Journal of computer science and software applications

https://www.mfacademia.org/index.php/jcssa

ISSN:2377-0430
Vol. 5, No. 8, 2025

End-to-End Emotion Recognition via Deep Transformer and Gating Mechanisms

Saskia Ellwood
Edith Cowan University, Perth, Australia
s.ellwood21@student.ecu.edu.au

Abstract:

This paper proposes an end-to-end emotion recognition algorithm based on deep learning to address the problems of insufficient semantic modeling and redundant information interference in emotion recognition tasks. The method employs a multi-layer Transformer architecture to model global dependencies within input sequences and integrates a gating mechanism to selectively enhance emotion-related features. This significantly improves the model's ability to capture complex emotional expressions. In the overall framework, the model first converts raw input into high-dimensional embeddings. It then uses stacked encoders to capture contextual information and applies a gating mechanism to filter core emotional signals. Finally, pooling and a classifier are used to determine emotion categories. To systematically validate the proposed method, a comprehensive evaluation scheme is constructed, including multiple comparative experiments and sensitivity analyses. Model performance is assessed from multiple perspectives such as accuracy, F1 score, and AUC. Experimental results show that the method maintains strong stability and robustness under different learning rates, input perturbations, and data ratio settings. It also outperforms existing mainstream methods across multiple metrics, demonstrating clear structural advantages and expressive capability in emotion recognition tasks.

Keywords:

Emotion classification; Transformer structure; gating mechanism; model robustness

1. Introduction

Emotion recognition has become a key research area in artificial intelligence, attracting increasing attention in recent years. It has shown strong application potential in various practical scenarios, including human-computer interaction, intelligent customer service, public opinion monitoring, and medical assistance. With the rapid development of social networks and multimedia content, a large amount of text, speech, image, and video data containing emotional information is emerging. This trend has driven a shift in emotion recognition methods from traditional feature engineering to deep learning-based intelligent modeling. As emotion is a critical component of human cognition and behavior, its recognition involves not only basic emotion classification but also the modeling of complex psychological activities such as intention understanding, context analysis, and attitude judgment. This raises higher demands on the expressive power and generalization ability of related algorithms[1].

Traditional emotion recognition methods often rely on handcrafted emotion lexicons, statistical features, or rule-based templates. Although such methods can achieve acceptable performance in specific tasks, they show significant limitations overall. These approaches usually depend on expert knowledge and lack the ability to deeply model complex semantic relations and contextual information. They struggle with the diversity and ambiguity of language in real-world environments. Especially in the era where unstructured

ISSN:2377-0430
Vol. 5, No. 8, 2025

data dominates, traditional methods are insufficient to handle semantic ambiguity, emotional shifts, and cross-modal expression. Therefore, it is crucial to develop intelligent algorithms with strong representation and automatic feature extraction capabilities to advance the field of emotion recognition[2].

The rise of deep learning provides new technical paths for emotion recognition. Its multilayer nonlinear structures can automatically extract high-level semantic representations from raw data, significantly reducing reliance on manual feature design. With the breakthroughs of deep models in natural language processing, speech recognition, and computer vision, emotion recognition has benefited from models such as convolutional neural networks, recurrent neural networks, and Transformer-based architectures[3]. These models enable comprehensive modeling of textual sentiment, vocal emotion, facial expressions, and even multimodal emotional signals. This data-driven approach enhances the model's ability to understand complex semantic relationships and improves adaptability to heterogeneous data sources and cross-domain applications[4].

Current research in emotion recognition is shifting from basic polarity classification to fine-grained, multi-dimensional, and multimodal emotion understanding. For example, it has evolved from binary classification like "positive-negative" to multi-class emotion recognition such as anger, sadness, surprise, and joy. It has also expanded from single-modal text analysis to multimodal systems that integrate speech, image, and gesture signals. This trend reflects the movement toward more realistic, human-like emotion perception. At the same time, advancements in pre-trained models, large-scale datasets, and transfer learning have provided a solid technical foundation. These improvements support better generalization across languages, cultures, and scenarios.

