IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 December 2024, accepted 28 December 2024, date of publication 30 December 2024, date of current version 15 January 2025.

Digital Object Identifier 10.1109/ACCESS.2024.3524499

RESEARCH ARTICLE

Local-Global and Multi-Scale (LG-MS) Mixer Architecture for Long-Term Time Series Forecasting

ZHENNAN PENG, BOYONG GAO, ZIQI XIA, AND JIE LIU
College of Information Engineering, China Jiliang University, Hangzhou 310018, China

Corresponding author: Boyong Gao (gaoby@cjlu.edu.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2021YFF0600100.

ABSTRACT Although deep learning models dominate time series forecasting, they still struggle with long-sequence processing due to the challenges of extracting dynamic fluctuations and pattern features as input length increases. To address this challenge, we propose a framework – LG-MSMixer—to enhance long-term time series forecasting through three key steps: multi-scale dual decomposition, local-global information extraction, and fusion prediction. Specifically, we first conduct multi-scale dual decomposition of the long input sequence to derive a seasonal-trend component combination. To capture a more comprehensive effective information within the components, we then utilize a customized patch-based triple attention local-global information extractor that models both temporal feature information and variable dependencies, alongside an MLP-based feature interaction iterator facilitating interactions among multi-scale information to guide macro-level predictions. Finally, we integrate the predictions from the multi-scale sequences to leverage their complementary advantages. In our experiments, we demonstrate the effectiveness of LG-MSMixer across various real-world long-term forecasting tasks, significantly outperforming previous baselines.

INDEX TERMS Deep learning, long-term time series forecasting, information extraction, local-global, multi-scale decomposition.

I. INTRODUCTION

Long-term time series forecasting, a technique that relies on historical observation data to predict several steps into the future [1], has demonstrated extensive practicality in crucial fields such as energy applications [2], transportation planning [3], climate modeling [4], and financial decision-making [5]. Significant progress has been made in time series forecasting through the application of deep learning models such as Convolutional Neural Networks (CNNs) [6], [7], Recurrent Neural Networks (RNNs) [8], [9], Transformers [10], [11], [12], and Multi-layer Perceptron s (MLPs) [13], [14]. Among them, transformers have attracted research interest in this field due to their unique self-attention mechanism, which can deeply extract the feature information of time series data [15]. In addition, many linear models have provided researchers with new perspectives on time series analysis [16].

However, one fatal weakness, that has hindered models based on the above studies from being applied in broader application scenarios, is the lack of the model's ability to effectively handle the complex dynamic fluctuations present in real-world time series.

To enhance the capability of deep models in processing long sequences, prior research has primarily concentrated on two approaches: multi-scale analysis [17] and local-global modeling [6]. The objective of multi-scale analysis is to leverage unique fluctuations across various temporal scales to enrich feature representations and boost predictive performance. However, the most effective scale partitioning is often task-specific, which complicates the optimization process. Furthermore, some studies indicate that multi-scale analysis may detrimentally affect the prediction fusion stage due to suboptimal feature extraction [12].

The associate editor coordinating the review of this manuscript and approving it for publication was Olarik Surinta.

Another approach for long time series processing is local-global modeling, which enhances predictive accuracy by integrating local features with global contextual information while capturing dependencies across diverse temporal scales [8]. Although this strategy has demonstrated substantial effectiveness, a critical challenge lies in addressing its high reliance on finely-tuned adaptable architectures.

Additionally, sequence decomposition methods represent an effective strategy for time series analysis. Deep learning models that integrate decomposition techniques can partition complex temporal patterns into distinguishable components [18]. By capturing distinct information from each component, these models enhance predictive accuracy. Researchers have also employed sophisticated decomposition methods to separate mixed temporal variations into independent components with different periodicities, thereby extracting richer feature representations [1], [19].

The aforementioned theoretical methods, each as a conceptual framework rather than a fixed method implementation, have been widely applied in the field of time series forecasting. However, to the best of our knowledge, no work has yet achieved a comprehensive integration within a single model framework, resulting in incomplete feature extraction for time series.

To comprehensively extract and leverage patterns and feature information from time series, addressing performance limitations in long-term forecasting, we propose an innovative local-global multi-scale pattern feature extraction architecture – LG-MSMixer. This model decomposes the sampled multi-scale observation sequences into seasonal and trend components, thereby enriching the feature space. Each component employs tailored information extractors, followed by fusion prediction utilizing feature representations across multiple temporal scales.

For seasonal pattern extraction, we implement a patch-based triple attention local-global information extractor, effectively modeling temporal features and variable interdependencies. For trend pattern extraction, we introduce a multi-layer perceptron (MLP)-based iterative interaction structure, promoting interaction among multi-scale representations to guide macro-level predictions.

In the forecasting phase, we integrate prediction layers based on these multi-scale observations, maximizing their complementary predictive strengths. Experimental results indicate that LG-MSMixer consistently outperforms existing benchmarks in long-term time series forecasting tasks, demonstrating exceptional performance across diverse benchmark tests. The contributions of this paper are as follows:

- Building on previous work, we propose an innovative framework for time series forecasting that aims to achieve accurate predictions through a profound analysis of the intrinsic features of time series data. The framework effectively integrates the advantages of various forecasting strategies, including multi-scale analysis, seasonal and trend pattern recognition, patch division, and local-global modeling theory, through an implementation method that differs from existing works, rather than just one or two. This is rare throughout the entire time series forecasting field.
- We propose LG-MSMixer as an effective model for long-term time series forecasting, aiming to enhance forecasting performance. To the best of our knowledge, our work represents the most comprehensive extraction and utilization of various temporal information and pattern features in time series. In our approach, dual decomposition at multiple scales enables the model to capture both seasonal and trend patterns. The local-global mixer, based on patch division and a triple-attention mechanism, captures short-term local dynamics and long-term global trends while simultaneously modeling temporal dependencies and cross-variable relationships. The feature iteration interactors based on MLP adjust the macro long-term trends, which enhances forecasting accuracy. The final fusion prediction structure integrates rich information across multiple scales, further improving the prediction performance.
- Our method has been extensively evaluated on various real-world datasets, consistently achieving state-of-the-art performance in long-term time series forecasting tasks. It also demonstrates exceptional efficiency across a wide range of benchmark assessments.

II. RELATED WORK

A. LONG-TERM TIME SERIES FORECASTING

The prediction of long-term time series, a crucial aspect in time series analysis, has garnered extensive research attention in recent years.

Diverging from traditional TCN methodologies [7], modern CNN approaches, characterized by their unique designs, continue to hold prominence in the field of time series forecasting. The MICN model conducts decomposition processing on input sequences, followed by the integration of local-global information for modeling and prediction [6]. Meanwhile, ModernTCN employs a pure convolution structure with large convolution kernels to augment receptive fields, consequently significantly enhancing the performance of TCN models [20].

