DESD: A Dual-Encoder Model with Adaptive Training for TCM Syndrome Differentiation

Han Guan, Chongjun Wang*
State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210023, China
502023330017@smail.nju.edu.cn, chjwang@nju.edu.cn

***Abstract*—Accurate syndrome differentiation is essential for diagnosis and treatment in Traditional Chinese Medicine (TCM), yet real-world clinical data often contain complex and heterogeneous cases that challenge automated systems. To address this, we propose DESD, a novel dual-encoder model that separately processes clinical texts and detection results, improving feature extraction while maintaining architectural simplicity. Furthermore, we introduce an adaptive training strategy that identifies hard samples using an evaluation model, verifies their label confidence via a large language model (LLM), and quantifies sample difficulty using information entropy. A nonlinear mapping function then assigns dynamic weights to emphasize high-confidence hard samples during training. Experiments on the TCM-SD dataset show that DESD achieves state-of-the-art performance, with an accuracy of 84.85% and significant gains in Macro-F1 and Macro-Precision, demonstrating its effectiveness in enhancing both classification accuracy and robustness in real-world TCM syndrome differentiation.**

***Index Terms*—TCM Syndrome Differentiation, Dual-encoder Architecture, Sample Weighting, Large Language Model (LLM)**

I. Introduction

Traditional Chinese Medicine (TCM) constitutes a unique medical knowledge system and methodology derived from long-term practical summarization, with syndrome differentiation and treatment as its core principle. Accurate syndrome classification, known as TCM syndrome differentiation, is a critical step in TCM diagnosis and treatment, providing the theoretical foundation for developing individualized therapies and herbal formula compositions. TCM practitioners collect clinical manifestations through the four diagnostic methods—inspection, auscultation and olfaction, inquiry, and palpation—to determine the patient's syndrome type. In interdisciplinary research combining computer science and medicine, this syndrome classification process is often modeled as a text classification task in natural language processing (NLP) [1], [2], aiming to achieve automatic mapping from symptom descriptions to syndrome categories.

In recent years, intelligent TCM diagnosis and treatment research based on machine learning (ML), deep learning (DL), and NLP techniques has garnered significant attention. Machine learning methods are extensively applied in TCM, including syndrome classification modeling [3], analysis of tongue and pulse features [4], [5], and the generation of TCM prescriptions [6]. The application of some deep learning methods has significantly enhanced the capability to extract information from TCM. For instance, Convolutional Neural Networks (CNNs) [7] have been employed for entity relation extraction tasks in TCM [8]. Pre-trained language models based on the Transformer architecture [9], such as BERT [10] or LLaMA [11], have significantly enhanced the processing capability of TCM texts. The domain-specific adaptation of these models through pre-training on TCM corpora enables the effective extraction of features and identification of syndromes from clinical narratives. Recently, numerous studies have explored the application of LLMs in TCM inquiry and syndrome-based prescription, demonstrating superior performance in human-computer interaction and interpretability compared to previous methods [12]. However, their performance in clinical syndrome differentiation remains suboptimal.

The scarcity of high-quality TCM syndrome differentiation datasets constrains research on intelligent TCM syndrome differentiation. The TCM-SD dataset provides standardized training samples and evaluation criteria for NLP-based TCM syndrome differentiation tasks [13]. ZY-BERT, through pre-training and domain adaptation on TCM corpora, has achieved notable results on the TCM-SD dataset, establishing a reproducible baseline for subsequent studies. CCSD is designed to capture the intrinsic relationships between syndromes and symptoms in patient complaints through fine-grained, token-level matching. To address the long-tail distribution in the TCM-SD dataset, the model employs a focal loss. However, this approach also led to a decrease in the model's overall accuracy. The DCKA model advances previous research by integrating a CNN and Bi-LSTM dual-channel architecture with a knowledge attention mechanism, a combination that improves both syndrome differentiation accuracy and inference efficiency [14]. Nevertheless, this model still exhibits limited recognition capability for syndromes with scarce samples or those containing numerous rare terms.