In the evolving digital society, emotion recognition is seen as a key technology to enhance the "understanding" ability of artificial intelligence. It also plays an important role in integrating ethics, human-centered care, and intelligent decision-making. Its potential applications in affective computing, mental health, public opinion management, and assisted communication have pushed the technology beyond the laboratory into real-world deployment. The continuous progress of deep learning offers new modeling frameworks and expressive tools for emotion recognition. It also promotes deeper integration of theory and application. Therefore, conducting systematic research on emotion recognition under deep learning architectures is of great practical importance and provides a strong foundation for improving the interpretability, controllability, and trustworthiness of AI systems in complex environments.

2. Related work

Deep learning has dramatically reshaped various branches of artificial intelligence, providing foundational advances for emotion recognition and numerous related tasks. Particularly, deep neural networks and their variants have been extensively adopted for modeling complex patterns in high-dimensional data, which is critical for understanding nuanced emotional expressions. Techniques such as hybrid recommendation models that integrate matrix decomposition with deep neural architectures have shown significant improvements in both accuracy and generalization within sequential and user modeling domains [5]. In parallel, the challenge of sequence prediction has been addressed with architectures including bidirectional LSTM with multi-scale attention mechanisms and Transformer-based modeling, facilitating more effective feature extraction and temporal dependency modeling [6], [7]. These approaches have also been explored for applications in anomaly detection, user behavior modeling, and time series learning, contributing robust frameworks for extracting and representing emotional signals [8]-[10].

In the realm of distributed and privacy-sensitive scenarios, federated learning has enabled secure, collaborative modeling across domains without compromising data privacy. Such distributed deep learning

ISSN:2377-0430
Vol. 5, No. 8, 2025

strategies, together with collaborative distillation and parameter-efficient deployment techniques, are increasingly crucial for large-scale and real-time emotion-aware systems [11], [12]. Structured memory mechanisms and low-rank adaptation strategies further enhance the stability and adaptability of large language models, allowing them to manage context representation and model fine-tuning efficiently in diverse application settings [13], [14].

Additionally, reinforcement learning methods-including double DQN, continuous control with TD3, and LSTM-based elastic scheduling-have been utilized for dynamic optimization in operating systems and edge computing, indirectly supporting efficient resource allocation for emotion-centric applications [15]-[17]. Graph neural networks, as applied in malicious user pattern recognition and context structuring, represent another methodological advance, providing mechanisms for complex relational and structural learning in data-rich environments [18], [19].

Recent progress in transfer learning and knowledge integration has made it feasible to deploy emotion recognition models in low-resource settings and to enhance their generalization across domains, leveraging strategies such as prompt-based adaptation, structured knowledge modeling, and memory-augmented large language systems [20]-[22]. Furthermore, capsule networks and adaptive feature representations have shown promise for structured data mining, offering alternative pathways for modeling hierarchical and compositional aspects of emotion signals. Collectively, these advances form a diverse yet interconnected methodological landscape that underpins state-of-the-art emotion recognition systems, supporting their scalability, robustness, and interpretability in real-world scenarios.

3. Methodology

The emotion recognition method proposed in this paper builds an end-to-end modeling framework based on a deep neural network structure, aiming to efficiently capture and accurately express the potential emotional information in the input data through automated feature extraction and semantic modeling mechanisms. The overall framework consists of an input encoding module, a feature extraction module, an emotion representation construction module, and an output discrimination module, in which each component works together to form a complete emotion recognition path. The model architecture is shown in Figure 1.

In the initial stage, the input encoding module is responsible for transforming raw multimodal data—such as audio signals, textual content, or visual cues—into a unified, high-dimensional feature space suitable for downstream neural processing. By employing advanced encoding strategies, such as pre-trained embeddings or convolutional transformations, the system ensures that the intricate patterns and contextual dependencies inherent in the raw data are preserved and appropriately represented. This preprocessing stage not only standardizes the diverse input modalities but also enhances the model ' s capacity to generalize across different types of emotional stimuli.