Benefiting from its robust modeling capability for long sequences, Transformer models have recently demonstrated exceptional performance in long-term time series forecasting [21]. Informer enhances the attention mechanism by selectively choosing key components to improve computational efficiency and predictive performance [22]. Autoformer proposes an auto-regressive mechanism to replace traditional self-attention, providing a more effective approach for modeling temporal dynamics [23]. FEDformer effectively captures long-term patterns in time series using frequency

domain enhancement techniques [17]. PatchTST introduces patch segmentation and channel-independent methods, paving the way for new directions in Transformer models [21]. Some generative pre-trained models have also demonstrated exceptional capabilities in this field [24].

Notably, some simple models have also performed well in this field [25]. In particular, Multi-layer Perceptrons (MLPs) have gained widespread recognition for their straightforward structure and superior performance [26]. LightTS converts time series data into a two-dimensional format using lightweight sampling techniques, thereby employing MLP architectures to extract temporal features crucial for modeling purposes [13]. TiDe implements a dense MLP encoding-decoding structure, ensuring optimal performance while substantially mitigating computational complexities [27].

Furthermore, certain Graph Neural Networks (GNNs) methodologies, particularly those modeling spatial dependencies, have also played a pivotal role in advancing the frontier of long-term time series forecasting [28], [29].

B. INFORMATION EXTRACTION OF TIME SERIES

In tasks necessitating high precision in time series forecasting, the effective extraction and utilization of feature information have been proven to be paramount [30].

Autoformer integrates a time series decomposition module, dissecting intricate temporal patterns into components abundant in seasonal and trend information, subsequently subjecting these patterned data to modeling [23]. TimesNet transmutes one-dimensional temporal inputs into two-dimensional spatial constituents via multi-period analysis, thereby capturing multi-period features through convolution [1]. Scaleformer proposed a highly generalized multi-scale time series forecasting framework, enhancing predictive performance by integrating multi-scale temporal feature information [12]. Pyraformer introduced pyramid attention to extract sequence information at different time resolutions [11]. FED-former utilized frequency-domain enhancement methods, employing Fourier transformation to remove redundancies in time feature information, for refining the extraction of long-term regularity features in time series data [17]. MICN extracted temporal dynamic information from the perspective of local-global comprehensive modeling. Crossformer employed dual-stage attention, modeling capturing both temporal correlations and inter-variable dependencies [31].

These methods achieve good predictive performance, yet they do not comprehensively extract and utilize feature information beneficial for prediction. Our proposed model, based on multi-scale decomposition and local-global modeling, aims to address these limitations by offering a more comprehensive approach to information extraction.

III. METHODOLOGY

The precision of time series prediction crucially hinges upon the effective extraction and utilization of intricate patterns and feature information inherent within the data. Herein, we introduce LG-MSMixer model, a novel framework embedding a multi-scale local-global pattern feature extraction architecture. This model adeptly extracts and amalgamates intricate mixed temporal features, harnessing the synergistic predictive potentials of multi scale sequences to achieve superior forecasting performance. In Table 1, we provide a summary of the abbreviations and terms involved in this study.

TABLE 1. The summary of abbreviations and terms involved in this study.

X: Raw Input Sequence	R*: Data Dimension
H: Length of Raw Input Sequence	C: Number of Variables (Channels)
M: Number of Scales	L: Number of HL-GIE Blocks
dm: Expansion Dimension	P: Number of Patches
W: Patch Size	Z: Number of Routing Information
x_m: Input of the m-th Scale	X^l: Input of the l-th HL-GIE Block
S: Seasonal Component	T: Trend Component
Mid: Routing Information Receiver	Pred: Model Final Prediction
s_m^l / t_m^l : Seasonal / Trend Component in the l-th HL-GIE Block (m-th Scale)	
HL-GIE: Local-Global Information Extraction	
MSFP: Multi-Scale Fusion Prediction	
TMG: Trend Mixing Guidance	
L-GE: Local-Global Information Extraction	

As depicted in Figure 1, LG-MSMixer predominantly comprises multiple down sampling stages, stacked Historical Local-Global Information Extraction (HL-GIE) blocks, and Multi-Scale Fusion Prediction (MSFP) blocks. As noted earlier, time series data at varying scales manifest diverse attributes, where finer scales predominantly capture local intricacies, while coarser scales emphasize macroscopic trends [5]. Thus, harnessing the distinct predictive capacities of multi-scale time series often proves more efficient in modeling complex temporal dynamics [32].

Initially, we conduct average pooling on the observation sequence $X \in R^{\boldsymbol{H}\times\boldsymbol{C}}$ to generate M scales, resulting in a set of multi-scale time series, $X = \{x_1, \ldots, x_M\}$, $x_m \in R^{\frac{H}{2^m}\times\boldsymbol{C}}$ for $m \in \{1, \ldots, M\}$. Here, H represents the original input sequence scale, and C signifies the number of variables. Subsequently, we employ a stacked architecture of HL-GIE blocks to extract feature information from each scale of the observation sequence. For the l-th layer, with X^{l-1} as the input, the operation of HL-GIE can be formalized as follows:

$$\mathrm{X}^l = \mathrm{HL-GIE}(\mathrm{X}^{l-1}), \quad l \in \{1, \ldots, L\}. \tag{1}$$

where L represents the total number of layers, $\mathrm{X}^l = \{x_1^l, \ldots, x_M^l\}$, $x_m^l \in R^{\frac{H}{2^m}\times C}$ denotes the output representation with C channels. For the prediction phase, we employ the MSFP block to merge the extracted multi-scale temporal feature information X^L and then producing our prediction results, formalized as:

$$\mathrm{Pred} = \mathrm{MSFP}(\mathrm{X}^{\mathrm{L}}). \tag{2}$$

Here, $\mathrm{Pred} \in \mathrm{R}^{\mathrm{F}\times\mathrm{C}}$ represents the final prediction results, where F denotes the prediction length of the sequence.

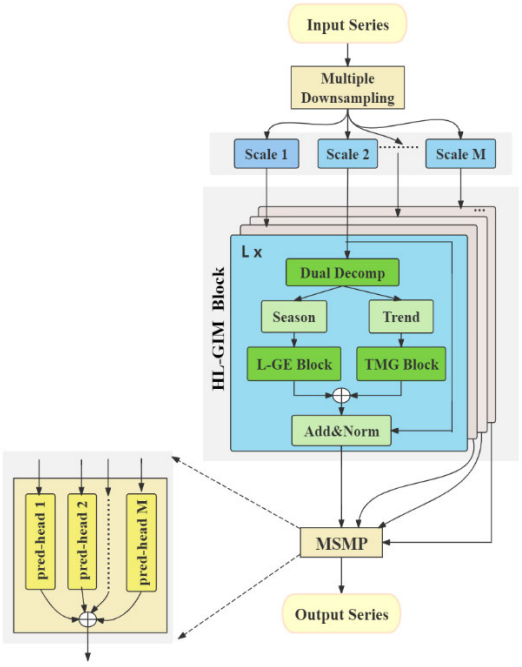

FIGURE 1. Overall architecture of LG-MSMixer, which consists of HL-GIE and MSFP.

