ISSN 2616-5775 Vol. 9, Issue 1: 48-55, DOI: 10.25236/AJCIS.2026.090106

NHiTS-MSFB: Dynamic Local–Global Feature Fusion for Time Series Forecasting

Tiancai Zhu[1,a], Jiangtao Liu[1,b,*], Jiahao Duan[1,c], Wei Wang[1,d], Yonghao Wu[1,e], Tongzhu Zhao[1,f]

[1]*School of Information Science and Technology, Yunnan Normal University, Kunming, China*
[a]*2324100067@ynnu.edu.cn,* [b]*ljt@ynnu.edu.cn,* [c]*2324100012@ynnu.edu.cn,* [d]*2324100048@ynnu.edu.cn,* [e]*2324100051@ynnu.edu.cn,* [f]*2324100063@ynnu.edu.cn*
[*]*Corresponding author*

Abstract: *The core challenge in time series forecasting lies in effectively modeling long-term dependencies and multi-scale patterns. Although the NHiTS model has made progress in long-term forecasting through its multi-scale framework, its core multilayer perceptron (MLP) building blocks have limitations in feature representation capability, making it difficult to jointly capture local fine-grained patterns and global longterm dependencies. To address this, this paper proposes an improved model architecture, with its core innovation being the design of a novel Multi-Scale Fusion Block (MSFB) to enhance multi-period feature representation. This module explicitly models local temporal patterns and global dependencies through parallel multi-scale 1D convolutions and block-sparse attention mechanisms, respectively, and introduces a learnable dynamic fusion gating mechanism to adaptively integrate heterogeneous feature streams. Experiments were conducted on four benchmark datasets—ETTm2, Traffic, Weather, and Exchange—for training, validation, and testing. The results show that the improved model achieved average reductions of 11.20% and 7.79% in MAE and MSE metrics, respectively, compared to the original NHiTS model, and significantly outperformed mainstream comparative models such as TimesNet, Autoformer, and FEDformer. This validates the effectiveness of the proposed module in enhancing temporal representation learning and improving forecasting accuracy.*

Keywords: *Time Series Forecasting; NHiTS Model; Multi-Scale Convolution; Block-Sparse Attention*

1. Introduction

Time series forecasting, as a core task in data analysis, plays a crucial role in key fields such as power load forecasting, meteorological prediction, and financial market analysis [1-3].Real-world time series data often contain complex multi-scale temporal patterns (e.g., daily, weekly, and seasonal variations) and non-stationary dynamic characteristics, posing severe challenges to forecasting models: they must simultaneously possess high sensitivity to local fine-grained patterns, robust modeling capabilities for long-term trends and dependencies, and high efficiency in practical deployment.

In recent years, deep learning models, leveraging their powerful nonlinear fitting capabilities, have become mainstream methods for time series forecasting. Architectures represented by Transformer [4] and Temporal Convolutional Networks (TCN)[5]have demonstrated advantages in long-range dependency modeling and local pattern extraction, respectively. However, these models often face issues of high computational complexity or severe error accumulation when dealing with extremely long forecasting horizons. To address this, the NHiTS model proposed by Challu et al[6]. constructs an efficient pyramid analysis framework through multi-scale[7] hierarchical sampling and interactive downsampling mechanisms[8]. It achieves a good balance between forecasting accuracy and computational efficiency in long-term forecasting tasks.

Nevertheless, the core building block of the NHiTS model—the Multilayer Perceptron (MLP)—has inherent limitations in feature representation capability[9]: it struggles to effectively capture fine-grained intra-period patterns; it inefficiently establishes long-range dependencies by stacking multiple network layers, which can lead to gradient-related problems; and its feature transformation is simplistic, failing to adaptively balance local details and global contextual information.

To address the limitations of the MLP block, researchers have explored methods to enhance temporal feature extraction from different perspectives. For instance, Convolutional Neural Networks (CNN) [10-

ISSN 2616-5775 Vol. 9, Issue 1: 48-55, DOI: 10.25236/AJCIS.2026.090106

[11], with their local receptive fields and weight-sharing mechanisms, have been proven effective in capturing local patterns and multi-scale features in time series. Sparse or localized attention mechanisms[12]can maintain modeling capability for key long-term dependencies while reducing computational complexity. These works provide important insights for designing more powerful temporal feature extractors. However, how to organically embed and adapt such capabilities into the efficient multi-scale framework of NHiTS, rather than simply stacking or replacing the entire architecture, remains a problem warranting further exploration.