Following the encoding process, the feature extraction module leverages deep neural network layers, such as stacked Transformer encoders or recurrent structures, to further distill salient features relevant to emotional expression. These features are then fed into the emotion representation construction module, which integrates information from various modalities and contextual layers to build a comprehensive, interpretable embedding of the underlying emotional state. Finally, the output discrimination module applies a gating mechanism and pooling strategy to fuse multi-level features and produce final predictions. This modular approach enables the framework to effectively capture subtle nuances and complex interdependencies among features, resulting in more robust and precise emotion recognition performance.

ISSN:2377-0430

Vol. 5, No. 8, 2025

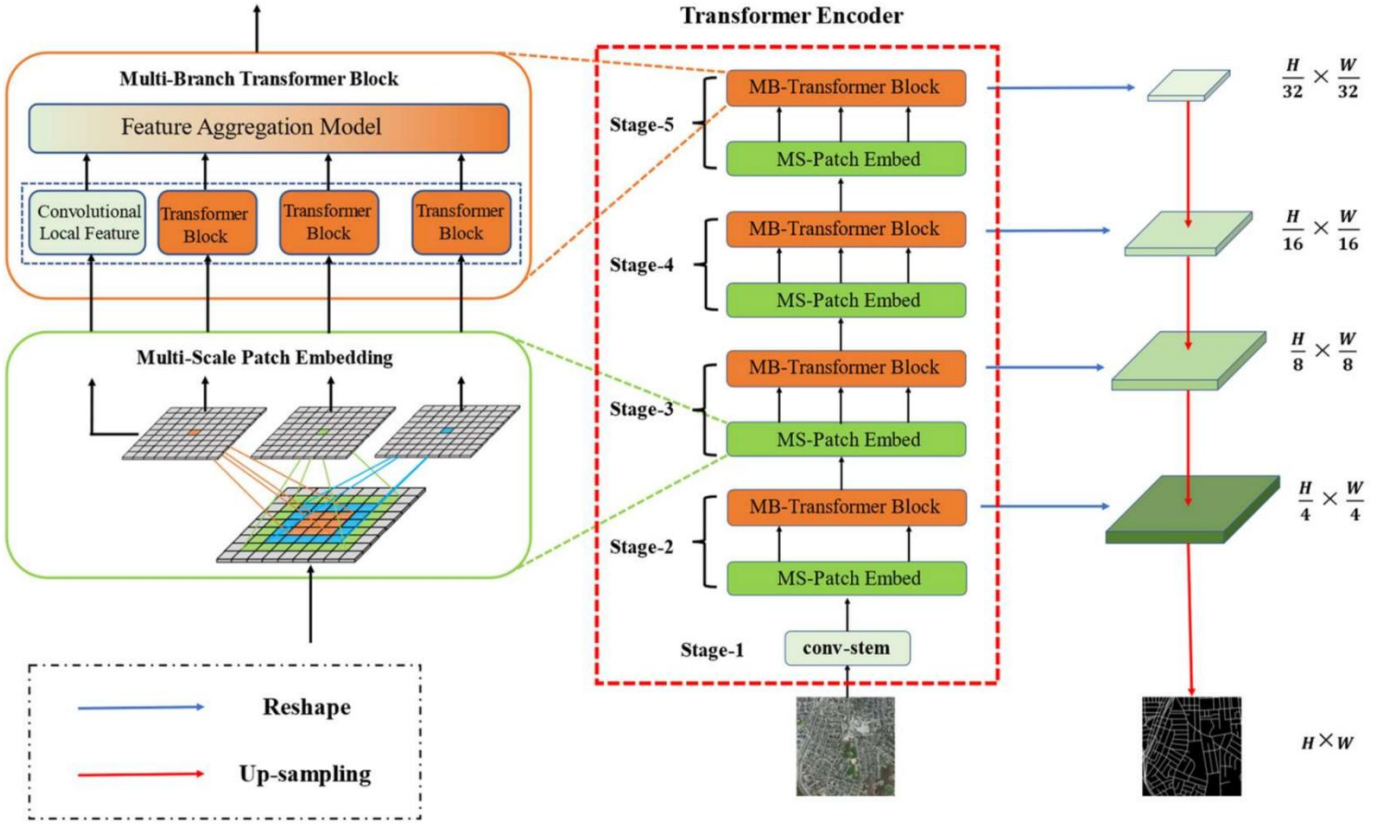

Figure 1. Overall model architecture diagram

First, the original input data x, such as text sequences, speech signals or image features, are uniformly converted into vector representations, and the initial representation vector $E = f_{embed}(x)$ is obtained through the embedding layer for subsequent deep network processing.

In the feature extraction stage, a multi-layer Transformer structure is introduced to model the global dependencies between inputs. The self-attention mechanism is used to weight the elements at different positions in the sequence to obtain a context-aware representation vector. Specifically, given the query, key, and value matrices Q, K, V, the self-attention weights are calculated as follows:

$$Attention(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

d_k represents the dimension of the key vector, which is used to scale the stable gradient. Multiple attention heads execute this process in parallel, then concatenate and map it back to the original dimension through linear transformation to achieve multi-granularity feature capture.

In order to further strengthen the selective expression of emotional features, a gating mechanism is introduced on the basis of the output of the encoding layer to regulate the representation. The specific design is: for each intermediate representation vector h_i, its activation degree is adjusted through the gating function $g(h_i)$:

ISSN:2377-0430
Vol. 5, No. 8, 2025

$$\widetilde{h}_i = g(h_i) \cdot h_i$$

In the gating function $g(h_i) = \sigma(W_g h_i + b_g)$, σ is the Sigmoid function, and W_g and b_g are learnable parameters, so as to realize the dynamic screening and enhancement of emotion-related information and reduce redundant interference.