To further enhance the performance of TCM syndrome differentiation models, we propose a dual-encoder based model that significantly simplifies the architecture while maintaining high performance. Addressing the hard samples within the TCM-SD dataset, we leverage the extensive TCM knowledge and powerful reasoning capabilities of LLMs to verify label credibility and employ information entropy [15] to quantify sample difficulty and importance. The main contributions of

*Corresponding Author.

this paper are summarized as follows:

(1) In the Traditional Chinese Medicine syndrome differentiation task based on the TCM-SD dataset, we propose for the first time an adaptively trained Dual-Encoder model for TCM Syndrome Differentiation (DESD). This model processes the heterogeneous components within patient information through two independent Chinese RoBERTa encoders and aligns them via late fusion [16]–[18], effectively enhancing the model's performance in TCM syndrome identification.

(2) In the context of TCM syndrome differentiation, we leverage the generalization capabilities of large language models to verify the label confidence of hard samples within the training set. By increasing the weight of high-confidence cases and suppressing that of low-confidence ones, the model's performance in identifying syndrome types for challenging cases is effectively enhanced.

(3) We introduce information entropy as a metric for sample difficulty and employ a non-linear mapping to dynamically adjust the weight of hard samples during training. This focuses the model more on learning samples with high label confidence and high difficulty, further boosting the overall accuracy of TCM syndrome differentiation.

II. THE PROPOSED METHOD

We propose the Dual-encoder based Syndrome Differentiation (DESD) model for TCM, which leverages both patient clinical information and detection results from the TCM-SD dataset. The paper is structured as follows: Section A details the DESD architecture and its advancements over prior models; Section B explains the identification of hard samples, the use of LLMs for label verification, and weight adjustment based on information entropy; Section C introduces strategies for mapping entropy to sample weights; and Section D describes the training set reconstruction and procedure.

A. Dual-Encoder Model for TCM Syndrome Differentiation

The DESD model architecture, as illustrated in Fig.1, adopts a dual-encoder design comprising two Chinese pre-trained RoBERTa modules that process textual information from distinct sources: Encoder1 is specialized for clinical information, while Encoder2 handles detection results. Based on insights from ZY-BERT studies, the core information for TCM syndrome differentiation predominantly resides in patients' chief complaints and medical history descriptions. We therefore concatenate these two text components and feed them into Encoder1, which leverages self-attention mechanisms to enable deep semantic interaction and extract crucial pathological features.

Although detection results of patient signs provide valuable references for syndrome differentiation, their effective information density remains relatively low. To mitigate noise interference and enhance focus on discriminative features, we employ a separate Encoder2 for processing this information. The encoding process for both types of inputs is formally represented as:

$$\mathbf{h}_c = \text{Encoder1}(\mathbf{x}_c) \in \mathbb{R}^{d_e} \tag{1}$$

$$\mathbf{h}_d = \text{Encoder2}(\mathbf{x}_d) \in \mathbb{R}^{d_e} \tag{2}$$

where $\mathbf{x}_c$ denotes the clinical text comprising chief complaints and medical history, $\mathbf{x}_d$ represents the detection results, while $\mathbf{h}_c \in \mathbb{R}^{d_e}$ and $\mathbf{h}_d \in \mathbb{R}^{d_e}$ are their corresponding feature representations with hidden dimension $d_e = 768$.

Unlike the single-encoder CCSD and dual-channel DCKA, DESD employs a dual-encoder framework to process heterogeneous patient data separately. For instance, patient complaints and medical history are more information-dense and semantically rich than detection, making them challenging for a single encoder to model effectively. Although DCKA incorporates TCM knowledge, it utilizes ZY-BERT merely as a static encoder without adapting to data heterogeneity. In contrast, DESD adapts to such heterogeneity through two trainable encoders, fuses their extracted features, and maps them into a unified vector space for classification, as formalized below:

$$\hat{y} = \text{Softmax}(\text{MLP}(\mathbf{h}_c \oplus \mathbf{h}_d)) \tag{3}$$

where $\mathbf{h}_c \oplus \mathbf{h}_d$ denotes the concatenated feature vector with dimension $\mathbb{R}^{2d_e}$, and $\hat{y}$ represents the predicted probability distribution over syndrome categories.

For the loss function, DESD employs Cross-Entropy Loss (CELoss) to prioritize overall accuracy in the syndrome classification task. The objective function is defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij}) \tag{4}$$

where N is the batch size, C denotes the number of syndrome categories, yij indicates the ground-truth label, and $\hat{y}_{ij}$ represents the predicted probability for the j-th category of the i-th sample.