Therefore, this paper proposes an improved NHiTS model, with its core innovation being the design of a Multi-Scale Fusion Block (MSFB). Targeting the multi-scale hierarchical characteristics of NHiTS, this module combines multi-scale convolution and block-sparse attention mechanisms to achieve efficient modeling of local temporal patterns and global dependencies, respectively. It enhances basic feature representation capability through a dynamic gating mechanism for adaptive feature fusion.

The main contributions of this paper are as follows:

(1) We systematically analyze the specific limitations of the MLP building block in the NHiTS model in terms of temporal feature representation, explicitly pointing out its inadequacy in jointly modeling local fine-grained patterns and global long-term dependencies, thereby providing a theoretical basis for improving such models.

(2) To address the above limitations, we propose a Multi-Scale Fusion Block (MSFB), which achieves unified and efficient modeling of local features and global dependencies in time series through the collaborative design of parallel multi-scale convolution and block-sparse attention, along with a dynamic gating fusion mechanism.

(3) Experimental results on multiple public datasets demonstrate that the proposed improved model significantly enhances prediction performance through module-level replacement while maintaining the multi-scale pyramid architecture.

2. Related Work

2.1. Baseline Model and Related Methods

The NHiTS model constructs a pyramid structure through multi-scale hierarchical sampling and interactive downsampling mechanisms, capturing short-term patterns at shallow layers, modeling long-term trends at deep layers, and fusing predictions from different scales[6]. However, its fundamental MLP building block has inherent limitations in capturing complex nonlinear patterns within multi-periods: it lacks an explicit modeling mechanism for the local continuity of time series; it relies on stacking multiple network layers to indirectly establish long-range dependencies; its feature transformation is simplistic. These limitations make the basic feature extraction capability of the original NHiTS a performance bottleneck on sequences with complex periodic structures, which is the direct motivation for the improvements in this paper. In the field of time series forecasting, TimesNet captures multi-period patterns via 2D convolution but incurs high computational cost[13]; Autoformer introduces seasonal-trend decomposition and auto-correlation mechanisms[14]; FEDformer performs global modeling in the frequency domain. While these methods have their respective advantages, they mostly focus on overall architectural innovation[15]. This paper focuses on lightweight module improvements within the NHiTS framework. Inspired by the multi-period modeling idea of TimesNet, this paper designs a lightweight Multi-Scale Fusion Block (MSFB), employing 1D convolution and block-sparse attention to collaboratively model local and global features, and achieving adaptive fusion through dynamic gating, thereby enhancing the model's ability to model complex temporal patterns without excessively increasing computational burden.

2.2. Improved Model Design

Addressing the specific limitations of the MLP building block's temporal feature representation in the original NHiTS model analyzed in Section, this paper proposes an improved model. Its core idea is to embed the self-designed Multi-Scale Fusion Block (MSFB) into the NHiTS model to enhance local pattern capture and global dependency modeling capabilities. The overall model architecture is shown in Figure 1, consisting of three core parts:

(1) Input Projection Layer: The model first maps the input to a high-dimensional space via a linear

ISSN 2616-5775 Vol. 9, Issue 1: 48-55, DOI: 10.25236/AJCIS.2026.090106

projection layer, obtaining $X^{(0)} \in R^{L \times d\_model}$. This projection is added to a sinusoidal positional encoding to obtain the processed data $X^{(1)}$.

(2) Multi-Scale Fusion Processing Layer: Consists of N_s stacked stacks. Each stack contains N_b serially connected MSFB modules (see Figure 2), responsible for feature extraction and transformation at the current scale. The final output of each stack is a prediction sequence $\hat{Y}_i \in R^{H \times 1}$ (where H is the forecast length).

(3) Gated Fusion and Output Layer: Inspired by the gated fusion mechanism introduced by Kocak et al[16]. for prediction outputs $\hat{Y}_1, \hat{Y}_2, \ldots \ldots, \hat{Y}_{N_s}$ from all stacks are passed through a shallow neural network to generate normalized weights $g \in R^{N s}$, dynamically fusing the results of each stack: $\hat{Y}_{fused} = \Sigma_{i=1}^{N_s} g_i \cdot \hat{Y}_i$. Finally, $\hat{Y}_{fused}$ is passed through an output projection layer to obtain the model's final prediction $\hat{Y} \in R^H$.