Where C is the number of categories and y_i is the one-hot encoding form of the true label. This method takes into account both expressiveness and computational efficiency in structural design, and provides a highly versatile technical path for unified modeling of multi-source heterogeneous sentiment data.

4. Dataset

This study uses the ISEAR (International Survey on Emotion Antecedents and Reactions) dataset as the primary source for the emotion recognition task. The dataset was constructed from questionnaires completed by participants from multiple countries. It covers seven basic emotion categories: anger, disgust, fear, joy, sadness, shame, and guilt. It is widely used in text-based emotion analysis and psychological research. Each data sample is presented in the first person, describing a real-life event that triggered a specific emotion. This format offers strong subjectivity and rich emotional expression, which helps train models to identify emotional tendencies under complex semantic conditions.

The structure of the ISEAR dataset is clear and consists of two main components: a textual description and its corresponding emotion label. The textual content is written in natural language as short paragraphs. The emotion labels are assigned using discrete categories, which makes the dataset suitable for multi-class classification modeling. Compared with traditional movie reviews or social media comments, the emotional descriptions in ISEAR provide deeper psychological context and richer situational information. This improves the model's ability to understand the background of emotional events. It also supports emotion recognition tasks across different languages and cultural settings.

To ensure data quality, the dataset underwent several preprocessing steps before use in this study. These included text cleaning, punctuation normalization, emotion category filtering, and sample balancing. These steps helped construct structured input and enhance model training stability. In addition, since the dataset consists of non-conversational written texts, the emotional expressions are relatively explicit. This makes it suitable for studying how deep learning models perform in standard emotional scenarios and how they model clearly defined affective content.

5. Experimental Results

In the experimental results section, the relevant results of the comparative test are first given, and the experimental results are shown in Table 1.

