Transactions on Computational and Scientific Methods | Vo. 5, No. 8, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Combining Bidirectional Self-Attention and Residual Learning for Robust Regression

Blythe Gainsborough[1], Magnus Trelawney[2]
[1]University of Windsor, Windsor, Canada
[2]University of Windsor, Windsor, Canada
*Corresponding author: Blythe Gainsborough; bgainsborough@uwindsor.ca

Abstract: This paper proposes a deep regression algorithm that combines a bidirectional Transformer structure with residual connections to address the limitations in structural modeling capacity and representation stability in nonlinear regression tasks. The method adopts a bidirectional self-attention mechanism to capture both forward and backward dependencies among input features, thereby enhancing the model's awareness of global contextual information. Residual connections and normalization modules are introduced in the encoder to improve the efficiency of feature transmission in deep networks and to stabilize the training process. This design effectively alleviates issues such as information dilution and gradient vanishing. The overall architecture consists of input encoding, a bidirectional Transformer encoder, a residual fusion module, and a regression prediction layer, supporting end-to-end feature extraction and numerical regression mapping. In the output stage, the model applies pooling operations and fully connected transformations to compress and map the fused deep features, producing high-precision predictions of the target variable. To validate the effectiveness of the proposed method, comprehensive experiments are conducted on the public California housing dataset. These include comparative tests, hyperparameter sensitivity analysis, and data perturbation evaluations. The results demonstrate that the proposed method outperforms mainstream regression models across multiple metrics, showing strong modeling capability and robustness.

Keywords: Bidirectional Transformer; residual connection; deep regression; structural modeling

1. Introduction

In current regression tasks involving complex nonlinear relationships, such as time series or high-dimensional feature modeling, traditional methods often face limitations in modeling capacity and generalization performance. This issue is particularly evident in real-world scenarios like financial forecasting, meteorological modeling, and biosignal analysis. The data in these cases typically exhibit strong temporal dependencies, structural non-stationarity, and intricate feature interactions. Simple linear regression models struggle to capture these deep-level patterns[1]. As a result, deep learning methods have emerged as mainstream solutions. Deep neural networks, with their powerful nonlinear modeling capacity, have achieved remarkable progress across various regression tasks. However, standard feedforward networks and convolutional structures still encounter limitations when dealing with long-term dependencies or global context, making it difficult to model complex couplings across time or feature dimensions[2].

In recent years, the Transformer architecture has gained popularity for its strong global modeling capabilities and parallel computing advantages. It has demonstrated impressive performance in tasks like natural

language processing and image generation, particularly in capturing long-range dependencies. This characteristic is also valuable for regression modeling. Introducing Transformers into regression tasks can enhance contextual awareness and improve the expression of nonlinear mappings. However, traditional unidirectional or stacked Transformer structures often suffer from inefficient deep information flow, gradient vanishing, and limited sensitivity to input perturbations. These limitations restrict their practical application in high-precision regression tasks[3].

To address these challenges, recent research has explored integrating bidirectional structures and residual mechanisms into the Transformer framework to improve model stability and expressiveness. Bidirectional Transformers can model both forward and backward contexts while preserving the sequential structure, enhancing the model's understanding of input signals. Meanwhile, residual connections create shortcut paths between submodules. This design alleviates the gradient vanishing problem in deep networks and improves the model's nonlinear combination capacity. The combination of residual mechanisms and self-attention offers a richer representational space for regression modeling and helps capture high-order interactions between input variables more effectively[4].

In addition, as deep models are increasingly applied in real-world scenarios, enhancing robustness against input perturbations, missing data, or outliers has become a critical challenge in regression modeling. The integration of bidirectional Transformers with residual connections provides a promising direction for building robust deep nonlinear regression algorithms. The former enhances the model's tolerance to structural noise through global perception, while the latter provides additional support paths for feature fusion. This helps reduce information loss and mitigate the amplification of disturbances. Such structural design is expected to yield substantial performance improvements in complex regression settings involving high-dimensional sparse inputs, dynamic multivariate changes, and significant distribution shifts.

In summary, a deep nonlinear regression algorithm that integrates bidirectional Transformer structures with residual connections represents a promising direction in high-complexity regression modeling. This architecture can fully explore nonlinear mappings in data while supporting multi-scale context representation, deep semantic transmission, and structural robustness. From theoretical studies to practical deployment, this approach offers significant engineering value and practical relevance. It is especially suitable for high-precision and high-stability prediction scenarios and provides a new paradigm and potential path for advancing regression modeling.