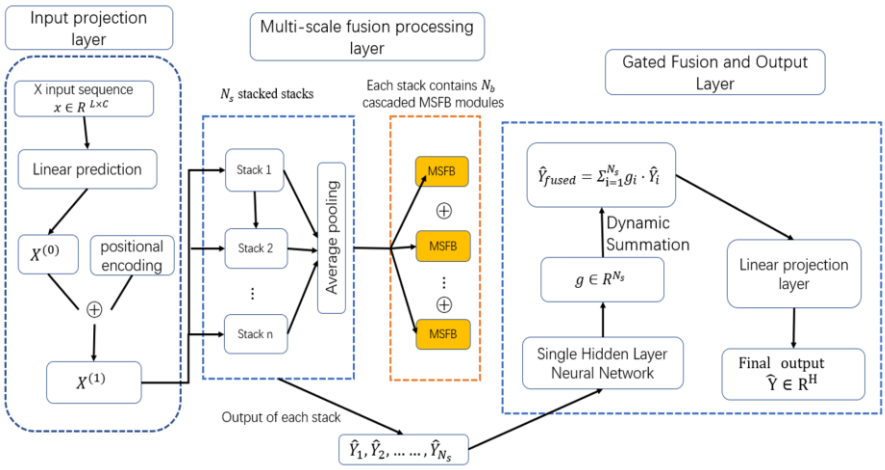

Figure 1: model structure diagram

Through this processing flow, the overall model ensures backward compatibility of the framework while significantly enhancing basic feature representation capability.

2.3. MSFB Module Design

The MSFB module is the core feature extraction unit of the model. Its structure is shown in Figure 2. Input X enters two parallel pathways:

Multi-Scale Convolution Pathway: The input passes through multiple parallel Conv1dk layers, undergoes batch normalization and GELU activation, and is then averaged to obtain the local feature F_{conv}

$$F_{conv} = \frac{1}{|K|} \sum_{k \epsilon K} Dropout(GELU(BatchNorm(DepthwiseConv1d_k(X)))) \quad (1)$$

where K is the set of convolution kernel sizes.

Sparse Attention Pathway: The input sequence is divided into blocks of length block size.Attention is computed independently within each block ($Q_i, K_i, V_i \rightarrow softmax$). The outputs A_j are concatenated and passed through a linear projection layer to obtain the global feature F_{attn}. For the j -th block:

$$Q_i, K_i, V_i = X_{block}[j]W_q \, , X_{block}[j]W_k \, , X_{block}[j]W_v \quad (2)$$

$$A_j = Softmax\left(\frac{Q_j K_j^T}{\sqrt{d_k}}\right) V_j \quad (3)$$

All block outputs $A\,j$ are concatenated and passed through a linear projection layer to obtain the final output F attn of this pathway.

Dynamic Fusion and Output: Adaptive weights $\alpha = Softmax(w \,/\, \tau)$ are computed using learnable weights $w = [w_{conv}, w_{attn}] \epsilon R^2$, and the features from the two pathways are weighted and fused:

ISSN 2616-5775 Vol. 9, Issue 1: 48-55, DOI: 10.25236/AJCIS.2026.090106

$$F_{fused} = \alpha_1 F_{conv} + \alpha_2 F_{attn} \tag{4}$$

The fused features undergo GLU nonlinear transformation, are then combined with the input X via a residual connection, and output after layer normalization:

$$X_{out} = LayerNorm\left(X + Dropout\left(GLU\left(F_{fused}\right)\right)\right) \tag{5}$$

The outputs of the two pathways are weighted and fused in the dynamic fusion gating layer, then undergo nonlinear transformation via a Gated Linear Unit (GLU), and finally undergo a residual connection (Add) with the input X , and are output after layer normalization.

