Dual-Channel Prompt-Enhanced BERT-RCNN with Adapters for Schizophrenia Detection from Dialogue Text

Runzhe Zhang, Bowei Pan, Zhou Yuan, Yan Ling, Lianghuai Yang, and Keji Mao

Zhejiang University of Technology, Hangzhou, Zhejiang 310023, China
maokeji@zjut.edu.cn

Abstract. Schizophrenia is a severe mental disorder that imposes significant burdens on individuals, families, and society, making early detection critical for improving patient outcomes. However, current diagnosis relies predominantly on subjective clinical assessment. AI-assisted screening presents a promising alternative. In particular, automated analysis of patient language, a direct reflection of thought disorder, offers an objective and highly informative data modality. Despite this potential, research and practice are currently hindered by a scarcity of Chinese textual datasets and an over-reliance on non-linguistic data. To address these limitations, we construct the Chinese Schizophrenia Text Dataset (CSTD). Furthermore, we propose a novel framework designed to decouple text into "general" and "visual and auditory hallucination" features. This framework utilizes prompt templates to inject clinical prior knowledge, employs BERT for feature extraction, and integrates dual sets of Adapters after each Transformer layer to learn the decoupled features in parallel. The resulting outputs are fused and processed by a Recurrent Convolutional Neural Network (RCNN) for final classification. Experiments demonstrate that this framework achieves a leading accuracy of 95.58% on the CSTD, outperforming other strong models (including BERT and SBERT-BiLSTM) by at least 7.68%. Crucially, this robust performance highlights the strategic advantage of the parameter-efficient Adapter mechanism, which was deliberately chosen to restrict the number of trainable parameters and effectively prevent overfitting on the limited clinical dataset. These results validate the model's efficacy in recognizing linguistic patterns associated with schizophrenia and support its viability as an auxiliary screening tool in clinical outpatient settings.

Keywords: Schizophrenia Detection, CSTD, Visual and Auditory Hallucination Features, Dual-Channel Adapter Feature Learning.

1 Introduction

1.1 Research Background and Significance

Schizophrenia is a severe mental disorder characterized by cognitive, perceptual, and behavioral impairments that profoundly disrupt daily life [1, 2]. Currently, diagnosing schizophrenia involves detailed psychiatric interviews to assess symptoms like functional impairment and speech disorders [3]. However, the subjective nature of these interviews heightens the risk of misdiagnosis [4]. Furthermore, reliance on manual diagnosis by experienced clinicians is constrained by shortages of practitioners, high costs, and long diagnostic cycles. To address the lack of reliable objective biomarkers, there is a critical need for more sensitive screening tools that can facilitate faster, more precise diagnosis and improve treatment outcomes.

In recent years, computational psychiatry has introduced objective quantitative analysis using multimodal data as a new pathway for screening [5]. Among various data types, dialogue text has emerged as a crucial tool due to its accessibility and the rich psychological cues it contains. Because patients often exhibit specific linguistic features, such as reduced coherence, loose semantic associations, and increased self-referential content, these language patterns present a promising opportunity for automated identification [6].

1.2 Related Works

The early screening and objective diagnosis of schizophrenia represent core challenges in psychiatry and computational psychiatry. Recently, the rapid advancement of artificial intelligence has driven researchers to explore machine learning and deep learning techniques to assist in identifying such mental disorders.

Building reliable AI models fundamentally relies on high-quality, large-scale datasets. However, despite a global trend toward open data sharing has made neuroimaging databases publicly available, datasets specifically covering schizophrenia remain scarce compared to other conditions. For instance, across 42 shared global neuroimaging databases, schizophrenia samples (675 cases) are drastically outnumbered by those for dementia (5,865 cases) or depression (1,245 cases) [7]. Furthermore, existing databases are heavily biased toward neurophysiological data, such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and electrocardiograms (ECG).

This scarcity of such resources is not merely a matter of research focus, but a direct consequence of profound, objective barriers inherent to clinical psychiatry. Foremost among these are strict patient privacy protections and rigorous ethical constraints, which heavily restrict the recording, storage, and public dissemination of clinical interactions. Furthermore, the unique and highly sensitive nature of mental health disorders—particularly schizophrenia, where patients may actively experience severe paranoia, delusions, or disorganized thought—makes standard data collection procedures exceptionally challenging. Securing informed consent and ensuring patient well-being during data collection require highly specialized, restrictive protocols. Due to these systemic hurdles, publicly available, finely annotated clinical dialogue datasets remain exceedingly rare on a global scale. Within the Chinese linguistic and cultural context, this scarcity is amplified to a near-total void, establishing a formidable barrier to training and validating domain-specific AI models.

Methodologically, neurophysiological signals like EEG and ECG have been widely utilized with machine and deep learning to classify various mental disorders [8-10]. However, these approaches frequently encounter challenges regarding small sample sizes, data heterogeneity, and limited model generalizability across different clinical scenarios [8-10].