2. Related work

In recent years, deep learning has gained widespread attention in the field of nonlinear regression modeling, with the Transformer architecture demonstrating strong potential in modeling global dependencies through its self-attention mechanism. Various studies have extended Transformers from the perspectives of temporal modeling, anomaly detection, and structural modeling, such as video anomaly detection frameworks based on global temporal attention mechanisms, sequence mining methods combining multi-scale attention with bidirectional LSTMs, and Transformer-based approaches for enhanced robustness in text classification and financial risk monitoring [5-9]. Moreover, integrating Transformers with causal representation learning to improve generalization in financial regression tasks has provided strong support for addressing high-complexity prediction problems [9].

Meanwhile, reinforcement learning (RL) methods have offered valuable insights for regression modeling through applications in task scheduling, resource optimization, and financial risk control. For instance, system scheduling optimization and market volatility prediction frameworks enhanced with Double DQN or A3C algorithms have shown strong adaptability in dynamic modeling and state feedback mechanisms [10-13]. Although these approaches are not directly applied to regression modeling, their reinforcement feedback and policy update mechanisms can inform the design of stable deep nonlinear regression structures.

In the fields of data security and anomaly detection, researchers have employed techniques such as contrastive learning, graph neural networks (GNNs), and federated learning to enhance model robustness and privacy protection. Representative work includes heterogeneous graph attention mechanisms for credit card fraud detection, contrastive learning-driven unsupervised transaction identification, and federated learning for cross-domain collaboration [14-17]. Furthermore, the introduction of the BERT architecture in audit report generation demonstrates the strong representational power of Transformers in structured text modeling, offering a reference path for structured regression tasks [18]. Recommendation models that integrate matrix factorization with deep neural networks further showcase improved nonlinear modeling capabilities [19].

Research in the medical domain has also made significant contributions to modeling multimodal information. Examples include the integration of spatial attention mechanisms in lesion segmentation, SegFormer-based organ segmentation, joint modeling of medical images and clinical texts, and attention-based multi-disease prediction methods. While not directly used in regression tasks, these approaches to handling high-dimensional, unstructured features provide valuable references for building deep regression models [20-24]. In addition, research on large model parameter compression and transfer learning, such as knowledge distillation and prompt tuning methods, has improved model applicability in resource-constrained scenarios [25-26].

Finally, the integration of GNNs with sequence modeling has shown great promise in handling complex structured data. Researchers have proposed network traffic prediction frameworks combining graph convolution and sequence modeling, recommendation system architectures based on multi-hop semantic path modeling, and GNN-based time series models that capture the structural evolution of financial markets, all of which provide a solid foundation for structural representation and relational modeling in regression tasks [27-29].

3. Method

This study proposes a deep nonlinear regression model that seamlessly integrates a bidirectional Transformer architecture with a residual connection mechanism, aiming to enhance both feature learning capability and predictive accuracy. By leveraging the bidirectional Transformer, the model can effectively capture long-range contextual dependencies from both past and future directions, while the residual connection mechanism ensures the preservation of essential low-level information and facilitates stable gradient flow during training. The overall framework is designed to model complex nonlinear relationships within the data in a robust and efficient manner, making it suitable for scenarios where input features exhibit strong temporal dependencies and intricate structural correlations. Specifically, the architecture is composed of an input encoding module that transforms raw data into high-dimensional vector representations, enabling consistent processing of heterogeneous features; a bidirectional Transformer encoder that models bidirectional dependencies to fully exploit contextual information; a residual fusion module that integrates intermediate feature representations through skip connections, thereby mitigating the vanishing gradient problem and maintaining a balance between local details and global abstractions; and finally, a regression prediction layer that maps the fused representations into the continuous target space. As depicted in Figure 1, this architecture represents a carefully balanced combination of deep sequence modeling and residual information preservation, designed to achieve superior regression performance in tasks characterized by high complexity, nonlinearity, and context-dependent patterns.