B. Adaptive Training Strategy

The TCM-SD dataset is collected from real clinical environments, where patient symptoms exhibit high heterogeneity and disease progression varies complexly, significantly differing from theoretical scenarios of TCM syndrome differentiation. To enhance the model's ability to recognize syndromes in hard samples, we increase the weight of such samples during training, enabling the model to focus more on learning their effective features. However, directly increasing the weight of hard samples may lead to a decline in the model's overall performance, as the label confidence of hard samples inherently requires more cautious evaluation. To address this issue, we introduce LLMs to validate the label confidence of hard samples and adjust sample weights quantitatively based on information entropy theory. To identify hard samples, we first train a dual-encoder model (with only 220M parameters) on the original training set to obtain an evaluation model capable of TCM syndrome differentiation. This model is then used to

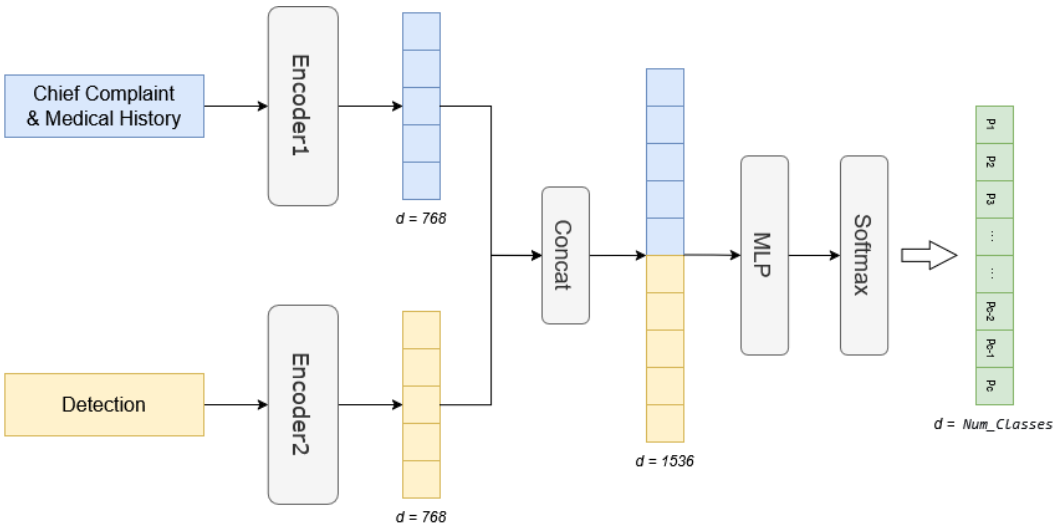

Fig. 1. Dual-Encoder Model for TCM Syndrome Differentiation

Algorithm 1 TCM Syndrome Differentiation with Hard Sample Reweighting

Require: Training set $D = \{(x_i, y_i)\}_{i=1}^{N}$, LLM, dual-encoder model M

Ensure: Trained model M_{final}

Phase 1: Hard Sample Identification
Train M on D for T epochs
$D_{hard} \leftarrow \{x_i \mid \hat{y}_i \neq y_i, \text{ where } \hat{y}_i = M(x_i)\}$
Phase 2: LLM Verification
$D_{hc} \leftarrow \{x_i \in D_{hard} \mid \text{LLM}(x_i) = y_i\}$
$D_{lc} \leftarrow \{x_i \in D_{hard} \mid \text{LLM}(x_i) \neq y_i\}$
Phase 3: Entropy-based Weighting
for each $x_i \in D$ **do**
  if $x_i \in D_{hc} \cup D_{lc}$ **then**
    $P = M(x_i)$ ▷ Get probability distribution
    $H_i = -\sum_{j=1}^{C} P_j \log P_j$
    $w_i = \begin{cases} f_{hc}(H_i) & \text{if } x_i \in D_{hc} \\ f_{lc}(H_i) & \text{if } x_i \in D_{lc} \end{cases}$
  else
    $w_i = 1.0$
  end if
end for
Phase 4: Final Training
Train M_{final} using weighted loss:
$L = \frac{1}{|B|} \sum_{(x_i, y_i) \in B} w_i \cdot \text{CE}(M_{\text{final}}(x_i), y_i)$
return M_{final}

repredict the syndromes of all training samples. A sample is identified as a hard sample if its prediction from the evaluation model disagrees with the original label:

$$D_{hard} = \{x_i \mid \hat{y}_i \neq y_i\} \tag{5}$$

Here, D_{hard} denotes the set of hard samples, $\hat{y}_i$ represents the syndrome category predicted by the evaluation model, and y_i is the original label.