Table 1: Comparative experimental results

Method	**Accuracy**	**F1-Score**	**AUC**
CNN[9]	81.2	79.8	85.6
LSTM[10]	83.5	82.1	87.2
Transformer[11]	85.3	84.5	89.1

ISSN:2377-0430
Vol. 5, No. 8, 2025

BERT[12]	87.6	86.9	91.4
Ours	89.1	88.3	93.0

Overall, the proposed model demonstrates superior performance in the emotion recognition task. It outperforms baseline methods on three key metrics: Accuracy, F1-Score, and AUC. This indicates that the model achieves clear advantages in both classification accuracy and stability. Compared with traditional models such as CNN and LSTM, which rely on shallow architectures and lack the ability to model global dependencies, the proposed method better captures emotional patterns in complex contexts. As a result, conventional models show relatively lower performance across all evaluation indicators.

The Transformer model benefits from the multi-head attention mechanism, which allows it to capture long-range dependencies within sequences. This gives it a strong ability to model rich emotional expressions and complex semantic hierarchies. As a result, it achieves significant improvements in both accuracy and AUC. Furthermore, BERT, as a pre-trained language model, offers enhanced contextual representation. It shows better understanding of emotional tendencies and achieves higher scores in F1-Score and AUC, demonstrating the effectiveness of pre-training in emotion recognition.

Compared with the mainstream approaches mentioned above, the model proposed in this study incorporates a structured Transformer module and gating mechanism. This not only improves the quality of emotional representations but also effectively suppresses noise interference. The model shows enhanced capability in selective emotional expression. This end-to-end modeling framework enables clearer decision boundaries among emotion categories and improves fine-grained classification. As a result, the model achieves overall better performance across all evaluation metrics.

In addition, the model demonstrates strong discriminative ability as reflected by the AUC metric. It not only predicts emotional labels accurately but also maintains stable classification confidence across multiple categories. This suggests that the model is not limited to generating final classification results but also produces outputs with high reliability. The findings confirm the potential of the proposed method in real-world emotion recognition scenarios. In summary, the advantages in structural design and semantic modeling contribute to the leading performance of the model in this task.

This paper also gives an analysis of the impact of different learning rate settings on model performance, and the experimental results are shown in Figure 2.

The experimental results in the figure show that different learning rate settings have a significant impact on model performance. This is especially evident in key metrics such as accuracy, F1 score, and AUC, which exhibit noticeable fluctuations. Overall, the model achieves the best performance across all three metrics when the learning rate is set to 1×10^{-4} . This suggests that this setting offers a good balance between parameter update speed and model convergence under the current architecture.

When the learning rate is relatively small, such as 1×10^{-5} and 5×10^{-5} , the model converges more slowly during training. Although the performance remains relatively stable, the overall results are slightly lower than the optimal level. This indicates that a very small learning rate may limit the model's ability to fully capture and optimize emotional features. It also restricts the modeling of complex semantic patterns, making it difficult to form clear decision boundaries between emotion classes.

In contrast, when the learning rate increases to 5×10^{-4} and 1×10^{-3}, model performance drops significantly. This may be due to unstable gradient updates caused by larger step sizes, which make the model more likely

ISSN:2377-0430
Vol. 5, No. 8, 2025

to miss the optimal solution during training. The effect is especially evident in the AUC metric, indicating that the model has greater difficulty distinguishing between emotional categories under high learning rate conditions.

In summary, the emotion recognition task shows high sensitivity to learning rate settings. Both overly small and overly large learning rates can negatively affect the model's performance in complex emotional environments. Therefore, it is essential to carefully tune and validate the learning rate during training. This ensures a stable optimization state that balances representational accuracy, class separation, and semantic consistency. As a result, the model can achieve stronger emotional perception and better generalization.