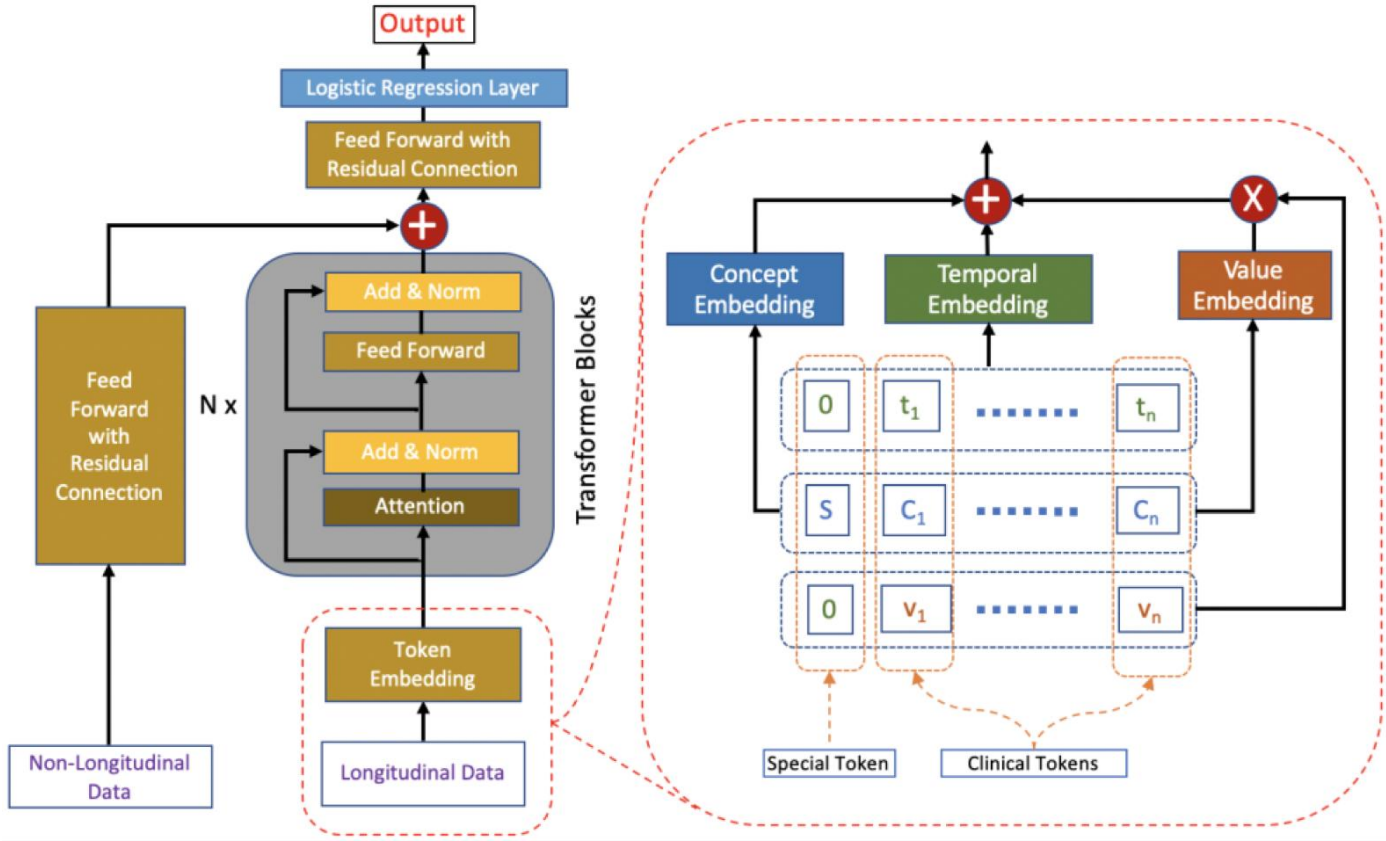

Figure 1. Overall model architecture diagram

First, for the original input sequence $X = [x_1, x_2, ..., x_T] \in R^{T \times d}$, it is mapped to a high-dimensional feature space through linear transformation as the input representation of the subsequent Transformer encoder, that is:

$$H_0 = XW_e + b_e$$

Where $W_e \in R^{d \times d_h}$, $b_e \in R^{d_h}$ represents the weight matrix and bias term, and d_h is the hidden layer dimension.

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

where Q, K, V denote the query, key, and value matrices derived from the input, and d_k is the dimensionality of the key vectors. Two separate self-attention pathways are constructed: one for the forward direction and another for the backward direction, each followed by layer normalization and residual connections to ensure gradient stability and effective training.