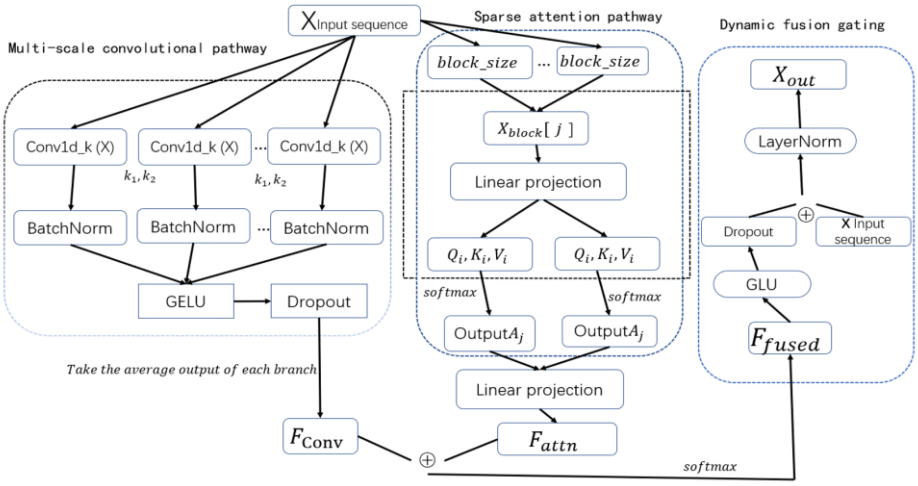

Figure 2: Structure of the Multi-Scale Fusion Block (MSFB)

This module, while maintaining the lightweight nature of NHiTS, enhances the modeling capability for local patterns and global dependencies through the collaborative design of multi-scale convolution and block-sparse attention, and achieves adaptive feature fusion through dynamic gating, demonstrating the value of module-level innovation within an efficient multi-scale framework.

The effectiveness of the proposed MSFB will be validated through extensive experiments in Section4.

3. Experimental Setup

To comprehensively evaluate the performance of the improved model, we conducted rigorous experiments on multiple public datasets and compared it with a series of advanced baseline models. To evaluate the model's forecasting performance from multiple dimensions, the experiments employed Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2) as evaluation metrics.

3.1. Datasets

The experiments used four public time series datasets from different domains. All datasets were sourced from the Tsinghua University Time Series Forecasting Bench mark Library (Time-Series-Library)[17]. ETTm2: Electricity Transformer Temperature dataset, exhibiting clear daily (96 points) and weekly (672 points) cycles, making it an ideal dataset for testing multi-period modeling capability. Traffic: Highway sensor occupancy rate data, containing daily cycles and weekend patterns. Weather: Multivariate meteorological data containing multiple periodicities like temperature and humidity, making it a dataset with cross-scale cycles. Exchange: Multi-country exchange rate data, with strong trends and no obvious fixed periodicity, making it a non-stationary sequence dataset lacking explicit periodicity.

3.2. Data Processing

To ensure the fairness and reproducibility of the experiments, a unified data preprocessing pipeline was adopted. The data processing flow is shown in Figure 3, mainly including four steps: data acquisition, missing value handling, feature engineering, and standardization:

ISSN 2616-5775 Vol. 9, Issue 1: 48-55, DOI: 10.25236/AJCIS.2026.090106

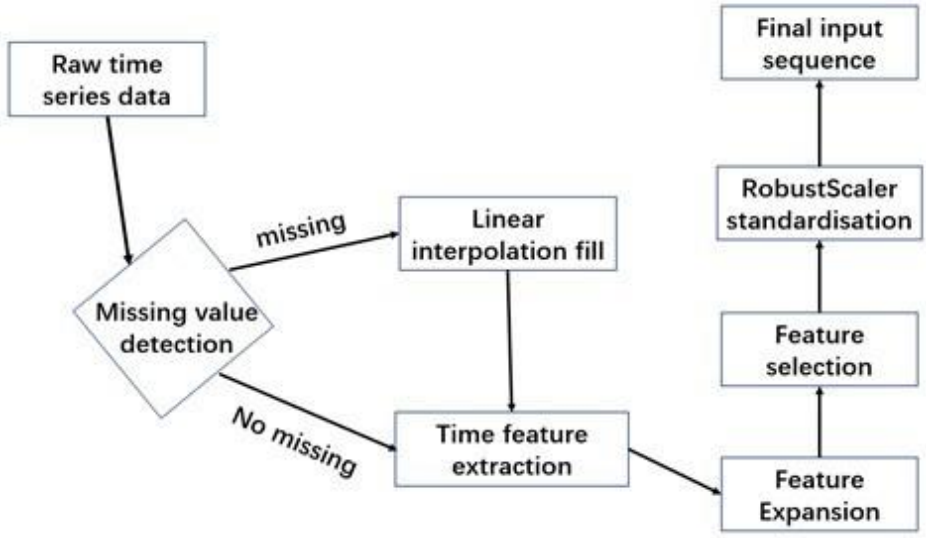

Figure 3: Data Processing Flowchart

Linear interpolation was first applied to handle missing values. Temporal features were then encoded using sine–cosine transformations to preserve their cyclical properties. For high-dimensional datasets such as Traffic, mutual information–based feature selection was performed to identify key predictors, following the findings of Covert et al.[18]. Finally, RobustScaler was used for normalization due to its robustness to outliers, as recommended by Arefi et al.[19-20]. All preprocessing steps were fitted exclusively on the training set to avoid data leakage.

3.3. Experimental Parameter Configuration

All experiments were conducted under the same hardware and software environment to ensure fair comparability of results.All models are trained with the same early stopping strategy.The specific configuration is as follows: the processor is an Intel Core i5-13400F, the graphics card is an NVIDIA GeForce RTX 3060 12GB, and the operating system is Windows 11. The experiments were implemented based on Python 3.11 and the PyTorch 2.5 framework. Model training utilized the Adam optimizer, with a batch size of 32, a learning rate of 3e-4, and was run for 50 epochs.

4. Experimental Results and Analysis

4.1. Ablation Study

To verify the necessity of each sub-component within the MSFB module and its contribution to overall performance, we conducted a systematic ablation study on the ETTm2 dataset. The experimental setup is shown in Table 1. By sequentially removing key modules while keeping the experimental configuration consistent (total samples: 69,680, split 44,595/11,149/13,936 for train/validation/test, 24-step multi-step forecasting task), we observed the changes in model performance.

Table 1: Ablation study of model components

Component	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
Multi-scale fusion	×	✓	✓	✓	✓
Sparse attention	✓	×	✓	✓	✓
Dynamic fusion	✓	✓	×	✓	✓
GLU gating	✓	✓	✓	×	✓
MAE	0.317	0.173	0.175	0.179	**0.159**
MSE	0.403	0.215	0.219	0.228	**0.173**
R^2	0.803	0.939	0.936	0.931	**0.964**
Params (M)	0.27	1.49	1.48	1.29	1.49

The experimental results demonstrate that each component within the improved module is

ISSN 2616-5775 Vol. 9, Issue 1: 48-55, DOI: 10.25236/AJCIS.2026.090106

indispensable: removing the multi-scale fusion module leads to a 133.0% surge in MSE, highlighting its importance in capturing local patterns; removing sparse attention, dynamic fusion, or GLU gating results in 24.3%, 26.6%, and 31.8% increases in MSE, respectively, confirming the critical role of each component in feature extraction and fusion.

4.2. Model Comparison Experiment

To comprehensively evaluate the performance of the improved model, we conducted systematic comparative experiments with multiple advanced time series forecasting baseline models,including: TimesNet, NHiTS, FEDformer, and Autoformer. All comparative experiments were conducted on the ETTm2 dataset under the same data preprocessing pipeline and evaluation metrics (total samples: 69,680, split 44595/11149/13936 for train/validation/test, 96-step multi-step forecasting) to ensure fairness of the results. Detailed quantitative results are shown in Table 2.

Table 2: Comparative experiments of different models

Dataset	Ours		TimesNet		NHiTS		FEDformer		Autoformer	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
ETTm2	**0.165**	**0.179**	0.333	0.291	0.176	0.185	0.305	0.349	0.327	0.371
Traffic	**0.313**	**0.394**	0.336	0.620	0.402	0.431	0.376	0.610	0.379	0.628
Weather	**0.148**	**0.189**	0.287	0.259	0.158	0.205	0.360	0.309	0.382	0.339
Exchange	**0.196**	**0.215**	0.443	0.416	0.218	0.243	0.429	0.519	0.539	0.613

A comprehensive analysis of the comparative experimental results presented in Table 2 reveals that the improved model significantly outperforms the original NHiTS model across all four public datasets. Specifically, compared to NHiTS, the improved model reduced the MAE metric by 6.25%, 22.14%, 6.33%, and 10.09%, respectively, with an average reduction of approximately 11.20%. It reduced the MSE metric by 3.24%, 8.58%, 7.80%, and 11.52%, respectively, with an average reduction of approximately 7.79%. Furthermore, the improved model demonstrated outstanding performance on datasets with prominent periodic patterns, such as ETTm2 and Weather, indicating that the MSFB module effectively enhances the model's ability to capture complex intra-period patterns. It also achieved significant improvements on non-stationary, trend-strong datasets like Exchange, suggesting that the multi-scale fusion mechanism improves the modeling stability for long-term dependencies.

