Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 3, No. 6, 2024

A Hybrid CNN-LSTM Model for Enhancing Bond Default Risk Prediction

Jianhua Yao[1], Jiajing Wang[2], Bingxing Wang[3], Bingyao Liu[4], Mohan Jiang[5]

[1]Trine University, Phoenix, USA

[2]Columbia University, New York, USA

[3]Illinois Institute of Technology, Chicago, USA

[4]University of California, Irvine, Irvine, USA

[5]New York University, New York, USA

Correspondence should be addressed to Mohan Jiang; mj2589@nyu.edu

Abstract: This paper explores the importance of credit risk management in the global financial market environment, especially for the prediction of bond default risk. With the advent of the big data era, the amount of market information has surged, covering multi-dimensional data from traditional financial statements to social media comments. However, traditional credit rating methods rely mainly on structured data and ignore the value of unstructured data, resulting in limited prediction accuracy. The development of deep learning technology provides a new way to process such data. By introducing a combined model of convolutional neural network (CNN) and long short-term memory network (LSTM), we propose a novel algorithm to predict bond default risk. The model uses CNN to process unstructured text data to extract key features and uses LSTM to process time series data to capture the trend of data changing over time. Experimental results show that the model performs well in terms of accuracy, surpassing other common models. The research in this paper provides new ideas for the application of deep learning in the financial field.

Keywords: Credit risk, deep learning, bond default prediction, big data, convolutional neural network

1. Introduction

In the global financial market environment, credit risk management has become one of the major challenges faced by various financial institutions. As a type of fixed-income security widely traded in the financial market, the accurate prediction of the default risk of bonds is not only the key to investors' asset allocation and risk management but also an important factor in maintaining the stability of the financial market [1-2]. With the advent of the big data era, the amount of information available in the market has exploded, covering a range from traditional financial statements to real-time market dynamics, and even extending to multiple dimensions such as comments and emotional expressions on social media. However, traditional credit rating methods, especially those that rely on financial ratio analysis, are overwhelmed when dealing with such diverse data, because they are mainly limited to structured data and ignore the potential value in unstructured data, which undoubtedly limits the accuracy and comprehensiveness of their predictions.

At the same time, the development of deep learning technology provides a new perspective and tools for financial data processing. With its powerful data processing capabilities, deep learning algorithms can discover implicit laws and patterns in complex and large data sets, especially when dealing with time series analysis and image recognition tasks [3]. For example, through convolutional neural networks (CNN), we can efficiently process image data and extract useful visual features [4]; while long short-term memory networks (LSTM) are good at processing sequence data and capturing the trend of data changes over time [5]. These technological advances have brought unprecedented opportunities for data analysis in the financial field, especially in the sub-field of bond default risk prediction [6]. Deep learning models can better adapt to the variability and complexity of data and provide more sophisticated risk assessment results than traditional methods.

The accuracy of bond default risk prediction models directly affects the quality of investment decisions and is of great significance for maintaining financial market stability. Although traditional statistical models can reflect the possibility of default to a certain extent, they usually only consider structured data and ignore the influence of unstructured information such as market sentiment and industry trends, resulting in unsatisfactory prediction results. In contrast, models based on deep learning can effectively integrate multiple types of data sources and capture the complex patterns hidden behind the data through automatic feature learning mechanisms. This not only helps to improve prediction accuracy but also provides financial institutions with more comprehensive risk management strategy support. In addition, the application of such models can also help regulators identify systemic risks in a timely manner and take preventive measures to avoid financial crises [7].

This paper proposes a hybrid architecture that combines Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) to predict bond default risk. LSTM is good at capturing long-term dependencies in time series due to its ability to process sequence data; while CNN is good at extracting local features from two-dimensional data such as images or text [8]. In order to fully utilize the advantages of these two networks, this study designed a two-stage framework: first, CNN is used to process unstructured text data to extract key semantic features; second, these features are input into LSTM together with structured financial data for time series modeling. In this way, the model can comprehensively consider historical financial performance and the latest market feedback, thereby realizing dynamic monitoring of bond default risk.

