10]. This method has an advantage over previous approaches as it automatically filters out irrelevant information, making it more suitable for large-scale financial text analysis. Additionally, this model fine-tunes pre-training tasks to accurately convey semantic information from original financial text. Bernhard [11] proposes another model that identifies emotional values in financial text data using a pre-trained bi-directional LSTM network combined with transfer learning to compute sentiment. Experimental results show that this model consistently outperforms traditional machine learning models for sentiment classification.

2.2. Series-based risk identification

Time-series mining focuses on the distribution of features and behavioral performance of the data in terms of time series, such as the time period in which the behavior is generated. Si [12] improves the deep neural network (DNN) algorithm to predict bitcoin prices, which extracts and uses bitcoin-related features to achieve price prediction. Loannis [13] proposes a new deep learning prediction model that accurately predicts gold prices and trends. This model supports financial investors and central banks in making informed decisions about their investment policies while reducing potential risks. The model uses a convolutional neural network (CNN) layer to learn the patterns of time-series data and a long and short-term memory network (LSTM) layer to identify the long and short-term dependencies of modeled time-series data. Similarly, Hasan [14] applies the CNN-LSTM network structure to the task of identifying electricity theft. The model can accurately identify electricity theft and is able to classify most classes (common users) and few classes (power theft users).

2.3. Graph-based risk identification

Yulong [15] proposes a framework for detecting subgraph anomalies in financial transaction networks to identify collective fraudulent behavior. The circular structure of the subgraph poses a risk of money laundering, while the tree-like structure poses a risk of pyramid schemes. Yingtong [16] designs the CARE-GNN model by combining label-aware similarity-based measures and reinforcement learning techniques with graph neural network GNN. This enhances the process of GNN information aggregation and demonstrates its effectiveness in fraud detection on real user-review datasets. Qiwei [17] discovers a pattern in fund transfer within credit card payment businesses and proposed a multi-view heterogeneous attribute network detection model. Experiments conducted on Alibaba's real dataset have demonstrated that this model outperforms others, making it crucial for identifying and managing risks associated with Internet financial institutions.

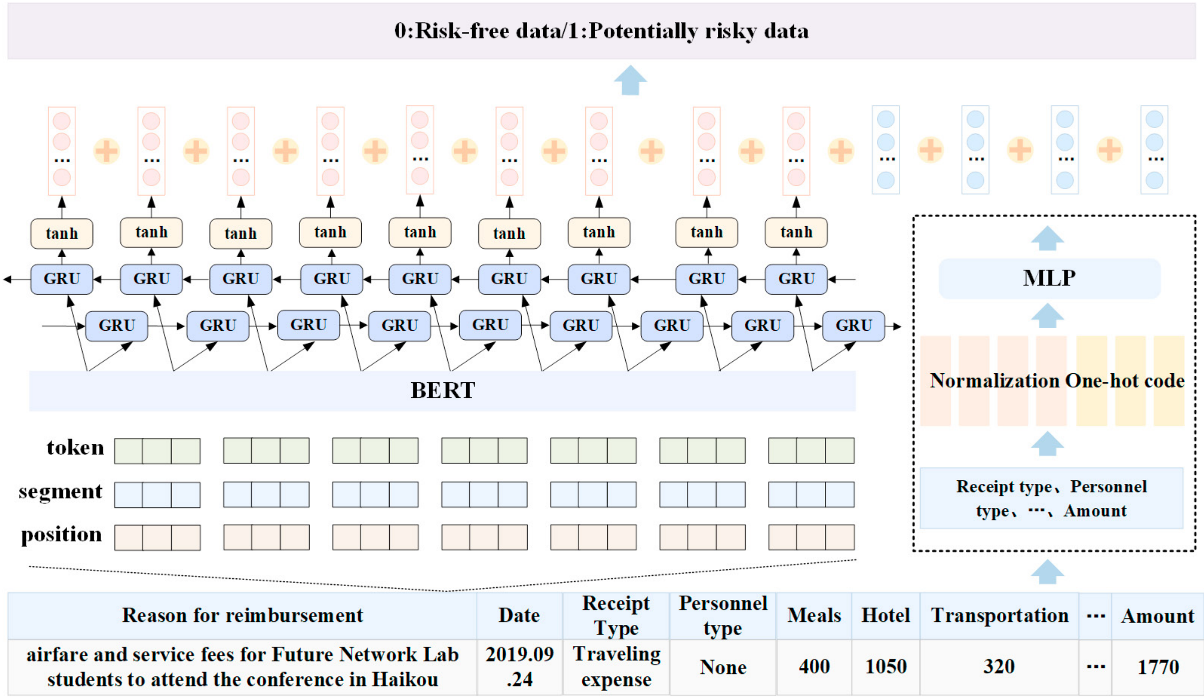

Fig. 1. The overall framework of our proposed model for scientific economic activities from hybrid data.