4.3. Visualization Analysis of Forecasting Results

To further analyze the forecasting behavior of different models beyond quantitative metrics, we visualize the averaged prediction results on the ETTm2 dataset. Specifically, Figure 4 illustrates the mean forecasting curves over 30 randomly selected test samples with a prediction horizon of 96 steps. Compared with single-sample visualization, this averaged representation effectively reduces the influence of extreme fluctuations and provides a more robust reflection of the overall predictive characteristics of each model.

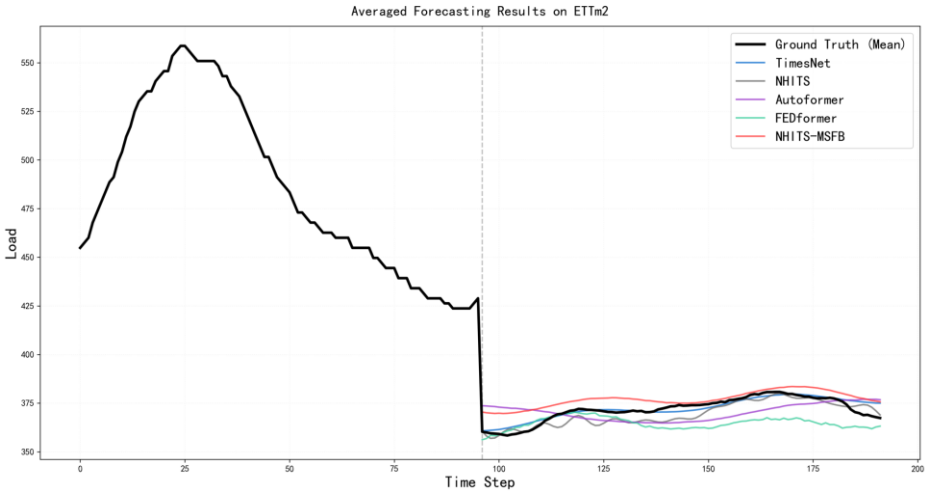

Figure 4: Averaged Forecasting Results on ETTm2

As shown in Figure 4, all models are able to capture the general trend of the ground truth to varying

ISSN 2616-5775 Vol. 9, Issue 1: 48-55, DOI: 10.25236/AJCIS.2026.090106

degrees. TimesNet exhibits relatively fast responses to short-term variations but also introduces noticeable oscillations in the averaged curve, indicating higher sensitivity to local fluctuations. Autoformer and FEDformer tend to underestimate the overall load level, suggesting limited capability in modeling fine-grained temporal dynamics under long forecasting horizons.

In contrast, the proposed NHiTS-MSFB model produces smoother prediction curves with strong trend consistency. Although its averaged prediction exhibits slightly reduced sensitivity to short-term amplitude variations, it closely follows the overall trajectory of the ground truth and avoids abrupt oscillations. This behavior can be attributed to the design of the Multi-Scale Fusion Block, which enhances cross-scale feature consistency by jointly modeling local temporal patterns and global dependencies while suppressing isolated noise.

It is worth noting that the observed smoothing effect in the averaged visualization does not indicate inferior predictive performance. Instead, it reflects a deliberate bias–variance trade-off introduced by the MSFB module, favoring robustness and stability across different samples. This characteristic is particularly desirable in long-term forecasting tasks such as power load prediction, where stable trend estimation is often more critical than fitting transient fluctuations. The visualization results therefore complement the quantitative improvements reported in Section 4.2 and further demonstrate the effectiveness of the proposed model.

5. Conclusion

This paper addresses the limitations of the NHiTS model in temporal feature representation by proposing an improved architecture based on a Multi-Scale Fusion Block (MSFB). This module models local patterns and global dependencies through parallel multi-scale 1D convolutions and block-sparse attention mechanisms, respectively, and achieves adaptive integration of heterogeneous features through a dynamic gating fusion mechanism. Experimental results demonstrate that the improved model outperforms the baseline model in predictive performance across multiple public datasets.