To address these limitations, language-based analysis has emerged as a promising avenue for clinical translation. Language disturbance is a core symptom of schizophrenia, with features such as speech coherence, semantic density, and syntactic structure closely linked to formal thought disorder. Automated language analysis, combined with AI, has demonstrated significant potential in predicting symptoms across various mental disorders [11]. Semantic analysis, for example, effectively identifies psychosis-related markers; studies using part-of-speech tagging [12] and SBERT-based word embeddings [13] have classified schizophrenia spectrum disorders with over 86% accuracy.

Building on these linguistic approaches, Large Language Models (LLMs) are increasingly utilized in psychiatric research, but challenges remain. Pugh et al. [14] found a trade-off between accuracy and consistency when using LLMs to assess thought disorder. Similarly, Gargari et al. [15] demonstrated that while GPT-3.5 and GPT-4 outperform traditional NLP models in interpreting DSM-5 criteria, they struggle with specific disorder nuances, underscoring the need for domain-specific fine-tuning. Conversely, Botelle et al. [16] reported high precision and recall (ranging from 89% to 98%) using a fine-tuned BioBERT model for identifying interpersonal violence in records, highlighting the efficacy of pre-trained Transformers for specialized tasks.

Ultimately, while neurophysiological analysis is advanced, it is constrained by sample size and heterogeneity. Conversely, language-based analysis is rapidly evolving, offering the ability to quantify semantic and syntactic features via techniques ranging from word vectors to LLMs. However, a critical bottleneck remains: the near-total absence of dedicated Chinese dialogue datasets for schizophrenia. This lack of data, combined with the difficulty of adapting general LLMs to the mental health domain, severely hinders deeper research in this field.

1.3 Main Work

To mitigate the reliance on subjective assessment and the lack of Chinese textual data, this paper introduces a novel framework for the automated screening of schizophrenia. Our approach bridges the gap between clinical psychiatry and natural language processing. Our main works are:

- Construction of the CSTD: The Chinese Schizophrenia Text Dataset (CSTD) is a rigorously annotated and ethically curated corpus. This dataset specifically isolates visual and auditory hallucination markers, effectively addressing the critical data scarcity in Chinese computational psychiatry.

- Clinical-Prior Prompt Learning: A novel prompt-learning mechanism is proposed to inject clinical priors into the model. By wrapping input text in diagnostic templates, we guide the model to explicitly attend to both general linguistic coherence and specific perceptual abnormalities.
- Parameter-Efficient Decoupling Architecture: A parameter-efficient architecture is designed to decouple text features using independent Adapters embedded within a shared BERT backbone. This allows for the parallel learning of "general semantic" and "hallucinatory" representations, reducing parameter overhead while maximizing model interpretability.
- RCNN-Based Feature Fusion: An RCNN fusion module is used to process the decoupled features. This module captures both local semantic nuances and global contextual dependencies, significantly enhancing classification accuracy.

Structure of the Paper: Section 2 outlines the existing landscape of datasets and methodologies. Section 3 elaborates on the proposed framework, including the prompt design and Adapter architecture. Section 4 discusses the experimental results and performance benchmarks. Section 5 summarizes the findings and future directions.

2 Dataset

The dataset was collected from a specialized psychiatric hospital in Hangzhou and was originally stored as audio files using a Newsmy Digital Recorder RD07. The dataset was converted from the collected audio using Automatic Content Recognition (ACR) software. This software was independently developed by the author zhaoyun007 based on Baidu's open-source Paddle Speech framework. Due to factors such as environmental noise, device interference, and patient accents in the original data, a certain error rate occurred during transcription, necessitating post-processing manual correction.

After obtaining the conversion results, a secondary review of the data was conducted, ultimately forming the CSTD dataset. The detailed information of this dataset is shown in **Table 1**. The CSTD dataset contains a total of 240 text entries, divided into 114 schizophrenia samples and 126 normal control samples. Text lengths range from 10 to 88 characters, with an average length of 35 characters per entry. As discussed in Section 1.2, acquiring clinical dialogue data from schizophrenia patients involves immense ethical and practical challenges. Therefore, while our finalized CSTD contains 240 stringently annotated samples, it represents a substantial and exceedingly rare resource for Chinese computational psychiatry.

To align with the dual-channel architecture of our proposed model—which processes general linguistic features alongside specific audiovisual hallucination features—the dataset was partitioned accordingly, as outlined in **Table 2**. For normal controls, the Universal channel uses all 126 samples, while the audiovisual hallucination channel uses only 72. For schizophrenia patients, the Universal channel uses all 114 samples, whereas the audiovisual hallucination channel uses a subset of 65 samples that exhibit clear hallucinatory features.

Table 1. Composition of CSTD dataset.

Item	Schizophrenia	Normal
Minimum Length	10	15
Maximum Length	88	80
Average Length	32	37
Test Subjects	34	45
Sample Size	114	126

Table 2. Distribution of training samples.