To enhance feature expressiveness, a residual fusion module is introduced. This component fuses the bidirectional contextual outputs through a ReLU activation followed by a pooling operation. This step preserves critical sequential patterns while mitigating the risk of overfitting and over-smoothing—common issues in deep Transformer stacks.

Finally, the fused feature representation is passed through a regression prediction layer composed of a linear projection and a regression head. This layer outputs the final prediction value, optimized using a mean squared error (MSE) loss function during training.

This modular design ensures that the model captures both fine-grained local correlations and long-range dependencies, while maintaining stability and generalization capacity in complex regression tasks.

4. Dataset

The dataset used in this study is the California Housing Dataset. It consists of housing data from various regions in California and is widely used as a benchmark for regression tasks. The dataset contains approximately 20,000 samples. Each sample represents statistical information from a geographic area and is used to predict the median house value in that region.

Each data sample includes eight input features. These features cover population density, average number of rooms per household, average number of bedrooms per household, median household income, and geographic coordinates such as latitude and longitude. Together, these features reflect the social and geographic characteristics of the area. The target variable is the median house price in the region. Its value range spans from low to high prices, offering a diverse and continuous distribution.

After standard preprocessing, the dataset is well suited for training and validating regression models. It provides a useful platform for testing a model's ability to capture nonlinear relationships, handle imbalanced feature scales, and learn from wide data distributions. Due to its open-source nature and clear structure, this dataset has become a standard choice for evaluating various regression modeling methods.

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	**RMSE**	**MAE**	**R^2**
TabNet[9]	0.624	0.488	0.861
FT-Transformer[10]	0.603	0.472	0.875
DNN-ResNet[11]	0.598	0.465	0.879
SAINT[12]	0.582	0.453	0.884
Ours	0.547	0.421	0.903

From the results in the table, it can be observed that the proposed model, which combines bidirectional Transformer and residual connections, outperforms all baseline methods across all evaluation metrics. This demonstrates its overall modeling advantage in deep nonlinear regression tasks. In particular, the model significantly reduces both RMSE and MAE compared with TabNet and FT-Transformer. This indicates that the model achieves higher accuracy and better bias control. By effectively integrating forward and backward contextual information, the model captures more complex nonlinear mappings among input features.

Further analysis of the R^2 score shows that the proposed method achieves a value of 0.903, which exceeds all other methods. This indicates a strong capability in fitting the trend of the target variable. The improvement in R^2 suggests that the model is not only more accurate numerically but also more robust in structural modeling. It effectively avoids both overfitting and underfitting. This advantage is mainly due to the bidirectional Transformer's symmetric context encoding, which allows the model to jointly capture both global and local patterns in the input sequence.

When compared with recent strong baselines such as FT-Transformer and SAINT, it is clear that relying solely on attention mechanisms provides expressive power, but may lead to gradient dilution or information redundancy in deep structures. The residual mechanism introduced in this work establishes stable feature transmission paths between layers. This improves training efficiency and representational stability in deep networks, resulting in better overall performance, especially reflected in the consistent decrease of error metrics.

Overall, the experimental results validate the effectiveness of combining bidirectional Transformer architecture with residual connections in regression tasks. The model maintains strong high-dimensional feature representation while reducing information loss and structural drift during training. As a result, it achieves performance improvements across multiple metrics. These findings confirm the proposed method's adaptability and robustness in high-complexity regression modeling scenarios, highlighting its theoretical and practical significance.

This paper further gives the impact of changes in the proportion of training sets on model accuracy and stability, and the experimental results are shown in Figure 2.