In future work, we plan to further investigate the performance of the proposed model under longer forecasting horizons and more diverse real-world scenarios. Moreover, extending the proposed approach to more complex and highly non-stationary time series remains an interesting direction for future research. Beyond forecasting, exploring more efficient implementations of the proposed module to enhance scalability also deserves further study.

References

[1] Li Yunfeng, Cai Ziwen, Zhao Yun, et al. Improved time series network for short-term power load forecasting method [J]. Journal of Hunan University (Natural Sciences), 2025, 52(10):205-216.
[2] Li Yong, Wu Hanxin, Li Chanxiao, et al. Short-term load forecasting for distribution network considering multi-dimensional meteorological indicators [J]. Smart Power, 2025, 53(06):116-123.
[3] Zhu Xiaotong, Lin Peiguang, Sun Mei, et al. Multivariate time series forecasting of financial data based on APDFinformer model [J]. Journal of Nanjing University (Natural Science), 2024, 60(06):930-939.
[4] Wang C, Chen Y, Zhang S, et al. Stock market index prediction using deep Transformer model[J]. Expert Systems with Applications, 2022, 208: 118128.
[5] Fan J, Zhang K, Huang Y, et al. Parallel spatio-temporal attention-based TCN for multivariate time series prediction[J].Neural Computing and Applications, 2023, 35(18): 13109-13118.
[6] Challu C, Olivares K G, Oreshkin B N, et al. Nhits: Neural hierarchical interpolation for time series forecasting[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(6): 6989-6997.
[7] Guo C, Fan B, Zhang Q, et al. Augfpn: Improving multi-scale feature learning for object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 12595-12604.
[8] Van Der Donckt J, Van Der Donckt J, Van Hoecke S. tsdownsample: High-performance time series downsampling for scalable visualization[J]. SoftwareX, 2025, 29: 102045.
[9] Taud H, Mas J F. Multilayer perceptron (MLP)[M]//Geomatic approaches for modeling land change scenarios. Cham: Springer International Publishing, 2017: 451-455.
[10] Sun Y, Zhou Q, Sun L, et al. CNN–LSTM–AM:A power prediction model for offshore wind turbines[J]. Ocean Engineering, 2024, 301: 117598.
[11] Wu H, Hu T, Liu Y, et al. Timesnet: Temporal 2d-variation modeling for general time series

ISSN 2616-5775 Vol. 9, Issue 1: 48-55, DOI: 10.25236/AJCIS.2026.090106

analysis[J]. arXiv preprint arXiv:2210.02186, 2022.
[12] Wang J, Zhang L, Li X, et al. ULSeq-TA: Ultra-long sequence attention fusion transformer accelerator supporting grouped sparse softmax and dual-path sparse Layer Norm[J].IEEE Transactionson Computer Aided Design of Integrated Circuits and Systems, 2023, 43(3): 892-905.
[13] Wu H, Xu J, Wang J, et al. Autoformer: Decomposition transformers with auto-correlation for longterm series forecasting[J].Advances in neural information processing systems, 2021, 34: 22419-22430.
[14] Zhou T, Ma Z, Wen Q, et al. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting[C]//International conference on machine learning. PMLR, 2022: 27268-27286.
[15] Yang M, Wang D, Zhang W. A short-term wind power prediction method based on dynamic and static feature fusion mining[J]. Energy, 2023, 280: 128226.
[16] Kocak A, Erdem E, Erdem A. A gated fusion network for dynamic saliency prediction[J]. IEEE Transactions on Cognitive and Developmental Systems, 2021, 14(3): 995-1008.
[17] Wu H, Hu T, Liu Y, et al. Time-Series-Library: A comprehensive benchmark for deep time series forecasting[J/OL]. GitHub Repository, 2022.
[18] Covert I C, Qiu W, Lu M, et al. Learning to maximize mutual information for dynamic feature selection[C]//International Conference on Machine Learning. PMLR, 2023: 6424-6447.
[19] QianH,Wen Q,Sun L,et al.Robustscaler:Qos-aware autoscaling for complex workloads[C]//2022 IEEE 38th International Conference on Data Engineering (ICDE). IEEE, 2022: 2762-2775.
[20] Arefi M, Khammar A H. Nonlinear prediction of fuzzy regression model based on quantile loss function[J]. Soft Computing-A Fusion of Foundations, Methodologies and Applications, 2024, 28(6).