2. Background

2.1. Related Work

Deep learning models, particularly Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM), have become key tools for handling complex datasets, including temporal and unstructured data. For financial applications, particularly in predicting bond default risk, such models are advantageous due to their ability to process diverse data sources like financial indicators, market sentiment, and time-series data. Various studies have contributed to improving the predictive power of such models through innovative combinations of deep learning techniques.

One relevant work proposed a Hybrid LSTM-GARCH framework for financial market volatility prediction, which demonstrated the benefits of combining LSTM's sequence modeling with GARCH's volatility estimation techniques to handle time-series data effectively. This framework laid the groundwork for modeling financial markets, aligning with the use of LSTM in the bond default risk model introduced in this paper, where temporal financial data is critical for prediction accuracy [9].

Optimization techniques in deep learning have also been explored to enhance model training. One study introduced a method to optimize gradient descent in neural network training, offering improved convergence rates and accuracy. Such approaches are crucial for training complex models like the hybrid CNN-LSTM framework, ensuring that large financial datasets can be processed efficiently and effectively [10].

In the context of financial time-series data, spatiotemporal feature representation has proven effective for mining multidimensional patterns. Research into spatiotemporal features has shown how such techniques can uncover hidden relationships in time-series data, which is highly relevant for models designed to predict financial trends or risks. This aligns with the LSTM's role in processing sequence data within the hybrid model [11].

Moreover, the use of multimodal data fusion techniques has emerged as a powerful method for integrating multiple types of data to improve prediction outcomes. Recent research has demonstrated that combining different data modalities, such as text and structured data, can significantly enhance model performance in tasks like object recognition. This multimodal approach resonates with the CNN-LSTM hybrid model, where structured financial data is combined with unstructured textual data to provide a more comprehensive assessment of bond default risk [12]. In addition, studies have shown that using lightweight neural network architectures can reduce computational complexity without sacrificing model performance. This approach is particularly useful when dealing with large-scale financial datasets, where processing efficiency is a priority. Lightweight architectures, like those proposed in recent research, can help in real-time financial risk monitoring and fast computations [13].

Recent advancements in contrastive learning have shown potential for improving feature extraction in large language models, which is applicable to the unstructured text data processing capabilities of CNNs. Contrastive learning can be adapted to enhance the extraction of relevant financial insights from unstructured data sources like market comments and analyst reports, thereby improving the performance of hybrid models for predicting financial outcomes [14]. While techniques for modeling time-series data and generating accurate predictive models are equally applicable to financial risk prediction, demonstrating the versatility of deep learning techniques [15-16]. In conclusion, this paper builds upon these foundational works by proposing a hybrid CNN-LSTM model specifically designed for bond default risk prediction. By integrating structured and unstructured data sources, the model addresses key challenges in predicting bond defaults, achieving higher accuracy and more comprehensive financial risk assessments.

3. Method

3.3. Model Structure

The First, we need to define the input data set. For the bond default risk prediction task, the input data usually consists of two parts: structured data and unstructured data. Structured data may come from corporate financial statements, market conditions, etc., while unstructured data may come from news reports, comments on social media, etc.

For structured data, suppose we have a time series dataset $X = \{x_1, x_2, ..., x_T\}$, where x_t is a vector of financial data at time t, and each vector x_t contains multiple features, such as net profit, total liabilities, etc.

As for unstructured data, suppose we have a series of text data $Y = \{y_1, y_2, ..., y_T\}$ where y_t is the text data at time t, which may be news reports or comments on social media. For unstructured data, we first need to convert it into a form suitable for machine learning algorithms to process. Here we use word embedding technology to convert each text fragment y_t into a vector form. Then, these vectors are input into one or more convolutional layers to capture local features. The convolution operation can be expressed as:

$$z_i = f(w * x_i + b)$$

Where W is the convolution kernel (filter), x_i is part of the input vector, * represents the convolution operation, and b is the bias term. After a series of convolution and pooling operations, we will get a fixed-length vector z, which contains the key features of the text data.