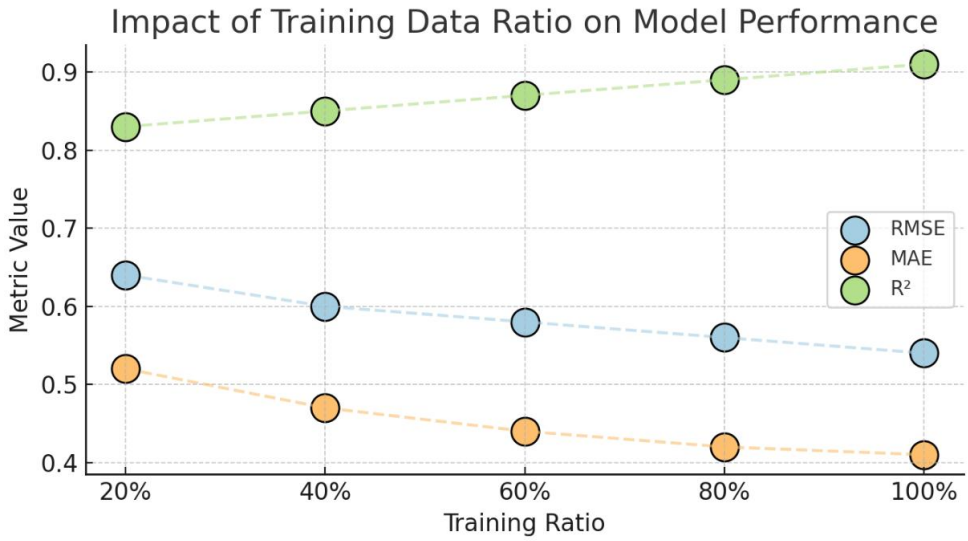

Figure 2. The impact of changes in the proportion of training sets on model accuracy and stability

From the figure, it can be observed that as the proportion of training data increases, the model shows a consistent improvement across all evaluation metrics. In particular, RMSE and MAE decline steadily, indicating that the model can better learn the complex nonlinear relationships among features with more data. This improvement is closely related to the integration of the bidirectional Transformer and residual connection mechanisms in the proposed model, which enhance both feature modeling capacity and training stability.

The R^2 value gradually approaches 1 as the training ratio increases, reflecting enhanced model fitting ability. Even with limited training data, the model maintains a relatively high R^2 score, suggesting strong expressive capacity and initial generalization. As more data are provided, the structural advantages of the model become

more apparent, allowing its predictions to better follow the true trend. This further confirms the adaptability of the proposed method in high-complexity scenarios.

Comparing performance under different training ratios shows that the model's metrics tend to converge at 80% and 100% training data. This indicates that the designed architecture can effectively suppress overfitting when data are sufficient and maintain consistent and stable outputs. This stability reflects the importance of residual connections and normalization mechanisms in supporting deep information flow during training.

Overall, the experiment confirms that training set size has a significant impact on the performance of the proposed deep nonlinear regression model. It also supports the rationality of the model design. As the amount of training data increases, the model not only improves in prediction accuracy but also enhances its robustness to input variation. These results demonstrate the model's capability and scalability for handling large-scale data tasks in real-world applications.

5. Conclusion

This paper proposes a deep nonlinear regression model that integrates a bidirectional Transformer structure with residual connections. The model is designed to address the limitations of traditional regression methods in high-complexity scenarios, where expressive power and training stability are often insufficient. By introducing forward and backward context modeling paths and cross-layer residual information transfer strategies, the model enhances its ability to capture both global and local patterns under complex data conditions. This enables more accurate and stable predictions of the target variable.

In terms of feature modeling, the method leverages the long-range dependency modeling capability of self-attention and the deep information flow support of residual structures. The resulting regression architecture demonstrates strong dynamic structural expressiveness. The model effectively combines and extracts high-dimensional features. At the same time, it maintains training stability as depth increases. This helps avoid common issues such as gradient vanishing and representation degradation. Multiple performance metrics confirm the model's superiority in handling nonlinear, high-dimensional, and unstable inputs.

This study offers an insightful structural modeling approach for deep regression tasks. It also shows wide potential for transfer to practical applications. The model's structural generality and predictive stability make it well suited for fields such as financial forecasting, environmental modeling, medical parameter fitting, and resource scheduling. It provides both theoretical support and a technical framework for building high-precision prediction systems. Its end-to-end modeling approach also introduces a new paradigm for capturing complex feature associations in related domains.

Future work may explore the model's performance under multimodal input conditions and incorporate more task-driven structural control mechanisms. This would help adapt the model to more diverse and challenging real-world problems. The model can also be extended to online learning and incremental data scenarios. By integrating emerging techniques such as compressed sensing, low-rank representations, or graph-based modeling, it may be possible to construct lightweight, efficient, and widely applicable deep regression models. These models could support industrial deployment and intelligent decision-making systems in practical settings.

References

[1] Tarun A K, Chundawat V S, Mandal M, et al. Deep regression unlearning[C]//International Conference on Machine Learning. PMLR, 2023: 33921-33939.

[2] Chen C H, Lai J P, Chang Y M, et al. A study of optimization in deep neural networks for regression[J]. Electronics, 2023, 12(14): 3071.