Dataset	Universal Channel		Audiovisual Hallucination Channel	
	Schizophrenia	Normal	Schizophrenia	Normal
Train	82	90	35	40
Validation	16	18	14	14
Test	16	18	16	18
Total	114	126	65	72

To practically illustrate the linguistic distinctions among these categories, **Table 3** provides representative textual excerpts. These examples clearly highlight the characteristic variations between general schizophrenia speech, speech marked by audiovisual hallucinations, and normal control dialogues, providing an intuitive context for the model's classification task.

Table 3. Text data example.

Category	No.	Text Example
Schizophrenia with Audiovisual Hallucinations	1	Unreal people and scenes appear in my mind, but I have seen too many of them, and I have started to believe that person is real.
General Schizophrenia	1	I'm going mad, and it's spread overseas. My current behavior shows I'm going mad. I want to be hospitalized—let me stay in the hospital.
Normal	1	Currently, the situation is good. I sleep about seven to eight hours a day. My mood is quite positive, and I feel very confident.

3 Model

Observing that most schizophrenia samples exhibit audiovisual hallucinations, this study proposes a corresponding fusion semantic model (**Fig. 1**). The model employs a

BERT-based dual-channel architecture sharing the same backbone. The Universal channel uses Adapters after each BERT layer to extract general features, while the audiovisual hallucination channel uses separate Adapters to extract hallucination-specific features. Features from both parallel channels are fused and fed into an RCNN for higher-level feature extraction and classification. To incorporate diagnostic prior knowledge, a Prompt Learning strategy is adopted. These prompted inputs guide the subsequent dual-channel Adapters to focus on distinct symptom categories from the outset. The following sections detail the dual-channel modules and the global feature extraction module (RCNN) to elucidate the model's construction principles.

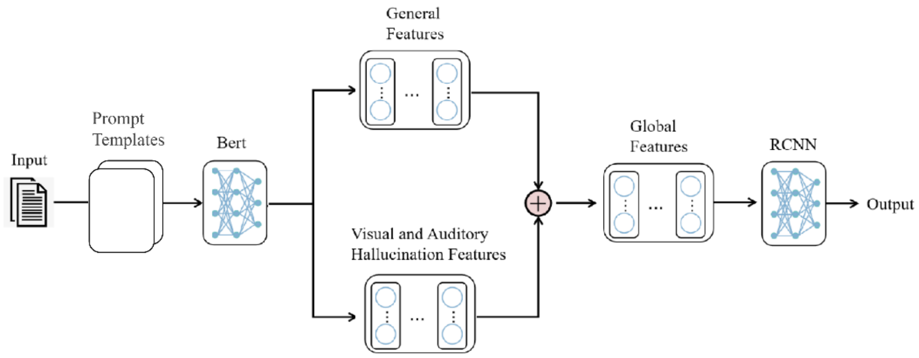

Fig. 1. The structure of semantic model for audio visual illusion fusion.

3.1 Prompt

To incorporate clinical prior knowledge, we employ a dual-template prompt injection strategy that reframes classification into a masked prediction task. This involves designing distinct natural language prompts for "general symptom recognition" and "audiovisual hallucination symptom recognition," transforming the classification task into a BERT-friendly masked prediction format.

Based on clinical diagnostic logic, this paper designs a pair of concise prompt templates:

General symptom prompt template: [Text] The overall psychological state reflected in the above dialogue is [MASK].

Hallucination symptom prompt templateSchizophrenia is a chronic: [Text] Does this description contain abnormal audiovisual perceptions (e.g., auditory or visual hallucinations)? [MASK].

For each text sample, we create two prompt-formatted sequences. These are encoded independently by a shared BERT model to produce differentiated initial features. The general-prompt features are then processed by the universal channel Adapter, while the audiovisual hallucination-prompt features are routed to the hallucination channel Adapter, ensuring specialized refinement within the dual-path architecture.

3.2 Universal channel module

The Universal channel extracts general features from schizophrenia samples. Crucially, this module establishes the foundational shared mechanism for the entire architecture: the vast majority of the BERT backbone parameters are entirely frozen and shared across both channels. We employ a partial fine-tuning strategy where only the last four BERT layers, the Embedding layer, and the layer-specific Adapter modules are updated. This shared foundation prevents on our limited dataset while enabling high-level semantic adaptation, preserving BERT's general expressiveness and gaining task-specific learning capacity.

Built upon the BERT Transformer Encoder [17], the Universal channel utilizes Adapters to capture general language disturbances, such as speech incoherence, while keeping the majority of the backbone frozen to prevent overfitting [18].

The workflow diagram of the Universal channel module is shown in **Fig. 2**. The network architecture and the model's input and output will now be detailed as follows:

- Network Architecture: Within the Transformer Encoder, an Adapter module is added after the Feedforward and Add&Norm operations to efficiently extract General Features. The structures are shown in **Fig. 3** and **Fig. 4**.
- Model Input: The input consists of input_ids, attention_mask, and token_type_ids. To distinguish between the prompt template and the original patient text, we set the token type for the prompt part to 0 and for the original text part to 1. This design helps the model understand the structured semantics of the input.
- Model Output: The input text is processed layer by layer through the shared BERT Transformer encoder, and the output of each Transformer layer is fed into the corresponding Adapter for feature adaptation. the Universal channel outputs the feature representation from the final Adapter layer, resulting in the final general feature representation as the network output. This can be denoted as $c \in R^{T_A \times n}$, where T_A and n represent the output sequence length and feature dimension, respectively. The specific dimensions of the general features are 512 × 768.