The structured data $X = \{x_1, x_2, ..., x_T\}$ will be directly input into the LSTM model. The LSTM model can effectively capture the long-term dependencies in time series data. For each time point t, the LSTM unit will update its internal state based on the input x_t at the current moment and the states h_{t-1} and c_{t-1} at the previous moment. The LSTM update equation is as follows:

$$i_t = \sigma(W_{xi} x_t + w_{hi} h_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf} x_t + w_{hf} h_{t-1} + b_f)$$
$$o_t = \sigma(W_{xo} x_t + w_{ho} h_{t-1} + b_o)$$

Where σ represents the Sigmoid function. The feature vector z after unstructured data processing is fused with the LSTM hidden state h_t at each time step. This can be done by a simple splicing operation:

$$H_t = [h_t; C]$$

In this way, we obtain the time series feature vector H_t that combines structured and unstructured information.

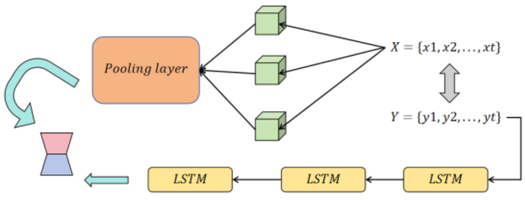

Figure1. Overall architecture diagram

Finally, we project H_t to the default probability estimate through a fully connected layer. Assuming that we use a binary classification task to predict whether a default occurs, the output layer can be a node with a Sigmoid activation function, and its output value p represents the probability of default.

The loss function can choose Binary Cross-Entropy Loss to measure the difference between the model prediction value and the true label:

$$L = -[y\log(p) + (1-y)\log(1-p)]$$

4. Experiment

4.1. Dataset

The data sets used in this article are all from Bloomberg API, covering both structured data and unstructured data. Structured data mainly comes from the company's financial statements, including the following variables: Total Assets, which is the total value of all assets owned by the company at a certain point in time, including fixed assets and current assets [17]; Total Liabilities, which is the total amount of all debts undertaken by the company at a certain point in time, including short-term debts and long-term debts; Current Assets, which is the assets that the company can realize within one year or one operating cycle, such as cash, accounts receivable, inventory, etc.; Current Liabilities, which is the debt that the company needs to repay within one year or one operating cycle, such as accounts payable, short-term loans, etc.; Net Income, which is the net income of the company after deducting total expenses from the total income in a certain period; Revenue, which is the total income obtained by the company through the sale of goods or services in a certain period. Unstructured data involves financial articles, which usually include news reports (covering corporate operations, management changes, major contract signings, legal proceedings, etc.), analyst reports (industry analysis reports issued by investment banks or consulting firms, including market trends, competitor conditions, corporate strategies, etc.) and market comments (investors, analysts or ordinary users' views and comments on corporate performance). These data provide comprehensive information and help build a comprehensive bond default risk prediction model.

Table 1: Dataset Example

Data Types	Data name
Structured Data	Total Assets
Structured Data	Total Liabilities
Structured Data	Current Assets
Structured Data	Current Liabilities
Structured Data	Net Income
Unstructured Data	Revenue
Unstructured Data	Operational Status
Unstructured Data	Analyst Reports
Unstructured Data	Market Comments

4.2. Data cleaning

For the dataset used in this paper, detailed data cleaning was carried out to ensure the accuracy and reliability of model training. First, the structured financial data was carefully checked to remove missing values and outliers, such as filling missing financial indicators with industry averages or medians, and extreme values were corrected or deleted to avoid adverse effects on the model. At the same time, the data was standardized to ensure that different financial indicators were compared at the same level, so as to avoid the situation where some variables dominated the model prediction results due to excessive values. In addition, the time series data was smoothed, and moving averages and other methods were used to eliminate the noise interference caused by short-term fluctuations, so that the model can better capture long-term trends. For unstructured data, the text was first preprocessed by natural language processing technology, including removing HTML tags, filtering stop words, removing punctuation marks, etc., to reduce the interference of irrelevant information. Subsequently, the text was segmented and stemmed, long sentences were broken down into meaningful vocabulary units, and different forms of the same vocabulary were unified into base forms to improve the accuracy of feature extraction. On this basis, the word frequency-inverse document frequency (TF-IDF) method was used to vectorize the text and convert the unstructured text into numerical features for further processing by the model. Through these steps, the quality and availability of unstructured data are ensured, so that it can be effectively integrated with structured data to serve the construction of the bond default risk prediction model.