As mentioned above, the design of LG MSMixer was implemented. The specific structure and details are detailed in the subsections below.

A. HL-GIE BLOCK

Existing research has demonstrated that time series in real-world settings commonly display intricate mixed-mode characteristics irrespective of their scale. Notably, seasonal and trend components exhibit unique attributes in time series forecasting tasks [33], corresponding to stationary and non-stationary dynamics, respectively.

Thus, we introduce stacked HL-GIE blocks to segregate the decomposed seasonal and trend components for tailored information extraction based on their inherent characteristics, followed by subsequent integration. Specifically, for the l-th HL-GIE block, we decompose the multi-scale time series X^L into seasonal components $S = \{s_1, \ldots, s_M\}, s_m \in R^{\frac{H}{2^m} \times C}$ and trend components $T = \{t_1, \ldots, t_M\}$, using the DualDecomp block.

We then perform dimensional expansion and deep projection on the seasonal components S^l to make $S_m^l \in R^{\frac{H}{2^m} \times C \times dm}$, followed by inputting S^l into the Local-Global Extraction (L-GE) block. The trend components are embedded and projected into deep features $T^l = \{t_1^l, \ldots, t_M^l\}, t_m^l \in R^{\frac{H}{2^m} \times dm}$, which serve as the input to the Trend Mixing Guidance (TMG) block. The final output of the HL-GIE block is the sum of these two component outputs. In summary, the l-th HL-GIE block can be formalized as:

$$s_m^l, t_m^l = \text{DualDecomp}(x_m^l), \quad m \in \{1, \ldots, M\},$$
$$S_{out} = \text{L-GE}(\left\{s_m^l\right\}_{m=1}^{M}),$$
$$T_{out} = \text{TMG}(\left\{t_m^l\right\}_{m=1}^{M}),$$
$$\mathrm{X}^l = \mathrm{X}^{l-1} + Proj_S(S_{out}) + Proj_T(T_{out}). \quad (3)$$

Here, $Proj_S(\cdot)$ and $Proj_T(\cdot)$ denote linear mapping functions for the seasonal and trend components, respectively, with output dimensions of $R^{\frac{H}{2^m} \times C}$.

We will provide a detailed description of the aforementioned structural modules. To simplify notation, we will omit the layer index in the relevant formulas, with 'l' assumed as the default layer.

1) DUAL-DECOMPOSE BLOCK

To more effectively extract and utilize the intricate features of time series, we employ a dual decomposition block to decompose them into seasonal and trend components and process them separately, rather than directly modeling them.

Regarding the seasonal decomposition, we utilize the Discrete Fourier Transform (DFT), denoted as DFT $(\cdot)$ [34], to convert the time series from the time domain to the frequency domain, thereby extracting detailed periodic characteristics. The input X is decomposed into Fourier bases, and we selectively retain the top K bases with the highest amplitudes to preserve sparsity in the frequency domain. Subsequently, our seasonal component S is obtained through inverse DFT, denoted as IDFT $(\cdot)$. This process can be formalized as follows:

$$S = \text{IDFT}(f_1, \ldots, f_k, A, \phi). \quad (4)$$

Here, ϕ and A represent the phase and amplitude, respectively, while $(f_1, \ldots, f_k)$ denotes the frequencies corresponding to the selected K largest amplitudes.

The trend decomposition involves applying moving average pooling with various kernel sizes to the input X to extract trend pattern information. The trend component is obtained through weighted aggregation of the results obtained from different pooling kernels:

$$T = \text{SoftMax}(\text{L}(\text{X}))\,(Avgpool\,(X)_{kernel_1}, \ldots, Avgpool\,(X)_{kernel_N}). \quad (5)$$

in this formulation, $Avgpool(X)_{kernel_i}$ denotes the function associated with the i-th pooling kernel, with a total of N pooling kernels. The function SoftMax(L$(\cdot)$) computes the weights corresponding to the outcomes produced by different pooling kernels.

2) L-GE BLOCK

The seasonal component of time series data harbors significant periodic information critical for accurate forecasting. The extraction and utilization of these insights often dictate the success of forecasting endeavors. To tackle this challenge effectively, we introduce the L-GE Block—Local-Global Information Extraction block. This block conducts comprehensive extraction and processing of historically significant

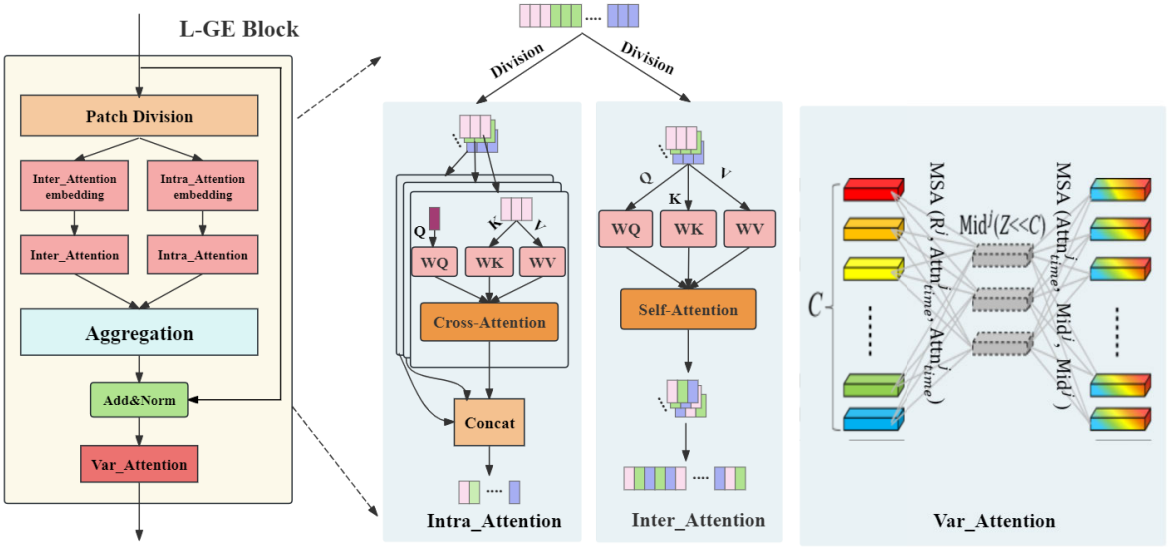

FIGURE 2. The architecture of the L-GE block, illustrated in the structural diagram, employs three distinct attention mechanisms to extract feature information. Intra-Attention focuses on extracting local details within individual patches, while Inter-Attention captures global dependencies between patches, collectively capturing temporal dependencies. Var-Attention models the interactions between variables.

information, incorporating both local details and global perspectives, to cater to our prediction tasks.