3. Problem definition

Given a set of scientific economic activities $\mathcal{D} = \{d_1, d_2, \cdots, d_n\}$, which consists of n instances, each instance $d_i = (text_i, numerical_i, category_i)$ contains multiple types of characteristics for each scientific and economic activity, where $text_i$ denotes textual features, $numerical_i$ denotes continuous features, and $category_i$ denotes discrete features. $y_i \in (0, 1)$ is a binary value, which takes value 1 if the corresponding economic activity d_i is at risk and 0 otherwise. Risk identification for scientific economic activities aims to detect risk behaviors that significantly differ from the normal behaviors with similar purposes, which can be formulated as an imbalanced binary classification problem:

$$\mathcal{F} : d_i = (text_i, numerical_i, category_i) \rightarrow y_i \tag{1}$$

4. Methodology

Figure 1 illustrates the architecture of the proposed model, HY-RISE, which aims to identify risks in scientific economic activities from hybrid data. It first applies BERT to learn semantic information from reasons for reimbursement, which is typically a short text. Next, a BiGRU module is incorporated to augment features extracted by the BERT module to learn local patterns and generate more expressive representations. Then, one-hot encoding and normalization are used to preprocess structured data such as receipt type, personnel type and amount, which are subsequently fed into a neural network. Finally, we concatenate the semantic representation and the other structure data representation and feed them into a fully connected layer to generate the prediction.

4.1. BERT Layer for Word Embedding

In BERT, the embedding layer consists of three types of embeddings, namely token embedding, segment embedding and position embedding. Given that BERT is a pre-trained model, it imposes specific requirements for the input data format. Therefore, the raw data need to be tokenized and represented as a 768-dimensional vector in the token embedding layer at first. In addition, [CLS] is generally inserted at the beginning of the sentence, whereas [SEP] is placed at the end. The [CLS] symbol is prepared for downstream tasks, which summarizes and records the semantic information of words in the textual data, and [SEP] works as a segmentation symbol for cutting sentences. Segment embedding is used to distinguish which sentence the semantics of each word belongs to, and position embedding is a vector representation of the corresponding position and order of words in the text. Segment embedding is to distinguish the semantics of each word based on which sentence it belongs to, while position embedding is a vector representation of the respective position and order of words in the text. The default input sentence length in BERT, i.e., max_sentence_length, is set to 128. To improve the efficiency of the model processing, we set the max_sentence_length to 64 in HY-RISE as reimbursement reasons are typically short sentences with a limited number of words. The crucial component of the BERT model is its bidirectional transformer encoding layer. It extracts text features through the Encoder of the Transformer, which comprises a self-attention layer, feedforward neural network, and add-and-normalize layer. Assume that the text vector after passing through the embedding layer is $X = \left(x_1^k, x_1^k, \ldots, x_n^k\right)$, where n is the maximum length of the text and k is the embedding dimension. By defining Q,K,V characteristic matrices for each word in the sentence and making the characteristic matrices of different words cross multiply, the characteristic vector is calculated as follows.

$$Z = softmax\left(\frac{Q \times K^T}{\sqrt{d_k}} \times V\right) \tag{2}$$

where Q represents the index, K stands for the key, and V refers to the value, which represents the dimension of K. Save the vector generated by BERT as the input to the BiGRU.

4.2. BiGRU Layer for Feature Extraction

In this layer, BiGRU will be used to process the vectors obtained by the pretrained BERT model. Gated recurrent units (GRUs) simplify the model of LSTM. Specifically, they merge the forgetting and inputting gates of LSTM into an update gate. The GRU network receives the text embedding representation vector $F_{cls} = \{x_1, x_2, \cdots x_t\}$ processed by the BERT model, where x_t represents a word vector. The update gate, z_t, determines the extent to which the information from the previous state is forgotten and what information from the new content is to be added. The reset gate, r_t, controls how the extent of the previous hidden state and the current input is ignored. The t-th update gate, reset gate, and cell state are calculated as follows:

$$z_t = \sigma\left(w_z * [h_{t-1}, x_t]\right) \tag{3}$$

$$\tilde{h} = tanh\left(w_h * r_t \mu_t h_{t-1}, x_t\right) \tag{4}$$

$$h_t = (1 - z_t)\,\mu h_{t-1} + z_t \mu \tilde{h}_t \tag{5}$$

$$r_t = \sigma\left(w_r * [h_{t-1}, x_t]\right) \tag{6}$$

where $\tilde{h}_t$ is the new memory content, which is obtained from h_{t-1}, h_t is the new cell state. In Eq.(4), the information is output to the current hidden layer $\tilde{h}_t$ through the tanh function, and then, h_t is obtained by using Eq.(5).