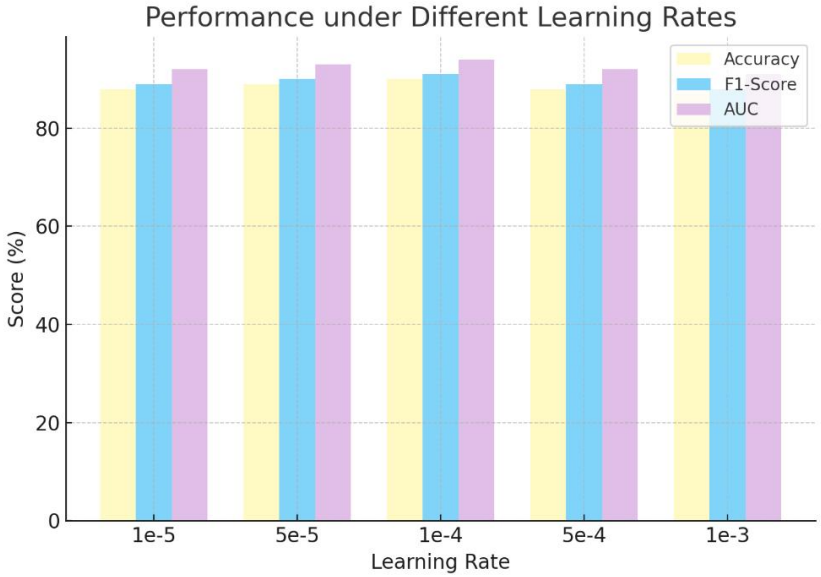

Figure 2. Analysis of the impact of different learning rate settings on model performance

6. Conclusion

This study focuses on the modeling advantages of deep learning in emotion recognition tasks and proposes a structurally optimized end-to-end framework. The model integrates a multi-layer Transformer encoder with a gating mechanism to achieve effective extraction and high-quality representation of emotional features. This method fully leverages the capabilities of deep models in semantic modeling, contextual understanding, and information filtering. It can handle various types of emotional inputs and forms clear decision boundaries between different emotional categories. As a result, the model significantly improves the accuracy, stability, and robustness of emotion recognition.

In terms of model design, structural awareness and dynamic control modules are introduced. These components allow the model to learn emotional tendencies within complex semantic relations without relying on handcrafted features. They also suppress redundant or irrelevant information. From the perspective of comparative experiments and sensitivity analysis, the model achieves leading performance across multiple evaluation metrics. This demonstrates its adaptability to diverse task requirements. Notably, the model maintains consistent output and structural stability under different learning rates, noise levels, and data ratios, showing strong environmental tolerance and application scalability.

This research provides a general and efficient design approach for future emotion recognition tasks. It also offers technical support for intelligent emotion perception in real-world applications such as human-computer

ISSN:2377-0430
Vol. 5, No. 8, 2025

interaction, public opinion analysis, and mental health monitoring. The end-to-end nature of the model makes it easy to deploy and iterate. It has strong engineering feasibility and integration potential. The method can be applied in fields such as social media content analysis, intelligent customer service, and virtual assistant behavior control. It supports the development of emotion-aware systems toward greater intelligence and personalization.

Future work may further extend the model's recognition capabilities in cross-modal, cross-lingual, and open-domain settings. By integrating multi-source perceptual signals such as speech, images, and physiological indicators, a more comprehensive emotion modeling system can be built. Enhancing the interpretability of the model is also a promising direction. This would help maintain performance while improving traceability and controllability, which are essential for high-sensitivity and safety-critical applications. As emotion computing continues to evolve alongside deep learning, emotion recognition is expected to play an increasingly important role in key areas such as education, healthcare, and social governance.

References

[1] Ameer I, Bölücü N, Siddiqui M H F, et al. Multi-label emotion classification in texts using transfer learning[J]. Expert Systems with Applications, 2023, 213: 118534.

[2] Liu X, Shi T, Zhou G, et al. Emotion classification for short texts: an improved multi-label method[J]. Humanities and Social Sciences Communications, 2023, 10(1): 1-9.

[3] Ahmed M Z I, Sinha N, Phadikar S, et al. Automated feature extraction on AsMap for emotion classification using EEG[J]. Sensors, 2022, 22(6): 2346.

[4] Khateeb M, Anwar S M, Alnowami M. Multi-domain feature fusion for emotion classification using DEAP dataset[J]. Ieee Access, 2021, 9: 12134-12142.