Structurally, as depicted in Figure 2, the L-GE Block primarily integrates three distinct attention mechanisms for information extraction from time series data: Intra-Attention layer, modeling dependencies between time steps within each patch; Inter-Attention layer, capturing dependencies between patches; and Var-Attention layer, facilitating the interaction of information among variables. Implementation-wise, our approach is inspired by patch partitioning, a widely adopted practice in contemporary studies. For brevity, we illustrate this process using univariate time series data in this section, with the understanding that it readily extends to multivariate scenarios. For the seasonal component sequence $S = \{s_1, \ldots, s_M\}$, $s_m \in R^{\frac{H}{2^m} \times dm}$, m $\in \{1, \ldots, M\}$, at any given scale, the sequence is divided into P patches of size W ($P = \frac{H}{2^m \times W}$). Thus, $s_m \in R^{P \times W \times dm}$. The processed S is then concurrently fed through Intra-Attention and Inter-Attention. The outputs from tow attentions are then captured and fused by Var-Attention to yield the final output of the L-GE block. The process can be formalized as:

$$\begin{aligned} Attn_{intra} &= \text{Intra-Attention}(\{s_m^l\}_{m=1}^{M}), \\ Attn_{inter} &= \text{Inter-Attention}(\{s_m^l\}_{m=1}^{M}), \\ S_{out} &= \text{Var-Attention}(Attn_{intra} + Attn_{inter}). \end{aligned} \tag{6}$$

The final output $S_{out} \in R^{\frac{H}{2^m} \times dm}$.

The specific implementation details of the multi-attention mechanisms will be presented in subsequent sections. We will use $s_m \in R^{P \times W \times dm}$, m $\in \{1, \ldots, M\}$, from the patch-divided S, as the focus for this introduction.

3) INTRA-ATTENTION LAYER

The Intra-Attention layer captures the temporal dependencies within each patch. Specifically, for any scale of $s_m \in R^{P \times W \times dm}$, we focus on the i-th patch $Y_{intra}^i \in R^{W \times dm}$,. Initially, we subject it to trainable linear transformations to derive the keys and values for attention computation, denoted as K_{intra}^i, $V_{intra}^i \in R^{P \times W \times dm}$, respectively. Additionally, we employ a trainable query matrix $Q_{intra}^i \in R^{1 \times dm}$ to integrate contextual information. Subsequently, the cross-attention between Q_{intra}^i, K_{intra}^i and V_{intra}^i, is computed to model the local details within the i-th patch ($i \in \{1, \ldots, P\}$):

$$\begin{aligned} Attn_{intra}^i &= \text{SoftMax}(Q_{intra}^i (K_{intra}^i)^T \Big/ \sqrt{dm}) V_{intra}^i, \\ Attn_{intra} &= \text{Concat}(Attn_{intra}^1, \ldots, Attn_{intra}^P). \end{aligned} \tag{7}$$

Here, $Attn_{intra} \in R^{1 \times dm}$ denotes the output after undergoing the Intra-Attention process for the i-th patch. The concatenation of outputs from all patches results in the final output $Attn_{intra} \in R^{P \times dm}$ for this block, characterizing the local details between neighboring time steps in the time series.

Furthermore, to facilitate integration with the output of Inter-Attention layer, we perform a linear transformation on the patch length dimension of this output, ranging from 1 to W, thereby reshaping it into $Attn_{intra} \in R^{P \times W \times dm}$.

4) INTER-ATTENTION LAYER

The Inter-Attention layer establishes relationships between patches to capture global correlations. For $s_m \in R^{P \times W \times dm}$, to aggregate all time steps within the same patch, we combine the dimensions of patch length (W) and feature embedding (W), resulting in $s_{inter}^m \in R^{P \times D}$, $D = W \times dm$. Subsequently, we apply a learnable linear mapping to s_{inter}^m to obtain Q_{inter}, K_{inter}, $V_{inter} \in R^{P \times D}$, and employ a self-attention mechanism to model dependencies between patches, thus representing

the global correlations in the time series:

$$Attn_{inter} = SoftMax(Q_{inter}(K_{inter})^T \Big/ \sqrt{dm})V_{inter}. \quad (8)$$

Here, $Attn_{inter} \in R^{P\times D}$ and it will become $Attn_{inter} \in R^{P\times W\times dm}$ after flattening the dimensions.

We perform an element-wise addition between $Attn_{intra}$ and $Attn_{inter}$, resulting in a variable $Attn_{time} \in R^{\frac{H}{2^m}\times dm}$, thereby effectively amalgamating the local intricacies and global correlations of the time series. It can be formalized as:

$$Attn_{time} = Attn_{intra} + Attn_{inter}. \quad (9)$$

5) VAR-ATTENTION LAYER

In multivariate time series forecasting, the interplay between variables is crucial for accurate predictions [3], [20]. Consequently, we extend $Attn_{time}$ to accommodate a multivariate context, represented as $Attn_{time} \in R^{\frac{H}{2^m}\times C\times dm}$, and model the inter-variable dependencies by Var-Attention layer. To manage computational and memory demands in datasets with extensive variable dimensions, we implement the routing attention mechanism based on multi-head self-attention (MSA) as proposed in Crossformer [31], which can effectively mitigate the complexity of computation and memory usage. In Figure 2, we allocate a predefined set of learnable vectors ($Z \ll C$) as routers for each time step. Initially, these routers function as queries, while vectors across all variable dimensions (C) are designated as keys and values. The Multi-Head Self-Attention (MSA) mechanism aggregates information from all variable dimensions. Subsequently, the routers use variable vectors as queries, with the aggregated information serving as keys and values. The MSA then redistributes the integrated information across the C dimensions. The operation is conducted sequentially for each time step within $Attn_{time}$, with parameters being consistent across all time steps ($1 \le j \le \frac{H}{2^m}$). The formalization for the j-th time step is as follows:

$$\begin{aligned} Mid^j &= MSA_1(R^j, Attn^j_{time}, Attn^j_{time}), \quad 1 \le j \le \frac{H}{2^m}, \\ Attn^j_{var} &= MSA_2(Attn^j_{time}, R^j, R^j), \quad 1 \le j \le \frac{H}{2^m}, \\ \mathrm{S_{out}} &= \mathrm{Concat}\,(\mathrm{Attn}^1_{\mathrm{var}}, \ldots, \mathrm{Attn}^{\frac{\mathrm{H}}{2^{\mathrm{m}}}}_{\mathrm{var}}). \end{aligned} \quad (10)$$

Within this framework, $R^j \in R^{C\times dm}$ denotes the learnable vector, or router, specific to the j-th time step, which serves to mediate the attention mechanism. Here, $Mid^j \in R^{C\times dm}$ encapsulates the consolidated information derived from all variables. The output is denoted as a variable $S_{out} \in R^{\frac{H}{2^m}\times C\times dm}$. For integration with the processed trend component, we concatenate the variable dimension C and the feature dimension dm, culminating in the final output of the HL-GIE block: $S_{out} = \{\mathbf{s_1}, \ldots, s_M\}$, where $\mathbf{s}_m \in R^{\frac{H}{2^m}\times dm}$.

B. TMG BLOCK

When dealing with the trend component, capturing overly granular variations may introduce noise into the modeling of overarching trends. Therefore, we leverage macro-level information from coarser time scales to guide the trend modeling of finer time series. Coarser scales are typically considered to possess more comprehensive and clearer macro-level insights.