BiGRU network comprises two hidden layers, namely the forward GRU and the inverse GRU, both of which are simultaneously connected to the output unit. The bidirectional overlay of these hidden layers enables them to learn contextual information from the text data concurrently. Ultimately, the final vector representation of textual data is obtained as output from the BiGRU network.

4.3. Feature Fusion and Risk Identification Layer

The feature fusion layer combines the extracted text feature vector F_{text} with continuous features $F_{numerical}$ and categorical features $F_{category}$ to create a fusion feature vector, denoted as F_{usion}. The fused vector represents an instance of scientific economic activities. Continuous features are normalized before inputting into the model, while categorical features are converted to vectors using one-hot encoding. There are various methods for vector fusion after obtaining the feature vector representation of each type of data, such as addition, multiplication, and concatenation. We choose concatenation as our method since it maximizes information preservation between different vectors. The resulting fused vector is referred to as $Feature$.

$$Feature = F_{text} \oplus F_{numerical} \oplus F_{category} \tag{7}$$

$$F_{usion} = ReLU\,(MLP\,(Feature)) \tag{8}$$

Then, $Feature$ passes through three fully connected layers for feature fusion and dimension reduction, with ReLU serving as the activation function in each layer. The F_{usion} is obtained by using Eq.(7-8).

In the final step, we pass the F_{usion} into a fully connected layer. As stated earlier, the HY-RISE aims to identify potential risks by integrating various types of data in scientific economic activities, which essentially serves as a standard binary classification task. Therefore, the probability of risk is calculated as follows.

$$y = sigmoid\,(w \cdot F_{usion} + b) \tag{9}$$

where y represents the prediction result of the model, while W and b are the model parameters.

In this paper, the objective is to train the model by minimizing the cross-entropy loss as follows.

$$L\,(y, f\,(x)) = -\frac{1}{n}\,[y_i \ln f\,(x_i) + (1 - y_i) \ln (1 - f\,(x_i))] \tag{10}$$

where y_i denotes the real label of i-th data, $f\,(x_i)$ denotes the prediction probability of the model for the i-th data, and n is the total number of data.

5. Experiment

5.1. Experimental Setup

To verify the effectiveness of the proposed HY-RISE for risk identification, we conduct extensive experiments on a real-world dataset of financial reimbursement from an institute of Chinese Academy of Sciences. The dataset is divided into training, validation, and test sets in the ratio of 6:2:2. We constructed datasets with anomaly degrees of 1.25%, 2.5%, 5%, and 10% to evaluate the robustness of HY-RISE according to the experimental setting in a prior study [18]. The default anomaly degree of the dataset used for the comparison experiments is 5%. We compare HY-RISE with two classical machine learning methods, SVM and LR, and two deep learning methods TextCNN and LSTM. All models were implemented in Python using the scikit-learn and Keras libraries. To evaluate the performance of all models we adopt three widely used performance metrics: $Precision_{macro}$, $Recall_{macro}$, and $F1 - Macro$.

$$Precision_{macro} = \frac{Precision_{normal} + Precision_{risk}}{2},\; Recall_{macro} = \frac{Recall_{normal} + Recall_{risk}}{2} \tag{11}$$

$$F1 - Macro = \frac{2 \times Precision_{macro} \times Recall_{macro}}{Precision_{macro} + Recall_{macro}} \tag{12}$$

5.2. Overall Evaluation

Table 2 shows the performance of proposed HY-RISE and various baselines under the risk identification tasks for scientific economic activities. We have the following observations.

Firstly, HY-RISE outperforms both classical machine learning models and deep learning models under all metrics. Compared with the best-performing LSTM in the baseline, HY-RISE improves the precision, recall and F1-Macro by 2.25%, 4.34% and 3.91%, respectively. Regarding the classical machine learning models, although LR surpass SVM, it still falls below HY-RISE by 6.65%, 8.93% and 8.61% in terms of precision, recall, and F1-Macro. This is because classical machine learning methods can only fit the data based on given rules, which makes it challenging to learn the potential association relationships among multiple structural data. In contrast, the pre-trained BERT module in HY-RISE can better capture the semantic features of textual data. In addition to the powerful semantic representation of BERT, automated feature fusion can explore the association relationships among different structural data and extract more representative features from the reimbursement data, thus making the risk identification more effective.