To effectively distinguish between high-confidence and low-confidence samples, for each hard sample, we guide the LLM through predefined prompts to construct a chain of reasoning for TCM syndrome differentiation. The LLM is required to select the category that best matches the patient's clinical symptoms from the Top-K (defaults to 3) candidate syndromes predicted by the evaluation model as the final decision. If the original label of a hard sample is inconsistent with both the predictions of the evaluation model and the LLM, it is considered to have low label confidence. Based on this, hard samples can be further divided into high-confidence and low-confidence sample sets, formally defined as:

$$D_{hc} = \{x_i \in D_{hard} \mid \text{LLM}(x_i) = y_i\} \tag{6}$$

$$D_{lc} = \{x_i \in D_{hard} \mid \text{LLM}(x_i) \neq y_i\} \tag{7}$$

Here, D_{hc} denotes the set of high-confidence samples after LLM verification, and D_{lc} denotes the set of low-confidence samples.

Information entropy serves as a quantitative measure of system uncertainty: the higher the uncertainty, the greater the information entropy. We input hard samples into the evaluation model to obtain the predicted probability distribution over different syndromes, which serves as the basis for calculating information entropy. The specific calculation formula is as follows:

$$H(x_i) = -\sum_{j=1}^{C} P(y_j|x_i) \log P(y_j|x_i) \tag{8}$$

Here, C is the total number of syndrome categories, and $P(y_j|x_i)$ represents the probability that the model predicts sample x_i as belonging to the j-th syndrome.

A higher information entropy for a hard sample indicates greater uncertainty in the model's syndrome differentiation,

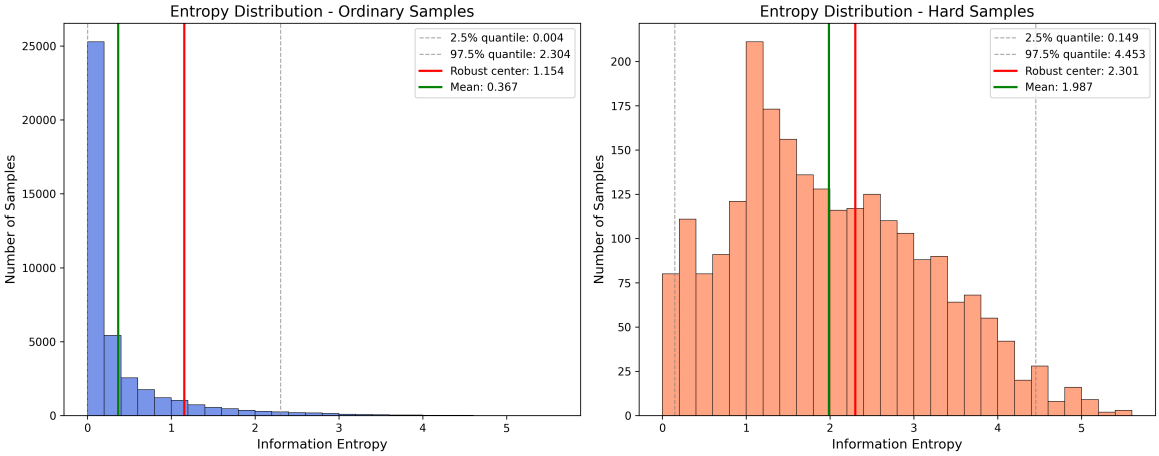

Fig. 2. Distribution of Information Entropy

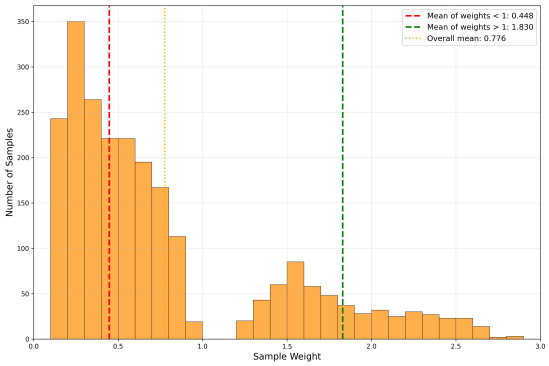

Fig. 3. Distribution of Hard Sample Weights

corresponding to a higher classification difficulty. Therefore, such samples should be assigned higher weights during training. We design an information entropy-to-weight mapping function to dynamically adjust sample weights: for high-confidence samples, their information entropy is mapped to an increased weight; for low-confidence samples, their information entropy is mapped to a reduced weight.The pseudocode for the sample reweighting strategy is presented in Algorithm 1.

