Chinese Scientific Paper Classification Based on BERT, Graph Convolutional Networks, and Ensemble Learning

Xin Luo
Faculty of Computer and Mathematical Sciences, Center of Computing Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia | School of Computer Information Engineering, Hanshan Normal University, Guangdong, China
2022201126@isiswa.uitm.edu.my

Syaripah Ruzaini Syed Aris
Faculty of Computer and Mathematical Sciences, Center of Computing Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
ruzaini@uitm.edu.my (corresponding author)

Sofianita Mutalib
Faculty of Computer and Mathematical Sciences, Center of Computing Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
sofianita@uitm.edu.my

Received: 21 July 2025 | Revised: 12 September 2025 and 25 September 2025 | Accepted: 28 September 2025

 DOI: https://doi.org/10.48084/etasr.13540

ABSTRACT

With the increasing number of published papers, improving the efficiency of researchers in finding relevant literature through high-precision automatic classification has become a key research focus. Currently, most studies rely on a single type of paper metadata, such as abstracts or titles, without comparing the importance of different metadata types or utilizing them comprehensively together. Additionally, many experiments utilize non-public datasets, which makes it difficult to compare model performances across different research. This paper examines and compares the classification performance of TextGCN, BERT, and BERT-GCN models on public Chinese paper datasets. It also investigates the role of abstracts, keywords, and titles in classification and identifies the data combination that yields the highest accuracy through experiments. Since the output of the BERT-GCN model is based on the weighted sum of the outputs of BERT and TextGCN, determining the optimal weight that results in the best classification performance is also a key focus. To further enhance the performance of the BERT-GCN model, ensemble voting is used to combine the prediction results of multiple models trained on different data sources. Compared to the latest baseline models, the proposed method significantly improves the classification accuracy in Chinese paper datasets, with the highest accuracy increasing from 83.6 to 87.2%.

Keywords-BERT-GCN; ensemble learning; Chinese scientific paper classification

I. INTRODUCTION

In recent years, as the number of scientific papers published continues to increase, automatic classification of papers can speed up online publishing and improve retrieval efficiency, which is of great significance in improving the work efficiency of researchers. The research on automatic classification of scientific papers mainly uses machine learning algorithms to establish a model between the metadata (abstract, title, keywords, etc.) of the paper and the category to achieve automatic classification. Using traditional machine learning techniques for paper classification involves two steps. First, obtain the document representation vector through Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, Global Vectors (GloVec) for word representation, FastText, and other methods, and then use them as input data for classification algorithms such as naive Bayes, decision tree, support vector machine, and neural networks [1, 2].

As deep learning and natural language processing technologies advance, pre-training language models have been proposed and applied to various scenarios. The basic concept is to utilize the large, unlabeled general corpus available online, such as data from Wikipedia or the Baidu Encyclopedia, to train the model. The model is then fine-tuned using downstream task-specific datasets, which can yield better results than traditional machine learning algorithms or other deep learning models that handle text classification, entity recognition, and text summarization. Related studies have evaluated the performance of pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers) against other models such as SVM (Support Vector Machine), RNN (Recurrent Neural Network), and CNN (Convolutional Neural Networks) in classifying scientific papers. The results show that BERT can achieve the best classification performance due to the advantage of pre-trained language models in acquiring sentence semantic information [3-8]. Graph Convolutional Networks (GCNs) can construct global heterogeneous graphs using corpora and utilize sentences or words as graph nodes for information transmission, thereby facilitating the acquisition of global information [9-12]. Combining BERT with GCN can achieve better text classification results.

Text classification is a core task in natural language processing, which can be understood as the process of assigning predefined labels. Text classification has a wide range of application scenarios, such as sentiment analysis and others [13]. Scientific paper classification typically uses metadata of papers, such as title, abstract, and keywords, as input data. In recent years, related studies have also utilized the content, images, citations, and other data from papers to improve classification performance.