In practice, as depicted in Figure 3, we process the multi-scale trend components $\mathrm{T} = \{t_1, \ldots, t_M\}$, where $t_m \in R^{\frac{H}{2^m}\times dm}$ for $\mathrm{m} \in \{1, \ldots, \mathrm{M}\}$, using the Trend Mixing Guidance (TMG) block in a residual framework. This process sequentially facilitates information exchange from coarser to finer scales across the multi-scale series, producing an output series guided by macro-level trend information:

$$\textbf{from}: \mathbf{M} \rightarrow \mathbf{1}\ \textbf{do}: t_{m-1} = t_{m-1} + \mathrm{TMG}(\mathrm{t_m}). \quad (11)$$

In this formulation, TMG$(\cdot)$ represents a dual-layer MLP incorporating a GELU activation function. The input dimension is $\frac{H}{2^m}$ and output dimension is $\frac{H}{2^{m-1}}$. Following the completion of the outlined process, the TMG block yields the final output $T_{out} = \{t_1, \ldots, t_M\}$, with $t_m \in R^{\frac{H}{2^m}\times dm}$.

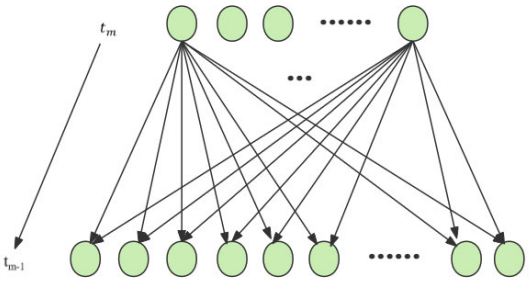

FIGURE 3. Function of the information interaction linear layer in the TMG block.

Overall, the HL-GIE block excels in capturing and integrating the local and global pattern information of the seasonal components in time series data by leveraging various attention mechanisms within the L-GE block. Concurrently, it facilitates cross-scale information exchange for trend components through the TMG block, resulting in trend sets that provide global guidance.

By merging these components, the model produces fused information that exhibits pronounced temporal characteristics and significantly enhances prediction accuracy.

C. MSFP BLOCK

After processing through L HL-GIE blocks, we obtain the output $\mathrm{X}^L = \{x^L_1, \ldots, x^L_M\}$, where $x^L_m \in R^{\frac{H}{2^m}\times C}$. To fully leverage multi-scale information, we integrate predictions from different scale sequences through the MSFP block, resulting in the final prediction. This process can be represented as:

$$Pred_m = Predictor_m(x^L_m), \quad \mathrm{Pred} = \sum_{m=1}^{M} Pred_m. \quad (12)$$

Here, $Pred_m \in R^{F\times C}$ represents the prediction for the m-th scale sequence, and the final output is $\mathrm{Pred} \in R^{F\times C}$.

TABLE 2. Summary of relevant information for datasets used in Our experiments.

Dataset	Count	Variate	Frequency	Predict Length	Domain
ETTh 1	17420	7	Hourly	96~720	Electricity
ETTh 2	17420	7	Hourly	96~720	Electricity
ETTm 1	69680	7	15mins	96~720	Electricity
ETT m2	69680	7	15mins	96~720	Electricity
WTH	52696	21	10mins	96~720	Weather
ELC	26304	321	Hourly	96~720	Electricity
Traffic	17544	862	Hourly	96~720	Transportation
Exchange	7588	8	Daily	96~720	Exchange rate
ILI	966	7	Weekly	24~60	Incidence rate

$Predictor_m(\cdot)$ denotes the predictor for the m-th scale sequence, composed of a two-layer MLP with $\frac{H}{2^m}$ hidden units and a GELU activation function. The MSFP block is a collection of these predictors, where each predictor's output corresponds to the temporal feature information extracted from observed sequences at various scales, enabling MSFP to integrate the complementary predictive abilities of multi-scale sequences for enhanced forecasting performance.

IV. EXPERIMENTS

A. EXPERIMENTS SETTING

1) DATASETS

We conducted comprehensive experiments on nine representative real-world datasets to evaluate the performance of the LG-MSMixer. The datasets encompass various domains such as power transmission, weather forecasting, and traffic management [20], [35], [36]. These datasets include ETT (ETTh1, ETTh2, ETTm1, ETTm2), WTH (Weather), ELC (Electricity), Traffic, ILI, and Exchange. We summarize the information, characteristics, and differences of these datasets in Table 2 to offer a clearer and more intuitive understanding.

2) BASELINES AND METRICS

We selected several state-of-the-art models in the field of long-term time series forecasting as baselines, including PatchTST [21], NLinear [16], TiDe [27], Scaleformer [12], Pyraformer [11], FEDformer [17], and Autoformer [23]. To ensure fairness in our experiments, we fixed the input length of all experiments to 96 (for the ILI dataset, the input length was set to 36). We used two commonly employed metrics in deep learning-based time series forecasting tasks to quantify the experimental results: Mean Absolute Error (MAE) and Mean Squared Error (MSE). These metrics are intuitive and computationally efficient, making them particularly suitable for large-scale deep regression tasks aimed at minimizing prediction errors. Additionally, we urge readers to recognize the importance of using information criteria, such as AIC and SBIC, to balance model complexity and performance.

3) IMPLEMENTATION DETAILS

We employed the pytorch framework for our experiments [37], which were conducted on a server with a single NVIDIA RTX 4090 24GB GPU. The learning rate was set to 10^{-3}, and the L1 Loss function was utilized. Moreover, the patch length W was judiciously adjusted based on the varying scales of the time series to achieve a balance between performance and efficiency.

B. RESULTS

Table 3 illustrates the comparative results of LG-MSMixer against other baseline models across all long-term time series forecasting benchmarks. Among the 72 experimental cases, LG-MSMixer achieved the top performance in 53 cases and ranked second in 12 cases. Compared to the second-best baseline, PatchTST, LG-MSMixer demonstrated significant advantages, reducing MSE by 15% for ELC and by 15.7% for ILI. The results fully demonstrate that our work is highly effective for long-term time series prediction, attributable to the reasonable extraction and utilization of various patterns and feature information. When compared to sequence decomposition-based models like FEDformer and Autoformer, LG-MSMixer exhibited marked performance enhancements, lowering MSE by 27.9% and MAE by 17.4%, underscoring the criticality of comprehensively utilizing decomposed pattern information. In addition, compared to multi-scale models such as Scaleformer and Pyraformer, LG-MSMixer reduced MSE by 31% and MAE by 20.7%, indicating the efficacy of time feature extraction based on local-global modeling theories and multi-scale fusion forecasting methods. Moreover, LG-MSMixer also showed favorable performance compared to the current robust linear models NLinear and TiDE, which underscores the necessity of capturing variable dependency information through an attention mechanism-based approach.

C. MODEL ANALYSIS

1) ABLATION STUDIES

To assess the contributions of different modules within the LG-MSMixer, we conducted ablation experiments on the dual decomposition, the L-GE block, and the TMG block.