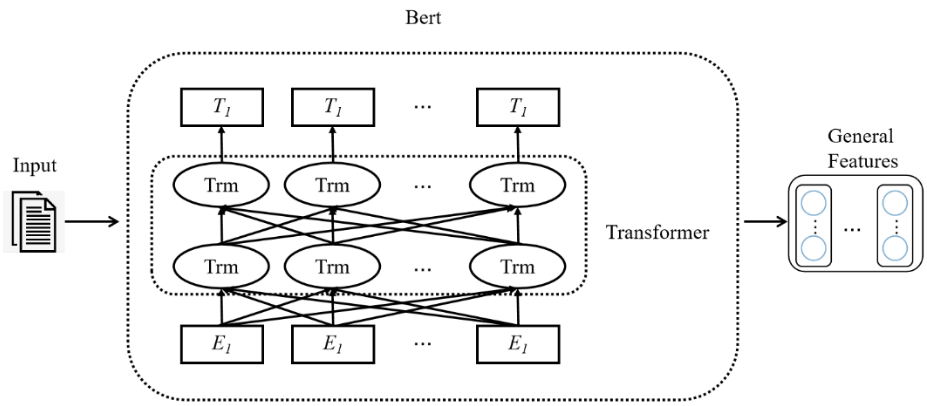

Fig. 2. Universal channel module.

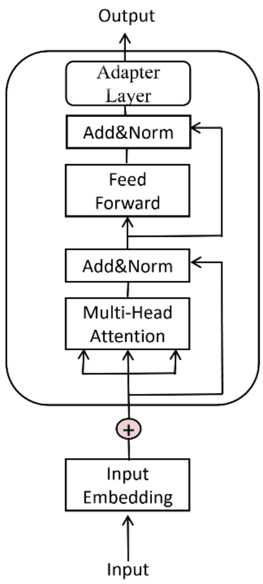

Fig. 3. The specific composition of Encoder.

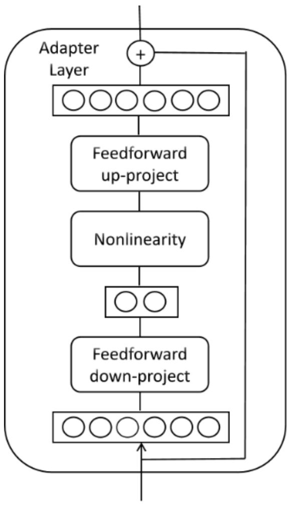

Fig. 4. The specific composition of Adapter.

3.3 Audiovisual hallucination channel module

Because the network input format and foundational backbone structure are identical to those detailed in Section 3.2, this section focuses exclusively on the unique design mechanisms and clinical rationale of this specialized channel.

The audiovisual hallucination channel extracts specific visual and auditory hallucination features from schizophrenia samples. Decoupling is essential because sparse hallucinatory signals can be easily diluted by general semantic features in a unified parameter space. To prevent this interference and ensure distinct representation, this separate channel provides an independent operational space specifically for the hallucination prior prompt. While it shares the BERT encoder with the Universal channel, the adaptation is handled via separate, independent Adapters.

A core mathematical mechanism of this module is its utilization of a bottleneck low-rank transformation within these dedicated Adapters. After each shared Transformer layer, the hallucination Adapter applies this transformation specifically to features related to hallucinations. Because hallucinatory signals are sparse within the text, this bottleneck low-rank transformation acts as a crucial mechanism for information filtering and feature enhancement. It seamlessly isolates targeted clinical cues from the broader linguistic noise, enabling targeted modeling of hallucinations in the shared representation space. This decoupling ensures that distinct clinical features are preserved and amplified without adding complex, entirely separate network branches.

Model Output: The output of this section is referred to as visual and auditory hallucination features, which can be represented as $h \in R^{T_A \times n}$, where T_A and n represent the output sequence length and feature dimension, respectively. The specific dimensions of the features are 512 × 768.

3.4 Global feature extraction module

The Universal channel module and the audiovisual hallucination channel module extract general features and visual and auditory hallucination features respectively. These two features are then fused using the following formula:

$$g = c + h \tag{1}$$

After concatenation, the global feature is obtained, which can be represented as $g \in R^{T_A \times n}$, where T_A and n denote the output sequence length and feature dimension, respectively. The specific dimensions of the global feature are 512 × 1536.