Many studies have compared the effectiveness of traditional machine learning algorithms and deep learning techniques in the classification of scientific papers. Deep learning techniques perform better in scenarios with large data sets and complex features. Pretrained language models have made significant progress in text classification by learning the semantic relevance of text from large corpora. Some researchers have combined pretrained language models with other models to enhance classification performance. In [14], BERT and GCN were combined to classify scientific papers. This study fine-tuned the BERT model through span masking, learning rate decay, and data augmentation to improve performance. The BERT model was used to generate embedding vectors for the tokens in the paper title, treating them as graph nodes, and a cosine similarity graph was constructed. Finally, GCN was used for aggregation and classification. In tests on the CSL (Chinese Scientific Literature) dataset and the ArXiv-2019 dataset, the BERT-GCN model outperformed TextCNN, TextRNN, DPCNN, and ERNIE models in classification performance. The study in [15] focused on the classification of research papers on Radio Frequency Electromagnetic Field (RF-EMF). Through experiments, it was found that the length of the input text affects the classification performance of the BERT-GCN model. The longer the text, the smaller the role of the GCN model. The accuracy of the BERT-GCN model in classifying abstracts was 10% higher than that of the GCN model in classifying citation information, but the dataset used was too small (only 1,407 papers). In [16], an automatic classification model for paper abstracts was proposed, based on BERT-GCN-ResNet. A residual network (ResNet) was added between graph convolutional layers to address the problem of gradient vanishing and exploding.

Table I compares the research on hybrid models of BERT and GCN across four dimensions: experimental dataset, paper metadata type, used model, and classification performance.

TABLE I. COMPARATIVE STUDY ON SCIENTIFIC PAPERS CLASSIFICATION

Ref.	Dataset	Paper metadata	Model	Accuracy
[14]	CSL/ Arxiv 2019	Title	BERT+GCN	86.31%/ 91.77%
[15]	1,407 papers	Abstract/ Citation Information	BERT+GCN	86% (λ=0.7)
[16]	Abstract (3,557 samples)	Abstract	BERT-GCN-ResNet	96.85%

Current research on scientific paper classification faces several limitations, including a lack of comparison across different metadata types, insufficient dataset sizes, poor model classification accuracy on large datasets, and the use of a single model. The main work of this paper includes:

- Comparison of the classification performance of the BERT, TextGCN, and BERT-GCN models on the CSL paper dataset. To examine the role of abstract, title, and keywords in paper classification, this study compares the classification performance of models that use only one or multiple types of paper metadata. The impact of weight factor settings for BERT and GCN and their effect on accuracy is also analyzed.
- Building on the original BERT-GCN model, an ensemble learning method is added to vote on the results of the best model after fine-tuning with different paper metadata, thus producing the final paper classification results. The improved model is compared with the baseline models in the CLUE Benchmark.

This research is innovative in that it not only analyzes the classification performance of the combined model using three types of paper metadata—abstract, keywords, and title—but also integrates the classification results of models trained using these various metadata using ensemble voting, ultimately improving classification accuracy.

II. METHODOLOGY

A. BERT-GCN Model for Classification

The steps for automatically classifying scientific papers using a model that combines BERT and GCN are as follows. First, preprocess the text data by removing punctuation, stop words, and other irrelevant elements. Then, perform word segmentation and build a vocabulary for the corpus. Next, create a heterogeneous graph of scientific papers, including word and document nodes, and set the edge weights between word nodes using Point-wise Mutual Information (PMI). PMI measures the association between words based on how often they co-occur. TF-IDF is used to assign weights to edges between word nodes and document nodes, where TF indicates how often a word appears in a document, and IDF reflects the number of documents containing that word, expressed as the inverse logarithm of the total number of documents. The edge weights between all nodes can eventually form an adjacency matrix A_U, as [17]:

$$A_{ij} = \begin{cases} PMI(i,j) & i,j \text{ are words}, PMI(i,j) > 0 \\ TF-IDF_{ij} & i \text{ is document}, j \text{ is word} \\ 1 & i = j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $PMI(i,j)$ is determined by the co-occurrence frequency of the two words and the frequency of each word, as expressed in:

$$PMI(i,j) = log\frac{p(i,j)}{p(i)p(j)} \quad (1)$$

The co-occurrence frequency of two words $P(i,j)$ is expressed in (3), where $\#W(i,j)$ indicates how many words co-occur in the sliding window, and $\#W$ represents the total number of sliding windows. Similarly, the frequency of occurrence of a word is expressed by (4).

$$p(i,j) = \frac{\#W(i,j)}{\#W} \quad (2)$$

$$p(i) = \frac{\#W(i)}{\#W} \quad (3)$$

Since positive PMI values better reflect the semantic relevance between words, edges are added to the graph only if the PMI values of the word pair are positive. PMI is suitable for measuring the relevance between words within a corpus, so it is used as the edge weight between word nodes in GCN. Conversely, TF-IDF measures the relevance between words and documents, so it is chosen as the edge weight between word nodes and document nodes in GCN.