4.3. Experiments

To We selected 5 different common models to analyze our data results, including RNN, LSTM, Resnet18, transformer, and our model. We first analyzed the accuracy. The results are shown in Table 2.

Table 2: Accuracy Experiment Results

Model	ACC
RNN	0.581
Resnet18	0.751
LSTM	0.779
Transformer	0.821
Ours	0.835

According to the experimental results provided, we observed significant differences in the accuracy of different models on this task, with the RNN model only achieving an accuracy of 0.581, which may be due to its limitations in processing long sequence data. In contrast, ResNet18, as one of the classic deep learning models, effectively alleviated the gradient vanishing problem in deep networks by introducing residual blocks, thereby achieving an accuracy of 0.751, showing its strong ability in feature extraction. The LSTM model, which is specially designed to capture long-term dependencies in sequence data, also proves its strong ability in handle such tasks with an accuracy of 0.779. With the rise of the attention mechanism, the Transformer model has increased its accuracy to 0.821 with its powerful parallel processing capabilities and efficient capture of long-distance dependencies, becoming a model architecture that has performed outstandingly in many fields in recent years. However, it is exciting that the model we proposed, a hybrid model that combines the advantages of CNN and LSTM, stands out with the highest accuracy of 0.835. This result shows that our model not only inherits the high efficiency of CNN in feature learning, but also integrates the unique advantages of LSTM in processing sequence data, making the overall model show excellent performance in dealing with the current task. It is worth noting that although Our model performs well, in order to ensure its robustness and generalization ability, it still needs to be further tested and cross-validated on a variety of data sets to consolidate its effectiveness. In practical applications, factors such as model complexity, training time, and computing resource requirements need to be considered to ensure that the model can maintain efficient and stable performance in different deployment environments. At the same time, future research directions can focus on optimizing the existing model structure, exploring new training strategies, or combining more advanced technologies, such as semi-supervised learning, transfer learning, and other methods, to further improve the performance and practicality of the model.

In order to further demonstrate the experimental results, we use a bar chart to show our experimental results, as shown in Figure 2.

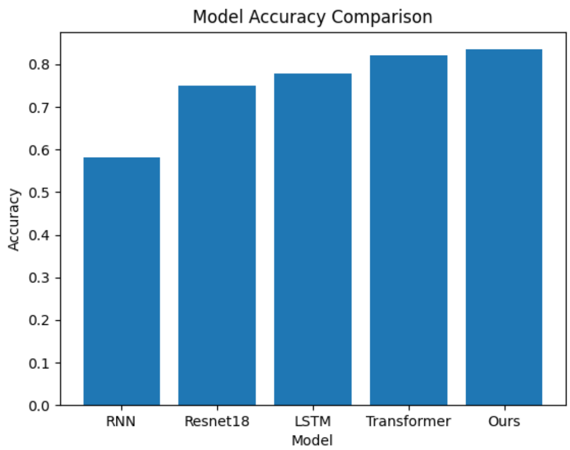

Figure 2. Accuracy results comparison chart

Furthermore, we use Recall and F1 values to show our experimental results, which are shown in Tables 3 and 4.

Table 3: Recall values Experimental Results

Model	Recall
RNN	0.565
Resnet18	0.749
LSTM	0.788
Transformer	0.817
Ours	0.821

Table 4: F1 values Experimental Results

Model	F1
RNN	0.553
Resnet18	0.750
LSTM	0.781
Transformer	0.819
Ours	0.831

From the recall rate in Table 3, we can see that the performance of each model on the recognition or classification task is different. Although traditional sequence models such as RNN and LSTM can capture certain temporal dependencies, their recall rates (0.565 and 0.788, respectively) are lower than ResNet18 (0.749), which indicates that ResNet18 using convolutional neural network (CNN) can more effectively reduce false negatives in this particular task. However, the most significant performance comes from the Transformer model and our proposed new model, which achieves recall rates of 0.817 and 0.821.