[5] Wang, R., Luo, Y., Li, X., Zhang, Z., Hu, J., & Liu, W. (2025, January). A Hybrid Recommendation Approach Integrating Matrix Decomposition and Deep Neural Networks for Enhanced Accuracy and Generalization. In 2025 5th International Conference on Neural Networks, Information and Communication Engineering (NNICE) (pp. 1778-1782). IEEE.

[6] Zhan, J. (2025). Elastic Scheduling of Micro-Modules in Edge Computing Based on LSTM Prediction. Journal of Computer Technology and Software, 4(2).

[7] Sun, X., Duan, Y., Deng, Y., Guo, F., Cai, G., & Peng, Y. (2025, March). Dynamic operating system scheduling using double DQN: A reinforcement learning approach to task optimization. In 2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE) (pp. 1492-1497). IEEE.

[8] Peng, Y. (2024). Structured Knowledge Integration and Memory Modeling in Large Language Systems. Transactions on Computational and Scientific Methods, 4(10).

[9] Zhang, Y., Liu, J., Wang, J., Dai, L., Guo, F., & Cai, G. (2025). Federated learning for cross-domain data privacy: A distributed approach to secure collaboration. arXiv preprint arXiv:2504.00282.

[10]Cheng, Y. (2025). Multivariate time series forecasting through automated feature extraction and transformer-based modeling. Journal of Computer Science and Software Applications, 5(5).

[11]Duan, Y. (2024). Continuous Control-Based Load Balancing for Distributed Systems Using TD3 Reinforcement Learning. Journal of Computer Technology and Software, 3(6).

[12]Dai, L., Zhu, W., Quan, X., Meng, R., Chai, S., & Wang, Y. (2025). Deep Probabilistic Modeling of User Behavior for Anomaly Detection via Mixture Density Networks. arXiv preprint arXiv:2505.08220.

ISSN:2377-0430
Vol. 5, No. 8, 2025

[13]Lou, Y. (2024). Capsule Network-Based AI Model for Structured Data Mining with Adaptive Feature Representation. Transactions on Computational and Scientific Methods, 4(9).

[14]Liu, J. (2025). Reinforcement Learning-Controlled Subspace Ensemble Sampling for Complex Data Structures.

[15]Yang, T., Cheng, Y., Ren, Y., Lou, Y., Wei, M., & Xin, H. (2025). A Deep Learning Framework for Sequence Mining with Bidirectional LSTM and Multi-Scale Attention. arXiv preprint arXiv:2504.15223.

[16]Wang, Y., Sha, Q., Feng, H., & Bao, Q. (2025). Target-Oriented Causal Representation Learning for Robust Cross-Market Return Prediction. Journal of Computer Science and Software Applications, 5(5).

[17]Gao, D. (2024). Graph Neural Recognition of Malicious User Patterns in Cloud Systems via Attention Optimization. Transactions on Computational and Scientific Methods, 4(12).

[18]Xing, Y., Yang, T., Qi, Y., Wei, M., Cheng, Y., & Xin, H. (2025). Structured Memory Mechanisms for Stable Context Representation in Large Language Models. arXiv preprint arXiv:2505.22921.

[19]Zheng, H., Ma, Y., Wang, Y., Liu, G., Qi, Z., & Yan, X. (2025). Structuring Low-Rank Adaptation with Semantic Guidance for Model Fine-Tuning.

[20]Wang, Y., Zhu, W., Quan, X., Wang, H., Liu, C., & Wu, Q. (2025). Time-Series Learning for Proactive Fault Prediction in Distributed Systems with Deep Neural Structures. arXiv preprint arXiv:2505.20705.

[21]Meng, X., Wu, Y., Tian, Y., Hu, X., Kang, T., & Du, J. (2025). Collaborative Distillation Strategies for Parameter-Efficient Language Model Deployment. arXiv preprint arXiv:2507.15198.

[22]Lyu, S., Deng, Y., Liu, G., Qi, Z., & Wang, R. (2025). Transferable Modeling Strategies for Low-Resource LLM Tasks: A Prompt and Alignment-Based. arXiv preprint arXiv:2507.00601.