A GCN network is a multi-layer neural network where each node can generate an embedded vector based on the attributes of its neighboring nodes. Each layer of the GCN can only access information from directly connected neighbors. By stacking multiple network layers, information from neighbors at greater distances can be gathered. For a multi-layer GCN network, the feature matrix L of the node can be expressed as:

$$L^{k+1} = \rho(\tilde{A}L^kW_k) \quad (5)$$

where k indicates the layer number, $\tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ denotes the normalized symmetric adjacency matrix, W_k represents the weight matrix of the k-th layer [18], and ρ denotes an activation function, such as the ReLU function. $L^{(0)} = X$, which represents the initial feature matrix of the node. For a two-layer GCN network, the size of the node embedding vector in the last layer matches the number of classification labels. When input into the Softmax classifier, the document classification can be realized as:

$$Z_{GCN} = \text{softmax}(\tilde{A}\text{ReLU}(\tilde{A}XW_0)W_1) \quad (4)$$

Since the nodes in the graph are categorized as word nodes and document nodes, the initial value for the input node feature matrix of the GCN model can be described by:

$$X = \begin{pmatrix} X_{doc} \\ 0 \end{pmatrix}_{(n_{doc}+n_{word})\times d} \quad (5)$$

The feature vector of the document node is created by the BERT model using the abstract of the scientific paper and other information. The final classification result of the combined BERT and GCN model is a weighted sum of the classification results from both models. This approach is used because BERT and GCN have different strengths in processing text. For different types of text, the best classification results can be obtained by adjusting the weights of the two models' classification outcomes. Let Z_{BERT} represent the classification result of the paper by the BERT model, which can be expressed by (8), where X represents the document embedding matrix generated by the BERT model [19].

$$Z_{BERT} = \text{softmax}(WX) \quad (6)$$

The final scientific paper classification result can be expressed as:

$$Z = \lambda Z_{GCN} + (1-\lambda)Z_{BERT} \quad (7)$$

where $\lambda \in (0,1)$. When $\lambda = 0$, only the classification results of the BERT model are used, and when $\lambda = 1$, only the classification results of the GCN model are used. Figure 1 shows the architecture diagram of the scientific paper classification model combining BERT and GCN. This study examined adding more GCN layers to improve the model's classification performance, but the results were disappointing. This is likely because increasing the number of layers can cause some feature information to be lost, making a two-layer GCN more effective. Other studies have reached similar conclusions.

B. Dataset

The CSL (Chinese Scientific Literature) dataset, a large-scale Chinese literature dataset, was used to evaluate the effectiveness of the BERT-GCN model in classifying scientific papers, which includes the abstract, keywords, title, category, and discipline of 396,000 documents [20]. The samples are distributed across 13 categories, including engineering, science, medicine, and others. The average title length across all categories is 18.62 characters, the average abstract length is 188.95 characters, and the average keyword length is 5.64 characters. Although the original CSL dataset consists of 13 categories, the number of samples varies between them, with the Engineering category having the most (177,600) and the Strategy category having the fewest (3,555). Therefore, when curating the experimental dataset, we did not randomly select samples from the entire dataset but randomly chose 3,000 samples from each category to reduce the effects of sample imbalance on model classification performance. After data cleaning, the final sample size was 38,997. Following word segmentation and stopword removal, the vocabulary size decreased from 151,550 to 135,813, and the average length of each category decreased from 239 Chinese characters to 138 tokens.

When using GCN to classify text, graph nodes and edges need to be constructed, where nodes represent documents and words, and edges show the relationships between words and documents, as well as between words themselves. Table II shows the constructed graph information from the experimental dataset.