The global feature extraction module performs secondary feature extraction and conducts binary classification. It first furthers processes and refines the features before classification. As shown in **Fig. 5**, the global feature extraction module utilizes an RCNN [19], which integrates recurrent and convolutional structures to capture both long-term dependencies and local patterns. The network architecture and input/output details are introduced next:

- Network Architecture: RCNN primarily consists of a recurrent structure and a convolutional structure. First, a bidirectional RNN is used to learn the left contextual representation $c_l(w_i)$ and the right contextual representation $c_r(w_i)$ of the current word w_i. These are then fused with the word's own representation $e(w_i)$ to form the input x_i for the convolutional layer. The specific formula is as follows:

$$x_i = [c_l(w_i); e(w_i); c_r(w_i)] \tag{2}$$

The specific formula after inputting into the convolutional layer is as follows:

$$y_i^{(2)} = \tanh\left(W^{(2)}x_i + b\right) \tag{3}$$

After passing through the convolutional layer, the representations of all words are obtained. These are then processed through a max-pooling layer and a fully connected layer to derive the representation of the entire text. Finally, classification is performed via a softmax layer. The specific formulas are as follows:

$$y^{(3)} = \max_{1 \leqslant i \leqslant n} y_i^{(2)} \tag{4}$$

$$y^{(4)} = W^{(4)}y^{(3)} + b^{(4)} \tag{5}$$

$$p_i = \frac{e^{y_i^{(4)}}}{\sum_{k=1}^{n} e^{y_k^{(4)}}} \tag{6}$$

- Model Input: The network input is the global feature g.
- Model Output: The network output is a binary classification result, representing the probabilities of the text data belonging to a normal individual and a schizophrenia patient, respectively, with a size of 1×2.

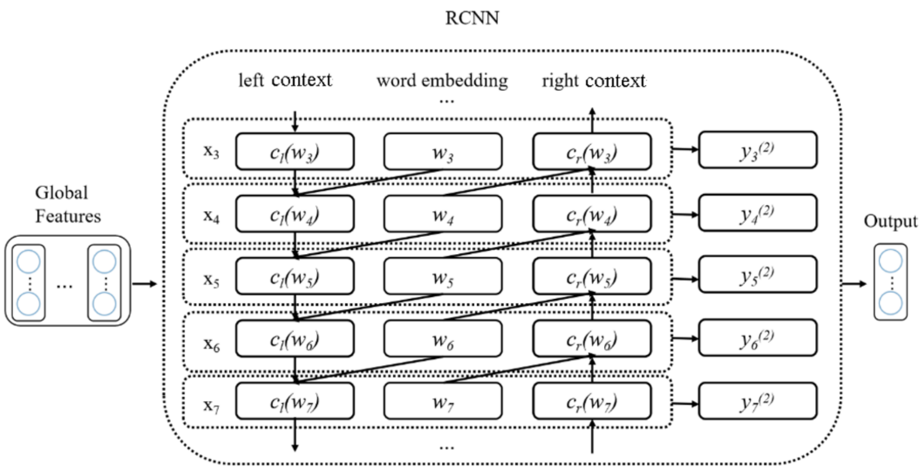

Fig. 5. Global feature extraction module.

4 Experiments

4.1 Evaluation Metrics

To validate the effectiveness of the model proposed in this paper, accuracy, precision, recall, and F1-score are used as evaluation metrics for the experimental results. Their calculation formulas are as follows:

$$Accuracy = \frac{TP+TN}{P+N} \tag{7}$$

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

$$F1 = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} = \frac{2TP}{FP+2TP+FN} \tag{10}$$

4.2 Experimental Environment and Parameters

The visual and auditory hallucination fusion semantic model proposed in this paper utilizes two network models: BERT and RCNN. After conducting a series of ablation experiments and comparative experiments, the optimal parameters for each network were determined. The specific parameter configurations are shown in **Table 4**.

Table 4. Experimental configuration.

Parameters	Value
Batch Size	4
Maximum Sentence Length	512
Training Epochs	26
Number of Classes	2
Learning Rate	0.0001
Dropout Rate	0.5
Hidden Unit Size	768
Number of Filters	100
Filter Sizes	[2,3,4]

4.3 Experimental Results and Analysis

1) Ablation Study.

The visual and auditory hallucination fusion semantic model primarily consists of the following three modules: the universal channel module, the audiovisual hallucination channel module, and the global feature extraction module. Both the universal channel module and the audiovisual hallucination channel module employ the BERT pre-trained model as their backbone but utilize different datasets during the fine-tuning stage. The global feature extraction module is implemented using the RCNN network. To validate the effectiveness of each module, this section first conducted experiments with the universal channel module alone, then added the audiovisual hallucination channel module for further experiments, finally incorporated the global feature extraction module for testing. The experimental results are presented in **Table 5**.

Ablation results confirm that the inclusion of the audiovisual hallucination channel improves discriminative capacity, and the RCNN module further enhances accuracy by extracting representative global features.

Table 5. Ablation experiment.

Module	Evaluation Indicators			
	Accuracy	Precision	Recall	F1
Universal	79.66 %	82.31 %	79.66 %	78.43 %
Universal+Audiovisual	83.29 %	85.02 %	84.14 %	82.79 %
Universal+Audiovisual+Global	95.58 %	96.77 %	93.75 %	95.23 %

2) Comparative Experiment on Global Feature Extraction Modules.