Furthermore, classical machine learning models perform worse than deep learning methods in risk identification for research and economic activities. For instance, SVM falls behind LSTM by 11.07%, 9.63% and 10.92% in terms of precision, recall, and F1-Macro, respectively. We argue that the classical machine learning methods

Table 2. Performance comparison for risk identification. The best performance is in bold.

Model	Precision	Recall	F1-Macro
SVM	0.7684	0.6315	0.6728
LR	0.8400	0.6916	0.7426
TextCNN	0.8602	0.7209	0.7717
LSTM	0.8791	0.7278	0.7820
HY-RISE	**0.8989**	**0.7594**	**0.8126**

have poorer ability in extracting semantic features from textual data and deep relational features among multiple features compared to deep learning methods.

Finally, the sequence model is more appropriate than the convolutional model for reimbursement matter text. LSTM performs better than TextCNN by approximately 1.89%, 0.69%, and 1.33% in precision, recall, and F1-Macro, respectively. Note that HY-RISE incorporates BERT and BiGRU modules for semantic extraction and augmentation, which are also sequence models. Overall, we infer that the sequence model can effectively capture the underlying semantic relations in short textual data.

5.3. Ablation Study

In addition, we conduct an ablation experiment to better examine the contribution of the introduced BiGRU module for risk identification. As shown in Figure 2, when using only the BERT module for semantic embedding representation of textual business data, the accuracy, recall and F1 scores of risk identification all decrease significantly. Specifically, the accuracy decreases by 2.01% and the F1 score decreases by 1.86% compared with the HY-RISE, which incorporates the BiGRU module to enhance the semantic representation extracted by BERT. We conclude that the introduction of the BiGRU module effectively improves the performance of HY-RISE for risk identification by semantic enhancement.

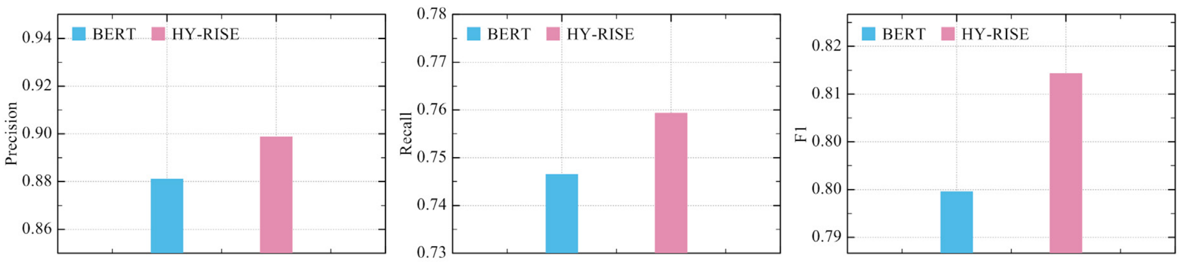

Fig. 2. The results(%) of ablation study for risk identification. 'BERT' is a variant of HY-RISE without the semantic enhancement module.

5.4. Impact of Anomaly Degree

To verify the robustness of the proposed HY-RISE, we conduct sensitivity analysis on the anomaly degree. Figure 3 compares the F1-Macro and Recall-Macro of the financial reimbursement dataset with different anomaly degrees. We can observe that the overall performance for risk identification of all models tends to improve as the anomaly degree increase, indicating that more risk instances is beneficial to detect risk behaviors. Moreover, the performance of HY-RISE consistently outperforms that of LR and LSTM, which demonstrates its robustness to various anomaly degrees in different financial reimbursement datasets.

6. Conclusion

In this paper, we propose a risk identification model for scientific economic activities that leverages multiple types of data to obtain comprehensive representations. Specifically, we use BERT to extract semantic information from textual terms and enhance it with a BiGRU module. To capture richer feature representations, we fuse the enhanced semantic representations with other features of structured terms using a shallow neural network. Finally,

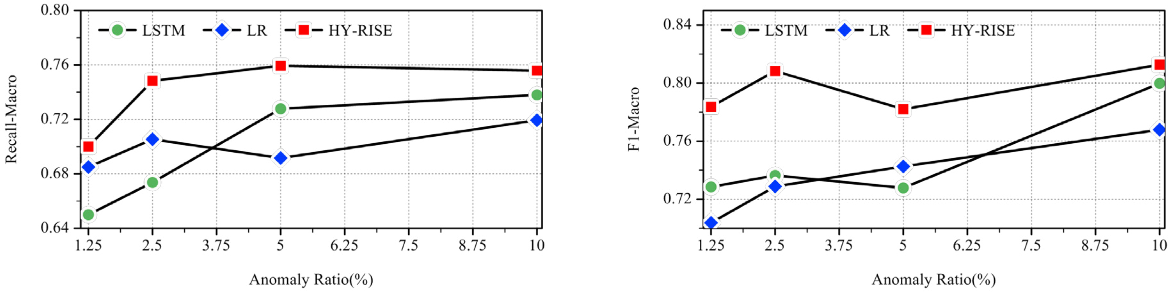

Fig. 3. Comparison of performance(%) in terms of Recall and F1-Macro with different anomaly degrees.