TABLE 3. Summary of results for long-term time series forecasting. We standardized the input length for all experiments to 96 (with the input length for the ILI dataset set to 36). Lower values of MSE or MAE indicate superior model performance. The best results are highlighted in bold, and the second-best results are underlined.

Method		LG-MSMixer		PatchTST		NLinear		TiDE		Scaleformer		Pyraformer		FEDformer		Autoformer	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.385	**0.392**	0.394	0.408	0.386	0.392	0.427	0.450	0.396	0.440	0.664	0.612	**0.376**	0.419	0.449	0.459
	192	0.434	**0.424**	0.446	0.438	0.440	0.430	0.472	0.486	0.434	0.460	0.790	0.681	**0.420**	0.448	0.500	0.482
	336	0.467	**0.441**	0.485	0.455	0.480	0.443	0.527	0.527	0.462	0.476	0.891	0.738	**0.459**	0.465	0.521	0.496
	720	**0.480**	**0.461**	0.495	0.474	0.486	0.472	0.644	0.605	0.494	0.500	0.963	0.782	0.506	0.507	0.514	0.512
ETTh2	96	**0.287**	**0.336**	0.294	0.343	0.290	0.339	0.304	0.359	0.364	0.407	0.645	0.597	0.346	0.388	0.358	0.397
	192	**0.356**	**0.380**	0.378	0.394	0.379	0.395	0.394	0.422	0.466	0.458	0.788	0.683	0.429	0.439	0.456	0.452
	336	0.396	0.415	**0.382**	**0.410**	0.421	0.431	0.385	0.421	0.479	0.476	0.907	0.747	0.496	0.487	0.482	0.486
	720	0.416	**0.432**	**0.412**	0.433	0.436	0.453	0.463	0.475	0.487	0.492	0.963	0.783	0.463	0.474	0.515	0.511
ETTm1	96	**0.316**	**0.346**	0.324	0.361	0.339	0.369	0.356	0.381	0.355	0.398	0.543	0.510	0.379	0.419	0.505	0.457
	192	**0.361**	**0.373**	0.362	0.383	0.379	0.386	0.391	0.399	0.428	0.455	0.557	0.537	0.436	0.441	0.553	0.496
	336	0.398	**0.392**	**0.390**	0.402	0.411	0.407	0.424	0.423	0.524	0.487	0.754	0.655	0.445	0.459	0.621	0.537
	720	0.469	**0.436**	**0.461**	0.438	0.478	0.442	0.480	0.456	0.558	0.517	0.908	0.724	0.543	0.490	0.671	0.561
ETTm2	96	**0.175**	**0.254**	0.177	0.260	0.177	0.257	0.182	0.264	0.182	0.275	0.435	0.507	0.203	0.287	0.255	0.339
	192	0.248	0.302	0.248	0.306	**0.241**	**0.297**	0.256	0.323	0.251	0.318	0.730	0.673	0.269	0.328	0.281	0.340
	336	**0.302**	**0.335**	0.304	0.342	**0.302**	0.337	0.313	0.354	0.340	0.375	1.201	0.845	0.325	0.366	0.339	0.372
	720	0.407	0.401	**0.403**	0.397	0.405	**0.396**	0.419	0.410	0.435	0.433	3.625	1.451	0.421	0.415	0.433	0.432
WTH	96	**0.150**	**0.190**	0.177	0.218	0.168	0.208	0.202	0.261	0.288	0.365	0.896	0.556	0.238	0.314	0.249	0.329
	192	**0.204**	**0.235**	0.224	0.258	0.217	0.255	0.242	0.298	0.368	0.425	0.622	0.624	0.275	0.329	0.325	0.370
	336	**0.261**	**0.283**	0.277	0.297	0.267	0.292	0.287	0.335	0.447	0.469	0.739	0.753	0.339	0.377	0.351	0.391
	720	**0.338**	**0.336**	0.350	0.345	0.351	0.346	0.351	0.386	0.640	0.574	1.004	0.934	0.389	0.409	0.415	0.426
ECL	96	**0.152**	**0.244**	0.180	0.264	0.185	0.266	0.194	0.277	0.182	0.297	0.386	0.449	0.186	0.302	0.196	0.313
	192	**0.163**	**0.254**	0.188	0.275	0.189	0.276	0.193	0.280	0.188	0.300	0.386	0.443	0.197	0.311	0.211	0.324
	336	**0.178**	**0.271**	0.206	0.291	0.204	0.289	0.206	0.296	0.210	0.324	0.378	0.443	0.213	0.328	0.214	0.327
	720	**0.205**	**0.293**	0.247	0.328	0.245	0.319	0.242	0.328	0.232	0.339	0.376	0.445	0.233	0.344	0.236	0.342
Traffic	96	**0.474**	**0.295**	0.492	0.324	0.645	0.388	0.568	0.352	0.558	0.343	2.085	0.468	0.576	0.359	0.597	0.371
	192	0.488	**0.298**	**0.487**	0.303	0.599	0.365	0.612	0.371	0.564	0.351	0.867	0.467	0.610	0.380	0.607	0.382
	336	**0.489**	0.323	0.505	**0.317**	0.606	0.367	0.605	0.374	0.570	0.349	0.869	0.469	0.608	0.375	0.623	0.387
	720	**0.539**	0.351	0.542	**0.337**	0.645	0.388	0.647	0.410	0.576	0.349	0.881	0.473	0.621	0.375	0.639	0.395
Ex-change	96	**0.081**	**0.200**	0.088	0.205	0.089	0.208	0.094	0.218	0.155	0.285	0.376	1.105	0.148	0.278	0.197	0.323
	192	**0.170**	**0.292**	0.176	0.299	0.180	0.300	0.184	0.307	0.274	0.384	1.748	1.151	0.271	0.315	0.300	0.369
	336	0.337	0.413	**0.301**	**0.397**	0.331	0.415	0.349	0.431	0.452	0.498	1.874	1.172	0.460	0.427	0.509	0.524
	720	0.883	0.710	0.901	0.714	1.033	0.780	**0.852**	0.698	1.172	0.839	1.943	1.206	1.195	**0.695**	0.447	0.941
ILI	24	**1.298**	**0.693**	1.724	0.843	2.725	1.069	2.154	0.992	2.678	1.071	3.420	2.012	2.624	1.095	2.906	1.182
	36	**1.493**	**0.735**	1.536	0.752	2.530	1.032	2.436	1.042	2.745	1.075	7.394	2.031	2.516	1.021	2.585	1.038
	48	**1.539**	**0.758**	1.821	0.832	2.510	1.031	2.532	1.051	2.748	1.072	7.551	2.057	2.505	1.041	3.024	1.145
	60	**1.574**	**0.770**	1.923	0.842	2.492	1.026	2.748	1.142	2.793	1.059	7.622	2.100	2.742	1.122	2.761	1.114

TABLE 4. Ablation studies on the ILI and ETTh2 datasets.