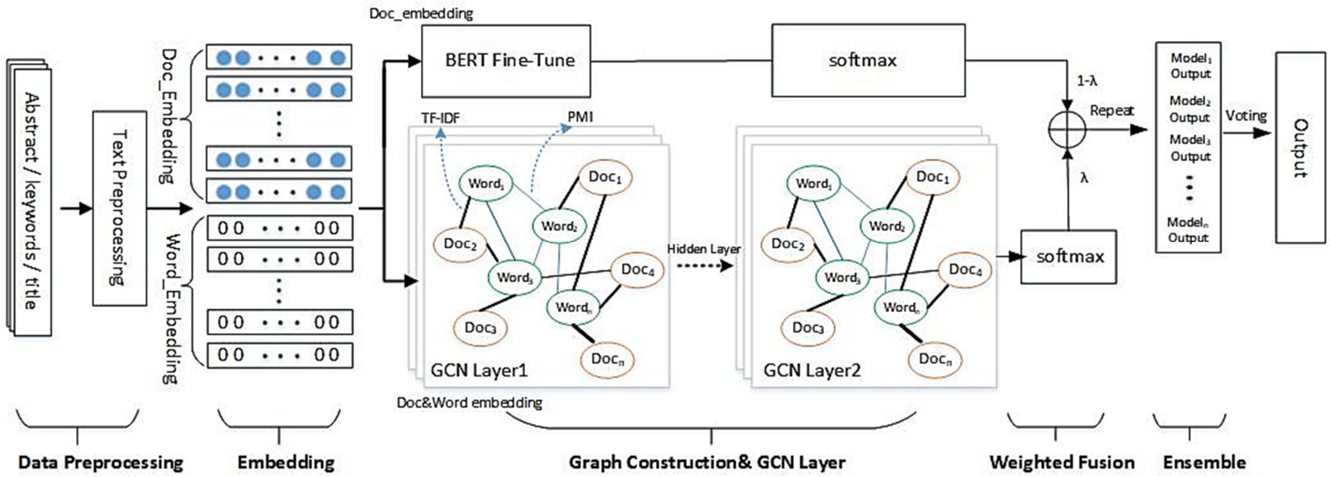

Fig. 1. The architecture of the paper classification model combining BERT and GCN.

TABLE II. GRAPH INFORMATION OF CSL DATASET

Graph properties	Value
#words	135,813
#nodes	174,810
#edges	29,559,468
#doc	38,997
#avg_doc_len	56.99
#categories	13

C. Data Preprocessing and Experimental Setup

To train and evaluate the model, an experimental dataset was created based on the number of documents and categories in Table II. The dataset was divided into training, validation, and testing sets in a 7:2:1 ratio. The splitting ratio is typically adjusted depending on the size of the dataset, with larger datasets having smaller proportions allocated to validation and testing. Since this dataset contains only 39,000 samples, a 7:2:1 ratio was selected based on data from other studies. In addition, an 8:1:1 ratio was examined, where the model's classification performance remained mostly consistent. This study mostly used simple cross-validation to assess the model's performance, randomly splitting the dataset into training and validation sets at each epoch and averaging the results of multiple model runs. The experiment was carried out on a workstation equipped with an Intel Xeon Silver 4216 2.8 GHz CPU, 128 GB of memory, and an RTX 3090 GPU. The models were developed using the PyTorch 1.13.1 framework.

The BERT_Chinese_Base pre-trained model was used to generate embedding vectors. The process of using BERT to produce token vectors for a text sequence involves first performing word segmentation. This is done character by character, resulting in a sequence of Chinese character tokens. [CLS] and [SEP] tokens are then added at the beginning and end of the sequence to serve as class and separator tokens, respectively. Each token's embedding is derived from BERT's vocabulary. The combined sum of token, segment, and position embeddings serves as input to the transformer layer, producing token vectors with global semantic context and a class vector.

To preprocess Chinese text, the Jieba library was used to segment words, then delete words with a length of 1, and finally use the Chinese stop word library to obtain a meaningful word list. The maximum epoch was set to 200, and if the classification accuracy in the validation set did not improve after 10 consecutive epochs, the training process was terminated early. Table III shows the learning parameters of the experimental model. This experimental design involves using the paper's abstract, keywords, and title as either individual or combined inputs to the model and comparing the classification accuracy across the TextGCN, BERT, and BERT-GCN models. Since the BERT and GCN weight settings in the BERT-GCN model significantly affect classification accuracy, the optimal weights for the model were experimentally determined. Finally, the prediction results of the models trained on different datasets were combined by ensemble voting, and the final classification accuracy was compared with state-of-the-art baseline models.

TABLE III. BERT-GCN MODEL PARAMETER SETUP

BERT_LR	GCN_LR	Input_Max_Len	Batch_Size	Dropout
1e-4	1e-3	128	128	0.5

III. RESULTS AND DISCUSSION

A. Comparison of the Role of Paper Meta-Information in Classification

The CSL paper dataset includes the abstract, keywords, and title of the paper. To analyze the role of different fields in classification and identify the best combination for optimal results, we created seven datasets: Abstract, Title, Keywords, Keywords+Abstract, Title+Abstract, Keywords+Title, and Keywords+Title+Abstract. The last four datasets combine multiple fields by concatenation. Figure 2 shows the classification accuracy of these seven datasets using the TextGCN and BERT models. It is clear that when using only one of the abstract, keywords, or titles as input, the BERT and TextGCN models achieve their highest accuracy with the abstract.