C. Entropy-to-Weight Mapping

For the weight mapping of hard samples, we design both linear and nonlinear functions. The linear mapping converts information entropy directly into sample weights via a scaling factor, ensuring simplicity and computational efficiency:

$$w_i = a + k \cdot (H(x_i) - H_{\min}) \tag{9}$$

where w_i is the sample weight and k is the scaling factor.

The nonlinear mapping employs a Sigmoid-based function to project normalized entropy smoothly into a specified weight range. This constrains weight magnitudes and improves robustness against entropy fluctuations:

$$w_i = a + \frac{b-a}{1+\exp\left(-\gamma \cdot (H(x_i) - \mu)\right)} \tag{10}$$

where γ and μ are tunable parameters. For high-confidence samples, we set $a = 1.0$ and b as the maximum weight to amplify their contribution; for low-confidence samples, $a = 0$ and $b = 1.0$ to suppress their influence.

D. Training Set Reconstruction

Based on the predictions of the evaluation model, the training set is repartitioned into hard samples and ordinary samples, with specific data summarized in Table I. Figure 2 illustrates a distinct contrast in the information entropy distributions. Ordinary samples exhibit entropy concentrated in the lower range, indicating low syndrome category uncertainty and relatively straightforward classification. In contrast, hard samples display broader dispersion and a higher mean entropy, reflecting greater predictive uncertainty and correspondingly higher classification difficulty.

TABLE I
DISTRIBUTION OF THE RE-PARTITIONED TRAINING SET

Category	Number of Samples
Ordinary Samples	40,829
Hard Samples	2,351
– *High-Confidence Subset*	558
– *Low-Confidence Subset*	1,793
Total	**43,180**

Subsequently, a large language model with strong reasoning capabilities, such as Qwen3 [19], is employed to to validate the label confidence of hard samples. These hard samples constitute only 5.44% of the original training data, amounting to 2,351 instances, resulting in a computational cost significantly lower than verifying the entire dataset. The samples are divided into two subsets: high-confidence and low-confidence samples. Using a nonlinear entropy-to-weight

TABLE II
PERFORMANCE COMPARISON FOR THE CLASSIFICATION TASK.

Method	Development Set				Test Set			
	Acc	Macro-F1	Macro-R	Macro-P	Acc	Macro-F1	Macro-R	Macro-P
Statistical Methods								
DT	59.42	20.68	21.33	21.52	59.10	21.67	22.38	22.20
SVM	77.63	32.13	29.56	43.10	78.53	36.37	32.98	49.35
Deep Learning Methods								
BiLSTM	69.30	17.53	15.08	14.76	69.65	15.15	15.65	17.08
BiGRU	73.57	19.53	20.12	21.81	74.43	20.93	21.90	23.76
CNN	77.56	31.79	30.39	37.99	78.58	32.83	31.29	39.19
BERT	79.44	34.18	34.12	38.00	80.17	35.45	34.99	42.00
distilBERT	79.09	36.07	36.62	38.13	80.46	40.24	39.99	45.84
ALBERT	79.62	37.88	37.65	41.94	80.51	40.50	39.57	46.54
LLMs Finetuned Methods								
TCM-BERT	79.48	37.84	37.60	42.00	80.55	41.58	40.91	48.47
ZY-BERT	81.43	49.47	48.89	54.08	82.19	51.01	49.42	57.70
CCSD	81.35	<u>53.69</u>	<u>52.56</u>	58.12	81.76	53.47	52.13	59.33
DCKA	<u>83.12</u>	53.53	**53.34**	57.41	<u>84.01</u>	54.25	<u>53.30</u>	58.86
DESD(w/o adapt)	82.68	52.99	51.74	<u>58.21</u>	83.88	<u>54.65</u>	52.87	<u>60.80</u>
DESD(full)	**83.30**	**54.48**	52.43	**60.65**	**84.85**	**56.70**	**54.59**	**63.75**

*Note: The best and the runner-up results are highlighted in bold and underline respectively.

mapping function, a corresponding weight is computed for each hard sample. The overall distribution of the assigned weights is shown in Fig.3. Specifically, weights for low-confidence samples are mapped to the interval (0, 1), while high-confidence samples are mapped to (1, 3).