Table 4 shows the results of the F1 score, which is the harmonic mean of recall and precision, providing a comprehensive perspective on the performance of the model. Consistent with the trend of recall, our model also achieved the highest F1 score (0.831), followed by the Transformer model (0.819). This further confirms that our model performs better than other models when considering the balance between precision and recall, especially in reducing false positives while maintaining high recall. Compared with traditional methods such as RNN and LSTM, these advanced models (including ResNet18) also have significant improvements in F1 scores, reflecting the importance of deep learning architectures for improving the performance of machine learning tasks. Similarly, in order to further demonstrate our experimental results, we also use two bar graphs to show our experimental results on Recall and F1 values, as shown in Figures 3 and 4.

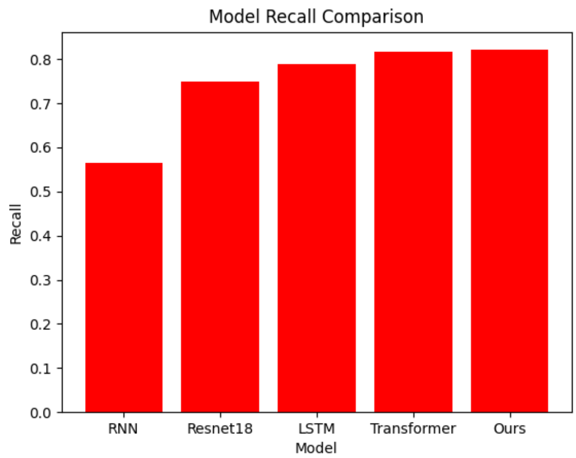

Figure 3. Recall results comparison chart

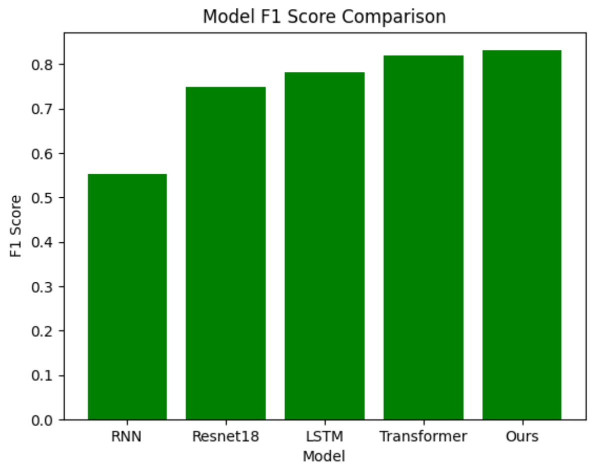

Figure 4. F1 results comparison chart

5. Conclusion

In this paper, we developed an innovative deep learning algorithm to predict the default risk of bonds. The algorithm combines the advantages of convolutional neural networks (CNN) and long short-term memory networks (LSTM). CNN is used to process unstructured data, such as analyst reports, market comments, etc., to extract useful visual features; while LSTM is used to process structured time series data, such as various indicators in financial statements. In this way, our model is able to process both structured and unstructured data, thereby improving the accuracy of bond default risk prediction. Experiments show that compared with commonly used models such as RNN, ResNet18, LSTM and Transformer, the model proposed in this study has an accuracy of 0.835, showing its superior performance in dealing with such problems. Despite the good results, more testing and verification are still needed in future work to further improve the robustness and generalization ability of the model. In addition, future research can also explore how to combine more external data sources to further improve the effect of the prediction model.

References

[1] Zhang N, Li W, Chen H, et al., "Predicting Credit Bond Default with Deep Learning: Evidence from China," *Journal of Social Computing*, vol. 5, no. 1, pp. 36-45, 2024.