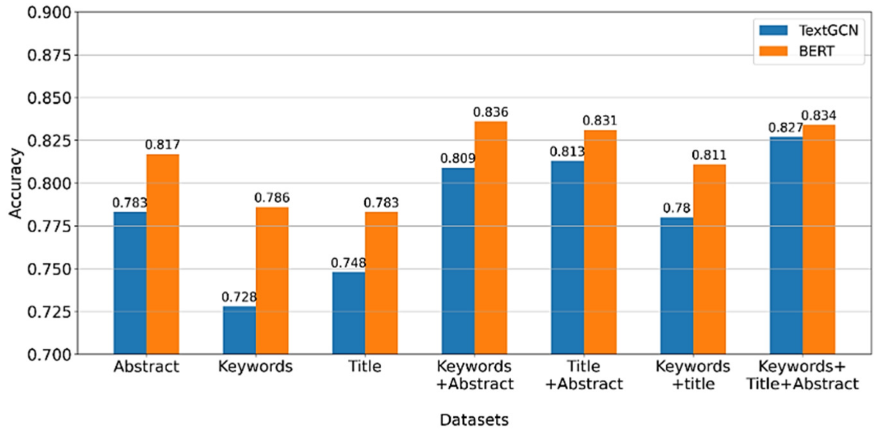

Fig. 2. Classification accuracy of TextGCN and BERT models.

When combining data, TextGCN performs best with Keywords+Title+Abstract concatenation, while the BERT model shows similar accuracy across Keywords+Abstract, Title+Abstract, and Keywords+Title+Abstract, and outperforms methods using Keywords+Title or only a single data type. Overall, across all seven datasets, the BERT model achieves a higher classification accuracy than the TextGCN model.

B. Best Weights for the BERT-GCN Model

The classification accuracy of the model was studied under different λ values for the seven datasets. When λ is between 0.6 and 0.7, the classification accuracy of the model on the six datasets reaches its highest value. For the Title+Abstract dataset only, the model attains the highest classification accuracy when λ=0.9. The performance of the BERT-GCN model depends on the weighted sum of the classification results from the BERT and GCN models. In most datasets, the GCN model has a dominant role, so the BERT-GCN model performs best when λ is between 0.6 and 0.7. When λ=0.9, the BERT-GCN model achieves slightly better classification results on the Title+Abstract dataset than when λ=0.6, but the difference is not significant. This shows that more meaningful word nodes can be obtained from the Title+Abstract dataset than from other datasets, enabling the GCN model to take a leading role.

The classification results of the TextGCN, BERT, and BERT-GCN models were compared across the seven datasets, as shown in Table V. On most datasets, BERT-GCN achieved a higher classification accuracy than both the TextGCN and BERT models. This shows that the BERT-GCN model effectively combines the strengths of BERT and TextGCN to achieve superior performance in paper classification. However, it has higher computational costs because it requires running both BERT and TextGCN simultaneously and combining their results. Additionally, more tests are needed to determine the optimal value of the weight parameter λ for different datasets. The computational costs of the BERT, TextGCN, and BERT-GCN models were compared on the CSL_Abstract dataset. Both TextGCN and BERT-GCN utilize a two-layer GCN network. The computational times for the three models were 53 minutes, 3 minutes, and 70 minutes, respectively. Abstracts are much longer than keywords or titles and therefore contain more valuable information. They also include some information from keywords and titles. The BERT model is more effective at classifying long texts, so various models perform better on abstracts than on titles or keywords.

To evaluate the scalability of the BERT-GCN model, a larger scholarly dataset was introduced, which included abstracts and titles of over 1.27 million English papers from the arXiv platform, categorized into six main groups, such as Mathematics and Physics. An experimental dataset was built by randomly selecting 100K papers from the original set. The classification accuracy for the TextGCN, BERT, and BERT-GCN models is 0.8786, 0.8733, and 0.8864, respectively. This shows that even with a relatively large dataset, the BERT-GCN model's classification accuracy remains higher than that of TextGCN or BERT, demonstrating its scalability.