III. EXPERIMENTS

A. Dataset

TCM-SD is the first public dataset for TCM syndrome differentiation, comprising anonymized real-world clinical cases [13]. Each sample includes chief complaints, medical history, detection results, and expert-annotated syndrome labels. The dataset is split into training, development, and test sets in an 8:1:1 ratio, which we follow in our experiments. During preprocessing, we tokenize and concatenate the chief complaints and medical history into a single segment, truncating or padding it to 512 tokens as input to Encoder1. Detection results are separately processed to the same length for Encoder2. We evaluate performance using Accuracy, Macro F1, Macro Recall, and Macro Precision.

B. Experiment Details

DESD used Chinese RoBERTa as the backbone encoder, with a hidden size of 768 per encoder. The evaluation model M was trained for 6 epochs. We adopted the nonlinear entropy-to-weight mapping, which performed best, with $\mu = 2.3$ (the 95% interval midpoint) and γ set as the reciprocal of the hard samples' entropy standard deviation. The weight bounds were $a = 1.0, b = 3.0$ for high-confidence samples and $a = 0, b = 1.0$ for low-confidence ones. DESD was optimized with AdamW [20] (initial learning rate 5×10^{-5}) and a linear decay schedule. All experiments were run on an NVIDIA 4090 GPU (24 GB) with a batch size of 8 and gradient accumulation steps set to 8.

C. Main Results

As shown in Table II, DESD outperforms state-of-the-art methods CCSD and DCKA on the TCM syndrome classification task across most metrics, with only Macro-R on the development set being an exception. On the test set, DESD achieves an accuracy of 84.85%, a Macro-F1 of 56.70%, and a Macro-P of 63.75%, representing the best overall performance. Specifically, DESD improves accuracy, Macro-F1, and Macro-P by 0.84%, 2.45%, and 4.89% over DCKA, and surpasses CCSD by 3.23% in Macro-F1 with comprehensive advantages. The results demonstrate that DESD, through hard sample selection and entropy-based weighting, effectively enhances classification accuracy and overall metric balance while maintaining high recall.

D. Ablation Study

To assess the effectiveness of the dual-encoder architecture, we performed ablation studies comparing single-encoder and dual-encoder models under the same training conditions and without sample weighting. As presented in Table III, the single-encoder model exhibits significantly lower performance, underscoring the benefits of the dual-encoder design. By processing heterogeneous data components through separate encoders, the dual-encoder architecture facilitates more thorough and effective feature extraction. Consequently, it enhances performance across all evaluation metrics in the task of TCM syndrome differentiation.

To evaluate the impact of different weighting strategies, we conducted an ablation study with the following three config-

TABLE III
PERFORMANCE COMPARISON OF DESD WITH SINGLE-ENCODER AND DUAL-ENCODER ARCHITECTURES.

Model	Development Set				Test Set			
	Acc	Macro-F1	Macro-R	Macro-P	Acc	Macro-F1	Macro-R	Macro-P
Single-Encoder	81.53	49.61	49.26	54.17	82.35	51.21	49.53	57.85
Dual-Encoder	**82.68**	**52.99**	**51.74**	**58.21**	**83.88**	**54.65**	**52.87**	**60.80**

TABLE IV
PERFORMANCE COMPARISON OF DESD WITH DIFFERENT WEIGHTING STRATEGIES.

Model	Weighting Strategy	Development Set				Test Set			
		Acc	Macro-F1	Macro-R	Macro-P	Acc	Macro-F1	Macro-R	Macro-P
DESD	Fixed	82.68	52.99	51.74	58.21	83.88	54.65	52.87	60.80
	Linear ($k = 0.25$)	**83.41**	52.18	50.54	57.91	**84.92**	54.03	51.57	61.71
	Nonlinear	83.30	**54.48**	**52.43**	**60.65**	84.85	**56.70**	**54.59**	**63.75**

urations for the DESD model: (1) fixed sample weights; (2) linear entropy-to-weight mapping; and (3) nonlinear entropy-to-weight mapping.