[2] Park D, Auh J K, Song G, et al., "Understanding Corporate Bond Defaults in Korea Using Machine Learning Models," *Asia-Pacific Journal of Financial Studies*, vol. 53, no. 2, pp. 238-276, 2024.

[3] Y. Li, X. Yan, M. Xiao, W. Wang and F. Zhang, "Investigation of Creating Accessibility Linked Data Based on Publicly Available Accessibility Datasets", Proceedings of the 2023 13th International Conference on Communication and Network Security, pp. 77-81, 2024.

[4] Dang, B., Zhao, W., Li, Y., Ma, D., Yu, Q. and Zhu, E. Y., "Real-Time Pill Identification for the Visually Impaired Using Deep Learning," Proceedings of the 2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE), pp. 552-555, 2024. doi: 10.1109/CISCE62493.2024.10653353.

[5] W. Dai, J. Tao, X. Yan, Z. Feng, and J. Chen, "Addressing unintended bias in toxicity detection: An LSTM and attention-based approach," Proceedings of the 2023 5th International Conference on Artificial Intelligence and Computer Applications (ICAICA), pp. 375-379, 2023.

[6] Shilbayeh S, Grassa R, "Creditworthiness pattern prediction and detection for GCC Islamic banks using machine learning techniques," *International Journal of Islamic and Middle Eastern Finance and Management*, vol. 17, no. 2, pp. 345-365, 2024.

[7] Nguyen L, "Machine Learning for Small Data: A Comprehensive Study with the Case of Sovereign Debt Default Forecasting," 2024.

[8] J. Yao, T. Wu, and X. Zhang, "Improving depth gradient continuity in transformers: A comparative study on monocular depth estimation with CNN," arXiv preprint arXiv:2308.08333, 2023.

[9] K. Xu, Y. Wu, M. Jiang, W. Sun, and Z. Yang, "Hybrid LSTM-GARCH Framework for Financial Market Volatility Risk Prediction," Journal of Computer Science and Software Applications, vol. 4, no. 5, pp. 22-29, 2024.

[10] Y. Ma, D. Sun, E. Gao, N. Sang, I. Li, and G. Huang, "Enhancing Deep Learning with Optimized Gradient Descent: Bridging Numerical Methods and Neural Network Training," arXiv preprint arXiv:2409.04707, 2024.

[11] X. Yan, Y. Jiang, W. Liu, D. Yi, H. Sang, and J. Wei, "Data-Driven Spatiotemporal Feature Representation and Mining in Multidimensional Time Series," arXiv preprint arXiv:2409.14327, 2024.

[12] Q. Wang, Z. Gao, T. Mei, X. Cheng, W. Gu, and H. Xia, "Deep Learning-based Multimodal Fusion for Improved Object Recognition Accuracy," Proceedings of the 2024 3rd International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC), pp. 471-474, July 2024.

[13] Z. Wu, H. Gong, J. Chen, Z. Yuru, L. Tan, and G. Shi, "A Lightweight GAN-Based Image Fusion Algorithm for Visible and Infrared Images," arXiv preprint arXiv:2409.15332, 2024.

[14] Z. Zhang, J. Chen, W. Shi, L. Yi, C. Wang, and Q. Yu, "Contrastive Learning for Knowledge-Based Question Generation in Large Language Models," arXiv preprint arXiv:2409.13994, 2024.

[15] H. Yang, Y. Hu, S. He, T. Xu, J. Yuan, and X. Gu, "Applying Conditional Generative Adversarial Networks for Imaging Diagnosis," arXiv preprint arXiv:2408.02074, 2024.

[16] B. Dang, D. Ma, S. Li, Z. Qi, and E. Zhu, "Deep learning-based snore sound analysis for the detection of night-time breathing disorders," Applied and Computational Engineering, vol. 76, pp. 109-114, Jul. 2024. doi: 10.54254/2755-2721/76/20240574.

[17] Meng B, Sun J, Shi B, "A novel URP-CNN model for bond credit risk evaluation of Chinese listed companies," *Expert Systems with Applications*, vol. 255, p. 124861, 2024.