TABLE IV. COMPARISON OF CLASSIFICATION ACCURACY ON VARIOUS CSL SUB-DATASETS

Dataset	**TextGCN**	**BERT**	**BERT-GCN (best result)**	**λ**
Abstract	0.7827	0.8168	0.8385	0.6
Title	0.7283	0.7858	0.8032	0.6
Keywords	0.7483	0.7827	0.8042	0.7
Keywords+ Abstract	0.8086	0.8355	0.8470	0.7
Title+Abstract	0.8134	0.8306	0.8511	0.9
Keywords+Title	0.7804	0.8106	0.8329	0.7
Keywords+Title +Abstract	0.8268	0.8337	0.8560	0.6

C. Improved BERT-GCN Model Based on Ensemble Learning

Ensemble learning creates a strong predictor by combining multiple base learners to enhance overall prediction accuracy. The main idea is to leverage the differences and complementarities among base learners to minimize generalization errors. Ensemble learning is primarily categorized into three types: Bagging, Boosting, and Stacking. Bagging trains several base learners simultaneously and combines their results through voting or averaging. This study utilized the Bagging method to combine the prediction results of the best-performing BERT-GCN model trained on various paper metadata datasets, and then determined the final classification result by voting.

Since paper metadata includes abstract, keywords, and title, different combinations of this metadata yielded seven datasets: abstract, keywords, title, title_keywords, title_abstract, keywords_abstract, and keywords_title_abstract. One model was trained for each dataset, and the results of multiple models were combined through ensemble learning to achieve the best classification performance. Table V shows the final classification accuracies for various model combinations. A represents a model trained on Abstracts, T a model trained on Titles, K a model trained on Keywords, K_A for a model trained on a combination of Keywords and Abstracts, A+T for the combined prediction results of the two models trained on Abstract and Title datasets, respectively, and so on.

TABLE V. CLASSIFICATION ACCURACY OF BERT-GCN USING ENSEMBLE LEARNING

Number of models	Model combination	Accuracy
7	A + K + T + K_A + T_A + K_T + K_T_A	0.8729
6	A + K + T + K_A + T_A + K_T_A	0.8690
5	A + T + K + K_A + K_T_A	0.8634
4	A + T + K + K_T_A	0.8598
1	K_T_A	0.8560

Based on the experimental results, compared to simply concatenating data such as keywords, title, and abstract as input for the BERT-GCN model, using ensemble voting on prediction results from models trained on seven datasets can boost classification accuracy from 85.6% to 87.29%. Ensemble voting was also tested on models trained on fewer than seven datasets, and their performance was also improved compared to a single model. This shows that employing ensemble voting on models trained with different datasets can effectively capture diverse features from the paper and improve the model's generalization ability.

An ablation study was conducted to analyze how various metadata elements influence ensemble learning by sequentially removing models trained on abstract, keywords, or title, as shown in Table VI. Keywords were found to be highly vital in ensemble learning, and removing them had the most significant impact on classification accuracy, while removing the title had the least effect. The model achieved the best classification results only when all three metadata elements were included.

TABLE VI. ABLATION STUDY FOR METADATA

Metadata	Ensemble learning models	Accuracy
Abstract/Keywords/Title	A + K + T + K_A + T_A + K_T + K_T_A	0.8729
w/o Abstract	K+T+K_T	0.8326
w/o Keywords	A+T+T_A	0.8229
w/o Title	A+K+K_A	0.8501

Although ensemble learning can improve the classification performance of the BERT-GCN model, it requires training multiple models on different datasets, which significantly increases the computational cost compared to training a single model. The most impactful models for ensemble learning can be selected based on the test results, thus balancing computational cost and classification accuracy.

D. Performance Comparison with Baseline Models

Four baseline models from the CLUE (Chinese Language Understanding Evaluation) benchmark [21] were used, which perform classification tasks on the CSL dataset and achieve the highest classification accuracy, as shown in Table VII. It can be seen that the BERT-GCN model, improved through an ensemble learning method, leverages the classification capabilities of multiple models trained using paper titles, keywords, and abstracts. This allows the proposed model to demonstrate better interference resistance and fault tolerance during classification, thereby improving its accuracy. Compared to the TextCNN model, the advantages of the proposed solution are even more evident.