After validating the performance improvement brought by the inclusion of the audiovisual hallucination channel and the global feature extraction module, this section focuses on investigating the specific structure of the global feature extraction module. The study compares the classification accuracy of the model when DPCNN [20], CNN [21], RNN [22], and RCNN [23] are used as the global feature extraction module.

Comparative results in **Table 6** show that while CNN-based models (CNN, DPCNN) outperform RNN by capturing local features, RCNN achieves the best performance by integrating both local patterns and long-term dependencies.

Table 6. Comparative experiment.

Module	Evaluation Indicators			
	Accuracy	Precision	Recall	F1
DPCNN	89.43 %	89.62 %	89.43 %	89.06 %
RNN	79.82 %	83.75 %	78.84 %	78.95 %
CNN	86.17 %	87.44 %	85.68 %	85.83 %
RCNN	95.58 %	96.77 %	93.75 %	95.23 %

3) Overall Model Comparison Experiment.

After introducing the complete model of this paper, this section will first compare it with other common text classification algorithms applied in mental disorder (primarily schizophrenia) classification research to validate the performance of the proposed model. The study compares several algorithms including LR [24], SVM [25], Naive Bayes [26], TextCNN [27], TextRCNN [28], BERT [29], and SBERT-BiLSTM [30].

As shown in **Table 7**, traditional machine learning (LR, SVM, NB) struggles with complex psychiatric patterns. While BERT-based models significantly improve accuracy, our proposed framework achieves a leading 95.58% by tailoring the architecture to specific schizophrenia symptoms.

Table 7. Overall model comparative experiment.

Module	Evaluation Indicators			
	Accuracy	Precision	Recall	F1

LR	73.50 %	70.60 %	75.00 %	72.70 %
SVM	70.60 %	66.70 %	75.00 %	70.60 %
Naive Bayes	73.50 %	76.90 %	62.50 %	69.00 %
TextCNN	79.40 %	76.50 %	81.20 %	78.80 %
TextRCNN	76.50 %	70.00 %	87.50 %	77.80 %
Bert	82.10 %	82.00 %	82.40 %	82.90 %
SBERT-BiLSTM	87.90 %	90.30 %	83.70 %	86.40 %
Ours	95.58 %	96.77 %	93.75 %	95.23 %

4) Cross-Validation Stability Analysis.

To further assess the stability and robustness of our proposed model given the limited dataset size, we conduct a 5-fold stratified cross-validation experiment on the CSTD dataset. The whole dataset is randomly split into five folds of equal size while preserving the original class distribution in each fold. In each iteration, one fold is used as the validation set and the remaining four folds form the training set. The complete model is trained from scratch for each fold using the same hyperparameter settings reported in **Table 4**.

The mean and standard deviation of accuracy, precision, recall, and F1-score across all five folds are summarized in **Table 8**.

Table 8. 5-fold cross-validation results.

Evaluation Metric	Mean (%)	Std (%)
Accuracy	94.01	1.82
Precision	95.12	1.65
Recall	92.89	2.10
F1-score	93.97	1.77

The model achieves a mean accuracy of 94.01% with a relatively low standard deviation of 1.82%, demonstrating consistently strong performance across different data splits. This confirms that the high performance reported on the original test set is not an artifact of a particularly favorable random split, and that the dual-channel architecture generalizes stably within the available data distribution.

It should be noted that cross-validation can only verify stability within the given dataset. The small size of CSTD remains a limitation, and the generalizability to external populations from different hospitals or linguistic backgrounds is yet to be examined. Nevertheless, these results significantly strengthen confidence in the model's reliability on the current data and serve as a solid foundation for future large-scale validation.

5) Prompt Sensitivity Analysis.

To ensure robustness against wording variations, we evaluated nine combinations of functionally equivalent prompt variants. Specifically, for the universal channel and the audiovisual hallucination channel, we designed multiple prompt variants that preserve the same semantic intention while only changing wording expressions. For the universal channel, the prompt variants are defined as follows:

T1: [Text] The overall psychological state reflected in the above dialogue is [MASK].

T2: [Text] Based on the dialogue, the patient's mental condition can be described as [MASK].

T3: [Text] From the conversation, the psychological status is [MASK].

For the hallucination channel, the prompt variants are defined as follows:

H1: [Text] Does this description contain abnormal audiovisual perceptions (e.g., auditory or visual hallucinations)? [MASK].

H2: [Text] Are there any hallucinatory experiences (such as hearing voices or seeing unreal things) in the description? [MASK].

H3: [Text] Does the patient report any abnormal sensory perceptions (auditory or visual)? [MASK].

All universal-channel prompts and hallucination-channel prompts were combined pairwise, resulting in nine prompt combinations. All experiments were conducted on the same test set under identical training settings. The detailed results are shown in **Table 9**.

Table 9. Prompt sensitivity analysis under wording-equivalent prompt variations.