Models		W/O DualDecomp		W/O TMG		W/O L-GE		LG-MSMixer	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	96	0.396	0.401	0.389	0.395	0.397	0.400	**0.385**	**0.392**
	192	0.443	0.429	0.440	0.429	0.446	0.429	**0.436**	**0.424**
	336	0.482	0.453	0.478	0.446	0.483	0.453	**0.469**	**0.441**
	720	0.491	0.463	0.488	0.462	0.498	0.472	**0.480**	**0.461**
ILI	24	1.474	0.703	1.488	0.712	1.506	0.725	**1.298**	**0.693**
	36	1.590	0.772	1.624	0.788	1.711	0.777	**1.493**	**0.735**
	48	1.678	0.794	1.614	0.787	1.757	0.801	**1.539**	**0.758**
	60	1.750	0.822	1.798	0.845	1.795	0.826	**1.574**	**0.770**

In Table 4, to highlight the unique contributions of each module, we use two challenging datasets, ETTm1 and ILI, in the time series forecasting task. These datasets differ in scale and characteristics, yet they are well-suited to the objectives of our work. The experimental results indicate these modules are crucial to our task. Particularly, the L-GE block, which is based on a multi-attention mechanism, demonstrates a notably significant impact. This finding highlights the importance and efficacy of comprehensively modeling both cross-variable dependencies and temporal dependencies in time series forecasting tasks. Additionally, the dual decomposition module's capability to decompose the observed series into seasonal and trend components effectively isolates and extracts complex pattern information, thereby enhancing the capture of the dynamic temporal characteristics of the input. The TMG block processes trend component interactions and plays a pivotal role in macro-level regulation during the subsequent prediction stages.

2) HYPER-PARAMETER SENSITIVITY STUDY

We evaluated the impact of two hyper-parameters on the ETTm1 dataset: the number of scales (the number of sequences of different lengths after down sampling, M) and the number of routers (Z in the MSA of Var-Attention layer). Number of Scales: As illustrated in Figure 4(a), we tested the MSE for different prediction windows with the number of scales ranging from 1 to 5. With increasing M, the performance improvement for shorter prediction windows quickly plateaued; however, for longer prediction windows, the model's performance continued to improve.

Hence, we set M to 2 for short-term predictions and 4 for long-term predictions to balance performance and efficiency. Number of Routers: As shown in Figure 4(b), we varied the number of routers (Z) from 3 to 18, and examined its effect on prediction performance across different prediction window lengths. When the prediction length was 96, model performance remained stable; for lengths of 192 and 336, performance exhibited significant fluctuations; and for a length of 720, the MSE was higher at Z=3, subsequently decreasing and stabilizing. Ultimately, we set Z to 10 to balance performance and efficiency.

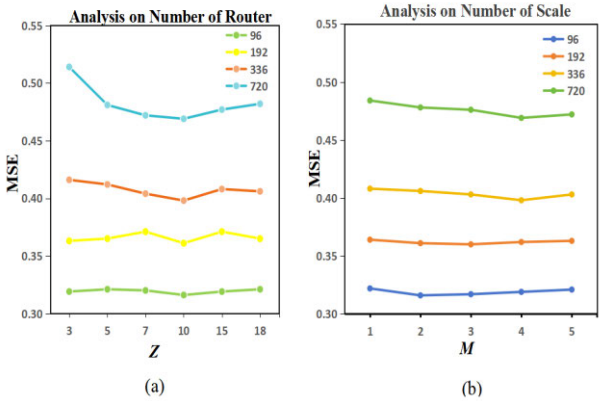

FIGURE 4. Hyperparameter sensitivity analysis. (a) MSE for the hyperparameter router number Z in Var-Attention on the ETTm1 dataset. (b) MSE for the hyperparameter scale number M on the ETTm1 dataset.

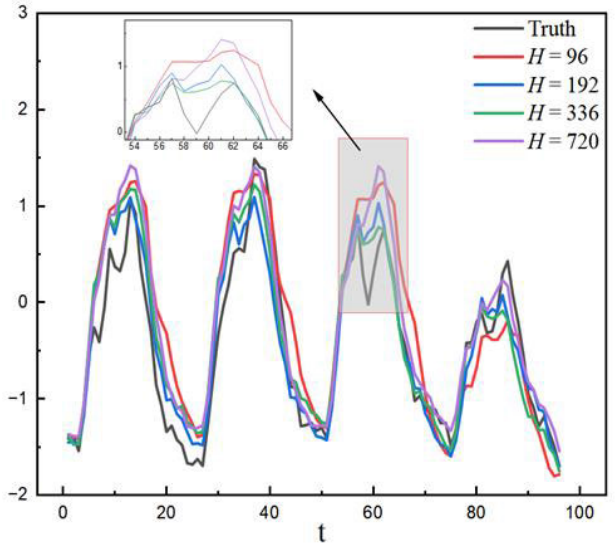

FIGURE 5. The experiment results of the prolonged look-back window on ELC. It illustrates the performance of the model across four different window length.

3) PREDICTIONS VISUALIZATION OF PROLONGED LOOK-BACK WINDOWS

In our long-term time series forecasting study, we systematically examined the effect of extending the look-back window length (H) on forecasting performance using the ELC dataset, while keeping the prediction window length constant at 96. The look-back window was varied from 96 up to 720 to assess its impact.

As illustrated in Figure 5, we observed that when H$\leq$ 336, the forecasting accuracy significantly improved with longer look-back windows. However, a notable decline in performance occurred at H = 720. We believe this could be attributed to the fact that longer lookback windows encompass richer dependency information, thus benefiting prediction. However, overly extended look-back windows introduce redundant pattern information and heightened fluctuations, leading to a decline in performance.

V. CONCLUSION

We introduce LG-MSMixer, an advanced model designed for time series forecasting, featuring a robust multi-scale local-global pattern extraction framework. By integrating multi-scale decomposition with patch-based local-global pattern modeling, the model proficiently captures intricate temporal features. Moreover, it harnesses the complementary predictive strengths of multi-scale sequences, leading to superior performance in long-term time series forecasting.

In 72 forecasting experimental cases, LG-MSMixer achieved the best performance in 53 cases and second-best performance in 12 cases. Compared to the second-best baseline, PatchTST, LG-MSMixer demonstrates a significant advantage: the MSE of ELC is reduced by 15%, and the MSE of ILI is reduced by 15.7%. Additionally, our model performed excellently in ablation studies and hyperparameter analysis, further validating the contributions of our work.

The comprehensive experiments demonstrate that LG-MSMixer consistently achieves state-of-the-art results across a spectrum of long-term forecasting tasks.

REFERENCES

[1] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "TimesNet: Temporal 2D-variation modeling for general time series analysis," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023, pp. 1–23.

[2] Z. Qian, Y. Pei, H. Zareipour, and N. Chen, "A review and discussion of decomposition-based hybrid models for wind energy forecasting applications," *Appl. energy*, vol. 235, pp. 939–953, Feb. 2019.

[3] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, and B. Yin, "Deep learning on traffic prediction: Methods, analysis, and future directions," *IEEE Trans. Intell. Transport. Syst.*, vol. 23, no. 6, pp. 4927–4943, Jun. 2022.