we feed the fused representations into a binary classifier to discriminate the risk of the corresponding activity. Experimental results demonstrate that our proposed HY-RISE achieves a significant improvement comparing with several baseline methods on a real reimbursement dataset collected from an institute of the Chinese Academy of Sciences. In the future, we will conduct fine-grained textual risk mining, such as detecting sensitive terms in reimbursement events by topic models and semantic understanding models, to further improve the comprehensiveness of risk identification in scientific economic activities.

7. Acknowledgment

This research is supported by a grant from the National Key Research and Development Program of China (No. 2022YFF0711801), the National Natural Science Foundation of China (Grant No. J2224012) and the CAS 145 Informatization Project CAS-WX2022GC-0301. Jianjun Yu is the corresponding author.

References

[1] M. Baumann, Improving a rule-based fraud detection system with classification based on association rule mining, INFORMATIK 2021.
[2] W. Szczepanik, M. Niemiec, Heuristic intrusion detection based on traffic flow statistical analysis, Energies 15 (11) (2022) 3951.
[3] Q. Zhang, Financial data anomaly detection method based on decision tree and random forest algorithm, Journal of Mathematics 2022.
[4] M. Nurek, R. Michalski, Combining machine learning and social network analysis to reveal the organizational structures, Applied Sciences 10 (5) (2020) 1699.
[5] R. Mansoor, N. D. Jayasinghe, M. M. A. Muslam, A comprehensive review on email spam classification using machine learning algorithms, in: 2021 International Conference on Information Networking (ICOIN), IEEE, 2021, pp. 327–332.
[6] M. Elhoseny, N. Metawa, G. Sztano, I. M. El-Hasnony, Deep learning-based model for financial distress prediction, Annals of Operations Research (2022) 1–23.
[7] O. Ajao, D. Bhowmik, S. Zargari, Fake news identification on twitter with hybrid cnn and rnn models, in: Proceedings of the 9th international conference on social media and society, 2018, pp. 226–230.
[8] A. M. Mubalaike, E. Adali, Deep learning approach for intelligent financial fraud detection system, in: 2018 3rd International Conference on Computer Science and Engineering (UBMK), IEEE, 2018, pp. 598–603.
[9] X. Dastile, T. Celik, Making deep learning-based predictions for credit scoring explainable, IEEE Access 9 (2021) 50426–50440.
[10] S. Ruan, X. Sun, R. Yao, W. Li, et al., Deep learning based on hierarchical self-attention for finance distress prediction incorporating text, Computational Intelligence and Neuroscience 2021.
[11] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, H. Prendinger, Deep learning for affective computing: Text-based emotion recognition in decision support, Decision Support Systems 115 (2018) 24–35.
[12] S. Chen, Cryptocurrency financial risk analysis based on deep machine learning, Complexity 2022 (2022) 1–8.
[13] I. E. Livieris, E. Pintelas, P. Pintelas, A cnn–lstm model for gold price time-series forecasting, Neural computing and applications 32 (2020) 17351–17360.
[14] M. N. Hasan, R. N. Toma, A.-A. Nahid, M. M. Islam, J.-M. Kim, Electricity theft detection in smart grid systems: A cnn-lstm based approach, Energies 12 (17) (2019) 3310.
[15] Y. Pei, F. Lyu, W. Van Ipenburg, M. Pechenizkiy, Subgraph anomaly detection in financial transaction networks, in: Proceedings of the First ACM International Conference on AI in Finance, 2020, pp. 1–8.
[16] Y. Dou, Z. Liu, L. Sun, Y. Deng, H. Peng, P. S. Yu, Enhancing graph neural network-based fraud detectors against camouflaged fraudsters, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 315–324.
[17] Q. Zhong, Y. Liu, X. Ao, B. Hu, J. Feng, J. Tang, Q. He, Financial defaulter detection on online credit payment via multi-view attributed heterogeneous information network, in: Proceedings of The Web Conference 2020, 2020, pp. 785–795.
[18] J. Tang, J. Li, Z. Gao, J. Li, Rethinking graph neural networks for anomaly detection, in: International Conference on Machine Learning, PMLR, 2022, pp. 21076–21089.