Prompt	Accuracy	Precision	Recall	F1
T1 + H1	95.58 %	96.77 %	93.75 %	95.23 %
T1 + H2	95.11 %	96.05 %	93.41 %	94.70 %
T1 + H3	94.82 %	95.64 %	93.12 %	94.36 %
T2 + H1	95.34 %	96.31 %	93.52 %	94.88 %
T2 + H2	94.97 %	95.84 %	93.26 %	94.53 %
T2 + H3	94.63 %	95.51 %	92.95 %	94.19 %
T3 + H1	95.21 %	96.14 %	93.33 %	94.72 %
T3 + H2	94.89 %	95.77 %	93.08 %	94.38 %
T3 + H3	94.54 %	95.32 %	92.87 %	94.03 %
Mean ± Std	95.01 ± 0.34	95.93 ± 0.42	93.25 ± 0.29	94.56 ± 0.36

Results in **Table 9** show consistent performance across different prompt wordings (std < 1%), confirming the framework's robustness to semantic variations and its reliance on stable clinical cues rather than specific templates.

4.4 Clinical Interpretability and Case Study

To improve the transparency and clinical viability of the proposed model, this section analyzes the internal feature attribution behavior of the hallucination channel and evaluates whether the model's predictions align with established psychiatric diagnostic principles.

Since schizophrenia screening relies heavily on identifying symptom-related linguistic expressions, the model must focus on clinically meaningful semantic patterns rather than irrelevant lexical artifacts. To verify this, we extract high-contribution keywords from correctly classified schizophrenia samples to evaluate feature attribution (**Table 10**), and conduct a case-based analysis on representative dialogue samples from the CSTD dataset (**Table 11**).

Table 10. High-contribution clinical keyword attribution analysis.

Keyword	Frequency	Symptom Type	Clinical Interpretation
voice	41	Auditory hallucination	Hearing unreal voices
auditory hallucination	35	Auditory hallucination	Explicit hallucination symptom
hearing	33	Auditory hallucination	Perceptual abnormality
talk	27	Auditory hallucination	Voice interaction symptom
scolding me	19	Persecutory delusion	Delusion of reference
ghost	14	Visual hallucination	Unreal visual perception
beside the bed	11	Visual hallucination	Hallucinatory scene perception
can't control	16	Disorganized behavior	Impulse dysregulation
bad words	18	Delusion	Paranoid cognition
The sound in my mind	13	Internal auditory hallucination	Intrusive hallucinatory thought

Table 11. Clinical case study analysis.

Case	Dialogue Sample	Prediction	Confidence	Clinical Interpretation
Case 1	I always hear someone talking to me at night.	Schizophrenia	0.973	Typical auditory hallucination
Case 2	I have been sleeping well lately, my emotions are stable, and my life is normal.	Normal	0.952	Healthy control
Case 3	Sometimes I feel like someone is watching me, but I don't know if I'm overthinking it.	Schizophrenia	0.624	Borderline paranoid symptom

As shown in **Table 11**, the model's predictions align with DSM-5 criteria, demonstrating high confidence in typical positive symptoms (Case 1) and realistic uncertainty in borderline paranoid scenarios (Case 3). In conclusion, this clinical interpretability analysis provides empirical evidence that the dual-channel architecture successfully separates general semantic representations from symptom-specific psychopathological features. By grounding its predictions in features that are highly consistent with established psychiatric diagnostic principles, the proposed framework demonstrates robust clinical validity and highlights its potential as a trustworthy decision-support tool in real-world clinical settings.

5 Conclusion and Limitations

In response to the reliance on subjective clinical assessment and the lack of linguistic datasets in China, this paper presents a complete pipeline for the intelligent screening of schizophrenia. We introduced the CSTD, filling a significant gap in Chinese computational psychiatry, and developed a dual-pathway model capable of distinguishing between general linguistic impairments and specific hallucination patterns. With an accuracy of 95.58%, our approach demonstrates that disentangling symptom-specific features yields superior performance compared to generic NLP methods.

However, the transition from a research model to a clinical tool requires further validation due to certain limitations. Primarily, our current sample size, while high-quality, is relatively small. Moving forward, the central focus of our next phase of research will be large-scale data acquisition through multi-center trials to ensure the model's reliability across different demographics. Future work will focus on multi-center trials for large-scale validation and the integration of acoustic and visual signals to develop a comprehensive multimodal digital phenotype for schizophrenia.

Acknowledgments. This research is supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province under No. 2026C02A1019, 2026C02A1222，2026C02A1018; Zhejiang Provincial Natural Science Foundation of China under Grant No: LTGG23F020002; Zhejiang University of Technology School-enterprise Cooperation Project under Grant No: KYY-HX-20250798,KYY-HX-20230493, KYY-HX-20250381, KYY-HX-20250423; Medical and Health Science Program of Zhejiang Province, Grant No: WKJ-ZJ-26092.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Shalbaf A, Bagherzadeh S, Maghsoudi A. Transfer learning with deep convolutional neural network for automated detection of schizophrenia from EEG signals[J]. Physical and Engineering Sciences in Medicine, 2020, 43(4): 1229-1239.
2. Ross C. Schizophrenia: Innovations in diagnosis and treatment[M]. Routledge, 2014.
3. Appaji A, Harish V, Korann V, et al. Deep learning model using retinal vascular images for classifying schizophrenia[J]. Schizophrenia Research, 2022, 241: 238-243.
4. Aslan Z, Akin M. Automatic Detection of Schizophrenia by Applying Deep Learning over Spectrogram Images of EEG Signals[J]. Traitement du Signal, 2020, 37(2): 235-244.