TABLE VII. COMPARISON OF BERT-GCN-BASED ENSEMBLE LEARNING VS. BASELINE MODELS

Model	Accuracy
BERT-GCN-based ensemble learning (Proposed)	0.8729
ALBERT-xxlarge [22]	0.8363
RoBERTa-wwm-large [23]	0.8213
RoBERTa-large	0.8136
XLNet-mid	0.8126
TextCNN	0.7806

IV. CONCLUSION

Current research on paper classification lacks a detailed analysis of how different types of metadata contribute and do not combine various metadata to improve model accuracy. In addition, many experimental datasets have too few samples and are not accessible to the public, making it harder to compare performance. This study used a large publicly available Chinese paper dataset to compare the effectiveness of the TextGCN, BERT, and BERT-GCN models in classifying scientific papers. It also offers an in-depth look at how paper abstracts, keywords, and titles help with classification. Finally, this study recommends combining multiple metadata types through ensemble learning to improve classification results.

The BERT-GCN model combines the strengths of TextGCN and BERT. When the appropriate weights are selected, its classification performance exceeds that of both TextGCN and BERT. For the CSL Chinese dataset, which includes abstracts, keywords, and titles, ensemble voting combines the prediction results of seven models, significantly boosting classification accuracy. Performance comparisons between the improved model and the advanced baseline demonstrate the effectiveness of the proposed approach. This research used only a Chinese dataset with 13 categories. Future work will explore datasets with more categories and those in non-Chinese languages. In addition, the model will also be applied to hierarchical multi-label classification of scientific papers to enhance practical applications.

The BERT-GCN model uses BERT-Base-Chinese as its pre-trained language model. Various BERT variants were also tested, such as Chinese-RoBERTa-WWM-EXT, which employs the Whole Word Masking (WWM) strategy to improve BERT. The learning rate, batch size, activation function, and dropout were set to 1e-4, 128, ReLU, and 0.5, respectively. It was observed that replacing the BERT model in

BERT-GCN with different variants kept the classification accuracy roughly the same. The confusion matrix showed that certain categories had higher classification error rates, which affects overall performance. The next step is to explore methods to handle an imbalanced dataset. For example, the SMOTEDNN model proposed in [24] used the Synthetic Minority Oversampling Technique (SMOTE) to greatly improve classification accuracy. Another plan is to explore combining the BERT model with other deep learning models, such as CDLSTM [25], to improve overall performance.

ACKNOWLEDGMENT

The authors acknowledge and thank the Faculty of Computer and Mathematical Sciences and IPSIS of Universiti Teknologi MARA for their valuable support of this research.

REFERENCES

[1] I. Jaya, I. Aulia, S. M. Hardi, J. T. Tarigan, M. S. Lydia, and Caroline, "Scientific Documents Classification Using Support Vector Machine Algorithm," *Journal of Physics: Conference Series*, *Medan, Sumatera Utara, Indonesia,* vol. 1235, no. 1, June 2019, Art. no. 012082, https://doi.org/10.1088/1742-6596/1235/1/012082.

[2] Y. Xiaohua and G. Haiyun, "Improved Bayesian algorithm based automatic classification method for bibliography,", *Computer Science*, vol. 45, no. 8, pp. 203-207, 2018.

[3] Q. Li *et al.*, "A Survey on Text Classification: From Traditional to Deep Learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 13, no. 2, pp. 1–41, Apr. 2022, https://doi.org/10.1145/3495162.

[4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, 2018, https://doi.org/10.48550/ARXIV.1810.04805.

[5] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, *China,* vol. 29, no. 1, Feb. 2015, https://doi.org/10.1609/aaai.v29i1.9513.

[6] R. Sarasu, K. K. Thyagharajan, and N. R. Shanker, "SF-CNN: Deep Text Classification and Retrieval for Text Documents," *Intelligent Automation & Soft Computing*, vol. 35, no. 2, pp. 1799–1813, 2023, https://doi.org/10.32604/iasc.2023.027429.

[7] M. I. Salih, S. M. Mohammed, A. Kh. Ibrahim, O. M. Ahmed, and L. M. Haji, "Fine-Tuning BERT for Automated News Classification," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 22953–22959, June 2025, https://doi.org/10.48084/etasr.10625.

[8] X. Luo, S. Mutalib, and S. R. Syed Aris, "Chinese paper classification based on pre-trained language model and hybrid deep learning method," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 14, no. 1, Feb. 2025, Art. no. 641, https://doi.org/10.11591/ijai.v14.i1.pp641-649.