[4] R. A. Angryk, P. C. Martens, B. Aydin, D. Kempton, S. S. Mahajan, S. Basodi, A. Ahmadzadeh, X. Cai, S. F. Boubrahimi, S. M. Hamdi, M. A. Schuh, and M. K. Georgoulis, "Multivariate time series dataset for space weather data analytics," *Sci. Data*, vol. 7, p. 227, Jul. 2020.

[5] M. A. R. Ferreira, D. Higdon, H. K. H. Lee, and M. West, "Multi-scale and hidden resolution time series models," *Bayesian Anal.*, vol. 1, no. 4, pp. 947–967, Dec. 2006.

[6] H. Wang, J. Peng, F. Huang, J. Wang, J. Chen, and Y. Xiao, "MICN: Multi-scale local and global context modeling for long-term series forecasting," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023, pp. 1–22.

[7] P. Hewage, A. Behera, M. Trovati, E. Pereira, M. Ghahremani, F. Palmieri, and Y. Liu, "Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station," *Soft Comput.*, vol. 24, pp. 16453–16482, Apr. 2020.

[8] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. W. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 2627–2633.

[9] S. D. Yang, Z. A. AliHyuna, and K. M. Wong, "Predicting complex Erosion profiles in steam distribution headers with convolutional and recurrent neural networks," *Ind. Eng. Chem. Res.*, vol. 61, no. 24, pp. 8520–8529, Jun. 2022.
[10] Y. Chen, K. Ren, Y. Wang, Y. Fang, W. Sun, and D. Li, "ContiFormer: Continuous-time transformer for irregular time series modeling," 2024, *arXiv:2402.10635*.
[11] S. Liu, H. Yu, C. Liao, J. Li, W. Lin, A. X. Liu, and S. Dustdar, "Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–20.
[12] M. A. Shabani, A. H. Abdi, L. Meng, and T. Sylvain, "Scaleformer: Iterative multi-scale refining transformers for time series forecasting," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023, pp. 1–23.
[13] T. Zhang, Y. Zhang, W. Cao, J. Bian, X. Yi, S. Zheng, and J. Li, "Less is more: Fast multivariate time series forecasting with light sampling-oriented MLP structures," 2022, *arXiv:2207.01186*.
[14] I. Tolstikhin, I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "MLP-Mixer: An all-MLP architecture for vision," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 24261–24272.
[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.
[16] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in *Proc. Assoc. Advancement Artif. Intell. (AAAI)*, Jun. 2023, pp. 11121–11128.
[17] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "FEDformer: Frequency enhanced decom-posed transformer for long-term series forecasting," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 27268–27286.
[18] T. Zhou, Z. Ma, X. Wang, Q. Wen, L. Sun, T. Yao, W. Yin, and R. Jin, "Film: Frequency improved legendre memory model for long-term time series forecasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 12677–12690.
[19] Z. Li, Z. Rao, L. Pan, and Z. Xu, "MTS-mixers: Multivariate time series forecasting via factorized temporal and channel mixing," 2023, *arXiv:2302.04501*.
[20] D. Luo and X. Wang, "Moderntcn: A modern pure convolution structure for general time series analysis," in *Proc. Int. Conf. Learn. Represent. (ICML)*, 2024, pp. 1–43.
[21] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023, pp. 1–24.
[22] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. Assoc. Advancement Artif. Intell. (AAAI)*, 2021, pp. 11106–11115.
[23] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 22419–22430.
[24] S. D. Yang, Z. A. Ali, and B. M. Wong, "LUID-GPT (Fast learning to understand and investigate dynamics with a generative pre-trained transformer): Efficient predictions of particle trajectories and Erosion," *Ind. Eng. Chem. Res.*, vol. 62, no. 37, pp. 15278–15289, 2022.
[25] S. S. Bahrainian, M. Bakhshesh, E. Hajidavalloo, and M. Parsi, "A novel approach for solid particle erosion prediction based on Gaussian process regression," *Wear*, vols. 466–467, Feb. 2021, Art. no. 203549.
[26] S.-A. Chen, C.-L. Li, N. Yoder, S. O. Arik, and T. Pfister, "TSMixer: An all-MLP architecture for time series forecasting," 2023, *arXiv:2303.06053*.
[27] A. Das, W. Kong, A. Leach, S. Mathur, R. Sen, and R. Yu, "Long-term forecasting with TiDE: Time-series dense encoder," 2023, *arXiv:2304.08424*.
[28] Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, and C. Zhang, "Connecting the dots: Multivariate time series forecasting with graph neural networks," in *Proc. Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2020, pp. 753–763.
[29] K. Zhao, C. Guo, Y. Cheng, P. Han, M. Zhang, and B. Yang, "Multiple time series forecasting with dynamic graph modeling," *Proc. VLDB Endowment*, vol. 14, pp. 7679–7693, Dec. 2023.
[30] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long- and short-term temporal patterns with deep neural networks," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 95–104.
[31] W. Wang, L. Yao, L. Chen, B. Lin, D. Cai, X. He, and W. Liu, "Crossformer: A versatile vision transformer hinging on cross-scale attention," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.
[32] M. C. Mozer, "Induction of multiscale temporal structure," in *Proc. Adv. Neural Inf. Process. Syst.*, 1991, pp. 1–8.
[33] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "STL: A seasonal-trend decomposition," *J. Off. Statist.*, vol. 6, no. 1, pp. 3–73, Mar. 1990.
[34] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Math. Comput.*, vol. 19, no. 90, pp. 201–216, 1965.
[35] H. Wu, H. Zhou, M. Long, and J. Wang, "Interpretable weather forecasting for worldwide stations with a unified deep model," *Nature Mach. Intell.*, vol. 5, no. 6, pp. 602–611, Jun. 2023.
[36] M. Liu, A. Zeng, M. Chen, Z. Xu, Q. Lai, L. Ma, and Q. Xu, "SCINet: Time series modeling and forecasting with sample convolution and interaction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 5816–5828.
[37] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–12.

ZHENNAN PENG is currently pursuing the M.A. degree with the School of Information Engineering, China Jiliang University. His research interests include artificial intelligence big models, time series analysis based on deep learning, and uncertainty quantification methods.

BOYONG GAO received the B.S. degree in forging from the Northeast Heavy Machinery Institute, Qiqihar, China, in 1994, the M.S. degree in metal plastic processing from the Huazhong University of Science and Technology, Wuhan, China, in 1997, and the Ph.D. degree in computer science and technology from Zhejiang University, Hangzhou, China, in 2011. He is currently an Associate Professor with the College of Information Engineering, China Jiliang University. His research interests include machine learning in time series and multimedia analysis.

ZIQI XIA is currently pursuing the M.A. degree with the School of Information Engineering, China Jiliang University. Her research interests include large language model (LLM), data mining, and deep learning.

JIE LIU is currently pursuing the M.A. degree with the School of Information Engineering, China Jiliang University. Her research interests include data transfer learning and machine learning algorithms.