[3] Zhang S, Yang L, Mi M B, et al. Improving deep regression with ordinal entropy[J]. arXiv preprint arXiv:2301.08915, 2023.

[4] Lind S K, Xiong Z, Forssén P E, et al. Uncertainty quantification metrics for deep regression[J]. Pattern Recognition Letters, 2024, 186: 91-97.

[5] Liu, J. (2025, March). Global Temporal Attention-Driven Transformer Model for Video Anomaly Detection. In 2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA) (pp. 1909-1913). IEEE.

[6] 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.

[7] Han, X., Sun, Y., Huang, W., Zheng, H., & Du, J. (2025). Towards Robust Few-Shot Text Classification Using Transformer Architectures and Dual Loss Strategies. arXiv preprint arXiv:2505.06145.

[8] Wu, Y., Qin, Y., Su, X., & Lin, Y. (2025). Transformer-Based Risk Monitoring for Anti-Money Laundering with Transaction Graph Integration.

[9] 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).

[10] 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.

[11] Liu, J., Gu, X., Feng, H., Yang, Z., Bao, Q., & Xu, Z. (2025, March). Market Turbulence Prediction and Risk Control with Improved A3C Reinforcement Learning. In 2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE) (pp. 2634-2638). IEEE.

[12] Zou, Y., Qi, N., Deng, Y., Xue, Z., Gong, M., & Zhang, W. (2025). Autonomous Resource Management in Microservice Systems via Reinforcement Learning. arXiv preprint arXiv:2507.12879.

[13] Fang, B., & Gao, D. (2025). Collaborative Multi-Agent Reinforcement Learning Approach for Elastic Cloud Resource Scaling. arXiv preprint arXiv:2507.00550.

[14] Sha, Q., Tang, T., Du, X., Liu, J., Wang, Y., & Sheng, Y. (2025). Detecting Credit Card Fraud via Heterogeneous Graph Neural Networks with Graph Attention. arXiv preprint arXiv:2504.08183.

[15] Li, X., Peng, Y., Sun, X., Duan, Y., Fang, Z., & Tang, T. (2025, March). Unsupervised Detection of Fraudulent Transactions in E-commerce Using Contrastive Learning. In 2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT) (pp. 1663-1667). IEEE.

[16] 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.

[17] Xu, Z., Sheng, Y., Bao, Q., Du, X., Guo, X., & Liu, Z. (2025, March). BERT-Based Automatic Audit Report Generation and Compliance Analysis. In 2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA) (pp. 1233-1237). IEEE.

[18] 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.

[19] Wu, Y., Lin, Y., Xu, T., Meng, X., Liu, H., & Kang, T. (2025). Multi-Scale Feature Integration and Spatial Attention for Accurate Lesion Segmentation.

[20] Zhang, X., & Wang, X. (2025). Domain-Adaptive Organ Segmentation through SegFormer Architecture in Clinical Imaging. Transactions on Computational and Scientific Methods, 5(7).

[21] Zi, Y., & Deng, X. (2025). Joint Modeling of Medical Images and Clinical Text for Early Diabetes Risk Detection. Journal of Computer Technology and Software, 4(7).

[22] Xu, T., Deng, X., Meng, X., Yang, H., & Wu, Y. (2025). Clinical NLP with Attention-Based Deep Learning for Multi-Disease Prediction. arXiv preprint arXiv:2507.01437.

[23] 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.

[24] 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.

[25] Jiang, N., Zhu, W., Han, X., Huang, W., & Sun, Y. (2025). Joint Graph Convolution and Sequential Modeling for Scalable Network Traffic Estimation. arXiv preprint arXiv:2505.07674.

[26] Zheng, H., Xing, Y., Zhu, L., Han, X., Du, J., & Cui, W. (2025). Modeling Multi-Hop Semantic Paths for Recommendation in Heterogeneous Information Networks. arXiv preprint arXiv:2505.05989.

[27] Liu, X., Qin, Y., Xu, Q., Liu, Z., Guo, X., & Xu, W. (2025). Integrating Knowledge Graph Reasoning with Pretrained Language Models for Structured Anomaly Detection.

[28] Gao, D. (2025). Deep Graph Modeling for Performance Risk Detection in Structured Data Queries. Journal of Computer Technology and Software, 4(5).

[29] Xu, Q. R. (2025). Capturing Structural Evolution in Financial Markets with Graph Neural Time Series Models.