[9] Z. Liu, F. Li, G. Hao, X. He, and Y. Zhang, "GCN-LSTM: multi-label educational emotion prediction based on graph Convolutional network and long and short term memory network fusion label correlation in online social networks," *Computer Science and Information Systems*, vol. 21, no. 4, pp. 1583–1605, 2024, https://doi.org/10.2298/CSIS240314049L.

[10] Y. Peng, W. Wu, J. Ren, and X. Yu, "Novel GCN Model Using Dense Connection and Attention Mechanism for Text Classification," *Neural Processing Letters*, vol. 56, no. 2, Apr. 2024, Art. no. 144, https://doi.org/10.1007/s11063-024-11599-9.

[11] H. Cui, G. Wang, Y. Li, and R. E. Welsch, "Self-training method based on GCN for semi-supervised short text classification," *Information Sciences*, vol. 611, pp. 18–29, Sept. 2022, https://doi.org/10.1016/j.ins.2022.07.186.

[12] H. Li, Y. Yan, S. Wang, J. Liu, and Y. Cui, "Text classification on heterogeneous information network via enhanced GCN and knowledge," *Neural Computing and Applications*, vol. 35, no. 20, pp. 14911–14927, July 2023, https://doi.org/10.1007/s00521-023-08494-0.

[13] N. A. Alabdulkarim, M. A. Haq, and J. Gyani, "Exploring Sentiment Analysis on Social Media Texts," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14442–14450, June 2024, https://doi.org/10.48084/etasr.7238.

[14] X. Zhang, X. Yu, X. Liu, and X. Lyu, "Scientific Paper Classification by Fusing BERT and GCN," in *2023 International Conference on Intelligent Education and Intelligent Research (IEIR)*, Wuhan, China, Nov. 2023, pp. 1–6, https://doi.org/10.1109/IEIR59294.2023.10391239.

[15] Y. Jang, K. Won, H. Choi, and S. Y. Shin, "Classification of Research Papers on Radio Frequency Electromagnetic Field (RF-EMF) Using Graph Neural Networks (GNN)," *Applied Sciences*, vol. 13, no. 7, Apr. 2023, Art. no. 4614, https://doi.org/10.3390/app13074614.

[16] Y. Guo, L. Lin, and Y. Liu, "Automatic Abstract Classification Method Based on BERT-GCN-Res Net," *Journal of Tianjin University of Science & Technology*, vol. 37, no. 2, pp. 51–56, 2022.

[17] M. Lu *et al.*, "EP-BERTGCN: A Simple but Effective Power Equipment Fault Recognition Method," in *2022 4th International Conference on Information Technology and Computer Communications (ITCC)*, Guangzhou, China, June 2022, pp. 64–68, https://doi.org/10.1145/3548636.3548646.

[18] Y. Zhang, Y. Xu, and Y. Zhang, "A Graph Neural Network Node Classification Application Model with Enhanced Node Association," *Applied Sciences*, vol. 13, no. 12, June 2023, Art. no. 7150, https://doi.org/10.3390/app13127150.

[19] Y. Lin *et al.*, "BertGCN: Transductive Text Classification by Combining GCN and BERT." arXiv, Mar. 21, 2022, https://doi.org/10.48550/arXiv.2105.05727.

[20] Y. Li *et al.*, "CSL: A Large-scale Chinese Scientific Literature Dataset." arXiv, Sept. 12, 2022, https://doi.org/10.48550/arXiv.2209.05034.

[21] L. Xu *et al.*, "CLUE: A Chinese Language Understanding Evaluation Benchmark," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain, 2020, pp. 4762–4772, https://doi.org/10.18653/v1/2020.coling-main.419.

[22] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." arXiv, 2019, https://doi.org/10.48550/ARXIV.1909.11942.

[23] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-Training With Whole Word Masking for Chinese BERT," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3504–3514, 2021, https://doi.org/10.1109/TASLP.2021.3124365.

[24] A. Attaallah and R. Ahmad Khan, "SMOTEDNN: A Novel Model for Air Pollution Forecasting and AQI Classification," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 1403–1425, 2022, https://doi.org/10.32604/cmc.2022.021968.

[25] M. A. Haq, "CDLSTM: A Novel Model for Climate Change Forecasting," *Computers, Materials & Continua*, vol. 71, no. 2, pp. 2363–2381, 2022, https://doi.org/10.32604/cmc.2022.023059.