The ablation results in Table IV validate the effectiveness of the dual-encoder architecture complemented by the nonlinear weighting strategy. Experimental findings demonstrate that integrating a nonlinear entropy-to-weight mapping function significantly enhances model performance. Specifically, DESD with nonlinear mapping achieves the best overall performance on the test set, obtaining a Macro-F1 of 56.70% and Macro-P of 63.75%, with consistent improvements over both fixed-weight and linear mapping configurations. Although the linear mapping yields optimal accuracy, it underperforms on other comprehensive metrics, falling significantly below both fixed-weight and nonlinear mapping approaches. These results indicate that the dual-encoder architecture provides an effective foundation for modeling heterogeneous inputs, while the nonlinear weighting mechanism further enhances the model's capability to perceive and process hard samples. The synergy of these components enables DESD to outperform existing state-of-the-art methods across multiple evaluation metrics, confirming the necessity and effectiveness of the proposed architectural components for TCM syndrome classification.

IV. CONCLUSION

In this paper, we propose a dual-encoder model (DESD) with an adaptive sample weighting strategy to address the challenges of TCM syndrome differentiation, particularly in handling hard samples. By separately encoding heterogeneous patient information, the model effectively improves the accuracy of TCM syndrome differentiation. We employ an LLM to verify the label confidence of hard samples and quantify sample difficulty using information entropy. A nonlinear mapping function is then applied to adaptively assign weights, focusing the training on critical samples. Experiments demonstrate that the proposed method leads to significant improvements across various metrics on clinical TCM syndrome differentiation tasks.

ACKNOWLEDGMENT

This paper is supported by the National Natural Science Foundation of China (Grant No. 62192783, 62376117), the National Social Science Fund of China (Grant No. 23BJL035), the Science and Technology Major Project of Nanjing (comprehensive category) (Grant No. 202309007), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118), and the Collaborative Innovation Center of Novel Software Technology and Industrialization at Nanjing University.

REFERENCES

[1] L. Yao, Z. Jin, C. Mao, Y. Zhang, and Y. Luo, "Traditional chinese medicine clinical records classification with BERT and domain specific corpora," *Journal of the American Medical Informatics Association*, vol. 26, no. 12, pp. 1632–1636, Sep. 2019.

[2] Z. Song, Y. Xie, W. Huang, and H. Wang, "Classification of traditional chinese medicine cases based on character-level BERT and deep learning," in *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, 2019, pp. 1383–1387.

[3] Y. L. Shi, J. Y. Liu, X. J. Hu, L. P. Tu, J. Cui, J. Li, Z. J. Bi, J. C. Li, L. Xu, and J. T. Xu, "A new method for syndrome classification of non-small-cell lung cancer based on data of tongue and pulse with machine learning," *BioMed Research International*, vol. 2021, p. 1337558, 2021.

[4] R. Kanawong, T. Obafemi-Ajayi, J. Yu, D. Xu, S. Li, and Y. Duan, "ZHENG classification in traditional chinese medicine based on modified specular-free tongue images," in *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops*, 2012, pp. 288–294.

[5] S. C. Hui, Y. He, and D. T. C. Thach, "Machine learning for tongue diagnosis," in *2007 6th International Conference on Information, Communications & Signal Processing*, 2007, pp. 1–5.

[6] Z. Liu, C. Luo, D. Fu, J. Gui, Z. Zheng, L. Qi, and H. Guo, "A novel transfer learning model for traditional herbal medicine prescription generation from unstructured resources and knowledge," *Artificial Intelligence in Medicine*, vol. 124, p. 102232, 2022.

[7] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[8] T. Bai, H. Guan, S. Wang, Y. Wang, and L. Huang, "Traditional chinese medicine entity relation extraction based on CNN with segment attention," *Neural Computing and Applications*, vol. 34, no. 4, pp. 2739–2748, Feb. 2022.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg,

S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[11] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," 2023.

[12] S. Yang, H. Zhao, S. Zhu, G. Zhou, H. Xu, Y. Jia, and H. Zan, "Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 368–19 376.

[13] M. Ren, H. Huang, Y. Zhou, Q. Cao, Y. Bu, and Y. Gao, "TCM-SD: A benchmark for probing syndrome differentiation via natural language processing," in *Proceedings of the 21st Chinese National Conference on Computational Linguistics*. Nanchang, China: Chinese Information Processing Society of China, Oct. 2022, pp. 908–920.

[14] B. Liu, H. Huang, X. Liu, J. Zhao, and J. Mo, "Dual-channel knowledge attention for traditional chinese medicine syndrome differentiation," *Scientific Reports*, vol. 15, Apr. 2025.

[15] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[16] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for Chinese natural language processing," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 657–668.

[17] 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.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019.

[19] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu, "Qwen3 technical report," 2025.

[20] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.