5. Baydili İ, Tasci B, Tasci G. Artificial intelligence in psychiatry: A review of biological and behavioral data analyses[J]. Diagnostics, 2025, 15(4): 434
6. Compton M T, Ku B S, Covington M A, et al. Lexical diversity and other linguistic measures in schizophrenia: associations with negative symptoms and neurocognitive performance[J]. The Journal of nervous and mental disease, 2023, 211(8): 613-620.
7. Tanaka S C, Kasai K, Okamoto Y, et al. The status of MRI databases across the world focused on psychiatric and neurological disorders[J]. Psychiatry and Clinical Neurosciences, 2024, 78(10): 563-579.
8. Jafari M, Sadeghi D, Shoeibi A, et al. Empowering precision medicine: AI-driven schizophrenia diagnosis via EEG signals: A comprehensive review from 2002–2023[J]. Applied Intelligence, 2024, 54(1): 35-79.
9. Zang X, Li B, Zhao L, et al. End-to-end depression recognition based on a one-dimensional convolution neural network model using two-lead ECG signal[J]. Journal of Medical and Biological Engineering, 2022, 42(2): 225-233.
10. Telangore H, Sharma N, Sharma M, et al. A novel ECG-based approach for classifying psychiatric disorders: Leveraging wavelet scattering networks[J]. Medical Engineering \& Physics, 2025, 135: 104275.
11. Olah J, Spencer T, Cummins N, et al. Automated analysis of speech as a marker of sub-clinical psychotic experiences[J]. Frontiers in psychiatry, 2024, 14: 1265880.
12. Çabuk T, Sevim N, Mutlu E, et al. Natural language processing for defining linguistic features in schizophrenia: A sample from Turkish speakers[J]. Schizophrenia Research, 2024, 266: 183-189.
13. Arslan B, Kizilay E, Verim B, et al. Automated linguistic analysis in speech samples of Turkish-speaking patients with schizophrenia-spectrum disorders[J]. Schizophrenia Research, 2024, 267: 65-71.
14. Pugh S L, Chandler C, Cohen A S, et al. Assessing dimensions of thought disorder with large language models: The tradeoff of accuracy and consistency[J]. Psychiatry research, 2024, 341: 116119.
15. Gargari O K, Fatehi F, Mohammadi I, et al. Diagnostic accuracy of large language models in psychiatry[J]. Asian Journal of Psychiatry, 2024, 100: 104168.
16. Botelle R, Bhavsar V, Kadra-Scalzo G, et al. Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: an applied evaluative study[J]. BMJ open, 2022, 12(2): e052911.
17. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
18. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.
19. Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification[C]//Proceedings of the AAAI conference on artificial intelligence. 2015, 29(1).
20. Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 562-570.
21. Li Z, Liu F, Yang W, et al. A survey of convolutional neural networks: analysis, applications, and prospects[J]. IEEE transactions on neural networks and learning systems, 2021, 33(12): 6999-7019.

22. Medsker L R, Jain L. Recurrent neural networks[J]. Design and applications, 2001, 5(64-67): 2.
23. Liu T, Liang Y, Yu Z. The influence of text length on text classification model[C]//International Conference on Green, Pervasive, and Cloud Computing. Singapore: Springer Singapore, 2020: 79-90.
24. Hua Y, Blackley S V, Shinn A K, et al. Identifying psychosis episodes in psychiatric admission notes via rule-based methods, machine learning, and pre-trained language models[J]. Translational Psychiatry, 2025.
25. Bae Y J, Shim M, Lee W H. Schizophrenia detection using machine learning approach from social media content[J]. Sensors, 2021, 21(17): 5924.
26. Rodela R R, Rahman M, Efty F T, et al. Analyzing Schizophrenic Texts From Social Media Through Machine Learning and Natural Language Processing[C]//2024 IEEE International Conference on Computing, Applications and Systems (COMPAS). IEEE, 2024: 1-6.
27. Amanat A, Rizwan M, Javed A R, et al. Deep learning for depression detection from textual data[J]. Electronics, 2022, 11(5): 676.
28. Meng X, Wang C, Yang J, et al. Predicting Users' Latent Suicidal Risk in Social Media: An Ensemble Model Based on Social Network Relationships[J]. Computers, Materials \& Continua, 2024, 79(3).
29. Lin P Y, Chen Y H, Chang Y J, et al. Deep learning for schizophrenia classification based on natural language processing—A pilot study[J]. Schizophrenia Research, 2024, 270: 323-324.
30. Wagay F A. Classification of Mental Illnesses from Reddit Posts Using Sentence-BERT Embeddings and Neural Networks[J]. Procedia Computer Science, 2025, 258: 1669-1676.